Memory Capacity Limits in Processing of Natural Connected Speech: The Psychological Reality of Intonation Units

Heather Elizabeth Simpson (hsimpson@umail.ucsb.edu)
Department of Linguistics, University of California – Santa Barbara
Santa Barbara, 93106 USA

Fermín Moscoso del Prado Martín (fmoscoso@linguistics.ucsb.edu)
Department of Linguistics, University of California – Santa Barbara
Santa Barbara, CA 93106 USA

Abstract

Many theories of memory propose some type of short-term store limited in capacity to a small number of information chunks. However, although short-term verbal memory is generally considered to be a crucial component of language processing, the relevant information chunk level that may define capacity limits in ecologically-valid spoken language has never been investigated. The Intonation Unit (IU), an intermediate-level prosodic phrase, has been theorized to be a fundamental unit of spoken language, the focus of a speaker's mental processing. This suggests that IUs might play a role as the relevant unit representing "chunks" of spoken language. We report the results of an experiment investigating the role of IUs in short-term memory in a serial recall task. We found a significant non-linear effect of stimulus size in IUs, but not clauses. We conclude that Intonation Units are the primary linguistic unit used for chunking spoken language input in memory.

Keywords: short-term memory; working memory capacity; information chunks; memory capacity; serial recall; verbal recall; sentence processing; prosody; attention

Introduction

Short-term memory is generally acknowledged to be a crucial part of processing verbal information (e.g. Miller, 1956; Baddeley & Hitch, 1974; McElree, 2000; Lewis, Vasishth, & Van Dyke, 2006). However, memory for spoken language naturally produced in context, the only type of language primary to development and universal across cultures, has gone virtually unstudied. Theories of short-term memory that propose a form of capacity-limited short-term storage (e.g. Miller, 1956; Broadbent, 1975; Mandler, 1985; Cowan, 2000, 2008; Baddeley, 2000; McElree, 2000; Jonides et al., 2008) acknowledge that in continuous complex input such as spoken language, capacity limits would be attested in terms of chunks of highly-associated items, but the nature of such chunks in natural spoken language has not yet been investigated.

The Intonation Unit (IU), an intermediate-level prosodic phrase defined by Chafe (1979, 1994) and DuBois, Cumming, Schuetze-Coburn, and Paolino (1992), is a strong candidate for a linguistic unit that may correspond to the relevant chunk level for a capacity-limited memory. The IU has been theorized to be a fundamental unit of spoken language, that represents the content of a speaker's focus of consciousness at the moment of verbalization, is restricted to representing one piece of new information, and is usually limited to a small number of words, about three or four in English (Chafe, 1994). This definition would seem to equate the IU to the spoken-language instantiation of a capacity-limited short-term store, but this claim has never been directly tested.

IUs represent coherent intonational contours, defined by a complex of prosodic cues. The major cues to IU boundaries include: changes in speech rate, certain types of abrupt pitch change, and pauses. The IU is comparable to the intermediate prosodic phrase in the hierarchy defined by Pierrehumbert and Beckman (1988). A sample transcript sentence, with IU boundaries indicated by line breaks with punctuation, is provided below.

> Anyway,
> this girl must only weigh like,
> a hundred and ten pounds.

Psycholinguistic accounts of memory have focused primarily on syntactic structure and written language, and as such have often implicitly assumed that the clause or sentence unit is the important higher-level unit in memory for connected discourse (e.g., Sachs, 1967; Bransford & Franks, 1971). However, sentence processing research has provided evidence that prosodic grouping can affect syntactic processing (e.g., Pynte & Prieur, 1996; Schafer, 1997; Kjelgaard & Speer, 1999). In addition, the few memory studies that have investigated both syntactic and prosodic grouping have found effects of both clauses and prosodic phrases on recall performance (Jarvalla, 1979; Marslen-Wilson & Tyler, 1976).

Jarvalla (1979) summarizes a set of experiments testing recall for portions of recorded readings of written passages. Each tested portion ended in two seven-word clauses, and the final clause either stood alone as a sentence or formed a sentence with the previous clause. When the passages were presented with normal prosody, the final clause was recalled nearly perfectly (96% by-word average recall) regardless of its sentence status, but the previous clause was recalled much better if it was

part of the same sentence (81% vs. 49%). When prosodic information was removed through a reading done in a monotone and controlled pace, the sentence boundary effect was weakened considerably, with the previous clause recall in both conditions at around 50%, but the average recall of the final clause remained high (88-91% vs. 96%). Therefore, it seems the effect of the sentence boundary was largely dependent on prosodic information, as would be expected due to its status as a combination of syntactic and prosodic information (Chafe, 1994), but there was a strong effect of the clause unit both with and without prosodic information.

