Individual Differences in the Processing of Complex Sentences

Wibke Hachmann (wibke@cognition.uni-freiburg.de)
Lars Konieczny (lars@cognition.uni-freiburg.de)
Daniel Müller (daniel@cognition.uni-freiburg.de)
Center for Cognitive Science, Friedrichstr. 50
79098 Freiburg, Germany

Abstract

To investigate the relationship of functions that underlie the processing of complex sentences, we conducted a self-paced-reading experiment on single and multiple center embedded subject- and object-extracted relative clauses, accompanied by three tests to measure individual differences. Our data suggest that inhibition and executive functions are better predictors for individual differences in sentence processing than working memory capacity.

Keywords: individual difference; sentence processing; executive functions; working memory; inhibition.

Introduction

According to a prominent view, individual differences in linguistic performance can be attributed to differences in working memory capacity (Gibson, 1998; Just & Carpenter, 1992). A widely accepted way of assessing the individual verbal working memory capacity is the reading span test of Daneman and Carpenter (1980) or a variant thereof. It is designed to tap processing capacity and memory span likewise, but has been subject to debates, first of all concerning the construct validity and second because of weak retest-reliability (Mendelsohn, 2002). The lack of relatedness to well established measures of working memory like digit or word span lead to very diverse interpretations of the results. „Given that the reading span task is the standard task in sentence processing for measuring working memory, these failures to find correlations with other working memory tasks pose serious questions about what specifically the reading span task is measuring" (ibid., p. 28).

A further objection is proposed by Baddeley (1996), who notes that the reading span test is unable to distinguish memory capacity from the use of different processing strategies and the ability to dissolve interferences.

In contrast to capacity-based approaches, the correlation of reading span and processing performance can alternatively be explained by individual differences in experience with language (MacDonald & Christiansen, 2002), or by variation in executive functions such as inhibition (Gernsbacher & Faust, 1991; Mendelsohn, 2002).

Good candidates to disentangle these issues and to gain further insight into the underlying processing mechanisms are center-embedded relative clauses:

(1) The reporter that the senator attacked admitted the error.
(2) The reporter that attacked the senator admitted the error.

Object-extracted relative clauses (ORC, 2) are more difficult to read than subject-extracted relative clauses (SRC, 1), and relative-clause type has been shown to interact with reading span (Just & Carpenter, 1992). These effects have been replicated several times in different languages, and thus provide a good field for evaluating the predictability of different measures of individual differences and their interaction with sentence complexity.

Another advantage is that there exist several elaborated attempts to explain these effects, varying in their claims as to which mechanisms are mainly responsible for the differences in processing complexity.

Capacity and experience

Based on Just and Carpenter's (1992) single resource assumption, Gibson proposed the *dependency locality theory* (DLT) as a detailed metric for estimating incremental processing difficulty. There are two crucial factors in this metric: (i) storage of pending predictions, and (ii) integration: the costs of integrating dependent elements across intermediate discourse referents, rising with the number of new discourse referents between the depending elements. According to this model, storage and integration draw on the same resource, and individuals differ with respect to its capacity. According to DLT, center embedding should increase both memory and integration cost during and shortly after the embedded clause. Since in English, there are more referents to be crossed in ORCs than in SRCs when the verb has to be integrated with its arguments, DLT correctly predicts the empirically observable differences in processing speed.

MacDonald and Christiansen (2002) on the other hand assume language processing as shaped by the individual's experience with language and base their theory on connectionist modeling. Thus, processing is highly sensitive to the distributional properties of the linguistic input. They reinterpreted the difference in performance on SRCs and ORCs as a function of frequency and regularity of linguistic forms: subject-extracted relative clauses are easier than object relative clauses because they display the regular S-V-O word order, in contrast to the unusual O-V-S word order

in ORCs. Under this perspective, the concept of a capacity-limited working memory resource is superfluous. Linguistic proficiency and the processes of language comprehension are a result of the interaction of statistical learning and processing mechanisms of the system with the properties of its input. So any measure of individual differences would only capture a result of this learning process, namely something that presents itself like a larger working memory, but could in fact be a very different way of processing.

