Computational Analysis of Referring Expressions in Narratives of Picture Books

Choonkyu Lee
Department of Psychology
Rutgers Center for Cognitive Science
Rutgers University – New Brunswick
choonkyu@eden.rutgers.edu

Smaranda Muresan
Library and Information Science Department
School of Communication and Information
Rutgers University – New Brunswick
smuresan@rci.rutgers.edu

Karin Stromswold
Department of Psychology
Rutgers Center for Cognitive Science
Rutgers University – New Brunswick
karin@ruccs.rutgers.edu

Abstract

This paper discusses successes and failures of computational linguistics techniques in the study of how inter-event time intervals in a story affect the narrator's use of different types of referring expressions. The success story shows that a conditional frequency distribution analysis of proper nouns and pronouns yields results that are consistent with our previous results – based on manual coding – that the narrator's choice of referring expression depends on the amount of time that elapsed between events in a story. Unfortunately, the less successful story indicates that state-of-the-art coreference resolution systems fail to achieve high accuracy for this genre of discourse. Fine-grained analyses of these failures provide insight into the limitations of current coreference resolution systems, and ways of improving them.

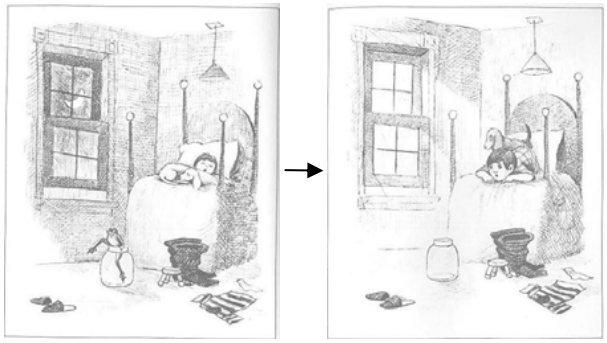
1 Introduction

In theories of information structure in extended discourse, various factors of discourse salience have been proposed as determinants of information 'newness' vs. 'givenness' (e.g., Prince, 1981). Based on evidence from speakers' choice of different types of referring expressions in referring back to a previously introduced discourse referent, scholars have discovered effects of (a) 'referential distance' (Givón, 1992), a text-based measure of distance between the antecedent and the re-mention in terms of number of intervening clauses; (b) topic-prominence of the referent in the previous mention (Brennan, 1995); (c) presence of another candidate referent ('competitor') in linguistic or visual context (Arnold and Griffin, 2007), among others. In re-mentioning individuals, one can, for example, simply repeat names or use anaphoric devices, such as definite descriptions and pronouns.

In our work, we have been investigating the role of mental representation of nonlinguistic situational dimensions of the storyline (e.g., Zwaan, 1999) as an additional factor of salience in discourse organization. From the five situational dimensions of the event-indexing model (Zwaan and Radvansky, 1998), we have focused on the time dimension. In a narrative elicitation study (Lee and Stromswold, submitted; Lee, 2012), we presented picture sequences from three wordless picture books in Mercer Mayer's "Boy, Dog, Frog series" (Mayer, 1969; Mayer, 1974; Mayer and Mayer, 1975), and had 8 adults estimate the *inter-event intervals in story time* between consecutive scenes with no linguistic stimuli, and had a different group of native English-speaking adults write stories to go along with the pictures. The 36 adults wrote a total of 58 written narratives, which consisted of 2778 sentences and 38936 word tokens (48 sentences and 671 word tokens per narrative on average). The use of wordless picture books allows fixed target content and clear visual availability of the characters and their actions.

