University of California
Santa Barbara

The Role of Intonation Units in Memory for Spoken English

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Linguistics

by

Heather E. Simpson

Committee in charge:

Professor Fermín Moscoso del Prado Martín, Chair
Professor Patricia Clancy (Committee Member)
Professor Stefan Gries (Committee Member)
Professor Michael Spivey (Committee Member)

June 2016

The Dissertation of Heather E. Simpson is approved.

Professor Patricia Clancy (Committee Member)

Professor Stefan Gries (Committee Member)

Professor Michael Spivey (Committee Member)

Professor Fermín Moscoso del Prado Martín, Committee Chair

May 2016

The Role of Intonation Units in Memory for Spoken English

by

Heather E. Simpson

To my partner Antonio, for supporting me, challenging me, and always keeping it real.

Acknowledgements

I wish to express my gratitude to the faculty, staff, and students in the Department of Linguistics for being there for me throughout my time at UCSB. I would like to thank my dissertation committee for their advice on this work. Thank you to my chair, Fermín Moscoso, for his help and guidance. I want to thank Pat Clancy for her detailed comments and emotional support. I am very grateful to Stefan Gries for his feedback and encouragement. I want to thank Wally Chafe for his inspirational ideas and his thoughtfulness in describing the inner workings of language. Thank you to Jack Dubois for his input and enthusiasm about this research. Thank you to Alicia Holm for her help and kindness.

I could not have gotten this far without my friends. Thank you to Joseph Brooks, Alex Wahl, Brendan Barnwell, Allison Adelman, and Brad McDonnell for all the great discussions and laughter. An additional thanks to Allison for helping me to discover linguistics and then UCSB Linguistics, I am forever in your debt! Thank you to Kevin Schaefer, Caroline Crouch, and Nate Sims for letting me invade your home and for making me laugh so hard. Thank you to Kazuaki Maeda for all your support in my first few years, I wish you the best in everything. I am eternally grateful to Spalding Lewis for her wise counsel and friendship, you will always be the Gonzo to my attorney. Thank you to Leah Weimann, you are my life mate and I love you with all of my heart.

I would like to thank my family for their love and support: especially my parents, Pamela (and David) Lussier and Scott (and Sue) Simpson, and my siblings Jen, Tim, Ben, Zach and Caroline. I love you all!

Finally, I want to thank my partner Antonio for all his support. Thank you for being there for me in the best and worst of times, may we continue to grow together for the rest of our days.

Curriculum Vitæ

Heather E. Simpson

Education

2016	Ph.D. in Linguistics, University of California, Santa Barbara. Interdisciplinary Ph.D. emphasis, Cognitive Science.
2013	M.A. in Linguistics, University of California, Santa Barbara.
2006	B.A. in Linguistics, Bryn Mawr College.

Publications

Simpson, H. E., Moscoso del Prado Martín, F. (2015) Memory Capacity Limits in Processing of Natural Connected Speech: The Psychological Reality of Intonation Units. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pp. 2206-2211, Pasadena, CA, Cognitive Science Society.

Simpson, H. E. (2013). Book note for: Polly E. Szatrowski, Storytelling across Japanese conversational genre. *Language and Society* 42(3).

Simpson, H. E., Strassel, S., Parker, R., McNamee, P. (2010). Wikipedia and the web of confusable entities: Experience from entity linking query creation for TAC 2009 Knowledge Base Population. *Proceedings of the 7th Conference on International Language Resources and Evaluation*, Valletta, Malta.

Simpson, H. E., Maeda, K. and Cieri, C. (2009). Basic language resources for diverse Asian languages: A streamlined approach for resource creation. *Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Singapore.

Simpson, H. E., Cieri, C., Maeda, K., Baker K. and Onyshkevych B. (2008). Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Abstract

The Role of Intonation Units in Memory for Spoken English

by

Heather E. Simpson

Comprehension and production of spoken language are very memory-intensive tasks, especially in real-time natural interactions. Yet, it is well-known that human beings have a very limited capacity for retention of newly-presented material, a phenomenon normally attributed to limitations on short-term memory. This dissertation provides evidence that the Intonation Unit (IU), an intermediate-level prosodic phrase, serves a critical role in processing of spoken English by carving up the continuous speech stream into bite-sized 'chunks' that can be easily fit into listeners' limited focus of attention. Three empirical studies are presented: a study of memory span in terms of IUs, employing data from a verbatim recall experiment; a study of association strength between and across IU boundaries, employing data from the same recall experiment; and a study of priming duration in terms of IUs, analyzing a corpus of spoken English. The implications of the findings with respect to Wallace Chafe's (1980,1987,1994) conception of Intonation Units and theories of short-term memory are explored.

Contents

Curriculum Vitae **vi**

Abstract **vii**

1 Introduction **1**
1.1 Limitations on STM 2
1.2 Chunking 3
1.3 Mechanism for Prosodic Chunking 11
1.4 Intonation Units 15
1.5 Outline of this Dissertation 17

2 Memory Capacity in Spoken English **21**
2.1 Goals of Study 22
2.2 Experiment 26
2.3 Results 29
2.4 Scoring 29
2.5 Analysis 36
2.6 Discussion 39

3 Association Strength in Spoken English **41**
3.1 Goals of Study 42
3.2 Methods 43
3.3 Results 45
3.4 Discussion 54

4 Lexico-Syntactic Priming between Intonation Units in Spoken English **55**
4.1 Goals of Study 56
4.2 Methods 60
4.3 Results 64
4.4 Discussion 68

5 Conclusions **70**
5.1 Role of clauses 72
5.2 Implications for theories of language processing 74
5.3 Implications for theories of language change 75
5.4 IUs and STM 75
5.5 Conclusion 77

A Stimulus Metadata **80**
References 83

Chapter 1

Introduction

Comprehension and production of spoken language are very memory-intensive tasks. Spoken language in normal interaction is produced as a more-or-less continuous stream of output and input, and its physical representation disappears almost as soon as it is created. On top of this, the pressures of real-time interaction often require that a listener be ready to begin their next utterance the moment the current speaker ends their turn (Stivers et al., 2009). Compare this to written language, which even in its most interactive sub-genres (e.g. text messaging) allows for revision during production, and re-reading during comprehension.

At the same time, it is well-known that human beings are sorely limited in the amount of information we can retain from newly presented material. Remembering the ten digits of a phone number, or fifteen items on a grocery list, usually requires writing the information down or repeating the sequence of words over and over until it is memorized. This phenomenon is generally attributed to limitations on our short-term memory (Miller, 1956; Baddeley & Hitch, 1974; Baddeley, 2000; Cowan, 2000; McElree, 2001; Jonides et al., 2008).

Nevertheless, we are able to accomplish the amazing feats involved in using spoken language with our limited short-term memory (STM), without constantly asking our interlocutors to repeat themselves. This dissertation will combine current theory on the nature of STM with research on spoken language to explain how this is possible.

1.1 Limitations on STM

There are multiple competing views on STM capacity. Some researchers argue for a capacity limit in terms of a specific number of items, e.g. the classic 'magic number' 7 Miller (1956), the newer 'magic number' 4 (Broadbent, 1975; Cowan, 2000; Cowan, Saults, Elliott, & Moreno, 2002; Cowan et al., 2005) or 1 (Baars, 1988; Garavan, 1998;

McElree & Dosher, 1989; McElree, 2000, 2001; Lewis, Vasishth, & Van Dyke, 2006; Jonides et al., 2008)). Others adhere to a temporal decay model (Baddeley & Hitch, 1974; Baddeley, 1986). Still others argue for no special STM capacity at all, but rather a limit on recall defined by interference and cue distinctiveness, the same factors restricting recall from long-term memory (Crowder, 1993; Nairne, 2002). In the empirical analyses presented in this dissertation, I will focus on the item-based capacity limit view, but in Chapter 5 I discuss the broader implications of my results for STM theories.

In most newer models, STM is equated with the focus of attention (Cowan, 2000; McElree, 2001), rather than a completely separate memory module (for discussion, see Jonides et al., 2008). In this dissertation, I use the terms STM and focus of attention interchangeably.

1.2 Chunking

Though the exact model for STM is disputed, all of these views of STM have in common a general recognition of the role of 'chunking' in increasing the amount of information that can be retained in short-term memory. A chunk is a 'coherent memory unit' (McLean & Gregg, 1967), a group of items that have a strong association between them, allowing them to be treated as a single item. Bybee (2010, p. 7) describes chunking as "the process by which sequences of units that are used together cohere to form more complex units." Chunking is by nature hierarchical, so larger chunks can be created from smaller chunks (Newell, 1990; Cowan, 2000).

McLean and Gregg (1967) identify three types of chunks that may be present in traditional word list recall studies:

- Type I: chunks that already exist as coherent units for the participants

- Type II: chunks that are created after or during presentation by the participants through deliberate means (e.g. rehearsal, mnemonic devices)

- Type III: chunks that are created at presentation through manipulation of stimulus grouping by the experimenter

A traditional example from the memory literature of Type I chunks comes from recall of alphabet letters compared to acronyms. A subject asked to recall a sequence of 12 random alphabet letters would almost certainly make some errors, but if the sequence was actually composed of familiar acronyms, such as *ciacbsirsbbc*, their performance will be greatly improved, as this sequence will be processed as a smaller number of chunks (in this example, four chunks of three letters each: *cia cbs irs bbc*).

Type II chunks are created through the conscious efforts of participants to retain the information. The most common strategy used for this is 'rehearsal', continuous repetition of the information, either silently or spoken out loud (Baddeley, 1986). Rehearsal in recall experiments can be prevented through various means, such as having participants count or repeat a word, or by giving them a large amount of information in a short amount of time (Cowan, 2000).

Type III chunks have been created in traditional recall experiments in various ways, such as by visually grouping stimuli on the same card or screen (McLean & Gregg, 1967), or temporally grouping stimuli using pauses (Ryan, 1969; Frankish, 1985). Pauses between groups of stimuli must be longer than the interval between stimuli in order to be perceived as a group (Nairne, 1988). McElree (1998) grouped stimuli semantically, by presenting sequences of words from the same semantic category (e.g. *horse*, *sheep*, *rabbit*) and then switching the category for the next sequence (e.g. *fireman*, *secretary*, *doctor*). Grouping stimuli in these ways results in an overall improvement in serial recall performance, and can also result in the appearance of primacy and/or recency effects

within groups, meaning better recall for the first and/or last item in a group (Ryan, 1969; Frankish, 1985, 1989; McElree, 1998; Saito, 1998; Cowan et al., 2002).

1.2.1 Chunking in Language

As in the *cia cbs irs bbc* example, we can consider words themselves to be a representation of the first type of chunk (associations which have been created over experience). Above the word level, Type I chunks can be identified as groups of words that frequently appear together and/or have other associating properties such as a unified meaning. Behavioral measures of association strength have been shown to reflect co-occurrence frequency (e.g. Lockhart & Martin, 1969; Shanks, 1995; Ellis, 2002). Highly frequent 'multi-word' expressions like *I don't know* may even be treated as single word-level chunks, as can be seen in their tendency for significant phonological reduction (e.g. *I don't know* [aɪ doʊnt no] → [aɪ dəno] and even [aʌno]) (Ellis, 2002; Bybee, 2010). Researchers who take a 'usage-based' approach to language, e.g. Bybee (2002, 2010), Tomasello (2000), and Ellis (2002), treat this type of chunking as a fundamental mental process that affects language learning and language change. Bybee (2010) describes the process of chunking as an integral part of language change, in which multi-word collocations undergo unifying changes to meaning and phonological form, and may eventually become single words or even grammatical morphemes.

The second type of chunking can be seen in situations where listeners consciously attempt to maintain a sequence of words verbatim, such as when taking notes on a lecture, or trying to memorize a grocery list or a phone number. Though rehearsal is an arguably 'natural' use of language, it is clearly not normally a part of spoken language interactions.

But what about the third type of chunking? Since chunking at presentation aids

in retention of information, it seems that Type III chunks would be a helpful feature to have in managing the processing challenge presented by spoken language. Is there some linguistic feature or combination of features that accomplishes the task of creating chunks in real-time during language production and comprehension? For likely candidate features, we look to the two major sources of high-level structure in language: syntax and prosody.

1.2.2 Evidence for syntactic chunking

Syntax is the focus of most psycholinguistic investigations of connected discourse, and has both explicitly and implicitly been given a privileged status in the field as the most important aspect of structure in language. The sentence or clause has often been implicitly assumed to be the relevant unit for measuring memory capacity, e.g. Sachs (1967); Bransford and Franks (1971); Potter and Lombardi (1990, 1998); Gurevich, Johnson, and Goldberg (2010).

Jarvalla (1971) found some evidence of a role for both sentence and clause boundaries in STM. He tested verbatim serial recall on prose passages that were read aloud to the participants. The participants were prompted to recall the final two sentences of the passage. These two sentences were formed from three clauses, containing 7, 6, and 7 words respectively. There were two configurations for the sentences: long-short or short-long. In the long-short condition, the first sentence consisted of two clauses, and the second sentence consisted of one clause. In the short-long condition, the sentence boundary was switched and the second (final) sentence contained two clauses. The wording of the first of the three clauses was varied to allow the structure to be changed sensibly, but the final two clauses contained the same sequences of words. An example is provided in 1.

(1) a. Kofach had been persuaded by the international to stack the meeting for

McDonald. The union had even brought in outsiders. *Long-Short*

b. The confidence of Kofach was not unfounded. To stack the meeting for McDonald, the union had even brought in outsiders. *Short-Long*

The results showed strong effects for both sentence and clause boundaries on recall, with only the final clause being recalled at a high level of accuracy (average of 96% words correct) in both conditions. When the pre-final clause was part of the final sentence (the short-long condition), it was recalled much more accurately than when it was part of the previous sentence (81% vs. 50%). When the first two clauses were part of the same sentence (the long-short condition), their recall percentage was very similar (47% and 50%), but when a sentence boundary was crossed (the short-long condition), they differed significantly (29% and 81%). Thus we can see that sentence boundary had a very strong effect, and the pre-final clause boundary had almost no effect on its own. The final clause boundary, however, did appear to have a significant effect on recall. In the short-long condition, in which the last two clauses did not cross a sentence boundary, recall for the pre-final clause was 81% and for the final clause 96%, meaning that recall increased by 15 percentage points overall for the final clause. Of course, an average boost could be the result of high recall on the last few items in the clause rather than an effect applying at the clause boundary. However, Jarvalla (1971) also provides a graph of average recall by serial position. In that graph, recall for the pre-final clause in the short-long condition exhibits the classic U-shape serial recall curve (primacy and recency boosts with poor performance in the middle), but the entire final clause exhibits flat, ceiling level performance, with average values by position almost exactly matching that of the long-short condition's final clause. Therefore, it is clear that the clause boundary within the final sentence affected recall performance.

Similar results were found for a version of the study using the visual modality, re-

ported in Jarvalla (1979). In this study, participants silently read the passages on a screen displaying a 'moving window' of text. The moving window effect prevented them from viewing previously read portions of the passage. The modality of the responses was reversed as well - participants were asked to orally recall the sentences, whereas in Jarvalla (1971) they wrote down their responses. The average percentage of words recalled correctly for the three target clauses was 4%, 39%, and 85% for the short-long condition, and 11%, 15%, 88% for the long-short condition. The biggest difference from Jarvalla (1971), other than the overall poorer performance, is that the effect of the final clause boundary appears to be much stronger when the participants read the passages. Average recall for the second clause was 46 percentage points lower than for the third clause in the short-long condition, compared to 15 percentage points lower when the participants listened to the passages. At the same time, the boost in recall at the final sentence boundary was not as strong for the read passages. Sentences, unlike clauses, cannot be differentiated purely by syntax, they are defined by a combination of prosodic and syntactic cues (Chafe, 1994). Jarvalla (1979) speculates that it is the weakening of the sentence recency effect, due to lack of prosodic information, that is causing a boost to the clause recency effect, i.e. a shift in prominence from prosodic to syntactic cues.

Jarvalla (1979) describes an additional experiment employing his long-short/short-long sentence recall paradigm that attempted to isolate the role of prosody in the clause and sentence boundary effects. As previously described, Jarvalla (1971) found that when the prose passages were presented with normal prosody, the final clause was recalled nearly perfectly (96% average recall) regardless of whether it was preceded by a sentence boundary, but the pre-final clause was recalled much better if it was part of the same sentence as the final clause (81% vs. 50%). When prosodic information was removed through a reading done in a monotone voice and with a controlled pace, the sentence boundary effect was weakened considerably, with the pre-final clause recall at around

50% regardless of whether it was followed by a sentence boundary; however, the average recall of the final clause remained high (88-91% vs. 96%). This result corroborates the findings for the visual modality discussed above, that without prosodic information, the clause boundary has a very strong effect on recall.

Overall the results in (Jarvalla, 1979) show that the effect of the sentence boundary on recall was largely dependent on prosodic information, but there was a strong effect of the final clause boundary both with and without prosodic information. Based on this previous research, it seems that the clause is the strongest candidate for a syntactic unit that may induce chunking of language in memory.

1.2.3 Evidence for prosodic chunking

A number of sentence processing models incorporate the idea that prosodic phrases serve to chunk the input for further processing (e.g. Marcus & Hindle, 1990; Pynte & Prieur, 1996; Schafer, 1997; Slowiaczek, 1981). Schafer (1997) argues that syntactic constituents attach to the most 'visible' syntactic node, where visibility is a gradient value determined by the node's distance in terms of prosodic phrases (i.e. if the node is in the same prosodic phrase it is most visible, with gradient decline for each intervening phrase). Using eye-tracking methodology in a visual-world paradigm, Snedeker and Trueswell (2003) and Kraljic and Brennan (2005) found that prosodic phrase boundary cues disambiguated noun reference in an ambiguous context at the very initial stages of processing, before the relevant syntactic ambiguity was actually uttered. However, since these models focus only on prosody in relation to syntax, none of them specify or test the relationship of prosodic units to short-term memory capacity.