Inhibition and sentence complexity

The ability to effectively inhibit strong prepotent answers and cope with interferences is also claimed to be crucial for correctly interpreting ambiguous expressions (Gernsbacher & Faust, 1991; Hasher & Zacks, 1988; Jonides, Smith, Marshuetz & Koeppe, 1998; Lustig, May & Hasher, 2001; May, Zacks, Hasher & Multhaup 1999; Mendelsohn, 2002). According to Baddeley's (1996) model of working memory, inhibition is a function of the *central executive*: Anytime a slave system, e.g. the phonological loop, is overloaded, the central executive takes over. Thus, linguistic performance is not only affected by the verbal working memory system, but influenced by numerous functions of cognitive processing, like *switching*, *inhibition* and the *capacity to timeshare* as well as *selective attention*.

As these functions, notably inhibition, influence language processing during ambiguity resolution and the understanding of complex sentences (Gernsbacher & Faust, 1991; Mendelsohn, 2002, Roberts, Hager & Heron 1994), it seems worthwhile to assume that a complex interaction of the above mentioned functions contributes to individual differences.

Our aim in this paper is (a) to gain further insights in the underlying mechanisms responsible for processing complexity, (b) to further investigate the role of inhibition as a general factor affecting syntactic processing and (c) to evaluate instruments that allow to access the various aspects of individual differences.

To address these issues, we conducted a self-paced-reading study with embedded sentences and measured individual differences with three different instruments. In the reading experiment, sentence complexity was manipulated by the type of the relative clause (SRC vs. ORC), as well as the type of embedding (simple, vs. double center embeddings, vs. double right attachments).

The reading experiment was accompanied by two behavioral tests and a questionnaire: (i) the verbal sorting test (Mendelsohn, 2002) designed to measure verbal inhibition skill, (ii) a version of the reading span test and (iii) a detailed questionnaire about reading habits.

We explored the correlations of reading times as reflecting the different dimensions of sentence complexity with the outcomes of the groupings provided by the different tests.

If inhibition is a crucial mechanism for language processing in general, the ability to inhibit conflicting linguistic material should predict performance on center embedded relative clauses. Since the reading span test could be interpreted as tapping several working memory processes or linguistic experience, it might not be sensitive to specific individual abilities that are crucial for language processing. We hypothesized that the VST would be a better predictor for the performance on center embeddings because it measures a more specific individual function. Correspondingly, if inhibition behavior or reading span are functions of exposure to language, they should correlate with the reading experience reported by subjects.[2]

Experiment and Tests

Verbal Sorting Test

In order to obtain a test that can measure inhibition in verbal processing behavior, Mendelsohn (2002) constructed a Verbal Sorting Test (VST) based on the Wisconsin Card Sorting Test (WCST). In the WCST, subjects are supposed to sort cards either by color, shape or by number of objects. After some time, the experimenter changes the sorting rule. Patients who suffer from dysexecutive syndrome have difficulties to switch to a new rule, so they make mistakes of repetition, so-called perseverative errors. In the computer based version of the VST of Mendelsohn, subjects were supposed to sort virtual cards by semantic category, word class or number of syllables.

In her study, the VST was a better predictor of performance in an ambiguity resolution task than the Waters and Caplan (1996) version of the reading span test. Mendelsohn argues that subjects had to process verbal material under constant ambiguity and still switch retrieval plans, while they had to actively suppress the familiar sorting rule.

We translated the Verbal Sorting Test into German and ran a pilot study with 7 participants. They rated items for plausibility in each sorting category. See Table 1Table 1 for a summary of categories and sorting rules.[3] The categories we used were the same as in Mendelsohn (2002). For semantic categories we used *machine*, *animal* and *plant*, for syntactic category *noun*, *verb* and *adjective* and for number of syllables words of *two*, *three* and *four syllables*. Each word had to fit one category of each of the three rules, leading to 27 combinations (3 rules x 9 categories).

[2] This method of self reports provides a rather vague measure, though.

[3] The table shows one set of words per combination. The experiment used two sets and each word appeared three times.

To add two words for each combination, we extracted the 54 most suitable target words out of a total of 270 after rating.[5] Three of the words with the highest scores for all categories formed the sorting references, meaning that those cards are visible throughout the experiment and represent the categories where the targets are supposed to be sorted to. They were:
Generator (*generator*, machine, noun, four syllables), *anpflanzen* (*to plant*, plant, verb, three syllables), and *haarig* (*hairy*, animal, adjective, two syllables).