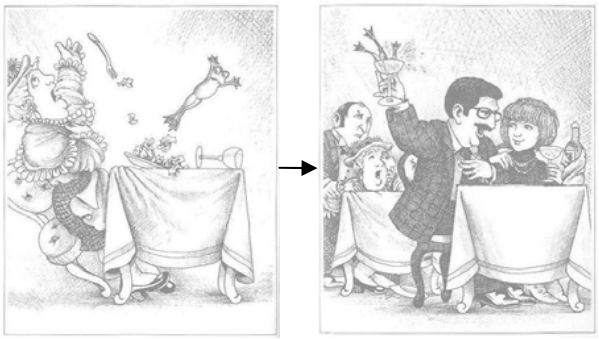
In our previous analysis (Lee and Stromswold, submitted) of the effect of inter-event time intervals on the narrator's referential choice in referring

Workshop on Computational Linguistics for Literature, pages 1–7,
Montréal, Canada, June 8, 2012.



S1) Finally though, the boy starts to get tired and decides to crawl into bed. His dog joins him and soon they are asleep. The boy forgot to put a lid on the bottle, and Mr. Frog is sneaking out!
S2) When the boy wakes up in the morning, he sees that Mr. Frog is gone. He is very upset that he lost his new friend.

Figure 1. Sample 'Long Interval' Between Scenes S1 and S2 (Mean Estimate: 6h 48m 45s).



S3) After staring at the frog for two minutes he says "Ribbitttttt" and she screams and throws her fork into the air, and falls back in her chair. Charles gets scared by her screaming and jumps off her plate into the air.
S4) Luckily, he lands safely into a man's drink. He is mid-conversation with a beautiful lady and doesn't feel the new addition to his martini.

Figure 2. Sample 'Short Interval' Between Scenes S3 and S4 (Mean Estimate: 3s).

back to characters, we manually annotated critical sentences selected on the basis of the eight longest (mean duration = 1 hour 7 minutes 2 seconds; henceforth, 'Long Intervals') and the eight shortest (mean duration = 10 seconds; henceforth, 'Short Intervals') estimated intervals. Examples of a Long Interval and a Short Interval between scenes are given in Figures 1 and 2, together with sample corresponding narratives. For each of the 58 narratives, we analyzed the first sentence after a Long and Short Interval. Our coding of *referring* expressions involved frequency counts (ranging from 0 to 3) of instances of each of our Referential Types – Proper Names (e.g., *Mr. Frog*), Definite Descriptions (e.g., *the frog*), and Pronouns (e.g., *he*) – per critical sentence. We found a significant interaction between Interval and Referential Type in both a chi-square test of association and an analysis of variance, and the effect generally held across participants. Our finding demonstrated that narrators used Proper Names more after Long Intervals than after Short Intervals in story time, and more singular-referent Pronouns after Short Intervals than after Long Intervals.

Addressing the issue of the effect of inter-event interval on referential choice on a larger scale requires accurate automatic methods for identification of Referential Types and coreference resolution for the narratives. In this paper we first present a simple computational method for analyzing the *entire scene* descriptions after the Long and Short Intervals to study how inter-event intervals affect referential choice, focusing on Proper Nouns and Pronouns. Our results from the automatic methods are consistent with the results obtained using manual coding of the *critical sentences*. Second, we present an annotation study of nine narratives with coreference chains, and also discuss the performance of two state-of-the-art coreference resolution systems on a sample of our data.

2 Inter-event Interval Effect on Referring Expressions: A Basic Computational Approach

In order to address the question of how inter-event intervals affect the choice of referring expressions, we analyzed the frequency of Pronouns and Proper Nouns in scenes following the Long and Short Intervals. The results in Table 1 are consistent with our previous results obtained based on manual coding of the critical sentences only: The 'Long Interval' (LI) scenes and the 'Short Interval' (SI) scenes diverge in relative frequencies of our target part-of-speech tags – Pronouns (nominal (PRP) and possessive (PRP$) forms) vs. Proper Names (NNP).