Prosodic units in recall

Marslen-Wilson and Tyler (1976) directly evaluated the effect of prosodic phrases on recall. They found that the effect of prosodic boundaries on recall remained even in the absence of syntactic information. Their stimuli were recordings of spoken prose passages ending in a sentence containing two eight-word clauses, with three conditions: normal prose, a semantically degraded condition, and a syntactically degraded condition. The speaker attempted to use the same prosody for all three conditions, by matching their prosody for the degraded conditions to the same word positions as the normal prose condition. In the semantically degraded condition, the words in the passage were replaced with randomly chosen frequency-matched ones from the same word class. In the syntactically-degraded condition, the passage used in the first condition was further scrambled through random re-ordering of the words, so it contained neither semantic nor syntactic information corresponding to the original passage.

The semantically and syntactically-degraded conditions induced lower average recall performance, with average recall of the final eight-word 'clause' at 86%, 75%, and 68%, for the normal, semantic, and syntactic conditions, respectively. The most dramatic effect was the reduction in recall for the pre-final clause, with performance at 79% (normal), 43% (semantically degraded), and 6% (syntactically degraded). Thus we can see that in the absence of syntactic or semantic information connecting the two clauses, there was an extremely strong chunking effect based on prosodic information, where participants appeared to only retain the chunk that they were currently focused on (i.e. the most recent 'clause' of the passage).

Of course, for natural language in use there will be syntactic and semantic information, so we would not expect such an extreme effect for prosody. Both this study and Jarvalla (1971) share a fundamental problem that does not allow us to tease apart the

effects of prosody and syntax, namely that the clauses and prosodic phrases in their stimuli always share a boundary.

Additional evidence for prosodic grouping comes from more traditional list recall paradigms. Pauses and pitch contours are both important boundary cues for prosodic phrases, and as mentioned in section 1.2, many word list recall studies have found that grouping list items with pauses seems to cause them to be treated as chunks, and leads to improved recall of the list (Ryan, 1969; Frankish, 1985, 1989; Saito, 1998; Cowan et al., 2002). The same kind of grouping effects have also been found for intonation-based grouping (Frankish, 1995; Saito, 1998). Frankish (1995) tested recall for digit lists created with a high-quality speech synthesizer, employing a pitch contour taken from a natural utterance to create three groups of three digits. He found an overall improvement in recall, along with a recency effect at the group level, i.e. high recall accuracy for the last item in each group.

1.3 Mechanism for Prosodic Chunking

The findings of Frankish (1995) help to address some important questions about the mechanism behind prosodic chunking and grouping effects more generally. We have stated that chunking can be done at presentation. Chunking by definition involves creating and/or strengthening associations between items in a group, and presumably improvements in recall could come from these associations. But how does grouping actually function to create chunks in memory?

One potential explanation is that the feature used to group adds an identical cue to all the items in a group that will facilitate retrieval of the set. This could explain grouping effects based on semantic category (McElree, 1998) and speaker identity (Frankish, 1989). Pitch contour grouping, on the other hand, involves variation in pitch across the group,

so this explanation falls short.

A possible explanation pointed out by Frankish (1995) is that the pitch contour is extracted and maintained in memory, and this attribute used to select the items belonging to that contour. To test this possibility, he measured the recall effect of pitch contours taken from familiar melodic structures, which should provide the same type of overall contour information. He found that the melodic contours did not improve recall significantly, indicating that there is some specific feature of natural pitch contours that gives rise to their strong grouping effects.

Frankish (1995) conducted one final experiment to identify that feature. He notes that the words at the group boundaries in the natural pitch contour from his first experiment contained a dramatic rise in pitch, and describes them as pitch-accented. Pitch accent is a term used in prosodic theory to refer to marked pitch changes, which in English and many other languages co-occur with increased duration and intensity to cue emphasis or stress on a word. Frankish tested the hypothesis that it is this pitch change that is responsible for the grouping effect, by creating a version of his stimuli with monotonic pitch on the first two words in each group, and identical pitch accent, copied from a single item in his first experiment, on the last word in each group. He again found significant grouping effects, and he concludes that it is the clear boundary cue provided by pitch accent, rather than overall intonation contour, that is key to obtaining grouping effects in recall.

Frankish (1995) proposes that grouping of auditory stimuli is accomplished through organization into "discrete events or perceptual 'objects"' in a separate auditory buffer store. The details of how this organization would aid retrieval from the buffer are not specified, but he states that it would increase the efficiency of auditory memory, and presumably these perceptual objects would be treated as a chunk.

A related alternative theory for prosodic grouping effects is that the focus of atten-

tion is responsible for chunk formation. The focus of attention can be thought of as a 'workspace' in which simultaneously-active items become associated with each other (Baars, 1988; Cowan, 1995, 2000). Cowan (2000) describes recall of a list of items as an attempt to reconstruct a series of prior STM states. Thus the mechanism for prosodic chunking may be that prosodic phrases serve to regulate the focus of attention. Prosodic boundary cues from a speaker could serve as a signal to the listener to shift their focus of attention. The associations between items in the phrase added by simultaneous activation in STM would then cause the phrase to be treated as a chunk in long-term memory. This explanation is not necessarily incompatible with an auditory buffer model; it could be that the perceptual division is done in a buffer, and then those segments are transferred to the focus of attention. Either way, the crucial point is that prosodic phrases are segmenting the speech stream into 'bite-sized' pieces that can fit into our limited focus of attention.

1.3.1 Prosody vs. Syntax

The few studies that have specifically investigated both syntactic and prosodic units in recall found effects of both clauses and prosodic phrases on recall performance (Jarvalla, 1971, 1979). However, it is important to note that in these studies, as well as in Marslen-Wilson and Tyler (1976), the researchers were not able to directly compare the effects of syntactic and prosodic grouping, because the stimuli were designed under the assumption that prosodic phrases would line up precisely with clause boundaries. In written-style language, this may very well be commonly the case, as the syntactic boundaries tend to be clear; we have conventions about the placement of prosody-regulating punctuation, and there is none of the 'messiness' involved in interaction such as restarts and interruptions. However, this should not be assumed for natural spoken language. Prosody has its own

hierarchical structure that is not isomorphic to syntax (Pierrehumbert & Beckman, 1988; Shattuck-Hufnagel & Turk, 1996).

Although clauses could be regulating the focus of attention in the way described in 1.3, given the additional evidence for prosodic chunking from serial recall studies (e.g. Frankish, 1995), and the demonstrably faster processing of prosodic boundaries over syntactic boundaries (e.g. Snedeker & Trueswell, 2003; Kraljic & Brennan, 2005), the evidence points toward prosodic phrases rather than clauses as the source of chunking at presentation.

In fact, the Intonation Unit (IU), an intermediate-level prosodic phrase defined by Chafe (1979, 1980, 1987, 1994) and further refined in Du Bois, Cumming, Schuetze-Coburn, and Paolino (1992); Du Bois (2014), is argued by Chafe to represent the contents of a speaker's focus of consciousness at the moment of verbalization. Chafe (1980) makes clear that he equates the focus of consciousness with the focus of attention, writing: "most of the information available to an individual is quiescent at any given time, only a small selection of it being activated in such a way that we would say we are paying attention to it, aware of it, or conscious of it." (p. 11) Chafe (1980) also states that this focus of consciousness has limited capacity, and that it moves jerkily from one thing to another rather than being a continuous stream of information. Overall, Chafe's conception of the focus of consciousness is remarkably similar to the limited-capacity STM described by Cowan (2000) and others, and in Chafe (1994) he explicitly equates that focus to the intonation unit. Similarly, Croft (1995), after finding that frequently used linguistic patterns ('constructions') are almost always produced within a single IU, suggests that the division of language into IUs may be a direct consequence of the limits of short-term memory storage. If IUs reflect a speaker's focus of attention, and thus STM limits, it is reasonable to expect that listeners, due to their experience as speakers, will have learned to process IU boundaries as a cue to shift the contents of their focus of attention (as

described above in section 1.3).

In this dissertation, I will test the hypothesis that IUs serve as the Type III chunking mechanism described in section 1.2, i.e. that they chunk the continuous speech stream into portions that can be incrementally processed in limited-capacity STM. I will compare the IU to the clause whenever possible, to address the possibility brought up in section 1.2.2 that the clause may be the more important chunking mechanism. In section 1.4 below, I describe the definition of the IU in more detail.

1.4 Intonation Units

IUs are segments of speech uttered with a coherent intonational contour. IUs often match up with clause boundaries (about 60% of the time in the conversational English speech analyzed by Chafe (1994)), so they are likely to be the closest analogue to clauses in the prosodic hierarchy. However, IUs and clauses are certainly not isomorphic. IUs can consist of a single word, as is often the case for discourse markers such as *well* and *okay*; a single clause can contain multiple IUs; and IUs can also contain more than one clause.

The examples below are excerpts from the IU-annotated transcripts in the Santa Barbara Corpus of Spoken American English (DuBois et al., 2000-2005). In 2, we can see that four of the five IUs in this excerpt are coextensive with clause boundaries, but IU 3 consists of a single noun phrase.

(2) The one that ... is b- .. blind now. *IU 1*

... And he was considered a killer. *IU 2*

... An unmanageable. *IU 3*

... And he's been perfectly lovely, *IU 4*

I give the little kids lessons on him. *IU 5*

In example 7, we see that a single clause can be spread out over multiple IUs.

(3) Anyway, *IU 1*

this girl must only weigh like, *IU 2*

a hundred and ten pounds. *IU 3*

And in example 4, we see two IUs, which each contain multiple clauses in the parses assigned to them by the Stanford Parser.

(4) And Bruge stands here half the day wanting to come in, *IU 1*

and then after he goes in he wants to go out. *IU 2*

The Stanford parser identifies two complete clauses (S nodes) in example 4 IU 1, *Bruge stands here half the day wanting to come in*, and the subordinate infinitive complement clause *to come in*. The parse for IU 2 contains three S nodes: *he goes in*, *he wants*, and *to go out*.

In section 1.4, I provide a brief discussion of the defining prosodic characteristics of IUs.

1.4.1 Definition of Intonation Units

IU boundaries are characterized by a complex of prosodic cues, generally representing significant shifts in the baseline or expected value of prosodic features (Du Bois, 2014). The major cues to IU boundaries include: pitch reset (abrupt change in baseline pitch level), anacrusis (accelerated speech rate) at the beginning of an IU, and prosodic lengthening of syllables at the end of an IU. Pauses are also a cue, but pauses are not a necessary or sufficient condition for an IU boundary (Chafe, 1980; Cruttenden, 1986).

The most basic defining property of an IU is that it is composed of a single coherent intonation contour. Though there are some differences in the specifics, the concept of a prosodic phrase composed of a single intonation contour is common to many prosodic theories, and may be referred as a tone unit (Quirk, Duckworth, Svartvik, Rusiecki, & Colin, 1964), tone group (Halliday, 1967), intonation group (Cruttenden, 1986), or intonational phrase (Selkirk, 1984; Nespor & Vogel, 1986). In the ToBi prosodic hierarchy defined by Pierrehumbert and Beckman (1988), the IU is comparable to the intermediate prosodic phrase, also called the phonological phrase (PPh). Like the IU, the PPh is also characterized as a coherent intonational contour, with boundaries indicated pause breaks and final syllable lengthening. A major difference between the two constructs is that the PPh construct is constrained by the number of nuclear pitch accents, namely there must be one and only one nuclear pitch accent (pitch change associated with a primary stressed syllable) in a PPh. The IU definition is more flexible; the relative strengths of multiple relevant cues can be taken into consideration in determining the boundary.

I am choosing to focus on the IU in this study due to the relevant previous claims about its function from Chafe (1979, 1980, 1987, 1994), and the availability of an IU-annotated corpus of naturally-produced spoken English recorded in a variety of contexts, the Santa Barbara Corpus of Spoken American English (SBC; DuBois et al., 2000-2005).

1.5 Outline of this Dissertation

Taken as a whole, prior research results support roles for both prosodic and syntactic structure in chunking of spoken language material. However, even in the few studies that investigate memory for spoken connected discourse (e.g. Jarvalla, 1979), the researchers' use of isomorphic clause and prosodic phrase boundaries in their stimuli does not allow for a clear conclusion to be drawn about the relative roles of syntax and prosody. In

addition, although spoken language in natural use ostensibly involves considerably higher demands on short-term memory than written language, the role of STM in natural spoken language remains notably understudied. This dissertation will address both of these issues by evaluating the relative effects of prosodic and syntactic structure on memory for naturally-produced spoken language.

In this dissertation, I present the results of three studies evaluating the proposal that Intonation Units represent chunks of spoken language that define the contents of our limited focus of attention. IUs are directly compared to clauses in the statistical model in the two studies for which that was possible (Chapters 2 and 3). In Chapters 2 and 4, which model memory span and priming duration, respectively, as a function of number of IUs, the shape of that function is predicted to be equivalent to that found for number of words in standard memory studies. In such studies, isolated words (common nouns, digits, letter names) are used as the stimuli, to avoid the effect of Type I (pre-existing) chunks on the results. Often, these words are presented individually to the participants. Therefore, my assumption is that in these studies, the word is the highest-level chunk available to participants, and also that in general, the contents of the participants' STM would be replaced every time a new word is presented, as the words are presented in isolation and are unrelated to each other.

The first study, discussed in Chapter 2, investigates the effects of IUs, clauses, and words on STM capacity by measuring recall for clips of spoken language that vary in IU count, clause count, and word count. Recall performance decay generally takes the shape of a logarithmic function, with a sharp decrease of recall accuracy from an initial ceiling level, followed by asymptote at a small number of words (Rubin & Wenzel, 1996; Cowan, 2000). This pattern is taken by some researchers to be evidence for an item-based STM capacity limit. The initial high-accuracy portion represents the active contents of STM, and the asymptote of accuracy represents where STM reaches its limits on capacity

(Broadbent, 1975; Cowan, 2000, 2008). Cowan (2000) states that excellent recall should occur for the active contents of the focus of attention, but when the information is no longer active, the former state of those contents must be reconstituted from long-term memory, a process which is prone to errors such as choosing the wrong prior state to recall, or selecting the wrong item from among the associated items in a prior state. Thus the recall performance decay function can be interpreted as a combination of high-accuracy STM recall for a small number of items, and lower-accuracy recall from long term memory (LTM). The shift to recall from LTM would be seen at the asymptote level of recall. If IUs - but not clauses or words - are serving to chunk information in memory, we should see this same discontinuous performance decay function at the IU level, but not at the clause or word level.

The second study, discussed in Chapter 3, evaluates the 'chunkhood' of IUs by measuring association strength between pairs of words in participants' memory within and across IU and clause boundaries. Intra-item association strength should be high within a chunk and low between chunks Cowan (2000). Wahl (2015) has provided some evidence that IUs respect the boundaries of Type I (pre-existing) chunks, finding that pairs of words with high association strength are very likely to occur within IU boundaries. Here, I investigate the hypothesis that processing IUs during comprehension creates new chunks. By definition, creating a new chunk will increase the association strength between the words in an IU. I measure association strength in memory in recall using the recall scores for pairs of words that were recalled from a clip of spoken language. Pairs of words which are highly associated should be treated as a unit, meaning they would either be remembered as a unit or forgotten as a unit. Therefore, the dependent measure used in this analysis is a binary value indicating whether the recall status for a word pair was matching (both remembered/both forgotten), or not matching (one remembered/one forgotten). I test the hypothesis that word pairs within an IU have a significantly higher

likelihood of matching recall status compared to word pairs which cross an IU boundary. The effect of pre-existing associations is controlled for by the inclusion of a corpus-based measure of collocation strength for each word pair.

The third study, discussed in Chapter 4, investigates the duration of lexical and syntactic priming effects in natural spontaneous interaction. The commonly observed decay function for priming is very similar to the recall function described in the Chapter 2 summary above, with a high level of priming that decays quickly, then stabilizes at a lower long-term level (Levelt & Kelter, 1982; McKone, 1995; Gries, 2005; Hartsuiker, Bernolet, Schoonbaert, Speybroeck, & Vanderelst, 2008; Reitter, Keller, & Moore, 2011; Pietsch, Buch, Kopp, & de Ruiter, 2012). Just as for recall, the shape of this function has been argued to reveal the capacity of short-term memory, with the location of the asymptote of the priming effect reflecting the maximum amount of items that can be contained in STM (Cowan, 2000). I evaluate the hypothesis that the priming decay function for IUs will match both that commonly observed decay function and the function observed for recall in Chapter 2. Following Moscoso del Prado Martin (2015), I use shared Shannon information between pairs of IUs to measure the amount of priming.

Finally, Chapter 5 concludes the dissertation with a discussion of the overall results and implications of the findings for theories of short-term memory and language.

Chapter 2

Memory Capacity in Spoken English

2.1 Goals of Study

This study investigates the notion that the IU represents a chunk of spoken language that may define the contents of the focus of attention/short-term memory (STM), by observing memory span for naturally-produced spoken English clips that vary in number of IUs, clauses, and words. The effect of IUs on recall will be compared with that of clauses and words, and we will look for a discontinuity in performance at some number of those units.

Discontinuity in memory performance at particular list lengths or stimulus set sizes has been argued to be an indicator of the capacity limit of STM (Broadbent, 1975; Cowan, 2000, 2008). Cowan (2000) states that there should be a flat performance function across list length or stimulus set size until three or four items, and he cites a number of studies that have found this function for various types of stimuli. For example, many studies of subitization, the ability to estimate how many objects are in a set without counting, have found essentially error-free performance up to four items but a decreasing performance function thereafter (Jevons, 1871; Atkinson, Campbell, & Francis, 1976; Mandler & Shebo, 1982). Similar results were found for visual tracking of a subset of moving dots (Yantis, 1992). Halford, Wilson, and Phillips (1998) investigated proactive interference (PI), the interference of older items with more recently processed items, for words from the same semantic category and rhyming words. They used a probed recognition task in which three lists of related items were presented sequentially, with probed recognition performed after the list. Thus, the recognition trial after the first list would be a low-PI trial, and the third trial would be a high-PI trial. They tested list lengths of four and six for rhyming items, and four and ten for semantically-related items. They found for the longer lists, there was PI, but for the list lengths of four, there was no PI. Cowan (2000) also provides a table summarizing the results from nine word list recall studies involving

articulatory suppression that appear to support an estimate for memory span of about 3-5 items.