Table 1: VST target examples for each combination

		animal	plant	machine
animal	noun	2 Katze 3 Elefant 4 Antilope	2 Rübe 3 Kartoffel 4 Artischocke	2 Motor 3 Teleskop 4 Druckerpresse
	verb	2 streicheln 3 erlegen 4 galoppieren	2 welken 3 absterben 4 kompostieren	2 schalten 3 einrasten 4 reparieren
	adj.	2 bissig 3 amphibisch 4 angriffslustig	2 blättrig 3 faserig 4 knollenförmig	2 künstlich 3 digital 4 automatisch

Reference words and the order of targets were held constant throughout the experiment. For a total of nine rules, targets were used three times and randomized into a fixed order to never appear with the same neighbors or with the same rule again.

Reading Span Test For the reading span test, we used the computerized German version conducted by Hacker, Handrick and Veres (2004), reproducing the procedure of the Daneman and Carpenter (1980) reading span test. Items were split into two lists to provide a set for retesting.

Reading Questionnaire All participants filled out a questionnaire about their reading habits. We asked for the amount of time spent reading and writing per day.

Mendelsohn (2002) reported a correlation of VST measures and early reading habits. This suggests that reading experience in childhood is more important for later efficiency in language processing than more recent reading habits. Second, it is unclear whether and to what degree everyday verbal communication is relevant for the processing of written language. To account for these arguments, we split our questions into early, middle and recent reading hours and into written and spoken language accordingly (see Table 2222). In the second part of the questionnaire subjects were asked at what age they started reading and when they became fluent.

Table 2: Sample questionnaire on reading habits.

I. Reading practice per day	
hours; minutes *now* (last half of a year):	
reading	listening
writing	speaking
hours; minutes *youth* (from puberty to majority):	
reading	listening
writing	speaking
hours; minutes *childhood* (until puberty):	
reading	listening
writing	speaking

Experimental materials and design We used 20 sentences of the self-paced reading experiment of Konieczny (2005). They were constructed in a 3x2 design with the factors *clause type* – SRC (1-3a) vs. ORC (1-3b) and *embedding* (simple center embedding – 1ab – vs. double center SC/RC embedding – 2ab – vs. right branching SC/RC – 3ab). Single center, double center and double right embeddings (1,2,3) were combined with either subject- or object relative clauses (a, b) as shown in the pattern below (1a-3b). The experiment was presented together with 100 filler sentences that consisted of other syntactic structures. Items were randomized in a latin square to show only one condition per sentence and all conditions in equal frequency across subjects.

1a) The producer who visited the actor, laughed and cheered.

1b) The producer who the actor visited, laughed and cheered.

2a) Georg's report, that the producer, who visited the actor, laughed and cheered, fascinated Nicole.

2b) Georg's report, that the producer, who the actor visited, laughed and cheered, fascinated Nicole.

3a) Georg reported that the producer, who visited the actor, laughed and cheered, and Nicole was fascinated.

3b) Georg reported that the producer, who the actor visited, laughed and cheered, and Nicole was fascinated.

Procedure

Participants completed the self-paced reading experiment first and then they filled out the questionnaire. The two tests were administered at the end in varying sequence. The experiment and both tests were presented on a 1.6 Ghz Pentium M Laptop running Windows XP. The reading span test was set up with the software of Hacker et al. (2004) and the VST was scripted in DMDX.[7] The whole experiment

[5] Items were rated for their suitability for each of the three rules and categories separately and chosen when they got only one unambiguous score for one category from all raters.

[7] Version 3.2.4.0; by Jonathan und Kenneth Forster (2002), Institut of Psychology, *University of Arizona*, Tucson, Arizona, USA

took 75 minutes. The experiment was implemented in Linger (Rhode, 2001)[8]

The experiment started with an introduction, then participants completed a test run with six test items. Sentences were displayed in a single word non-cumulative self-paced reading paradigm with moving window over hyphens indicating the length of words. Every next word was uncovered by pressing the space key. 25% of the sentences were followed by a forced choice comprehension question about a detail in the sentence, to assure that participants read carefully. Key press latencies of the space key were recorded as dependent variable.