In contrast, Marslen-Wilson and Tyler (1976) showed that the effect of prosodic boundary remained even in the absence of syntactic information. They presented stimuli ending in two eight-word clauses, with three conditions: normal prose, a semantically-degraded condition, and a syntactically-degraded 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. The stimuli in these two conditions were read with prosody matched to the normal prose condition. The degraded conditions had lower overall recall performance, with average recall in the final eight-word group at 86%, 75%, and 68%, for the normal, semantic, and syntactic conditions, respectively; but the most dramatic effect was the reduction in recall for the prior eight-word group, with performance at 79%, 43%, and 6%. This indicates that prosodic information was relied upon to group the stimuli into smaller chunks when other sources of information were not available.

There is also evidence for the influence of prosodic grouping on memory from more traditional short-term memory tasks such as serial recall of digits (Frankish, 1995; Saito, 1998). Frankish (1995) found that pitch patterns mimicking natural prosodic grouping contours significantly improved recall of nine-digit lists, but pitch contours taken from familiar melodic structures did not facilitate recall.

The results of prior research suggest important roles for both clause-level chunks and their prosodic analogues in memory for spoken language. However, none of these studies directly compared the effects of syntactic and prosodic grouping of the same stimuli, even though there is a wealth of evidence that language exhibits complex hierarchical prosodic organization that is not isomorphic to syntax (Beckman, 1996; Shattuck-Hufnagel & Turk, 1996; Cutler, Dahan, & van Donselaar, 1997). 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 prosodic organization, but they are clearly different from clauses. For example, the sample transcript sentence provided above contains three IUs, but only a single clause.

We conducted an experiment which addresses this issue by directly comparing number of words, IUs, and clauses as predictors of recall performance using naturally-produced spoken language stimuli. We assume that the complexity and unpredictable length of these stimuli made sub-vocal rehearsal during stimulus presentation improbable, and thus, following Cowan (2000, 2008), a 'pure' capacity limit should be observable as a discontinuity in verbatim recall performance at a certain number of chunks. This discontinuity may appear either as a recency effect or effect of total stimulus size, but the recency effect may not be reliable due to interference from recall of the earlier portions of the stimulus.

The experiment methodology is described below, followed by a discussion of the algorithm used to score participant responses, and finally the statistical modeling results and discussion. The results will shed light on the role of IUs and clauses in memory, and will provide a new source of evidence for the debate over short-term memory capacity limits.

Experiment

Methods

Materials Stimuli consisted of 54 audio clips selected from the Santa Barbara Corpus of Spoken American English (SBCSAE) (DuBois et al., 2000-2005). The SBCSAE corpus contains audio files of naturally-produced spoken English with accompanying detailed transcriptions. The Stanford Parser (version 3.2.0 [1]) was used to automatically derive syntactic parses for the transcripts of the SBCSAE corpus. The transcripts and their parses were then automatically processed using R scripts to extract IU, clause, and word counts for each continuous portion of uninterrupted speech from a single speaker. Clause counts were derived from the number of S nodes in the parse for that portion. IU counts were derived from the Intonation Unit boundaries provided in the SBCSAE corpus. 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, and the low, medium, and high ranges of clauses and words were selected from the available potential stimuli within that IU range. Other considerations, such as the general clarity of the stimulus, also constrained stimulus choice. Table 1 below shows stimulus length ranges for each of the three linguistic units.

[1] http://nlp.stanford.edu/software/lex-parser.shtml, accessed 06/20/2013

Table 1: Stimulus length ranges

	Low IU (n=18)	**Med IU** (n=18)	**High IU** (n=18)
IUs	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 verbal and written instructions explaining that their task would be to listen to a series of audio clips on a computer, and after each one, to transcribe what they remembered hearing into a text file. The instructions explicitly asked them to include everything that they remembered hearing, even if they thought it did not make sense or they weren't sure how to spell something. The text file used for transcription was separate from the experiment interface to allow full word processor-style manipulation of the text, which was not possible in the experiment design software used. Participants were explicitly instructed to minimize the transcription text file window while the clip was playing, so that they would not be tempted to start transcribing while listening to the clip. The 54 test stimuli were presented in random order, preceded by one audio clip used as a training phase.

The participant was 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.

Results

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 or missing information led to exclusion of data from 12 of the participants, so the total number of transcripts used in the analysis was 101.