Decay of recall accuracy as measured over time is also best described by a non-linear function such as the logarithmic function (Rubin & Wenzel, 1996). Generally, ceiling levels of recall accuracy are found for the most recently processed word, then recall performance decreases sharply, and then flattens out to a fairly stable long-term level (Rubin & Wenzel, 1996). Decay of recognition accuracy over list positions has been found to have a similar function (Wickelgren, T., & Dosher, 1980; McElree, 2001).

Connected meaningful language is generally assumed to contain chunks of information that are larger than the word. Any process that is sensitive to chunking, therefore, such as recall, is likely to have its observable patterns shifted to higher-level units in spoken language. If a higher-level unit in spoken language exhibits the same recall behavior as isolated words or other smaller units do in memory studies, that would be evidence for that unit being treated as a chunk in STM.

As discussed in section 1.2.2, Jarvalla (1971) found a discontinuous function for recall of the final portion of a spoken prose passage, but the clause boundary defined the discontinuity in performance. He found that recall performance was at ceiling levels for the final clause of the passage, with a sharp decrease at the clause boundary, and an even sharper decrease when that clause boundary was also the final sentence boundary.

This study evaluates the hypothesis that the IU, rather than the clause, segments spoken language into chunks for processing in STM, and makes the following three predictions: 1) the number of IUs in a stimulus will have a significant effect on recall, 2) recall will exhibit a non-linear decay function over number of IUs, specifically a sharp decrease with asymptote at a small number of IUs, e.g. 3-5 as would be predicted by Cowan (2000), 3) clauses and words will not have the same significant non-linear effect on recall.

As mentioned in section 1.5, I test this hypothesis using a verbatim recall task on clips from the Santa Barbara Corpus of Spoken American English (SBC, DuBois et al., 2000-2005) that vary in number of IUs, clauses, and words. The use of verbatim recall of connected discourse as a task may be somewhat controversial due to the commonly-held view that verbatim memory for meaningful language is highly limited, or even completely non-existent (Lombardi & Potter, 1992; Potter & Lombardi, 1990, 1998), so I will address this concern in the following section.

2.1.1 Verbatim Recall

The view that verbatim form is not retained in memory dates back to the classic study by Sachs (1967). She tested subjects' ability to discriminate between sentences they had read from a short prose passage, and meaning-related or form-related distractor sentences. She found that subjects had very good performance when the meaning was changed, but, with the exception of the last (i.e. most recently-processed) sentence in the passage, they had very poor performance when only the form was changed. Sachs concluded that syntactic form was only retained in memory for a very short interval and then discarded in favor of a 'gist'-based representation of the meaning. In an even stronger version of this view, Potter and Lombardi (1990, 1992, 1998) argue that form is not retained in memory at all, and the high recall performance for the most recently heard items is due solely to the priming of words from their activation in the immediately preceding sentence.

In contrast to this view, there is a great deal of evidence in the literature that more than the gist of connected discourse is retained. Retention of acoustic/phonetic details has been shown in studies of phonetic convergence between speakers in conversation (Pardo, 2006), phonetic convergence in word-shadowing (Goldinger, 1998), and effects of

talker voice on recognition memory (Luce & Lyons, 1998). Speer, Crowder, and Thomas (1993) found that the prosodic structure of a sentence was accessible from memory in a sentence recognition task. In addition, a wealth of evidence that corpus distribution of individual words and collocations has effects on language representation and processing (e.g. Bybee, 2010) makes it clear that exact wording does leave its mark on memory.

Most relevant for the current study, Gurevich et al. (2010) demonstrated that a significant amount of verbatim content was retained and retrieved from long-term memory, compared to alternative wordings with the same gist. One of their experiments showed that material from a confederate's verbal description of a short cartoon was re-used in participants' own descriptions of the same video over intervals as long as six days, even though participants were not warned there would be a memory task and were not asked to recall the description they had heard previously. Therefore, the claim that everything but the gist is immediately discarded is not tenable.

Furthermore, even if the explicit verbatim recall ability in such research is attributed to priming effects, it is difficult to specify exactly how that differs from recall itself. If a subject is successful in a conscious attempt to recall a portion of speech verbatim, and the reason for that is the lingering activation from the forms that the subject heard, how can we actually distinguish that from true recall? Indeed, Cowan (2000) cites the duration of a short-term repetition priming effect as evidence for his proposed STM capacity limit, and I investigate priming effects in the current study in Chapter 4. Accordingly, in this Chapter I do not make a distinction between verbatim recall arising from priming and verbatim recall in general.

2.2 Experiment

2.2.1 Methods

Materials

Stimuli were selected from the Santa Barbara Corpus of Spoken American English (DuBois et al., 2000-2005). The SBC corpus contains audio files of naturally-produced spoken English with accompanying detailed transcriptions. In order to avoid effects on recall from speaker change, only portions of continuous speech from a single speaker were considered as stimulus candidates. I will refer to such portions as speaker 'turns'. Turns were automatically processed to extract their IU, clause, and word counts.

The Stanford Parser (version 3.2.0 [1]) was used to create syntactic parses from which clause counts were extracted. The parser treated prosodic sentence boundaries as the root for the parse, with the prosodic sentence being defined as the following: within a single speaker's turn, an IU with final falling intonation (indicated with a "." in the transcript) or final rising intonation (indicated with a "?" in the transcript), together with any preceding IUs with continuing intonation (indicated with a "," in the transcript). Clause counts were equated with the number of S nodes in a parse. SBAR nodes were not included in these counts as they indicate an additional higher-level S node that normally has complete or nearly complete overlap with a corresponding S extent, such as a node for a relative clause marker, which is nearly entirely coextensive with its contained S (see figure 2.2.1 for SBAR example).

Figure 2.2.1 shows the Stanford parse for a portion of Stimulus 0, representing a prosodic sentence.[2] This excerpt contains two IUs, with boldface indicating the word that starts the new IU. The ".." indicates a pause occurred after the truncated word

[1] http://nlp.stanford.edu/software/lex-parser.shtml, accessed 06/20/2013
[2] See the Appendix for metadata on each stimulus

"*h*-". As we can see here, the Stanford parser has some trouble with the unique features of spoken language, incorrectly identifying the truncated *h*- as an adjective, but overall it does quite well in representing the structure.

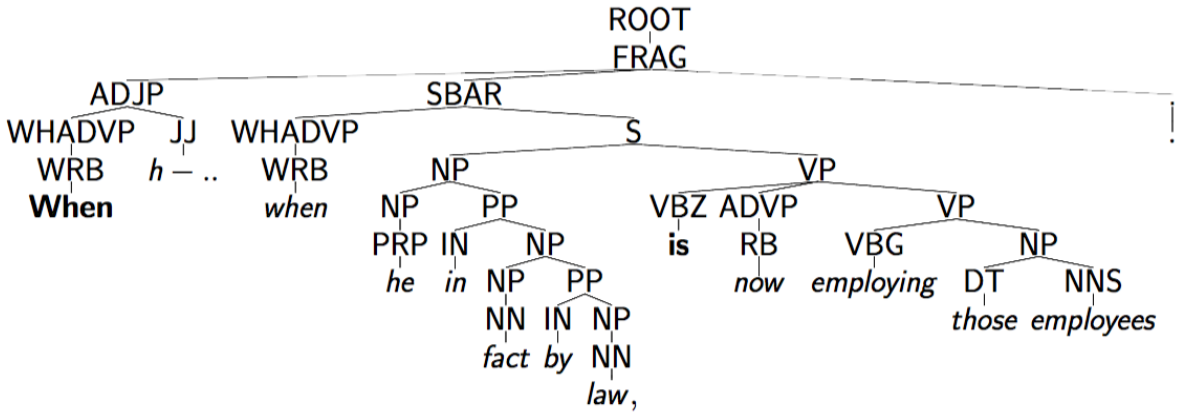

Figure 2.1: Stanford parse tree example taken from Stimulus 0

IU counts were derived from the Intonation Unit boundaries provided in the SBC. This information was used to select 54 stimuli, which were intended to represent a low, medium, and high range for number of IUs, clauses, and words, with two examples per combination. The number of IUs was used as a starting point, with a targeted low IU count range of 2-4 IUs (within Cowan's (2000) 3-5 item capacity limit), and medium and high ranges that represented an exponential increase from that range (around 8 IUs for medium, and 16 IUs for high). After suitable transcript portions were identified for each IU count range, stimulus candidates were chosen to represent the low, medium, and high ranges of clauses found within that IU range, and finally the stimuli were chosen to represent low, medium, and high word ranges found within the remaining pool of candidates. Other practical considerations, such as the general clarity of the speech in the stimulus, and avoidance of repetitions and truncated words, also guided stimulus choice. Table 2.1 below shows a summary of counts for each of the three linguistic units. A table with complete stimulus information is provided in the Appendix.

Table 2.1: Stimulus length ranges

	Low IU (n=18)	**Med IU** (n=18)	**High IU** (n=18)
	2-4	5-10	14-17
Clauses	1-8	2-7	8-28
Words	4-40	10-44	49-99

Participants

113 undergraduates recruited from introductory linguistics courses participated in the experiment. Students received extra credit in their courses as compensation for their participation.

Procedure

Stimuli were presented on a desktop computer over high-quality over-ear headphones. OpenSesame (Mathôt, Schreij, & Theeuwes, 2012) was used to create the experiment interface. Participants were given the following written instructions:

> Your task is to listen to a series of short audio recordings and then transcribe what you heard from memory. There are 54 recordings total of various lengths. It should take about 30 minutes to complete the experiment. Please feel free to take a break at any point if you are feeling mentally fatigued.
>
> When you begin the experiment, you will hear an audio recording play over your headphones, and after it finishes, you will see a text box prompting you to provide your transcription in the Notepad file. **Please type out everything you remember the person saying, even if it doesn't make sense or you aren't sure how to spell something.** You will only get to hear each clip once, and some are fairly long, so you won't always be able to remember

everything that you heard the person say, but just put down everything you remember to the best of your ability.

Then save and minimize the Notepad file, and press "Enter" to go on to the next recording.

You will start with one practice file, and you will get a prompt after that file to press any key to go on to the experiment. If there are any problems or you are confused about the task, please let me know after you do the practice file, before going on to the experiment.

As indicated in the instructions, the participants provided their transcriptions inside a Notepad text file. This was done to allow full word processor-style manipulation of the text, which was not possible in the experiment design software used. The text file contained a number for each stimulus, separated by a dashed line. The experimenter or research assistant gave verbal instructions to the participants to provide their transcriptions between the dashed lines. The 54 test stimuli were presented in random order, preceded by one short training clip (IU count = 3, clause count = 4, word count = 10). Participants were instructed to alert the experimenter or research assistant of any confusion or issues with the task after the training clip. This clip was the same for all participants and it was not included in the analysis.

2.3 Results

2.4 Scoring

Participants' transcripts were reviewed by a research assistant to fix formatting errors and unambiguous misspellings (e.g. "taht" instead of "that"). Transcripts were then

processed automatically to extract the transcript for each clip and link it to trial order information taken from the experiment interface. Problems with processing the text files due to remaining formatting errors led to exclusion of data from 12 of the participants, so the total number of transcripts used in the analysis was 101.

Participant transcripts were then scored against the original transcript from the SBC corpus. I will refer to the original transcript as the gold standard (GS) transcript. Scoring against the GS transcript was designed to be an exact serial recall measure.

An algorithm was created in the Python programming language to automate the scoring process. There were three passes over the data in this algorithm. In the first pass, participant transcript words with no exact match in the corresponding GS transcript were scored as incorrect. For words with exact string matches, the matching word position indices in the GS were identified and stored. In the second pass, the set of matching GS word positions were assigned to the participant transcript words in order (i.e. the first instance in the transcript was matched to the first instance in the GS). This would ensure that, for example, if there were three instances of *the* in the GS and *the* in the participant's transcript, they would be matched to the GS in the same relative order.

In the third pass, the algorithm reviewed relative position order information to correct any mistakes in the second pass assignment of GS positions. The assigned GS position for each word was compared to the GS position of the closest prior word with a GS match. The word's score was changed to incorrect if the previous match was more than 20 positions lower, because it was most likely either a spurious match, or correct recall but in the wrong order, which in a serial recall paradigm is scored as incorrect. There was one exception made: the word was still considered correct if the closest following word had a GS match that was also greater than 20 positions ahead, indicating that the participant omitted a large portion of the stimulus but is now correctly recalling a later portion. Spurious matches can occur for words with multiple matches in the transcript.

For example, if the word *the* was used instead of *a* in a particular noun phrase, it would be quite likely to have a string match to another instance of *the* from another part of the transcript. Requiring two sequential string matches to validate a jump in position match helps to minimize this type of error.

If the closest prior word with a GS match had a *higher* GS word position, suggesting that the current word's match is incorrect, the other GS match positions for the current word were reviewed, and the highest one closest to the prior match was chosen. If the highest remaining position index was still lower than the prior word's GS match position, the word was scored as incorrect. This third pass accounted for additional errors where the participant recalled part of the clip in a displaced order, and where a match was attributed incorrectly.

As was just mentioned, there is some potential for error using this algorithm when there are multiple instances of the same word in the GS transcript, for example if there are two instances of *the*, at positions 2 and 10 (where 0 = first word in stimulus), and the participant had three instances of *the* in his/her transcript, with the first and third being correct, the algorithm could judge the third instance of *the* in the participant's transcript as incorrect, because their second *the* would have 'taken' the last available index, position 10 (allowable since they are less than 20 positions apart). A total of 20.4% of words in the GS transcripts have more than 1 match within their own transcript, but only 5.7% have more than 2 matches, so that damage should be fairly minimal.

Below, I illustrate the behavior of the scoring algorithm with three examples of participant transcripts and corresponding GS transcripts. The words that were scored as correct by the algorithm are in bold. The overall score value indicates the percentage of words in the GS transcript that were marked correct for this participant.

(5) **Kind of.** A- a semi **bumped a** car, **and** then went and, went on two wheels,

and, just about **lost** it, **and** then **got back up** on all **all its wheels again**. *Gold Standard (Stimulus 50)*

Kind of. bumped a semi **and lost** control **and got back up** and lost **all its wheels again**. *Participant, Overall Score: 78%*

In example 5, *kind of* in the participant's transcript is matched to positions 0 and 1 in the GS. Next, *bumped a* is matched to positions 7 and 8 in the GS, just before *car*. When *semi* then appears, although it matches to the GS, it is scored as incorrect for being in the wrong position (skipping backwards to position 5). Then, *and* is matched to the next available *and* instance, at position 10 (after *car*). The next word, *lost*, matches to the GS as well, at position 24. The word *and* after *control* is matched to the GS in *lost it,* ***and*** *then*. Finally, two multi-word sequences, *got back up* and *all its wheels again* are scored as correct.

(6) **I always said, that** if I had children, I would always- **Especially if I had** a son, I would hug him, just so he knew **that I** loved **him**. *Gold Standard (Stimulus 47)*

I always said that especially if I had children **that** if **I** had a son I would hug **him** just so that he knew that I loved him. *Participant, Overall Score: 43%*

After the initial *I always said that*, the word *especially* is matched, jumping ahead to position 11 in the GS. Continuing in left-to-right order, *if I had* is matched with its occurrence after *especially*, and therefore *children* is scored as incorrect for being in the wrong position. Then the participant wrote *that if I*, and again, rather than going back to the text before *children*, the algorithm continues in left-to-right order and matches *that* and *I* to their next occurrence, near the end of the GS transcript in *that I loved him*. After this, because the GS match position is now at *that* ***I*** *loved him*, the correctly

recalled *had a son I would hug* sequence is scored as incorrect for being in the wrong position. After this, *him* is scored as correct, because there is a match in the GS after *loved*, and since *him* is the last word in the GS transcript, the rest of the participant's transcript is scored as incorrect.

(7) Anyway, this girl must only weigh like, a hundred and ten pounds. I mean, **she's just** a little shit. **And** she's out there, and she's got huge **arms**. I mean she's in shape like you can't believe. She's out there just, working away. And, those **guys** are so used to it, that they **do it all day long**. you know. *Gold Standard (Stimulus 2)*

She's just hundred **and** ten pounds this little shit. she has them big **arms** then there's these **guys** who **do it all day long**. *Participant, Overall Score: 17%*

In this example, the participant wrote *She's just* first, which immediately skips ahead to positions 18 and 19 in the GS, just before *a little shit*. Because of this, in *hundred and ten pounds*, *hundred* and *pounds* are marked as incorrect, because they occur before *she's just*. Only the *and* in this sequence is judged as correct, because there is an *and* in position 24, just after *little shit*, The next allowable match is *arms*, in position 33, which is followed by *then there's these*, which do not match any of the words in the gold standard so are incorrect. Finally, *guys* is marked correct, as well as the final phrase *do it all day long*.

When scoring algorithm may seem severe, but since I am evaluating serial recall, and looking for the boundary of ceiling level performance, it was deemed appropriate to choose a scoring method that would penalize out-of-order recall harshly. This helps to ensure that the results truly show the boundaries of capability for verbatim recall, rather than a 'gist' recall.

Admittedly, the left-to-right matching direction introduces some bias based on word position. If a match skips ahead in position, words after that match will be marked incorrect (as in *that...I* in example 6), but if a match skips backward in position (as in *semi* in example 5), only that match will be marked incorrect. However, although they were not restricted from writing the most recent part of the stimulus first, it is expected that participants would tend to write down their answers in temporal order, from the beginning of the stimulus to the end. Therefore, it seemed sensible to score their results in this direction. In addition, since the effect of interest in this study is the size of the stimulus, and word position can be controlled for by including it as an effect in the model, this bias is not expected to be problematic for the results.