One can observe that there are generally higher frequencies of Proper Names for the scenes after the Long Intervals compared to the Short Intervals, not only in absolute number but in relative proportion to Pronouns as well. A noticeable exception, Scene 3 of *One Frog Too Many* (Mayer and

Book	Scene#	PRP	PRP$	NNP
Frog Goes to Dinner	4 (LI)	62	56	106
	5 (LI)	54	37	96
	21 (LI)	87	60	120
	9 (SI)	45	22	27
	13 (SI)	50	44	50
	14 (SI)	40	21	40
One Frog Too Many	8 (LI)	33	33	55
	19 (LI)	63	42	90
	20 (LI)	60	29	88
	3 (SI)	70	65	158
	15 (SI)	69	50	73
	23 (SI)	1	2	2
Frog, Where Are You?	2 (LI)	89	70	143
	3 (LI)	70	65	158
	18 (SI)	64	56	86
	19 (SI)	63	42	90

Table 1. Scene-based Frequencies of Pronouns and Proper Names after the 16 Long and Short Intervals.

Mayer, 1975), is a very early scene in the picture book, with many character introductions and discourse-newness (Prince, 1981). Even with this exception included, the association between Interval (Long vs. Short) and Referential Type (Pronouns vs. Proper Names) was significant in a new analysis based on the entire scene descriptions, rather than just the first sentences for these scenes [$\chi^2(1) = 9.50$, $p = .0021$]. The significant effect of Interval reveals that Proper Names were more commonly used after Long Intervals than after Short Intervals, and Pronouns were more commonly used after Short Intervals than after Long Intervals.

The exception in Scene 3 of *One Frog Too Many* suggests, however, that excluding first few mentions in a coreference chain from analysis may reveal a stronger effect of Interval on referential type of re-mentions (although one mention for introducing a character does not always establish discourse-givenness from the narrator's perspective (Clancy, 1980)). Successful automatic coreference resolution would facilitate this analysis as well.

3 Annotation of Referring Expressions in Narratives of Picture Books

In order to provide descriptive statistics of referring expressions in our narratives of pictures books and to test the performance of coreference systems automatically in the future, we annotated 9 narratives manually with coreference chains (3 narratives for each of the 3 pictures books, with each narrative written by a different writer). Only animate entities, or characters in the stories, were considered. We used the MMAX2 annotation tool (Müller and Strube, 2006). A coreference schema is available from the Heidelberg Text Corpus (HTC, Malaka and Zipf, 2000) sample directory included in the MMAX2 package. The HTC schema allows marking a mention in terms of the discourse entity or coreference chain it corresponds to, as well as 'np_form' (what type of (pro)nominal it is), 'grammatical_role' (subject/object/other) and 'semantic_class' (abstract/human/physical object/other). We imported the HTC schema to annotate the mention level in terms of coreference, and also created a 'scene' level for our picture-book narratives.

The narratives were annotated by the authors of this paper independently in the initial version, and with adjudication for the final version. As the referents were very clear in the narratives for the picture books, there was only one case of initial disagreement in the authors' coreference decisions. Table 2 shows statistics related to these 9 narratives.

Narrative ID	# of Mentions	# of Chains	# of Words	Longest Chain	Average Chain Length	Density
1	65	8	280	22	8.13	.23
2	71	5	277	29	14.20	.26
3	52	7	268	15	7.43	.19
4	128	13	562	60	9.85	.23
5	62	12	256	20	5.17	.24
6	78	11	383	25	7.09	.20
7	271	23	1109	58	11.78	.24
8	111	21	514	38	5.29	.22
9	167	26	834	37	6.42	.20

Table 2. Descriptive Statistics for Each Narrative.

The density of referring expressions is very high (~22% of tokens/words in a story are referring expressions). Densities are also consistent across narratives: Narrative #7, which was by far the longest one with 1109 words, also showed a very high density (24%). Numbers of coreference chains are also consistent within each target picture book regardless of writer or narrative length: 8, 5, and 7 for *One Frog Too Many* (Mayer and Mayer, 1975); 13, 12, and 11 for *Frog, Where Are You?* (Mayer, 1969); and 23, 21, and 26 for *Frog Goes to Dinner* (Mayer, 1974). Table 2 also shows that the longest

chain contains 60 mentions, and the average chain has about 8 mentions.