VST Three white circles with the reference words were displayed in equal distance from a black dot in the middle of the screen. Clicking on the dot uncovered the next target word in another white circle of about 8.8cm diameter. The task demanded that participants clicked on one of the references to indicate where the target card should be sorted to. Contrary to a test-run, the rule was not provided but participants received feedback about whether they were correct.

Figure 1 shows the display screen after a trial during feedback presentation.

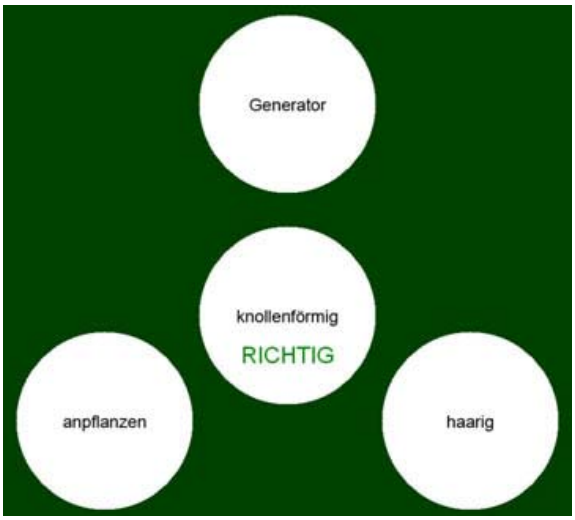

Figure 1: VST display during feedback presentation.

With three possible rules in the beginning it takes three clicks to infer the correct rule. Then they went on sorting until they could infer from the feedback that the rule had changed. The rules changed in a fixed order after 10 correct successive trials or after 18 trials of the same rule, which made a rule change rather unpredictable. The VST took 12 minutes to complete.

Reading Span Test The procedure was the same as described in Hacker, Handrick and Veres (2004). Participants read sentences of average 12 words length in blocks of two to eight sentences. After reading one block, they had to recall the last word of each sentence in correct order, and in addition they had to write down some key words of the sentences. This test took 15 to 20 minutes, depending on the highest block size the participant was able to remember correctly.

Participants

35 participants took part in the experiment either for course credits or for a compensation. Two data sets had to be excluded because one participant was no native speaker of German and the other had technical problems, leaving 33 subjects with a mean age of 23.9 years (SD=3.1 years), 16 of which were females.

Results

VST All trials except the first rule and the first two trials after each rule change were registered as *counted trials*. [9] Any mistake that would have been correct under the previous rule was counted as *perseverative error*. For each participant perseverative errors were set in proportion to all counted trials and yielded the individual measure *proportion perseverative*. There was an average of 95.32 counted trials per experiment, 10% of which had wrong answers. Half of these total errors where perseverative errors, which yielded a proportion of 5% (SD=4.1). Mean reaction times (rt) on wrong answers (rt=2686.83, SD=1198.23) were significantly slower than the average counted trials (rt=2078.36, SD=567.00, t-test: t=-5.390, p<.00001).

In addition, we collected mouse hovering times over any of the virtual cards during the test. The proportion of hovering over perseverative words to all hovering times over word circles formed the *proportion of perseverative hovering times*. The individual VST score was formed by the combined measure of *proportion of perseverative hovering times* and *proportion perseverative*. Average hovering times where 561.40ms (SD=243.09), while those above perseverative references where longer and more variable (689.704ms, SD=342.293). The proportion of perseverative hovering times made 44.5% (SD=15.1) and the measure combined perseverative was 49.5 (SD=15.3) on average. Data sets were split by the 33rd and 66th quantile of the combined measures into three groups of 16 participants respectively. Good inhibitors had combined perseverative scores of 0 – 41.3, medium of 41.4 to 51.2 and poor performers had scores of 51.3 and above.

[8] Version 2.94 from Rohde, D. (2007): Linger. A flexible platform for language processing experiments. Online: http://tedlab.mit.edu/~dr/Linger (April 27th, 2009).

[9] Rules never repeated directly, so any new rule could be derived with maximally two clicks.