2.4.1 Manual scoring analysis

To provide a sense of the accuracy of the scoring algorithm, in this section I describe the results of a manual scoring analysis conducted on a subset of the data. I trained three undergraduate research assistants to score the participant responses against the gold standard. I created a basic user interface for scoring using the PsychoPy library in Python. In this user interface, they were able to score one word at a time from a participant's transcript. For each word, they were shown the word in context in the participant's transcript, then selected the word in the gold standard that best matched the participant transcript, and indicate what type of match it was - an exact match, a replacement with similar meaning (e.g. *big* instead of *huge*), and a few other categories. If there was no match, they could indicate that as well. They provided scores for 44,539 words out of 107,052 total words in the set of participant transcripts (42%). The scored words covered some or all of the transcripts for 28 out of the 54 stimuli. Very little scoring was performed on the stimuli with the highest word counts, as they were very

difficult to score both due to the typically very poor performance of participants on these stimuli, and also due to physical limitations on screen size for viewing the list of words in the interface.

The human annotators were not asked to provide scoring based on word order, but instead were simply asked to score a word as an exact match if they felt intuitively that it had been recalled correctly by the participant. Unsurprisingly, they marked correct many of the words that were scored by the algorithm as incorrect due to word order errors. Of the words with a wrong position score from the algorithm, human annotators scored 75.7% as an exact match. This may seem like an alarmingly high rate, but this does not in itself indicate a problem with the scoring algorithm, it indicates that humans were not following the same rules as the algorithm. Based on my general experience supervising human annotation tasks, and my specific experience with testing out the scoring process on these transcripts, I believe that scoring these transcripts with explicit penalty for word order is a task better suited for an automated system than for human judgment. Unlike humans, the algorithm will be completely accurate in spotting every string match in the gold standard, and will be completely consistent in making a decision about them based on its rules of left-to-right position match order.

The most important metric for this scoring algorithm is precision of the 'correct' class - i.e., whether it allows words that should be marked as incorrect to slip through to the correct category. On this metric, it is appropriate to compare to manual scoring, as human annotators can be expected to perform with high precision, with only some small amount of error due to carelessness. The algorithm fares very well in this comparison. Of the words that the algorithm considered correct, human annotators judged 93.3% to be an exact match. This is very good performance, and thus supports the algorithm as a reliable way to measure verbatim recall performance.

2.5 Analysis

A generalized additive mixed-effects logistic regression with restricted maximum likelihood (REML) estimation was used to predict the likelihood of a correct match for each word in the stimulus based on the length of the stimulus in IUs, clauses, and words. Unlike linear or linear logistic models, generalized additive models do not force a linear relationship between the predictor and the dependent variable. Rather than choosing a single coefficient parameter for a given predictor, they allow the coefficient to be locally determined, i.e. to vary for different values of the predictor (Hastie & Tibshirani, 1986). Polynomial transformations of a predictor can approximate this effect, but they do so by changing the 'observed' values of the predictor variable, rather than actually allowing the coefficient of the model to vary. Essentially, generalized additive models minimize the assumptions made about the nature of the predictor function, so they are a good choice when the shape, rather than overall significance, of the predictor function is the primary result of interest.

The dependent variable was a binary value for each GS stimulus word for each participant, indicating whether that word was assigned a correct score for that participant in their transcript. The independent variables were: IU Count, Clause Count, Word Count, and Word Position.

Word Position indicated the position of the word in the stimulus, from most to least recent (0 = the last word in the stimulus). This variable was included to control for primacy and recency effects on recall, as well as bias introduced by the scoring algorithm as discussed in section 2.4. Random slopes and intercepts for participant and stimulus identity were also included to control for individual variability. Since I was primarily trying to identify the shape of the effects of the three stimulus length predictors, and the control variable (word position), I did not perform model selection based on significance.

The effects for length of stimulus in IUs and words, as well as the positional effects for word were found to be significant. The effect of length of the stimulus in clauses was nonsignificant, as the 95% point-wise confidence intervals computed by the model included 0 for all values of clause count.

The fixed effect and random effect summary statistics for the GAMM model are provided in Table 2.5. All four fixed-effect predictors mentioned above were fit with smoothing functions, so the only parametric term was the intercept. The significant non-parametric fixed effects from the model are plotted in Figure 2.2. The y-axis indicates the strength of the estimated effect of that variable on word recall. Values above 0 indicate a significant positive effect, and values below 0 indicate a significant negative effect.

Table 2.2: Summary statistics for GAMM model of the effects of stimulus size on likelihood of recall

Parametric Term	Estimate	Std. Error	t-value	P-value
Intercept	-0.3221	0.0755	-4.2654	<2e-16***
Smooth Term	Variance	Smooth Par.	DF	
sx(CLCount)	0.0000	23572.4000	1.1250	
sx(IUCount)	0.0022	453.9550	2.8967	
sx(WCount)	0.0002	4257.1000	1.5647	
sx(WordPosition)	0.2927	3.4162	18.0702	

Panel (a) of Figure 2.2 shows the size of the smoothed effect of IU count on word recall as a function of IU count values. There is one significant positive portion of the effect curve at IU counts of less than 3-6, with the rest of the IU count values having 95% confidence intervals that cover the 0 line. This indicates that having a low-IU stimulus has a significant positive effect on recall, but after the first 3-6 IUs, adding more IUs to the stimulus does not have a significant effect on recall.

Panel (b) shows the size of the smoothed effect of word count on word recall as a function of word count values. The effect for word count appears to be significant but linear, with high positive effect on recall for low word counts and a high negative effect

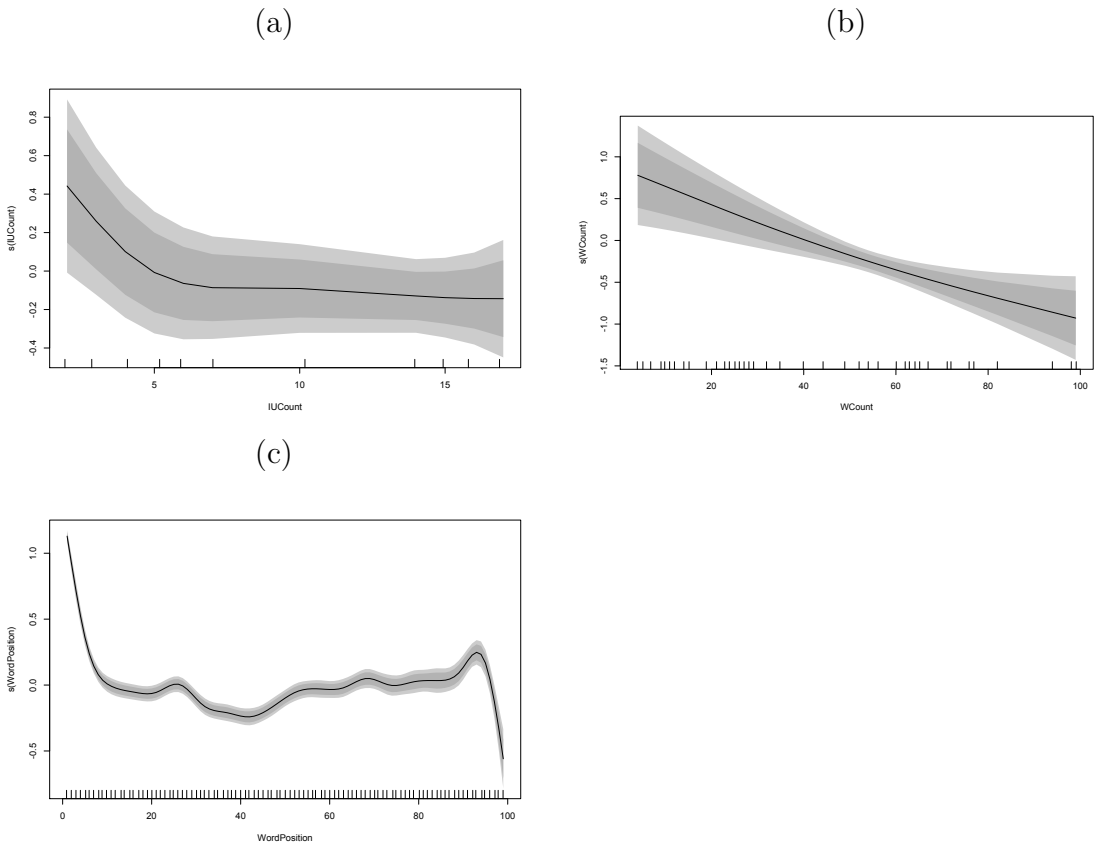

Figure 2.2: Plot of significant non-parametric estimated effects on recall including 80% and 95% point-wise confidence intervals for (a) Effect of IU count, (b) Effect of word count, (c) Effect of word position (0 = most recent)

on recall for high word counts.

Panel (c) shows the size of the smoothed effect of word position on word recall as a function of word position values. Word position exhibited a strong recency effect, as shown by the positive values for its effect on recall for words at the end of the stimulus (low word position values), with a fairly flat function after that, until a decrease at the highest word position values.

The significant negative effect at high word position values could indicate an anti-primacy effect, where the words at the beginning of the stimulus were less likely to be recalled. However, it should be noted that word position reflects raw positional values,

not normalized by the length of the stimulus, so this 'anti-primacy' effect is most likely due to relatively poor recall for the beginning of the longest stimuli as compared to stimuli having medium or low numbers of words.

To investigate this possibility, a generalized additive mixed-effects regression model was fit on only the low and medium IU length stimuli. The same overall pattern of results was found, supporting the stability of the findings described above. A small positive increase in the effect of word position at the very highest positions indicated a small primacy effect, supporting the stimulus length explanation for the word position function at the highest values.

2.6 Discussion

The results indicate that IUs play an important role in memory for naturally-produced spoken English. In accordance with the predictions in section 2.1, IU count had a significant, non-linear effect on recall, and word and clause count did not. The number of IUs was found to have a significant positive effect on recall performance when there is a small number of IUs in the stimulus, but to have no significant effect for stimuli with more than 3-6 IUs.

Clauses do not appear to play a significant role in recall of spoken English, with no significant effect found for number of clauses in the stimulus. Previous findings of significant effects of clause boundaries on recall (e.g. Marslen-Wilson & Tyler, 1976; Jarvalla, 1979) may be a side-effect of the overlap between clause and prosodic phrase boundaries in their stimuli.

Word count had a significant effect on recall, but the non-linear accuracy function that appears in various types of word list studies (Halford et al., 1998; Rubin & Wenzel, 1996; McElree, 2001) appears to be shifted to the IU level. The linearity of the word

count effect suggests that the negative effect from adding more words to the stimulus is entirely due to factors such as intra-stimulus interference, while the discontinuity for IUs can be taken as support for a privileged short-term capacity at the IU level. It should be noted, however, that the IU count function did decay quickly over the first few values, so it does not appear to support the flat performance function predicted by Cowan (2000) for list lengths within STM capacity. The IU effect function is more similar to commonly-observed functions for decay of accuracy over time or list position, exhibiting an initial ceiling level of performance which quickly decays and levels off to a long-term fairly stable baseline level (Rubin & Wenzel, 1996; McElree, 2001).

Chapter 3

Association Strength in Spoken English

3.1 Goals of Study

As stated in Chapter 1, this dissertation investigates the hypothesis that IUs represent Type III chunks - chunks created during processing by some grouping property of the input. A chunk is a group of items with strong inter-item associations in memory (McLean & Gregg, 1967; Newell, 1990). If IUs create new chunks, by definition this means that the inter-word associations within an IU are increased when the IU is processed.

Therefore, we should be able to test whether IUs are chunking spoken language by measuring the pre-existing association strength between words within and across IUs, and comparing this with their association strength after they are processed. If association strength was added for words within IUs during processing, words within IUs should have stronger associations than those across IUs, when pre-existing associations are taken into account.

To measure post-processing association strength for words in an IU, I analyze the participant responses from the same recall experiment used in Chapter 2. If words A and B are strongly associated, we can expect them to be treated as a unit in recall, i.e. if word A is remembered (or forgotten), word B will also be remembered (or forgotten). As an illustration of this, it would be very strange to find that a participant correctly recalled only one of the words in a highly formulaic word sequence like the discourse markers *I mean* or *you know*. We can then predict that an increase in association strength within IUs would be measurable as an increased likelihood that words in the same IU will match in their recall status, controlling for pre-existing associations, compared to words that cross an IU boundary.

Pre-existing associations between items can be approximated by measures of co-occurrence frequency (Shanks, 1995) For words, this means measuring their collocation strength in a corpus (Lockhart & Martin, 1969; Ellis, 2002). In this study, I employ

pointwise mutual information (PMI) to measure collocation strength for the bigrams in the recall experiment stimuli. There is some evidence that Type I (pre-existing) chunks tend to appear within IU boundaries, rather than be split across multiple IUs. Croft (1995) asserts that highly cohesive multi-word structures in conversational English are very unlikely to be split across multiple IUs, and Wahl (2015) supports that claim by showing that highly associated bigrams are very unlikely to by split by an IU boundary.

This study will test the hypothesis that IUs create new chunks during processing by modeling the effect of crossing an IU boundary on the likelihood that a two-word sequence (a bigram) will match in its recall status. PMI is included as an important control expected to affect the likelihood of matching recall status, and to account for a potential confound of pre-existing association strength with IU boundaries. The effect of clause boundaries is included in the model, to evaluate whether IUs have a privileged status as a chunking mechanism in spoken language.

3.2 Methods

The recall data for this study is taken from the experiment described in Chapter 2. See Chapter 2 for a full description of the experimental methodology and materials.

The IU boundary information was taken from the IU annotations for the stimuli in the Santa Barbara Corpus of Spoken American English (SBC; DuBois et al., 2000-2005). Clause boundaries were defined as the S nodes from the Stanford parsed version of the stimuli described in Chapter 2. Though the clause boundaries identified by the Stanford parser are not always correct, based on spot-checking by the author it has overall very good performance, in addition to being reproducible and representing a traditional views syntax (including its bias towards written language).

3.2.1 PMI

A measure of pre-existing association strength between bigrams, pointwise mutual information (PMI), was calculated from the SBC. The PMI for a pair of words x and y expresses the expected probability of their joint occurrence, normalized by their individual probabilities. PMI is related to mutual information (MI), which evaluates PMI for all potential values of two variables X and Y. For example, the MI of bigrams in a dataset would be calculated for all words appearing in first position and all words appearing in second position, resulting in a measure of how predictive first and second words in a bigram are of each other. PMI is defined as the log of the joint probability of x and y divided by the product of the probabilities of x and y. The equation for PMI is given in 3.1 below.

$$pmi = \log_e \frac{p(x, y)}{p(x)p(y)} \tag{3.1}$$

To calculate the PMI, the full SBC was treated as a single vector of words. The part-of-speech tags from the full Stanford-parsed SBC (parser v. 3.5.1) were used along with the WordNet lemmatizer implemented in the Natural Language Toolkit (NTLK) to change nouns and verbs with inflectional affixes (e.g. *churches*, *walked*) to their base forms (e.g. *church*, *walk*). This 'lemmatization' step ensured that all inflected and base forms of the same noun or verb would be treated as the same word when constructing bigrams and calculating their frequency. The bigrams and unigrams and their frequency distribution were derived from this vector, and the PMI calculated from that. The resulting data frame of bigram and PMI values was used as a lookup table for the (lemmatized) bigrams in the recall experiment stimuli.

3.2.2 Overlap of IU and clause boundaries

As mentioned in Chapter 1, IU boundaries often align with clause boundaries, about 60% in the analysis of English conversational data from Chafe (1994). Too much overlap between IU and clause boundaries could be problematic for this analysis, as I am aiming to provide a comparison between the two units. However, due to the careful selection of recall experiment stimuli for variation in IU, clause, and word counts, as described in Chapter 2, there are a significant number of non-aligned IU and clause boundaries in the dataset. The distribution of boundary types for the 570 bigrams spanning a boundary is divided roughly into thirds, with 188 (33%) containing only a clause boundary, 223 (39%) only an IU boundary, and the remaining 159 (28%) containing an overlapped IU/clause boundary. The breakdown of distinct and overlapping boundaries is shown in Figure 3.1.

3.3 Results

A binary logistic mixed effects regression was used to model the effects of IU and clause boundaries on matching of recall status for bigrams in the experiment stimuli. The dependent variable indicated whether the recall status for a bigram was matching (both words correct or both incorrect), or not matching (one incorrect, one correct). The following fixed effect predictors were included: two binary categorical predictors, IU Boundary and Clause Boundary, indicating whether the bigram crossed a boundary (=yes) or did not cross a boundary (=no), the interaction of IU and Clause Boundary, Word Count (number of words in the stimulus), and Word Position (relative position in the stimulus). Random slopes and intercepts for stimulus and participant identity were also included to control for individual variation along those dimensions.