4 Performance of Coreference Resolution Systems on Narratives of Picture Books

In computational linguistics, the increasing availability of annotated coreference corpora has led to developments in machine learning approaches to automatic coreference resolution (see Ng, 2010). The task of automatic NP coreference resolution is to determine "which NPs in a text […] refer to the same real-world entity" (Ng, 2010, p. 1396). Successful coreference resolution often requires real-world knowledge of public figures, entity relationships, and aliases, beyond linguistic parameters such as number and gender features.

In this paper, we have chosen two coreference resolution systems: Stanford's Multi-Pass Sieve Coreference Resolution System (Lee et al., 2011) (henceforth, Stanford dcoref) and ARKref (O'Connor and Heilman, 2011). Stanford dcoref consists of an initial mention-detection module, the main coreference resolution module, and task-specific post-processing. In this system, global information about the text is shared across mentions in the same cluster in the form of attributes such as gender and number. This system received the highest scores at a recent CoNLL shared task (Pradhan et al., 2011), which the authors attributed to the initial high-recall component (in mention detection) followed by high-precision classifiers in the coreference resolution sieves. ARKref is a syntactically rich, rule-based within-document coreference system very similar to (the syntactic components of) Haghighi and Klein (2009).

We analyzed in depth the performance of these systems on one of our narratives for *Frog Goes to Dinner* (Mayer, 1974). We expected automatic coreference resolution systems to show poorer performance when applied to our written narratives than that reported in the literature, because most of these systems have been trained on newswire, blog, or conversation corpora, which – though quite a heterogeneous set in themselves – are not similar to our written narrative data. Some of the most noteworthy particularities of our written narrative collection include (a) fictional content, in which animals occur frequently and are greatly anthropomorphized, (b) an imaginary target audience of a limited age range (six- to eight-year-olds), and (c) clear scene-by-scene demarcation in the writing process, with a new text input box for each new scene in a picture book. The first point, in particular, may limit the utility of named entity recognition (NER) and WordNet relations among nominals in the preprocessing steps prior to coreference resolution. As we discuss below, preprocessing errors in parsing and NER did in fact contribute to coreference precision errors.

Our written narratives had a lot of singleton mentions for secondary characters and plural combinations of characters. We thus evaluated the performance based on the B^3 measure proposed by Bagga and Baldwin (1998), rather than the link-based MUC (Vilain et al., 1995).

We computed the B^3 with equal weighting for all mentions. Stanford dcoref achieved B^3 scores of 0.78 Precision, 0.43 Recall and 0.55 F_1, while ARKref scores were 0.67 for precision, 0.45 for recall, and 0.54 for F_1. Stanford dcoref includes a post-processing module in which singletons are removed, which partially contributes to the low recall score for the system.

4.1 Qualitative analysis of coreference output

In this section, we discuss the errors from both ARKref and Stanford dcoref in depth. The coreference outputs from both ARKref and Stanford dcoref demonstrate that preprocessing errors can lead to errors downstream for coreference resolution. Misparsing is one of the serious issues. For example, in ARKref's output for our sample narrative (for *Frog Goes to Dinner)*, the third-person singular verb *waves* in *Billy waves goodbye* (Scene 6) and *Froggy waves goodbye* (Scene 7) was misparsed as a plural nominal and thus a headword of a mention for a discourse entity, and these two instances were marked as coreferent. Lee et al. also acknowledged misparsing as a major problem for Stanford dcoref.

A few surprising errors in the ARKref output include (a) marking *the woman* and *him* in the same clause as coreferent despite the gender mismatch, and (b) leaving *the lady* as a singleton and starting a new coreference chain for *her* in the same clause. It is strange that the explicitly anaphoric pronoun mention did not lead ARKref to link it to the identified mention *the lady*.