VST error variables did not correlate with any of the other test measures. However, reaction times in the VST correlated positively with the years it took the respective participant to learn how to read (r^2=.306, p<.05) and negatively with communication hours per day (r^2=-.296, p<.05). This means that subjects who reported much communication in their everyday lives and a short acquisition period until they could read fluently had fast reaction times in the VST.

Reading Span Test Scores were computed according to Hacker et al. (2004). Any block in which participants remembered all last words and the sentence content correctly yielded one point. Points were then matched by percentiles to an individual reading span as provided by Hacker et al. (2004). Span groups were built using the least difference in group size as in the VST, contrary to the grouping by span size reported in Hacker et al. (2004).[11] Since the granularity of this measure did not allow for equal group sizes in this sample, span scores were split so that the largest number of participants fell into the mid span group. Spans 1.5 to 2.1 had 8 participants, 2.5 to 3.1 had 15 and span scores 3.5 to 4.8 had 10 accordingly.

Reading Questionnaire The variables used were *age of acquisition*, *reading* (reading hours at present per day) and *communication* (speaking and hearing hours of the present collapsed).

Reading habit was correlated with reading times on V1 of *embedding*: *communication* correlated negatively with reading times on SRCs in the single embedding condition (r^2=-.466, p<.01) and with those in ORCs in double embedding (r^2=-.557, p<.01). *Reading* correlated negatively with ORCs in single embedding (r^2=-.490, p<.01) and reading span correlated positively with SRCs in double embedding (r^2=.453, p<.01). Though the groups of reading habits did not gain further results, single measures of reading, communication and age of acquisition proved to be worthwhile variables for further research.

Experimental results We computed *residual reading times* as the difference between actual reading times and the expected value from a linear regression predicting reading time by word-length. Residual reading times on the verbs of the main clause (V2) as well as the subordinate clause (V1) were entered as dependent measures into a general linear model (GLM, MANOVA) with the fixed conditions *embedding* and *sentence type*. There was a significant main effect of embedding on V1 ($F_{(2)}$=4.006, p<0.02), and a main effect of sentence type on V2 ($F_{(1)}$=7.451, p<0.01), but no interactions (all p>0.3).

[11] Hacker et al. grouped high spans by scores above 4, which would leave us with 3 subjects in the highest group.

As can be seen in Figure 2Figure 22, subjects needed most time for double embeddings and read right attachments as fast as single embeddings at the end of the relative clause (V1). This effect switches to a sentence type effect with longer reading times on the main verb (V2) in object relative clauses irrespective of embedding, mainly because the right attached ORCs are read much slower on V2.

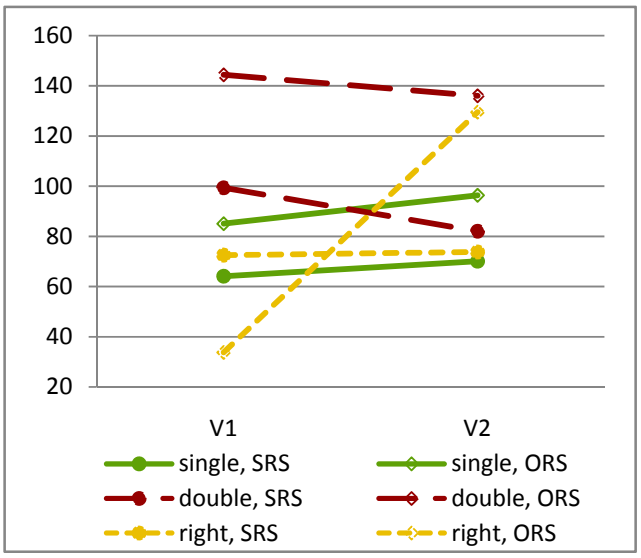

Figure 2: Residual reading times (ms) on V1 (main effect: embedding) and V2 (main effect: sentence type).

To compute results of group differences in reading times we added the group variables as fixed conditions into several linear models. Adding VST groups showed a significant three-way interaction of V1 with both effects on embedding and sentence type (F(4)=2.804, p<.05). Reading span interacted with embedding (F(4)=3.292, p<.05).

High proficient inhibitors showed both main effects, while the effect was modulated by a switch within the right attachments. In this group, right attached ORCs were read slower on V2, but not on V1. The group who made many perseverative errors showed high reaction times on V2 of right attached sentences.