Word Count and Word Position were included as controls, due to the expectation that certain value ranges for these factors would result in very good or very bad recall

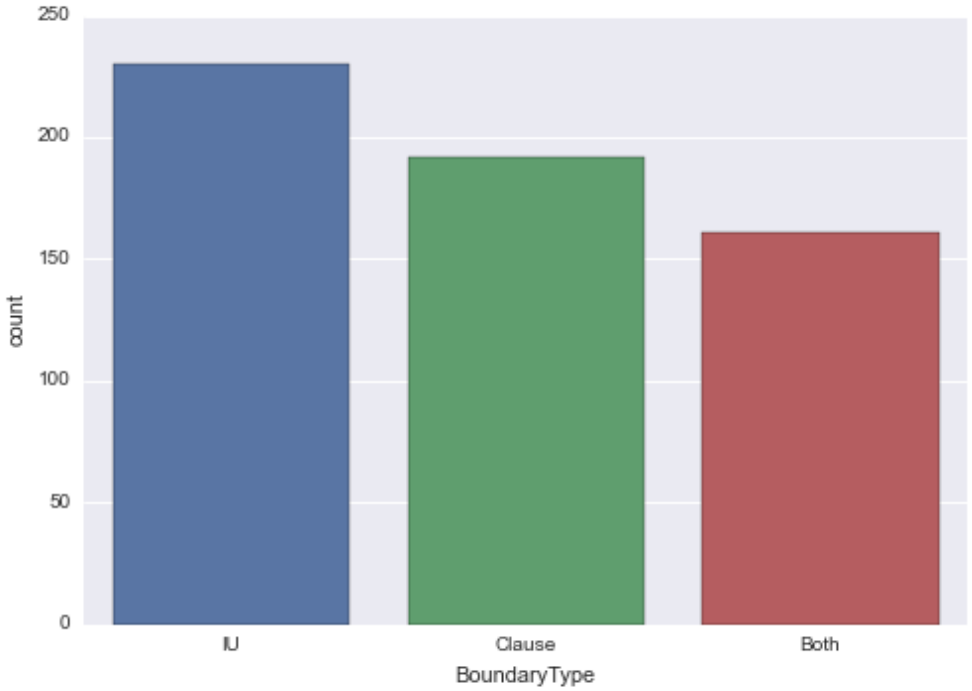

Figure 3.1: Frequency counts for IU-only, clause-only, and shared IU/clause boundaries in the stimuli

performance, inflating the number of matching bigrams. For example, low word count stimuli are likely to be recalled perfectly, in which case all bigrams would match in their recall status. Word Position was calculated as a relative value by dividing the bigram's raw word position value (the mean of the two words' positions) by the total word count. The resulting value was then subtracted from 1 to get percentages in the intuitive direction (small values = early in the stimulus, large values = later in the stimulus). Therefore, Word Position indicated the percentage of words in the stimulus that had not yet been heard at the point the current bigram appeared. The logarithm of both Word Position and Word Count were used in the model to normalize the scale of predictor values.

Model selection on the model was performed using p-values. The p-value of Word

Count was not significant (p >0.05) so it was discarded from the final model. This is likely because Word Count contains two opposite-polarity predictive relationships for Matching:yes - low values of Word Count would predict more correct bigrams (negative correlation with Matching), and high values of Word Count would predict more incorrect bigrams (positive correlation with Matching). So it is likely these effects canceled each other out. All other predictors had significant p-values.

The summary statistics for the final model are displayed in Table 3.3, and the fixed effects plotted in 3.3. The marginal R^2 value, estimating the amount of variance in the data explained by the fixed effects in the model, is 0.023, and conditional R^2, estimating the amount of variance explained by the fixed and random effects, is 0.093, indicating that the model explained about 2% of the variance without random effects, and about 9% of the variance with random effects included. Classification accuracy was high (0.819), but only .01% better than the baseline of assigning all bigrams to the most frequent class (Matching: yes). The C-statistic, also called concordance or AUC (area under the curve), a measure of classification accuracy that takes class frequency into account, was 0.64. Values for C around 0.8 or higher are generally considered good (Gries, 2013). Overall, R^2 was very poor and classification accuracy fairly poor, indicating that the model did not capture most of the variability in the dataset. However, my aim was not to create a model which represented all factors involved in determining whether a sequence of two words is remembered/forgotten. The aim was to evaluate the significance of the IU and Clause boundary effects, along with controls for known confounds, so the significance values and shapes of the functions are the important result.

Table 3.3 shows that PMI was a significant predictor, with a higher PMI value corresponding to a higher likelihood of Matching:yes (both words correct/incorrect). This validates the fundamental assumption of this study, that higher bigram association strength

Table 3.1: Summary statistics for GLMM predicting matching recall status for bigrams

Effect	Estimate	Std. Error	z-value	P-value
Intercept	1.482452	0.068453	21.657	<2e-16***
IU Boundary	-0.549458	0.017746	-30.963	<2e-16***
Clause Boundary	-0.184442	0.020659	-8.928	<2e-16***
log(Word Position)	-0.023534	0.006279	-3.748	0.000178***
pointwise MI	0.069096	0.003499	19.748	<2e-16 ***
IU Boundary:Clause Boundary	0.099563	0.031787	3.132	0.001735**

should result in higher likelihood of matching recall status[1]

We see in Table 3.3 that both IU and Clause Boundary are significant predictors, along with their interaction. The estimated coefficient value for IU boundary is stronger than that of Clause Boundary (-0.549458 vs. -0.184442). Although spanning a clause boundary (ClauseBoundary:yes) does lower the predicted likelihood of Matching:yes, it appears that spanning an IU boundary has a stronger effect. Bigrams with a clause boundary but no IU boundary (Figure 3.3c, left side of right panel) still have much higher predicted likelihood of Matching:yes than bigrams with an IU boundary but no clause boundary (Figure 3.3c, right side of left panel).

3.3.1 Both-correct model

As discussed in Chapter 2, the algorithm used to score the recall experiment responses strictly penalized out-of-order recall, and therefore the set of words with incorrect scores include words that were recalled correctly but in a different order. Based on the percentage of words marked incorrect by the algorithm that were marked correct by human scorers, about 44% of incorrect words represent correct out-of-order recall. Manual scoring was performed almost exclusively on the medium and low IU count stimuli, and the generally poor performance on the high IU stimuli would be expected to increase the pro-

[1]PMI as a measure of association strength may be criticized for its over-valuing of very low-frequency bigrams. The same model was run with the addition of t-score, an alternative measure that takes into account frequency of the bigram as well as strength, and there was no significant change to the results.

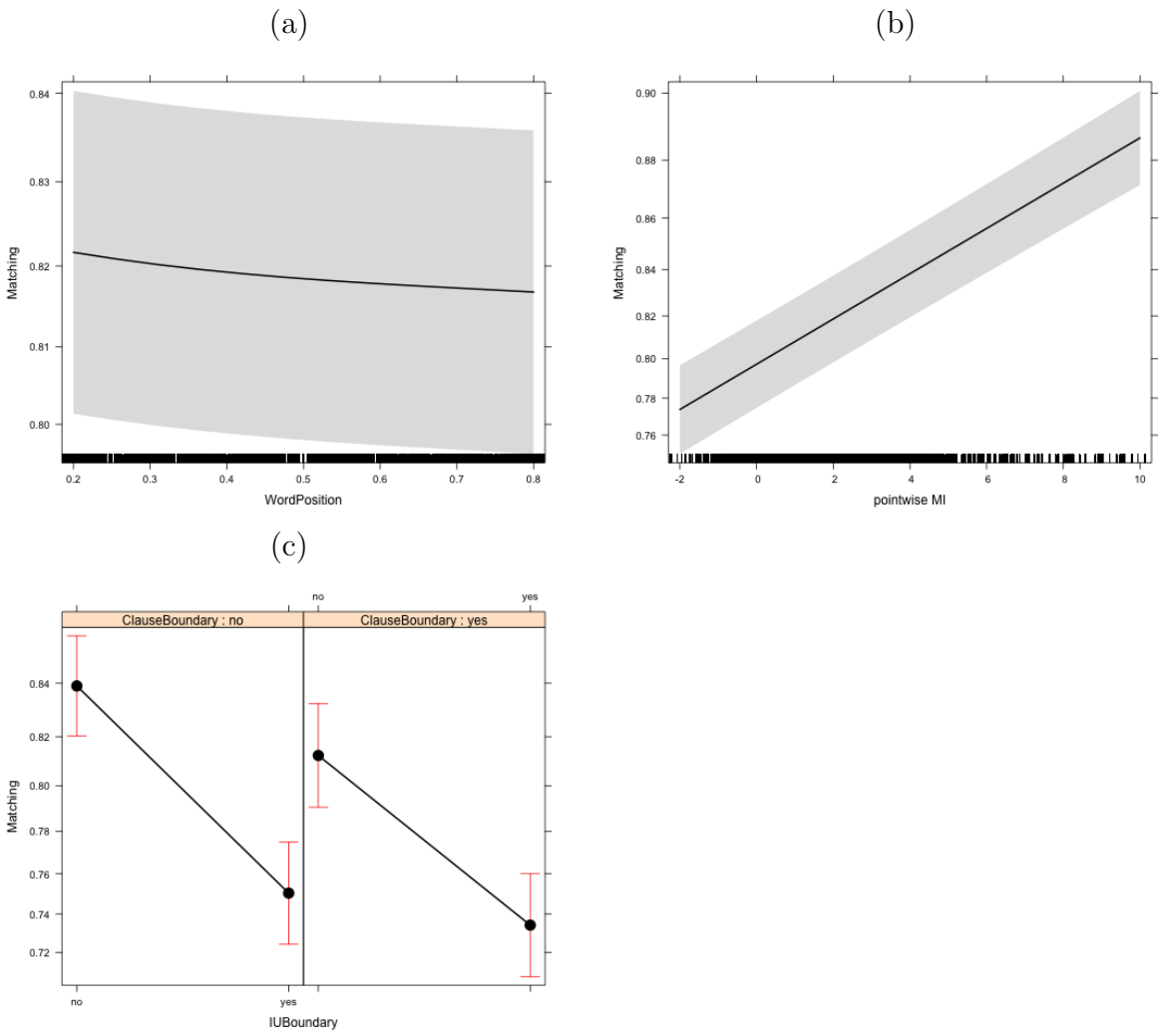

Figure 3.2: Significant fixed effects from GLMM model of bigram recall status (matching vs. non-matching) with y-axis representing likelihood of matching recall status as a function of (a) the relative position of the bigram in the stimulus (smaller values = earlier in the stimulus), (b) pointwise mutual information for the words in the bigram, (c) the interaction between clause and IU boundaries

portion of completely incorrect compared to out-of-order words, but clearly, out-of-order words represent a non-trivial proportion of the incorrect scores. Therefore, it is likely that some number of the both-incorrect bigrams represent cases in which one word was forgotten and one was remembered but in the wrong position. If there are many cases like this, it could potentially be problematic for the model because this type of bigram

is not actually representing a memory unit.

To make sure that this issue did not seriously affect the model, I ran another generalized linear mixed effects regression with the same fixed and random effects on a subset of the data with only the both-correct bigrams vs. non-matching bigrams. Model selection did not exclude any predictors, as all fixed effects were significant, including Word Count. The results of this model are summarized in Table 3.3.1, and the effects plotted in Figure 3.3.1.

Table 3.2: Summary statistics for both-correct GLMM predicting matching recall status

Effect	Estimate	Std. Error	z-value	P-value
Intercept	1.506949	0.161825	9.31	<2e-16***
IU Boundary	-1.080961	0.026700	-40.49	<2e-16***
Clause Boundary	-0.291487	0.027429	-10.63	<2e-16***
WCount	-0.029620	0.003476	8.52	<2e-16***
log(Word Position)	-0.017665	0.007674	-2.30	0.021344*
pointwise MI	0.102450	0.007674 x	22.27	<2e-16 ***
IU Boundary:Clause Boundary	0.176672	0.004601	3.76	0.000172***

Comparing Figure 3.3.1 with Figure 3.3, the functions for Word Position, PMI, and the interaction of IU and Clause boundary are nearly identical to that of the model run on the full dataset. The estimates and relative p-values in 3.3.1 are similar to that of the full model as well. The biggest difference between the two models is that Word Count is now a (highly) significant predictor, which is unsurprising since it now simply represents a positive correlation with correct responses.

The marginal R^2 value for the correct-only model is 0.179, and conditional R^2 is 0.290, indicating that the model explained about 18% of the variance without random effects, and 29% of the variance with random effects included, much more than for the all data model. Classification accuracy was lower than for the all data model (0.675), but the correct-only model performed much better in comparison to the baseline (13% higher). In line with this improvement compared to the all data model, the C-statistic

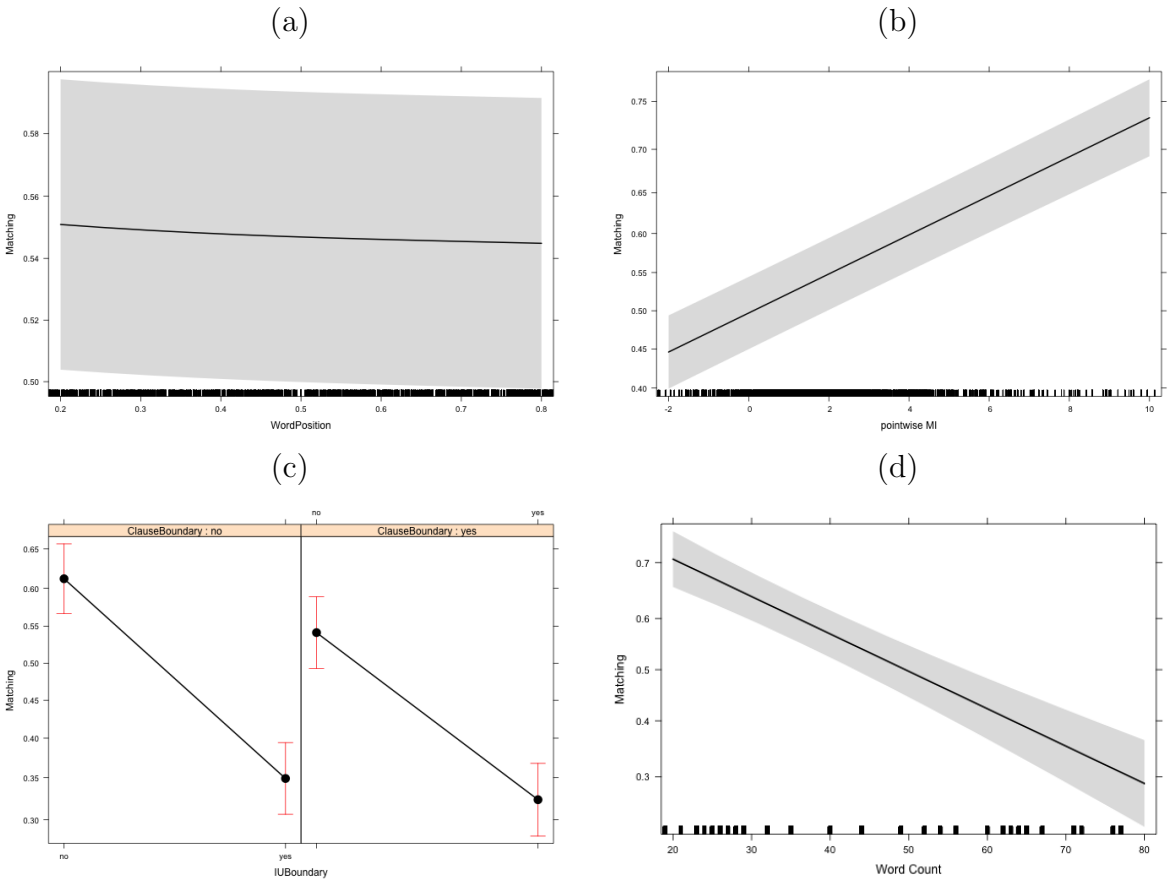

Figure 3.3: Significant fixed effects from both-correct GLMM model of bigram recall status (matching vs. non-matching) with y-axis representing likelihood of matching recall status as a function of (a) the relative position of the bigram in the stimulus (smaller values = earlier in the stimulus), (b) pointwise mutual information for the words in the bigram, (c) the interaction between clause and IU boundaries, (d) the number of words in the stimulus

is much higher for the correct-only model (0.74).

The improved metrics for the correct-only model reflect the lower variability in the comparison of non-matching with only both-correct bigrams. The effects of interest (PMI, IU and Clause Boundary) are nearly identical in function shape and significance in this model, supporting them as true representations of the relationship of those predictors with the data.

3.3.2 PMI analysis

The inclusion of PMI in the model was intended to account for effects of pre-existing associations between words in the stimuli due to collocation strength. If not included in the model, any correlation of IU and clause boundaries with pre-existing association strength would affect the results. As mentioned in section 3.1, (Wahl, 2015) analyzed the relationship of IU boundaries and bigram association strength, using several different measures of collocation strength. He found that IU boundaries had a significant correlation with collocation strength, tending not to break strongly-associated bigrams. If IUs have a stronger tendency than clauses to avoid breaking words with strong pre-existing associations, then the differential effects of IU and clause boundaries could mean that there was some remaining correlation with association strength that was not sufficiently captured by the PMI measure used.

To explore whether the current measure of PMI indicates a difference between clauses and IUs, I fit a linear mixed effects regression model predicting PMI for the bigrams in the recall experiment stimuli. The fixed effects included were IU Boundary, Clause Boundary, and their interaction. Random slopes and intercepts were fit for stimulus identity. All fixed effects were found to be significant. The interaction between IU and Clause boundary is plotted in Figure 3.4 below, and summary statistics provided in Table 3.3.2. The marginal R^2 value for this model is 0.05, and the conditional R^2 value is 0.062. These values are very low, which is expected given that IU and Clause boundaries account for only a small portion of the factors involved with bigram association strength.

Table 3.3: Summary of PMI model, p-values obtained through a type 3 Analysis of Variance with Satterthwaite approximation for degrees of freedom

Effect	Sum Sq	Mean Sq	NumDF	DenDF	F.value	P-value
IUBoundary	67.027	67.027	1	1941.3	19.414	1.110e-05 ***
ClauseBoundary	64.724	64.724	1	1941.8	18.747	1.569e-05 ***
IUBoundary:ClauseBoundary	34.533	34.533	1	1941.1	10.002	0.001588 **

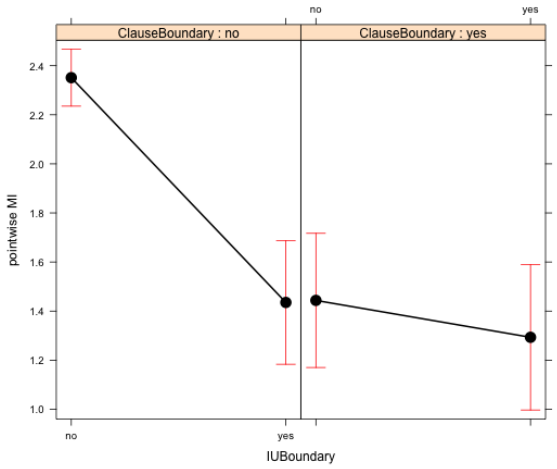

Figure 3.4: Significant interaction from linear mixed effects model predicting PMI of bigrams

Interestingly, predicted PMI seems to behave similarly for bigrams containing all three boundary types (IU and clause, IU only, clause only). Figure 3.4 shows nearly identical predicted PMI for IU-only boundaries (right side of left panel) and clause-only boundaries (left side of right panel), and only a slight drop when the two boundaries coincide. The only significant difference in predicted PMI is for bigrams that do not cross either type of boundary (left side of left panel), where predicted PMI is much higher. This result is alignment with the findings of Wahl (2015) using several different measures of bigram association within and across IUs in the full SBC. He found that weakly-associated bigrams did not have a strong correlation with IU boundary (i.e. they were fairly equally likely to cross or not cross a boundary), but strongly-associated bigrams were very likely to be within an IU. In other words, IUs were very unlikely to 'break' strongly-associated bigrams. Wahl (2015) did not investigate the effect of clause boundaries, but based on the current analysis, the results should be similar.