Scoring

A script was created using the Python programming language to automate comparison of participant transcripts to the gold standard (GS) transcript, the original transcript from the SBCSAE corpus representing the written equivalent of the audio stimulus. Scoring against the GS transcript was designed to be an exact serial recall measure. On the first pass, any word with no exact match in the GS was scored as incorrect. All indices of exact-string-matched words in the GS were identified and stored. In the second pass, for each transcript word in order, the GS word with the lowest serial position was taken to be the correct match, and then was removed from the list of indices. This would ensure that, for example, if there were three instances of the in the GS and in the participant's transcript, they would be matched to the GS in the same relative order. On the third pass, the algorithm reviewed relative position order information for matches. Each match was compared to the most recent previous match (if any), and scored as incorrect either if the previous match had a higher GS word position, or if the previous match was more than 20 positions lower, unless the following match 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. This third pass accounted for errors where the participant correctly recalled part of the clip but in a displaced order, and scoring errors where a match was attributed incorrectly (e.g. 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). The performance of this algorithm is very good as a measure of ordered exact word matches, based on spot-checking of the output by the author.

Analysis

A generalized additive mixed-effects logistic regression was used to predict the likelihood of retrieval for each word in the stimulus based on its length in Words, IUs, and Clauses. The dependent variable was a binary value for each stimulus word for each participant, indicating whether that participant correctly recalled that word in their transcript. Serial word position (i.e. number indicating it was the n-th word in stimulus) was included to control for primacy and recency effects on recall. IU position was not included in the final model due to high levels of collinearity. Clause position was not included due to the difficulty of defining clause membership for fragments (e.g. "*When h-* when he in fact"), discourse markers, (e.g. *y'know*) and filled pauses, (e.g. *um*).

The effects for length of stimulus in words and IUs, 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 random effects for subject and stimulus were found to have a normal distribution.

The significant non-parametric fixed effects from the model are plotted in Figure 1. The y-axis indicates the strength of the effect of that variable on by-word recall, values above 0 indicate a significant positive effect, and values below 0 indicate a significant negative effect.

Panel (a) of Figure 1 shows the size of the smoothed effect of IU count on by-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 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 by-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 high negative effects on recall for high word counts.

Panel (c) shows the size of the smoothed effect of word position on by-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 the highest word position values. The significant negative effect at high word position values would 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 serial position was computed on raw positional values, and since the stimuli were of different lengths, this is probably due to relatively poor recall for the beginning of very long stimuli as compared to medium or low stimuli.

Discussion

The results indicate that Intonation Units play an important role in the short-term retention of verbal information in spoken language. The number of IUs was found to have a significant positive effect on recall performance when there are a small number of IUs in the stimulus, but to have no significant effect for stimuli with more than 3-6 IUs. This result fits well with accounts of short-term memory capacity that assume a limit in terms of a small number of chunks, such as the 3-5 item limit in Cowan (2000).

Clauses do not appear to play a significant role in recall of naturally-produced spoken language, 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 IU boundaries in written-style language.

The word level findings were generally in line with findings from other serial recall studies. The shape of the word position function resembles the usual logarithmic function found for classic short-term memory tasks (Rubin & Wenzel, 1996), with a strong recency effect leveling off at around 7-10 words, the classic 'magic number' from (Miller, 1956) oft-cited as an estimate of short-term memory capacity. Since the current stimuli are expected to have strong associations above the word level, this effect is expected to represent what (Cowan, 2000) terms 'compound STM', a short-term memory containing multiple chunks of information, rather than a pure capacity limit estimate. In line with this assumption, the finding of a linear effect for word count indicates no such discontinuity in memory capacity in terms of stimulus size.

This linearity of the word count effect suggests that the negative effect from adding more words to the stimulus is due to factors such as intra-stimulus interference, rather than a privileged short-term capacity defined at the word level. Since the lowest word count included was 4 it is possible a discontinuity could be observed at lower word counts, but as discussed earlier, capacity limit theories would predict that items in short-term memory would consist of higher-level chunks in connected discourse.

The findings in the current study provide evidence, for the first time, that the IU plays a significant role in the memory representation of spoken language, and that the clause does not play a significant role. The results are most consistent with an explanation for short-term memory based on a capacity-limited focus of attention (e.g. Cowan, 2000), where IUs would serve as the unit of information defining that capacity, though they do not rule out interference-based explanations (e.g. Nairne, 1988, 2002) that have been used to account for performance discontinuities at the word level. The finding that prosodic grouping is significantly more important than syntactic grouping in predicting natural spoken language recall calls for a reassessment of the importance of prosody in language processing.