Reading span groups differed in their sensitivity to embedding, so that only mid span readers were less prone to having long reading times on double embedded sentences. The lowest span group had the longest reaction times on double embedded ORCs.

Discussion

The results show that the ability to correctly sort words into categories and switch between rules predicted behavior in reading and understanding embedded SRCs and ORCs. The three-way interaction of the VST with embedding and clause type supports the hypothesis that the VST is sensitive to individual differences of various processing components.

The reading span test interacted with both main effects, but yielded a two-way interaction only with center embedding. These data suggest that the VST is a broader measure of individual differences than the reading span test and that more specific functions than variation in one central resource capacity underlie individual differences. Reading span, meanwhile, still proved useful to predict embedding effects.

The results of the experiment on embedding support the hypothesis that center embedded clauses are more difficult to process than right attachments. The missing effect of clause type on the first verb makes a notable difference to this hypothesis, because at V1 subjects should integrate the missing verb arguments that had to be held active during the ORC. Instead, they seem to have longer reading times at V2 after the clause boundary of ORCs. This effect can be mainly attributed to right attached relative clauses, where reading times increase markedly from V1 to V2. The main clause in the beginning might give rise to anticipate a regular subject clause structure in the first subordinate complement sentence, for example: "Daniel erzählte, dass der Professor den Studenten begrüßte." (*_Daniel told, that the professor the student received._), which interferes with the relative clause that follows instead.

Conclusion

These results emphasize the importance of executive functions in accounting for inter-individual differences in language processing, counting in center embedding as well as sentence type. They represent a model for a group of mechanisms that interact in many aspects that requires high level cognitive processes and that still need further research to be firmly understood.

Acknowledgments

This work was granted by the German Research Foundation (DFG, grant KO 1932/3-1) We thank Sascha Wolfer and Peter Baumann for ceaseless help with data management and anonymous reviewers for their fruitful comments on earlier versions of this paper.

References

Baddeley, A. (1996). Exploring the central executive. *Journal of Experimental Psychology: Human Experimental Psychology, 49A*, 15-28.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences, 22*, 77 - 94.

Carpenter, P. A., Miyake, A., & Just, M. A. (1995). Language comprehension: Sentence and discourse processing. *Annual Review of Psychology, 46*, 91-100.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450-466.

Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17 (2)*, 245-262.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition, 68*, 1-76.

Hacker, W., Handrick, S., & Veres, T. (1996). *Lesespannentest*. Manuscript at the University of Dresden.

Jonides, J., Smith, E. E., Marshuetz, C., & Koeppe, R. A. (1998). Inhibition in verbal working memory revealed by brain activation. *Proceedings of the National Academy of Sciences of the USA, 95* (pp. 8410–8413). Washington: PNAS

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99 (1)*, 122-149.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The Role of Working Memory. *Journal of Memory and Language, 30*, 580-602.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research, 29 (6)*, 627-645.

Konieczny, L. (2005). *Storage, integration, and anticipation effects during sentence processing*. Talk at the 14th conference of the European Society for Cognitive Psychology (ESCoP). Leiden, 02. 09. 2005.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General, 130 (2)*, 199-207.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109 (1)*, 35-54.

May, C. P., Zacks, R. T., Hasher, L., & Multhaup, K. S. (1999). Inhibition in the processing of garden-path sentences. *Psychology and Aging, 14*, 304-313.

Mendelsohn, A. A. (2002). *Individual Differences in Ambiguity Resolution: Working Memory and Inhibition.* Dissertation presented at Northeastern University Boston, Massachusetts.

Roberts, R. J., Hager, L. D., & Heron, C. (1994). Prefrontal cognitive processing: Working memory and inhibition in the antisaccade task. *Journal of Experimental Psychology: General, 123 (4)*, 374-393.

Rummer, R. (1996). *Kognitive Beanspruchung beim Sprechen.* Weinheim: Beltz.

Waters, G. S., & Caplan, D. (1996). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review, 103 (4)*, 761-772.

Zacks, R. M., & Hasher, L. (1988). Capacity theory and the processing of interferences. In L. L. Light, & D. M. Burke, *Language, Memory, and Aging.* New York: Cambridge University Press.