3.4 Discussion

The results of this study support the hypothesis that IUs create new chunks at presentation, but they do not rule out clause-based effects as conclusively as the results in Chapter 2. It was found that words within IU boundaries are more strongly associated in memory than across IU boundaries, even when accounting for pre-existing associations, and the same is true for clauses. The fact that clauses and IUs both had significant effects on recall status in addition to the effect of PMI is strong evidence that both of these units have created new associations in memory during processing.

However, IUs seem to have a stronger chunking effect than clauses. Clause boundary had a smaller coefficient and z-score in the model summarized in Table 3.3, and the interaction between clause boundary and IU boundary showed that the presence of a clause boundary without an IU boundary had a much weaker effect on the likelihood of matching recall status than the reverse situation (presence of an IU boundary only). The post-hoc analysis of PMI presented in Figure 3.4 does not support attributing the stronger effect of IUs to a stronger correlation with pre-existing bigram association, as predicted PMI did not differ between bigrams with IU-only and clause-only boundaries. Therefore, the difference in strength between IU and clause boundary predictors in the model summarized in Table 3.3 should be attributable to the strength of newly-created inter-word associations.

Since IUs have a stronger chunking effect than clauses, overall this study continues to support the role of IUs as the primary chunking mechanism in spoken language processing.

Chapter 4

Lexico-Syntactic Priming between Intonation Units in Spoken English

4.1 Goals of Study

Repetition priming is a well-known effect by which production or comprehension of a linguistic form facilitates subsequent processing of that form, or influences a speaker to produce that form. Repetition priming effects have been found for both syntactic structures and lexical items, in studies of isolated word lists (Scarborough, Cortese, & Scarborough, 1977; Forster & Davis, 1984; McKone, 1995), experimentally-constrained sentence production (Bock, 1986; Bock & Griffin, 2000), and transcribed dialogues (Szmrecsanyi, 2005; Gries, 2005; Reitter, Moore, & Keller, 2006; Pietsch et al., 2012). Although measures of repetition priming duration have not always yielded consistent results, even within the same experimental paradigm (e.g. Bock & Griffin, 2000), one commonly-found pattern is exponential or logarithmic decay of priming over distance (Gries, 2005; Pietsch et al., 2012; Moscoso del Prado Martin, 2015), which can be broken down into two components: a strong, short-lived effect and a weaker long-term effect (Levelt & Kelter, 1982; McKone, 1995; Hartsuiker et al., 2008; Reitter et al., 2011). The strong short-lived priming effect is often attributed to a short-term or working memory process, such as time or interference-based decay of activation (Bock & Griffin, 2000; Hartsuiker et al., 2008; Pietsch et al., 2012), or storage in a temporary buffer (in combination with a decay process) (Reitter et al., 2011). The longer-term priming effect is generally considered to be a form of implicit learning, whereby representations are strengthened through repeated activation (Bock & Griffin, 2000; Chang, Dell, & Bock, 2006; Reitter et al., 2011).

Cowan (2000) argues that item-based STM capacity limits can be observed in the duration of short-term priming effects, citing the findings of McKone (1995) as support. McKone (1995) employed a lexical decision task in which participants provided a yes/no response to a stimulus, indicating whether it was an English word or not. The stimulus

set consisted of English words and phonotactically legal non-words. Repetition priming can be observed in lexical decision tasks as a significant decrease in reaction time for repeated words compared to new words. The lists of words in McKone's study contained repeated items at various lags, following 0, 1, 2, 3, 4, 5, 9, 23, and 1,050 intervening items. She observed a strong priming effect which decayed rapidly across lags of 0-4 intervening items, with an asymptote around lag 5. She concluded that the best-fitting function to describe this effect was an exponential decay function, which for low-frequency words (Experiment 1) was composed of a short-term effect overlaid on a steady long-term priming value. Cowan (2000) argues that the duration of the short-term strong priming effect, from 0-4 intervening items, provides support for his 3-5 item STM capacity limit. He does not provide an explicit discussion of the mechanism behind the short-term priming effect, but presumably it is an activation-based explanation, with the limitation on activation being the number of items rather than time or interference.

Reitter et al. (2011) propose a more clearly specified mechanism explaining short-term priming effects. They posit that semantic and lexical material (and potentially syntactic material as well) is stored temporarily in a memory buffer to allow integration of meaning and structure. In their model, spreading activation from the material in this buffer is the cause of strong short-term priming effects on syntactic structure. It creates the short-lived 'lexical boost' effect on syntactic priming, an effect documented by numerous researchers (Pickering & Branigan, 1998; Gries, 2005; Hartsuiker et al., 2008) in which repetition at the lexical level seems to 'boost' repetition at the syntactic level. Reitter and Moore (2014) explicitly identify the buffer in their model as being equivalent to short-term memory.

The results in the previous two chapters support the hypothesis that the Intonation Unit (IU) is an important processing unit in spoken language that defines the contents of short-term memory in comprehension, and presumably also in production although that

was not explicitly tested. If the duration of the strong short-term priming effect reflects the capacity of short-term memory, then the IU should be the relevant linguistic unit defining priming duration in spoken language. The analysis in this chapter will explore this hypothesis by measuring combined lexical and syntactic repetition priming across IUs in the SBC. Following the measure employed in Moscoso del Prado Martin (2015), I use the amount of shared Shannon information (Shannon, 1948) across pairs of IUs as the measure of priming strength.

It would be impossible to directly pit the IU against the clause to predict shared information values in the same model in the way that was done in Chapter 2, because shared information is calculated with respect to the IU unit, i.e. for the information contained in a pair of IUs, so it would not be the same for clauses. However, I was able to compare the effect of distance in IUs with another proposed source of STM limits, distance in time, by including the distance in seconds between each pair of IUs as a variable.

Decay of activation over time is the proposed mechanism for STM limits in the traditional multi-store memory model (Baddeley & Hitch, 1974; Baddeley, 1986), and is frequently assumed to be the cause of priming decay (e.g. Branigan, Pickering, & Cleland, 1999; Reitter et al., 2011). This explanation is appealingly intuitive, but has been criticized as lacking an actual mechanism to explain the decay. As pointed out by McGeoch (1932) in his classic analogy, rust accumulates on an iron bar over time, but the mechanism responsible is oxidation, not time itself. There may be some actual mechanisms of time-based decay such as neuronal fatigue, but a remaining related issue is how a time-based decay process can be reliably distinguished from interference-based decay, the alternative explanation provided for decay in many models of STM (e.g. Cowan, 2000; Nairne, 2002; Jonides et al., 2008). Models supposing no privileged STM capacity, e.g. Crowder (1993); Nairne (2002) assume that a constant process of gradual similarity-based

interference is the main factor responsible for limitations in STM. Models supposing a capacity-limited STM equivalent to the focus of attention, e.g. (Cowan, 2000; McElree, 2001), combine gradual interference with the sudden replacement of the contents of the focus of attention.

New information is continually received by the brain from external sensory input as well as self-generated content. Neuronal states continually change over time. Therefore, as time passes, there will necessarily be interference. As Jonides et al. (2008) conclude in their review of current STM theories, it may simply not be feasible to distinguish time-based decay from all types of interference. In the current study, the inclusion of distance in IUs can be assumed to reflect interference, but the distance in time variable may reflect either strictly time-based decay, or the gradual accumulation of other sources of interference.

I argue that if the IU is serving to segment spoken language into chunks for processing in STM, distance measured in number of IUs should be a significant predictor of priming strength between a pair of IUs, even with distance in time included in the model. Additionally, just as for recall effects in Chapter 2, I predict that the function describing the effect of distance in IUs on priming should strongly resemble the function usually found for distance measured in number of words in priming tasks employing isolated words (e.g. McKone, 1995), and other predictors should not have the same function shape. Therefore, this study will evaluate the following three predictions in support of the IU as a chunking mechanism in STM: 1) the distance between IUs will have a significant effect on priming, 2) priming will exhibit a non-linear decay function over distance between IUs, specifically a sharp decrease with an asymptote at a distance of about 3-5 IUs, mirroring the word-level priming function in McKone (1995), 3) distance in time should not have a significant non-linear effect on priming.

4.2 Methods

4.2.1 Shared Shannon Information

As in Moscoso del Prado Martin (2015), a measure of shared information was calculated using the set of phrase-structure production rules induced from syntactic parse trees created by the Stanford parser (version 3.5.1). In order to obtain production rules for each IU, the entire Santa Barbara Corpus of Spoken American English (SBC, DuBois et al., 2000-2005) was syntactically parsed using the Stanford parser. Each IU was treated as the root for the parse, meaning that no parse tree could span across multiple IUs. Figure 4.2.1 shows an example of a tree created by the Stanford parser for one IU in the SBC. As discussed in section 1.4, high-level syntactic units may span multiple IUs, however since we are comparing the content between IUs, it was necessary to use only the production rules that were contained within IUs. Using the IU as the root was the most straightforward way to accomplish the separation of production rules at IU boundaries.

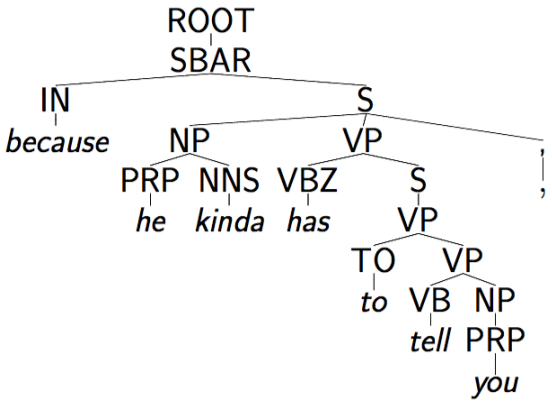

Figure 4.1: Stanford parse tree for one IU (SBC filename: SBC001)

Production rules express the hierarchical combination of syntactic units observed in some set of parse trees. They are generally represented formally as a series of symbols separated by an arrow (e.g. S $\rightarrow$ NP VP). A symbol on the left-hand side of the rule

indicates a higher-level unit, and the right side indicates the lower-level unit(s) contained in that unit. The right-hand side of a production rule can contain non-terminal nodes, right-hand sides that are themselves the left-hand side of another rule (e.g. PP → P NP), or terminal nodes (e.g. P → *into*). Shared information can be calculated separately by dividing production rules into rules with terminal right-hand sides (lexical), and rules with non-terminal right-hand sides (syntactic). Table 4.1 illustrates this difference with the list of lexical and syntactic production rules for the parse tree in Figure 4.2.1.

Table 4.1: Phrase-structure production rules derived from the tree in Figure 4.2.1

Lexical			**Syntactic**		
IN	→	*because*	ROOT	→	SBAR
PRP	→	*he*	SBAR	→	IN S
NNS	→	*kinda*	S	→	NP VP
VBZ	→	*has*	NP	→	PRP NNS
TO	→	*to*	VP	→	VBZ S
VB	→	*tell*	S	→	VP
PRP	→	*you*	VP	→	TO VP
			VP	→	VB NP
			NP	→	PRP

In this chapter I use a combined lexical and syntactic information measure. For each pair of IUs, the Shannon information (Shannon, 1948) was calculated for the set of all production rules derived from the Stanford parse trees for both IUs, excluding production rules appearing in the parse trees for the intervening IUs. The exclusion of shared rules from intervening IUs is important to ensure that the measure of priming at a particular distance is not confounded by more recent priming from an intervening IU (Moscoso del Prado Martin, 2015).

Shannon information, also known as 'self-information' or 'surprisal', quantifies the unexpectedness, and hence, informativeness, of the occurrence of a particular outcome. Highly unexpected outcomes have a high self-information value, capturing the intuition

that an occurrence of a very frequent outcome (e.g. the appearance of the word *the* in an English text), does not provide as much information as an occurrence of an infrequent outcome (e.g. the appearance of the word *sear*). The knowledge that *the* appeared in a particular English sentence tells a reader almost nothing about the rest of the sentence, whereas the knowledge that *sear* appeared would allow them to come up with some very specific guesses as to what that sentence is about (e.g. cooking of some kind of meat or fish).

The equation used here to calculate the estimated self-information of the occurrence of a particular production rule r is shown in 4.1. The value $\hat{p}(r)$ represents the observed probability in the corpus that the right-hand side of r is the expansion for its left-hand side symbol n. Therefore, the equation expresses that self-information for a production rule r is estimated as the base e logarithm of the observed probability of r.

$$\hat{I}(r) = -log_e \hat{p}(r) \tag{4.1}$$

The value for $\hat{p}(r)$ is calculated using the equation in 4.2.

$$\hat{p}(r) = \frac{f(r)}{f(n)} \tag{4.2}$$

This equation expresses that the observed probability $\hat{p}$ of a rule r is equal to the number of times the right-hand side of r occurs as the expansion of its left-hand-side symbol n, divided by the total number of times that n occurs as a left-hand-side in the list of production rules for the corpus. For example, if NP occurs 100 times in the corpus, and of those 100 instances, 99 are expanded as DT NN, and one is expanded as NN

PP, then $\hat{p}(\text{NP} \rightarrow \text{DT NN}) = \frac{99}{100} = .99$ or 99%, and its self-information $\hat{I}$ would be equal to $-log_e(.99) = .01$. The self-information for the single instance of NP $\rightarrow$ NN PP, however, would be much larger, equal to $-log_e(.01) = 4.61$.

Again following Moscoso del Prado Martin (2015), for each pair of IUs, self-information was calculated for the list of production rules shared between the pair and *not* with any intervening IUs. Self-information for a list of production rules was calculated simply as the sum of the $\hat{I}$ values for each rule instance.

4.2.2 Random Baseline

As in Moscoso del Prado Martin (2015), the shared information measure was corrected using a random baseline measure. This correction is intended to account for the amount of shared information that may happen purely by chance. The random baseline amount of shared information was derived by randomly sampling two trees from a probabilistic context-free grammar (PCFG) generated from the parsed SBC. A PCFG assigns probability weights to right-hand-side expansions of nodes based on their observed frequencies in a list of production rules. Pseudo-random trees can then be generated from the PCFG by starting at the root and iteratively choosing right-side expansions based on those probability weights until terminal nodes are reached. For each IU in the dataset, a tree was generated in this manner. The shared self-information measure described above was applied to the random-sample counterparts of each of the IU pairs, giving each pair its own random baseline.

4.2.3 Normalization

Both normal and random baseline shared information values were then normalized by the total amount of information in the IU pair. This was done to control for biasing

effects of IU size on the amount of shared information. IUs are often in the range of 3-5 words in English (Chafe, 1994), but can vary widely in size, as mentioned in 1.4, from a single word to multiple clauses, and naturally there would be more possibility for shared information in pairs containing large IUs.

4.2.4 Final Shared Information Measure

The final equation for total shared information between IU pairs is represented in 4.3.

$$\hat{I}_{correctedshared} = \frac{\hat{I}_{shared}}{\hat{I}_{total}} - \frac{\hat{I}_{sharedRandom}}{\hat{I}_{totalRandom}} \tag{4.3}$$

I will simply use $\hat{I}_{total}$ to refer to this value in the remainder of this chapter.

4.3 Results

Moscoso del Prado Martin (2015) did not find significant priming at distances of more than ten intervening 'phrasal units'; therefore shared information was only calculated for IUs across distance D $\leq$ 10. The full SBC corpus dataset contains 638,650 IU pairs for D = 1-10. Ten of the sixty files in the SBC corpus were excluded from analysis since they were monologues (e.g. lectures, sermons, speeches) rather than interactions[1], leaving 563,460 IU pairs. A histogram of the values for distance in time between IUs (DeltaT) in this dataset revealed a small number of extreme outliers, possibly due to errors in the timestamps recorded in the SBC. Therefore, the IU pairs with DeltaT values in the lower and upper .5% (n = 5591) were removed from the dataset (lower: <-0.92 s, upper: >28.14 s), yielding a total of 557,869 IU pairs in the final dataset.

[1]See http://www.linguistics.ucsb.edu/research/santa-barbara-corpus for file descriptions

A mixed-effects linear regression was run on this dataset, with the corrected normalized shared information measure for each IU pair as the dependent variable. The following fixed effect predictors were included: Speakers, a binary factor indicating whether the two IUs were spoken by the same speaker (Speakers = same) or different speakers (Speakers = different), distance between the pair in time (DeltaT, measured in seconds) and its quadratic and cubic polynomials (DeltaT^2 and DeltaT^3), and distance between the pair in number of IUs (D). Distance in IUs was included as an ordered factor, treated in the model as a set of orthogonal polynomials. The two-way interactions between Speakers and each of the two distance measures (DeltaT and D) were also included.

The polynomial transformations for DeltaT and D were included to test the fit of non-linear functions to the data. The transformation of distance in IUs into a set of orthogonal polynomials (up to D^9) means that for each distance value, at least one function was fit that allowed a curve at that value, with the highest polynomial (D^9) being a function where there is a curve fit at every level of D. Polynomials were used rather than a GAMM model, because the added complexity of the interactions between Speakers and the two distance measures made the GAMM model fit unreliable.