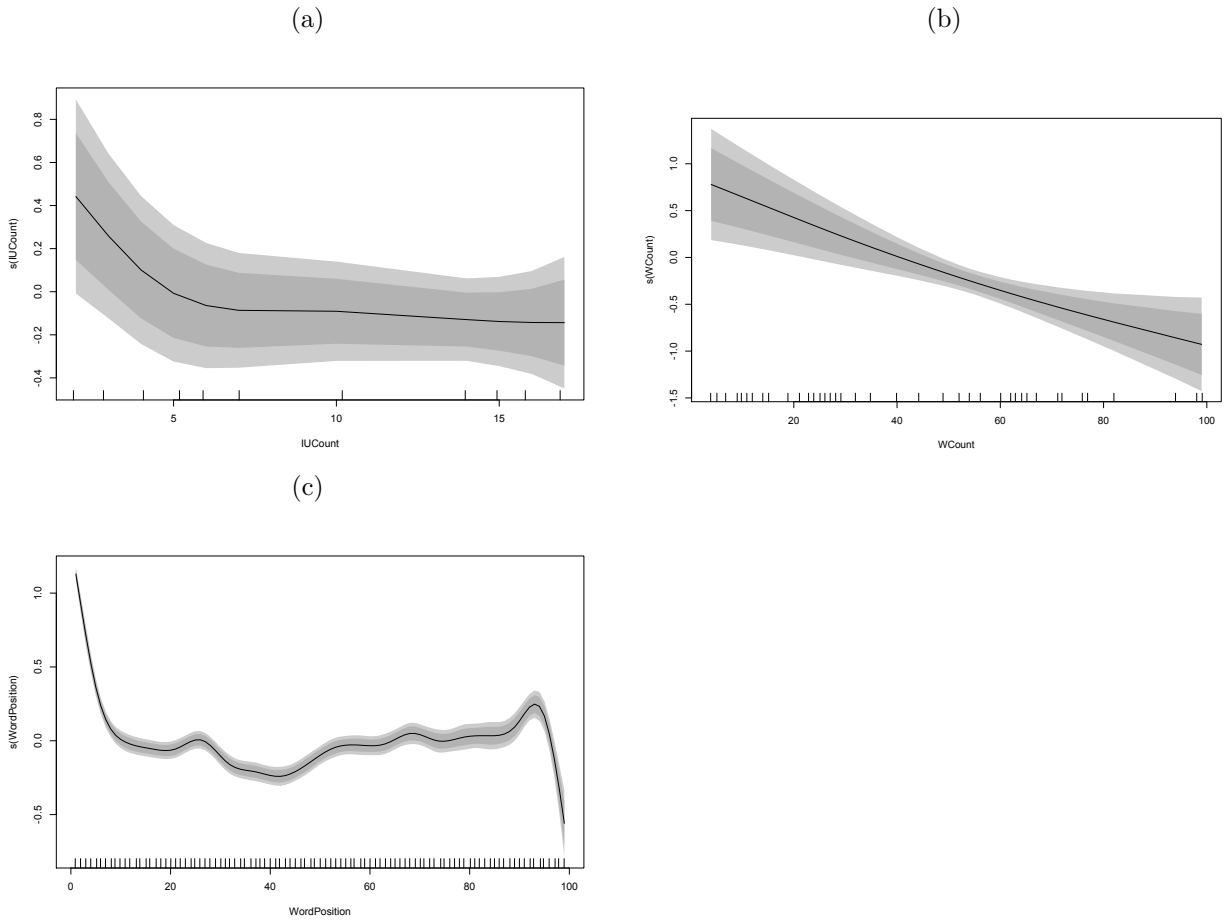

Figure 1: Plot of significant non-parametric estimated effects on recall including 80% and 95% point-wise confidence intervals for (a) Effect of size of stimulus in IUs, (b) Effect of size of stimulus in words, (c) Effect of word position in the stimulus (0 = most recent)

References

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.

Beckman, M. E. (1996). The parsing of prosody. *Language and Cognitive Processes*, *11*, 17-67.

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.

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

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language. *Language and Speech*, *40*(2), 141-201.

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.

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

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

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

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.

Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language*, *40*, 153-194.

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.

Mandler, G. (1985). *Cognitive psychology: An essay in cognitive science*. Erlbaum.

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. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111-123.

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.

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.

Pierrehumbert, J. B., & Beckman, M. E. (1988). *Japanese tone structure*. Cambridge: MIT Press.

Pynte, J., & Prieur, B. (1996). Prosodic breaks and attachment decisions in sentence processing. *Language and Cognitive Processes*, *11*, 165-191.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734-760.

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.

Schafer, A. J. (1997). *Prosodic parsing: The role of prosody in sentence comprehension*. Unpublished doctoral dissertation, University of Massachusetts Amherst.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193-246.