Random slopes and intercepts were included in the model for Conversation (SBC file id) and Speaker Name (the speaker name for the second IU in each pair) to control for individual variation along those dimensions. Speaker Name was nested in Conversation, because each speaker appeared in one file only.

The results of the model are summarized in Table 4.2, and the significant effects plotted in Figure 4.2. All fixed effects (same/different speakers, distance in time, distance in IUs, and interaction of speakers with distance measures) were found to be highly significant. Marginal R^2 for the model was 0.009, and conditional R^2 was 0.010, meaning that the model is estimated to account for about 1% of the variance in the data. This is very low, indicating that there are many factors beyond the predictors of interest in this

Effect	Sum Sq	Mean Sq	NumDF	DenDF	F.value	P-value
Speakers	15.8199	15.8199	1	130751	1102.14	<2.2e-16 ***
DeltaT	3.6041	1.2014	3	96182	83.70	<2.2e-16 ***
Distance in IUs	3.3503	0.3723	9	375548	25.93	<2.2e-16 ***
Speakers:DeltaT	0.2423	0.0808	3	526901	5.63	0.0007469 ***
Speakers:IUs	0.5743	0.0638	9	554043	4.45	7.564e-06 ***

Table 4.2: Summary of Total Shared Information model, p-values obtained through a type 3 Analysis of Variance with Satterthwaite approximation for degrees of freedom

study at play in shaping the amount of shared information between a given pair of IUs.

The significant effects are plotted in Figure 4.2.

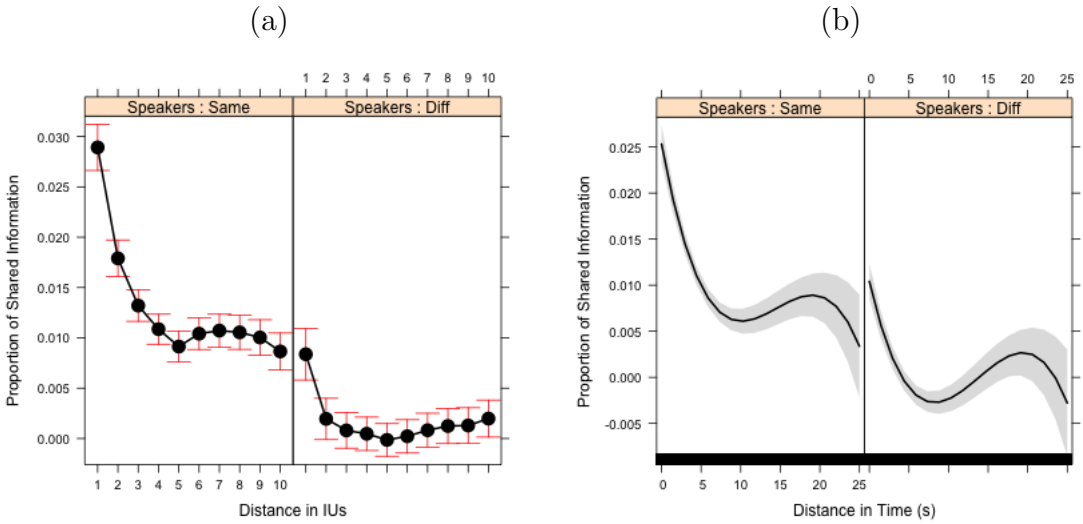

Figure 4.2: Plot of effects on Total Shared Information: Interaction of Speakers with (a) Distance in IUs and (b) Distance in time

Looking further into the polynomial effects using lmerTest's model summary function, we find that distance in time (DeltaT) is highly significant (p<.001) as a linear, quadratic, and cubic polynomial function. This indicates that the effect of DeltaT on shared information is well described by the cubic polynomial (i.e. a non-linear) function. The interaction of DeltaT with Speakers is significant, but only for the linear function. All three DeltaT polynomial levels and their linear and quadratic polynomial interac-

tions with Speakers have larger effect size estimates than any other predictor (>|1|, all others <|.02|). Distance in IUs is highly significant at the linear, quadratic, and cubic polynomial as both a main effect and in interaction with Speakers. Distance in IUs is also significant (p = 0.04) at the quartic (4th) polynomial as a main effect. This indicates that the function was changing significantly up until a distance of 4 IUs, which is where we can clearly visually see the asymptote for Speakers:Same in Figure 4.2.

Figure 4.2a demonstrates that combined lexical/syntactic priming is highest between adjacent IUs, with the highest $\hat{I}_{total}$ value predicted at D=1. The priming rate then drops dramatically and stabilizes quickly, at D=4 IUs for same-speaker pairs as was just mentioned, and at D=2 for different-speaker pairs. Within-speaker priming, i.e. self-priming, was found to be much stronger than cross-speaker priming, with the lowest predicted $\hat{I}_{total}$ values for Speakers:Same at around the same value as the highest $\hat{I}_{total}$ values for Speakers:Different. The duration of the strong priming effect is shorter for cross-speaker priming, as demonstrated by the faster asymptote (at D=2) compared to within-speaker priming (at D=4).

As seen in Figure 4.2b, lexical/syntactic priming exhibits a sharp decrease over time, which for both Speakers:Same and Speakers:Different reaches a low point between 5-10 seconds, followed by a smaller increase, peaking around 20 seconds, and then falling again. The significance of DeltaT to the 3rd polynomial indicates that these three segments of curvature are a good fit to the data, but the 95% confidence interval is much wider for the final two segments, indicating higher variability for values in those segments. It is unclear what might be causing the second peak, but its small size combined with wider variability makes it likely a result of overfitting to the dataset rather than a theoretically significant pattern. Unlike for distance in IUs, the shape of the DeltaT function is the same for both within-speaker and cross-speaker priming, but again within-speaker priming was found to be much stronger than cross-speaker priming, with the predicted maximum $\hat{I}_{total}$ values

for Speakers:Different in the low range of values for Speakers:Same.

4.4 Discussion

The results overall provide support for the importance of the IU in priming of spoken English. As predicted, the number of intervening IUs was found to have a significant, non-linear effect on the amount of shared information between a pair of IUs, that cannot be accounted for by the effect of decay over time. For within-speaker priming, there is a short-term priming effect that exhibits exponential decay and reaches asymptote levels at a distance of 4 IUs for shared total information, in a remarkably similar pattern to what McKone (1995) found for priming at the word level. Across speakers, the priming effect appears to decay at a similar rate, with the initial slope looking similarly steep, but it starts at a much lower value and reaches asymptote at a distance of only 2 IUs. This may be in part due to the extra load on STM of listeners planning their next utterance while speakers are still speaking. As Moscoso del Prado Martin (2015) and others have pointed out, second speakers must be doing some planning of their utterances during the final IUs of the previous speaker; otherwise they would not be able to achieve the median inter-turn interval of 0 ms documented for English and other languages by Stivers et al. (2009).

Contrary to the prediction in section 4.1, distance in time was also found to be a significant predictor of priming, and it was fit well by a non-linear function. This could be argued as providing support for a strictly time-based decay mechanism for priming, but there are some complications to this explanation. The duration of the short-term effect in Figure 4.2b is about 5 seconds, more than twice as long as the 2-second decay period for material in the 'phonological buffer' in the classic Baddeley model (Baddeley, 1986). As discussed in the introduction to this chapter, the time function may instead

reflect the gradual build-up of similarity-based interference, and could even include a separate function for word-level interference like that observed for recall in Chapter 2. In addition, other potential sources of interference (sensory input, self-generated thoughts) may be contributing to the decay function.

Though the findings in this study do not rule out the influence of other factors on priming, the important result in this chapter is that distance in IUs was found to be a significant predictor of priming strength, even with the effect of time and/or other sources of interference accounted for by its inclusion in the model. I am not aware of any other priming study that has directly compared a unit-based measure of priming with a time-based one, so it is possible that other units (e.g. clauses or sentences) would not be significant if time was included.

The predictive power of distance in IUs, and the fact that the shape of its function mirrors the word-level pattern found under conditions where words are likely to be the highest chunk level available, provides further evidence that IUs are processed as chunks. The particular shape of the function, with asymptote at around 4 IUs, matches the prediction of Cowan's (2000) capacity-limited STM theory, so can be taken as suggestive evidence for the IU as the unit defining the incremental contents of STM in spoken English.

Chapter 5

Conclusions

Chafe (1994) states that "Intonation units are hypothesized to be the linguistic expression of information that is, at first, active in the consciousness of the speaker and then, by the utterance of the intonation unit, in the consciousness of the listener" (p. 69). He asserts that speakers produce language in increments reflecting the incremental contents of their focus of attention, and those increments are IUs. I argue that for listeners, this indicates a crucial function for IUs in spoken language processing, that is, breaking up the continuous speech stream into chunks that can be processed efficiently within the limitations of STM. This dissertation presented the results of three studies investigating evidence for such a role of the Intonation Units in memory for spoken English.

Chapter 2 focused on the effect of IUs on memory span for spoken English clips of various sizes. It was found that the number of IUs had a strong effect on the likelihood of recall when the stimulus contained a small number of IUs (less than six), but if it contained a larger number, the number of IUs did not matter. Two other linguistic units tested, words and clauses, had a linear effect and no significant effect on recall, respectively. This result shows that number of IUs has a direct relationship with the duration of short-term high-performance recall, the portion of the recall decay function that many researchers consider to be a reflection of short-term memory limits (e.g. Cowan, 2000). Chapter 3 tested whether processing IUs created new chunks, by modeling the likelihood of a bigram pair to be remembered or forgotten as a unit. IUs may of course contain pre-existing chunks of strongly associated words, e.g. *I think so*, but the findings of Chapter 3 confirm the ability of IUs to create or strengthen associations between words over and above their pre-existing collocation-based associations, indicating that they are actually creating new chunks in the listener's memory. Chapter 4 investigated the role of IUs in determining short-term priming duration. Similar to recall, priming normally exhibits an initial high level that quickly decays over time and/or amount of intervening material, and then stabilizes at a lower long-term level. Also similar to recall, the

short-term ceiling-level component of the priming function has been linked to short-term memory by many researchers (Bock & Griffin, 2000; Cowan, 2000; Reitter et al., 2011). Priming levels were estimated in the Chapter 4 study using a measure of shared Shannon information applied to the lexical and syntactic information contained in grammatical parse trees of the IUs in the SBC. A linear mixed effects regression was used to predict amount of shared lexical and syntactic information shared across pairs of IUs at various distances. Two distance measures, time and distance in number of IUs, were included in the model. It was found that distance in IUs was a significant predictor of priming.

The remainder of this chapter discusses the comparative findings for IUs and clauses (section 5.1), the role of IUs in language processing (section 5.2), the role of IUs in language change (section 5.3), and the relationship of IUs and STM 5.4.

5.1 Role of clauses

Chapters 2 and 3 directly compared the effects of IUs and clauses. In Chapter 2, the discontinuous recall decay function observed in word list recall was found to appear only at the IU level in recall of connected speech, supporting the IU, and not the clause, as a chunking mechanism in spoken language. Chapter 3 offered a more complicated picture, with a significant effect found for both clauses and IUs on the association strength between bigrams. This result indicates that both clauses and IUs create chunks in spoken language during processing. However, the IU boundary effect in Chapter 3 was stronger than the clause boundary effect, as could be seen in the coefficient values as well as the behavior of the functions in their interaction.

Chunking is a general cognitive process (Newell, 1990) that occurs with all types of information and at various hierarchical levels. In addition, chunks are defined by their strength of intra-item association relative to items around them, and as such it should

be assumed that 'chunkhood' is inherently gradient. This dissertation is attempting to identify the linguistic feature responsible for a very important type of chunking in spoken language, that has a direct relationship with incremental focusing of information in short-term memory. The findings that clause-based chunking is weaker than IU-based chunking, and that number of clause units does not have a predictive relationship with STM span, as shown in Chapter 2, cast serious doubt on their suitability for this type of chunking.

Another source of evidence to support IUs over clauses is from processing speed. If listeners are relying on continuous real-time chunking of spoken language to deal with STM constraints, the relevant linguistic feature would need to be available during the initial stages of processing. The recall data used here comes from participant responses after hearing an entire stimulus, so it cannot be used to identify the point in processing at which the association is added, but prosodic grouping is generally thought to occur at the earliest stages of processing (e.g. Slowiaczek, 1981; Schafer, 1997; supported by the eye-tracking results of Snedeker, 2003 and Kraljic & Brennan, 2005). This would again point to the IU as the more likely candidate for a linguistic feature that carves up spoken language in service of the listener's limited STM.

Chapter 3 indicates that clauses do create chunks of spoken language during processing, but this could be a type of chunking that follows after prosodic processing, such as from an explicit clause unit identified during syntactic processing (as in sentence processing models like Schafer, 1997) or from processing semantic relationships encoded at the clause level (e.g. the relationship of verbs and their arguments).

Combining this with results in Chapter 2, it strongly suggests that the IU is the only chunk 'level' in spoken language that functions to regulate the contents of STM.

5.2 Implications for theories of language processing

Nearly all of the existing research on language and memory, even that involving auditory presentation, has been either on word lists or written language. However, in written language, short-term memory is not as important as in spoken language, since the stability of the visual signal and lack of temporal synchronicity between the participants allows the writer to review and revise their production, and the reader to revisit it as many times as needed to achieve comprehension. Experimental tasks such as self-paced reading can mimic the real-time linear processing demands of spoken language, but if the material used is stylistically recognizable as written language, that means it is employing written language structures, which have evolved without the pressure of real time processing. Indeed, Dabrowska (1997) found that the highly complex syntactic structures often employed in sentence processing research are extremely difficult to process even for highly educated native English speakers with the entire passage in front of them, and are impossible to process for less-educated speakers who have not had significant exposure to formal written language.

Although a number of sentence processing models incorporate the idea that prosodic phrases serve to chunk the input into domains for further processing (e.g. Marcus & Hindle, 1990; Pynte & Prieur, 1996; Schafer, 1997; Slowiaczek, 1981), assume that such prosodic groupings exist to serve syntactic processing, and line up with syntactic units. However, prosody in natural spoken language, such as that examined here, is not isomorphic with syntax. The findings in this dissertation documenting the important role of the IU in spoken language processing challenge the primacy of syntactic structure in models of language production and comprehension. I hope these results will encourage further research on how prosody and short-term memory can be incorporated into models of language processing.

5.3 Implications for theories of language change

Chunking in language has generally been assumed to result from associations developed over time from frequent collocation (Type I chunking). Many researchers have identified the process of chunking over time as a crucial factor in language change (e.g. Boyland, 1996; Bybee & Thompson, 2000; Bybee, 2010). The strengthening of associations between multi-word chunks creates formulaic constructions, which can eventually become single words or grammatical morphemes, in a language change process known as 'grammaticalization' (Hopper & Traugott, 2003; Bybee, 2010).

As discussed in Chapter 3, there is evidence that speakers tend to avoid splitting Type I chunks into multiple IUs (Chafe, 1994; Croft, 1995; Wahl, 2015). Croft (1995) specifically links IUs to language change, by asserting that multi-word constructions which have semantic and/or structural properties making them more likely to become more closely integrated units (i.e. to 'grammaticalize') in the future, are less likely than similar constructions without such properties to be split by an IU boundary. The studies in this dissertation, Chapter 3 in particular, suggest that rather than simply reflecting chunking levels, the IU is actually playing an important role in the process, by adding association strength to the words and multi-word chunks within, but not across, IU boundaries. Type III (at presentation) chunking that results from IU processing can thus reinforce ongoing grammaticalization of existing chunks, and begin the process of creating new chunks that may themselves become grammaticalized over time.

5.4 IUs and STM

As with many controversial issues, the debate over the nature of short-term memory limits often involves those with different viewpoints looking at the same evidence and

interpreting it in different ways. This study was not intended to adjudicate between various theories of STM, but primarily tested whether IUs are involved in those phenomena that have frequently been attributed to STM limits. The most important result is that whatever processes are applying to limit STM at the highest chunk level available in typical memory studies (i.e. the word level), they appear to apply at the IU level in spoken language. That being said, it is striking how well the results line up with the ideas of Chafe (1979, 1987, 1994) about IUs and with capacity-limit views of STM.

Chafe (1994) states that IUs are generally limited to five words or less in English (an average of 4.8 in his spoken conversational English corpus), and speculates that this may be the result of cognitive constraints on the amount of simultaneously active information in the focus of attention. The creation of a new chunk by a speaker should be limited by how many existing chunks (i.e. words or strong multi-word units) could fit simultaneously into their focus of attention, so a limit of around 5 pre-existing chunks would be in accordance with the 3-5 chunk capacity limit for STM proposed by Cowan (2000). In the Santa Barbara Corpus of Spoken American English, I found an average of 4.09 words per IU, with standard deviation of 2.9, so this is in a similar range. The upper boundary of that standard deviation (+2.9 words) means that 7 words per IU is not uncommon, but because word counts include strongly associated bigrams, even some that should really be considered a single word-level chunk, e.g. the discourse marker *you know*, we should expect them to skew to the high end of the true estimate.

In Chapters 2 and 4, memory span and same-speaker priming duration defined in IU units had a fast-decaying initial ceiling-level portion with asymptote at around 4 IUs. This is the same number at which discontinuous performance has been observed for individual stimulus items in recall, priming, and various other tasks in experiments in which the design minimized chunking above the stimulus item level, which Cowan (2000) cites as evidence for a 3-5 chunk STM capacity. In spoken language, with its complex

hierarchical structure, there are many possible sizes and types of chunks available, but the results in Chapter 2 showed that of the three linguistic units tested, only IUs were observed to have a discontinuous performance function on memory span, supporting the idea that they are the relevant linguistic unit for incremental processing within short-term memory limits.

It is intriguing that the number 4 continues to crop up in these results. However, I would caution against treating this as definitive support for a particular magical number, as the same results have been found in other studies and interpreted in different ways. The asymptote of the memory span and priming duration functions is at around 4 IUs, but it is not a flat function for that initial portion, it very sharply decreases. A support of the 1-item capacity model might point to this and say that the real capacity limit is 1, and the sharp decrease is due to a build-up of interference. Another possibility is to combine the models into a tripartite model of STM, with a narrow focus of attention of 1 item, and a larger set of activated items. Chafe (1980) advocates this view, writing, "Consciousness has a central focus and a periphery; that is, at any moment, an especially small amount of information is maximally activated, while there is also a larger amount of other information of which a person is to some extent conscious, but which is not being 'focused' on" (p. 12). The possibility for a larger-capacity set of activated but non-focused items is actually acknowledged by Cowan (2000) as well, and he considers it to be compatible with his own model, but he would still argue that the focus of attention has a capacity of 4.

5.5 Conclusion

It is generally accepted that the limitations of short-term memory shape the perception and production of language. As Farmer, Misyak, and Christiansen (2012) write,

"Language comprehension is a complex task that involves constructing an incremental interpretation of a rapid sequence of incoming words before they fade from immediate memory" (p. 353). Assuming that speakers hold planned speech in their focus of attention as they produce it, they must necessarily produce speech in increments that are not larger than the capacity of their own STM or that of their listener. It is reasonable to suppose that some identifiable linguistic unit has developed over time to represent those increments of speech, especially if there is consistency in the limitations of STM that could be exploited, such as the 3-5 chunk capacity proposed by Cowan (2000); Cowan et al. (2002, 2005). The hypothesis of Chafe (1994) is that IUs are that linguistic unit.

Prosodic grouping has been shown to improve recall of word lists (Frankish, 1995), and co-extensive clause and prosodic phrase boundaries have been employed to show some effects of prosody on recall of prose passages (Marslen-Wilson & Tyler, 1976; Jarvalla, 1979). However, there has not yet been an effort to empirically test the role of IUs in memory for naturally-produced spoken language. This dissertation provides evidence that the IU plays a significant role in memory. The fast-decaying non-linear function observed for both memory span (Chapter 2) and priming (Chapter 4) indicates that IUs are reflecting the limits of STM capacity in spoken language. The added association strength within IU boundaries found in Chapter 3 supports the conclusion that IUs are segmenting the speech stream into new chunks during processing, i.e. creating the Type III chunk described in section 1.2.

This dissertation provides evidence for the first time from statistical modeling of memory and priming in spoken English that supports the assertions of Chafe (1994) about IUs. If we look at Type I (pre-existing) chunks only, spoken language looks like a complex intertwined mass of different pre-existing chunks of various sizes and strengths. IUs carve up that continuous flow of information into portions that can be processed easily. The function of the IU to break up the continuous speech stream into processable

portions is a crucial missing piece in the puzzle of how we are able to successfully produce and comprehend spoken language in real-time interactions given our limited capacity STM. I hope that future research will expand the scope of inquiry to other languages, and that researchers working on memory and language will see that the crucial role of prosodic phrases must be taken into account in theories of language processing.

Appendix A

Stimulus Metadata

Table A.1: Metadata for experiment stimuli (Start/End/Dur are in seconds, MeanCor, MaxCor = Mean % of words correct, Max % of words correct across all participants in recall experiment described in 2.2)

Stim	File	Start	End	Dur	Speaker	Words	IUs	Clauses	MeanCor	MaxCor
0	sbc053	1055.53	1056.33	24.46	Mitchell	71	14	10	14.7	35.2
1	sbc056	1353.72	1354.94	21.99	Julie	76	14	11	14.2	35.5
2	sbc001	853.66	857.45	16.7	Lynne	60	14	13	23.2	46.7
3	sbc032	221.79	222.54	22.34	Tom2	82	14	16	14.9	37.8
4	sbc021	814.8	815.95	31.3	Walt	54	14	17	29.9	59.3
5	sbc014	656.25	658.58	16.74	Fred	63	14	17	14.5	39.7
6	sbc056	1496.97	1499.23	17.95	Julie	65	14	21	25.8	63.1
7	sbc049	205.36	206.43	21.74	Lucy	62	15	10	20.5	45.2
8	sbc014	254.71	255.67	22.16	Joe	56	15	13	17.0	39.3
9	sbc038	1191.03	1191.34	29.84	Ben	94	15	14	11.3	25.5
10	sbc039	447.41	447.89	25.39	Kirsten	72	15	16	22.2	63.9
11	sbc052	1369.35	1370.24	25.09	Darlene	67	16	14	15.3	43.3
12	sbc036	1435.78	1437.01	17.43	Marie	77	16	22	16.4	41.6
13	sbc041	370.76	372.85	26.16	Kristin	64	17	11	16.6	48.4
14	sbc056	1240.67	1243.52	24.85	Julie	99	17	24	14.8	34.3
15	sbc046	697.91	699.1	30.18	Reed	98	17	28	13.8	37.8
16	sbc055	861.34	862.39	30.48	Wood	49	17	8	28.1	71.4
17	sbc039	666.58	668.13	23.67	Kirsten	52	17	8	22.0	71.2
18	sbc031	1121.81	1121.98	2.91	Sherry	12	2	1	89.9	100.0
19	sbc006	1489.85	1490.4	1.01	Alina	4	2	1	77.2	100.0
20	sbc003	1526.73	1526.93	0.9	Pete	5	2	1	96.6	100.0
21	sbc006	645.82	647.01	1.67	Alina	7	2	1	91.4	100.0
22	sbc002	779.25	783.65	3.83	Miles	7	2	1	95.3	100.0
23	sbc007	566.12	567.09	6.53	Mary	10	2	4	72.3	100.0
24	sbc009	892.89	894.04	3.3	Kathy	9	2	4	92.3	100.0
25	sbc006	996.92	997.27	2.92	Alina	14	3	4	91.2	100.0
26	sbc008	1241.1	1242.01	3.1	Rebecca	15	3	5	83.2	100.0
27	sbc058	692.34	693.92	7.17	Sheri	28	3	5	42.8	89.3
28	sbc006	1327.74	1329.24	4.05	Alina	19	3	7	63.6	100.0
29	sbc022	406.01	407.06	4.02	Randy	11	4	2	55.1	100.0
30	sbc005	1099.95	1103.51	7.35	Darryl	25	4	4	60.8	100.0

31	sbc035	45.63	46.28	4.2	Stephanie	19	4	7	53.2	100.0
32	sbc007	509.5	511.33	8.5	Alice	27	4	7	31.6	70.4
33	sbc002	565.6	569.07	7.16	Miles	35	4	7	52.3	91.4
34	sbc053	398.1	398.68	8.35	Mitchell	40	4	7	35.3	80.0
35	sbc021	479.35	480.59	6.06	Walt	26	4	8	58.7	100.0
36	sbc003	1098.89	1099.14	5.37	Pete	27	5	3	29.4	74.1
37	sbc048	592.4	593.14	8.36	Lea	10	6	2	78.0	100.0
38	sbc006	1182.39	1184.39	5.43	Alina	25	6	3	55.8	84.0
39	sbc041	874.64	875.59	6.74	Kristin	19	6	4	56.2	94.7
40	sbc004	714.03	715.43	7.07	Sharon	32	6	4	44.2	84.4
41	sbc058	232.1	233.81	10.47	Steven	19	6	5	53.3	94.7
42	sbc029	889.14	890.84	9.18	Seth	24	6	5	44.2	87.5
43	sbc021	1557.46	1558.53	5.21	Walt	23	6	7	52.0	95.7
44	sbc051	766.68	767.71	7.24	Sean	25	6	7	41.9	88.0
45	sbc044	1344.01	1345.63	5.76	LaJuan	26	6	7	58.3	96.2
46	sbc019	484.84	486.58	5.86	Frank	27	6	7	26.7	66.7
47	sbc044	227.67	229.4	9.14	LaJuan	29	6	7	65.9	100.0
48	sbc009	1113.9	1115.22	17.84	Kathy	11	7	3	53.6	81.8
49	sbc005	875.96	876.46	7.19	Pamela	23	7	3	68.6	95.7
50	sbc007	457.35	458.1	13.08	Alice	32	7	3	34.0	71.9
51	sbc008	1181.81	1183.01	8.95	Rebecca	27	7	4	50.2	88.9
52	sbc056	356.55	357.6	6.86	Gary	21	7	5	65.7	90.5
53	sbc008	7.4	12.41	13.27	Rebecca	44	10	7	39.9	90.9

References

Atkinson, J., Campbell, F. W., & Francis, M. R. (1976). The magic number 4 plus or minus 0: A new look at visual numerosity. *Perception*, *5*, 327-334.

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Baddeley, A. D. (1986). *Working memory*. Clarendon Press.

Baddeley, A. D. (2000). The episodic buffer: A new component for working memory? *Trends in Cognitive Sciences*, *4*, 417-423.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8). Academic Press.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355-387.

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, *129*(2), 177-192.

Boyland, J. T. (1996). *Morphosyntactic change in progress: A psycholinguistic approach* (Unpublished doctoral dissertation). University of California, Berkeley.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (1999). Syntactic priming in language production: Evidence for rapid decay. *Psychonomic Bulletin and Review*, *6*, 635-640.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, *2*, 331-350.

Broadbent, D. E. (1975). The magic number seven after fifteen years. In A. Kennedy & A. Wilkes (Eds.), *Studies in long-term memory.* Wiley.

Bybee, J. (2002). Sequentiality as the basis of constituent structure. In T. Givón & B. F. Malle (Eds.), *The evolution of language from pre-language* (p. 109-132). Amsterdam: John Benjamins.

Bybee, J. (2010). *Language, usage, and cognition*. Cambridge University Press.

Bybee, J., & Thompson, S. A. (2000). Three frequency effects in syntax. In *Proceedings of the annual meeting of the berkeley linguistics society* (Vol. 23, p. 65-85).

Chafe, W. (1979). The flow of thought and the flow of language. In T. Givón (Ed.), *Discourse and syntax.* New York: Academic Press.

Chafe, W. (1980). The deployment of consciousness in the production of a narrative. In W. Chafe (Ed.), *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production* (p. 9-50). Norwood, NJ: Ablex.

Chafe, W. (1987). Cognitive constraints on information flow. In R. S. Tomlin (Ed.), *Coherence and grounding in discourse* (p. 21-51). Amsterdam: Benjamins.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: The University of Chicago Press.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234-272.

Cowan, N. (1995). *Attention and memory: An integrated framework* (Vol. 26). Oxford University Press.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87-185.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, *169*, 323-338.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42-100.

Cowan, N., Saults, J. S., Elliott, E. M., & Moreno, M. V. (2002). Deconfounding serial recall. *Journal of Memory and Language*, *46*, 153-177.

Croft, W. (1995). Intonation units and grammatical structure. *Linguistics*, *33*, 839-882.

Crowder, R. G. (1993). Short-term memory: Where do we stand? *Memory and Cognition*, *21*(2), 142-145.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Dabrowska, E. (1997). The lad goes to school: A cautionary tale for nativists. *Linguistics*, *35*(4), 735-766.

DuBois, J., Chafe, W., Meyer, C., Thompson, S. A., Englebretson, R., & Martey, N. (2000-2005). *Santa barbara corpus of spoken american english, parts 1-4*. Philadelphia: Linguistic Data Consortium.

Du Bois, J., Cumming, S., Schuetze-Coburn, S., & Paolino, D. (1992). Discourse transcription. *Santa Barbara Papers in Linguistics*, *4*.

Du Bois, J. W. (2014). *Representing discourse.*

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. In *Studies in second language acquisition* (Vol. 24, p. 143-188). Cambridge University Press.

Farmer, T. A., Misyak, J. B., & Christiansen, M. H. (2012). Individual differences in sentence processing. In M. J. Spivey, K. McRae, & M. F. Joanisse (Eds.), *The cambridge handbook of psycholinguistics* (p. 353-364). Cambridge University Press.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680-698.

Frankish, C. (1985). Modality-specific grouping effects in short-term memory. *Journal of Memory and Language*, *24*, 200-209.

Frankish, C. (1989). Perceptual organization and precategorical acoustic storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3), 469-479.

Frankish, C. (1995). Intonation and auditory grouping in immediate serial recall. *Applied Cognitive Psychology*, *9*, 5-22.

Garavan, H. (1998). Serial attention within working memory. *Memory and Cognition*, *26*, 263-276.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, *105*(2), 251-279.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, *34*, 365-399.

Gries, S. T. (2013). *Statistics for linguistics with r*. Walter de Gruyter.

Gurevich, O., Johnson, M. A., & Goldberg, A. E. (2010). Incidental verbatim memory for language. *Language and Cognition*, *2*(1), 45-78.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 723-802.

Halliday, M. A. K. (1967). *Intonation and grammar in british english*. The Hague: Mouton de Gruyter.

Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, *58*(2), 214-238.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297-318.

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. Cambridge University Press.

Jarvalla, R. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, *10*, 409-416.

Jarvalla, R. (1979). Immediate memory and discourse processing. *The Psychology of Learning and Motivation*, *13*, 379-420.

Jevons, W. S. (1871). The power of numerical discrimination. *Nature*, *3*, 281-282.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, *59*, 193-224.

Kraljic, T., & Brennan, S. E. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology*, *50*(2), 194-231.

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology*, *14*, 78-106.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447-454.

Lockhart, R. S., & Martin, J. E. (1969). Adjective order and the recall of adjective-noun triples. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 272-275.

Lombardi, L., & Potter, M. C. (1992). The regeneration of syntax in short-term memory. *Journal of Memory and Language*, *31*, 713-733.

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory and Cognition*, *26*(4), 708-715.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, *111*, 1-22.

Marcus, M., & Hindle, D. (1990). Description theory and intonation boundaries. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic*

and computational perspectives (p. 483-512). The MIT Press.
Marslen-Wilson, W., & Tyler, L. K. (1976). Memory and levels of processing in a psycholinguistic context. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 112-119.
Mathôt, S., Schreij, D., & Theeuwes, J. (2012). Opensesame: An open-source, graphical experiment builder for the social science. *Behavior Research Methods*, *44*(2), 314-324.
McElree, B. (1998). Attended and non-attended states in working memory: Accessing categorized structures. *Journal of Memory and Language*, *38*, 225-252.
McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111-123.
McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 817-835.
McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General*, *118*(4), 346-373.
McGeoch, J. (1932). Forgetting and the law of disuse. *Psychological Review*, *39*, 352-70.
McKone, E. (1995). Short-term implicit memory for words and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1108-1126.
McLean, R. S., & Gregg, L. W. (1967). Effects of induced chunking on temporal aspects of serial recitation. *Journal of Experimental Psychology*, *74*(4), 455-459.
Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81-97.
Moscoso del Prado Martin, F. (2015). *The time-course of lexico-syntactic resonance in natural dialogue: An information-theoretical study of english, german, and japanese.* Manuscript submitted for publication.
Nairne, J. S. (1988). A framework for interpreting recency effects in immediate serial recall. *Memory and Cognition*, *16*(4), 343-352.
Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology*, *53*, 53-81.
Nespor, M., & Vogel, I. (1986). *Prosodic phonology.* Dordrecht: Foris.
Newell, A. (1990). *Unified theories of cognition.* Harvard University Press.
Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, 2382-2393.
Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*, 633-651.
Pierrehumbert, J. B., & Beckman, M. E. (1988). *Japanese tone structure.* Cambridge: MIT Press.
Pietsch, C., Buch, A., Kopp, S., & de Ruiter, J. (2012). Measuring syntactic priming in dialogue corpora. In B. Stolterfoht & S. Featherston (Eds.), *Empirical approaches to linguistic theory: Studies in meaning and structure* (p. 29-42). Mouton de Gruyter.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*, 633-654.
Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language*, *38*(3), 265-282.
Pynte, J., & Prieur, B. (1996). Prosodic breaks and attachment decisions in sentence processing. *Language and Cognitive Processes*, *11*, 165-191.
Quirk, R., Duckworth, A. P., Svartvik, J., Rusiecki, J. P. L., & Colin, A. J. T. (1964). Studies in the correspondence of prosodic to grammatical features in english. In *Proceedings of the 9th international congress of linguists* (p. 679-691). The Hague: Mouton de Gruyter.
Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*, 587-637.
Reitter, D., Moore, J., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th annual conference of the cognitive science societyonference of the cognitive science society* (p. 685-690). Austin, TX.
Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29-46.
Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734-760.
Ryan, J. (1969). Grouping and short-term memory: Different means and patterns of grouping. *Quarterly Journal of Experimental Psychology*, *21*, 137-147.
Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics*, *2*(9), 437-443.
Saito, S. (1998). Effects of articulatory suppression on immediate serial recall of temporarily grouped and intonated lists. *Psychologica*, *41*, 95-101.
Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 1-17.
Schafer, A. J. (1997). *Prosodic parsing: The role of prosody in sentence comprehension* (Unpublished doctoral dissertation). University of Massachusetts Amherst.
Selkirk, E. O. (1984). *Phonology and syntax*. Cambridge, MA: MIT Press.
Shanks, D. R. (1995). *The psychology of associative learning*. New York: Cambridge University Press.
Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379-423, 623-656.
Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193-246.
Slowiaczek, M. L. (1981). *Prosodic units as language processing units* (Unpublished doctoral dissertation). University of Massachusetts Amherst.
Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of

speaker awareness and referential context. *Journal of Memory and Language*, *48*, 103-130.

Speer, S. R., Crowder, R. G., & Thomas, L. (1993). Prosodic structure and sentence recognition. *Journal of Memory and Language*, *32*, 336-358.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587-10592.

Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken english. *Corpus Linguistics and Linguistic Theory*, *1*(1), 113-150.

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*, 61-82.

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

Wickelgren, W. A., T., C. A., & Dosher, B. A. (1980). Priming and retreival from short-term memory: A speed-accuracy tradeoff analysis. *Journal of Verbal Learning and Verbal Behavior*, *19*, 387-404.

Yantis, S. (1992). Multi-element visual tracking: Attention and perceptual organization. *Cognitive Psychology*, *24*, 295-340.