Other noteworthy errors common to both systems' outputs were the following:

(1) inconsistent mention detection and coreference resolution for mentions of the frog character with *Froggy*;

(2) failure to recognize cataphora in *Without knowing Froggy's in [his]$_i$ saxophone, [the saxophone player]$_i$ tries to blow harder*… and linking the pronoun to *Froggy* instead;

(3) starting a new coreference chain at Scene 4 at the mention of *Billy* when the referent (the boy) has been already introduced as *Billy Smith* in Scene 1;

(4) the same type of error for another character (the frog) at an indefinite NP *a frog* in *She is so shocked that there is a frog in her salad*.

With regard to error (1), preprocessing results in the Stanford dcoref output reveal some NER errors in which *Froggy* was mislabeled as an 'organization,' which, along with the absence of *Froggy* in the name gazetteer for the system (Lee et al., 2011), would lead to both precision and recall errors for *Froggy*, as we observed.

Error (3) reveals the potential pitfall of overreliance on headwords for mention/discourse-new detection, which leads these systems to miss the internal structure to people's names – namely, [first name + last name] for the same person,[1] which then can be re-mentioned using just the first name. Although in news articles and other formal writing it is typical to mention a person by the last name (e.g., *Obama* rather than *Barack*) as long as the referent is clear, stories, conversations, and other less formal genres would make more frequent use of first names of individuals for re-mention compared to other genres. Because the importance of coreference resolution is not limited to formal writing, coreference resolution systems need to incorporate name-specific knowledge, either in preprocessing stages such as parsing and NER or in coreference resolution after the preprocessing.

Error (4) is not as undesirable as the other ones: Even for a human annotator, it is more difficult to make a coreference decision for a case like this one, in which the fact that the salad-eating lady was shocked would come about similarly for any frog, not just Froggy. Although there does not seem to be a rule for classifying an indefinite NP as denoting a new entity,[2] training on a large corpus would lead to such a tendency because indefinites usually do indicate discourse-newness introducing a new discourse referent.

In another narrative for the same picture book, there were two definite NPs (*the woman* and *the waiter*) for which the definiteness was due to the visual availability of the referent in the scene or a bridging inference (restaurant – waiter) rather than a previous mention. Definiteness may lead coreference systems to prefer assigning the mention in question to an existing coreference chain rather than creating a new chain, but ARKref processed both of these possibly misleading definite NPs successfully by creating a new coreference chain, and Stanford dcoref got one right and made a recall error for the other. On the other hand, referring to different secondary male characters similarly as *the man* did lead to a spurious coreference chain linking all of these mentions.

5 Conclusion and Future Directions

With the NLP tools discussed above, possibilities abound for interesting research on narratives. Based on scene-based segmentation of narratives written for fixed target picture sequences, one can collect various kinds of linguistic and nonlinguistic data associated with the picture sequences and conduct regression analysis to see which factor has the most predictive value for linguistic variation such as Referential Type choice. Important factors include temporal and thematic (dis)continuity in the target content (McCoy and Strube, 1999; Vonk et al., 1992), and discourse salience factors (Prince, 1981), for which we have collected measures in our previous work.

Our Interval Effect finding lends support to McCoy and Strube's (1999) intuition underlying their referring-expression generation system, for which they used reference time change in discourse as a major predictor of referential type. Gaining further insight into the impact of time change in content on referential choice in naturally occurring discourse can thus lead to a predictive model of referring expressions as well.

In the future, we plan to use 'semantic_class' attributes and features such as ANIMACY in the

[1] Application to East Asian languages would need to adjust to the opposite 'family name + given name' sequence, often even in English transliteration (e.g., *Kim Jong-il*).

[2] According to Lee et al. (2011), Stanford dcoref correctly recognizes coreference in appositive constructions with an indefinite NP *after* the first mention.

HTC schema as our task-specific filters for selecting just story characters. Moreover, we plan to explore other state-of-the-art coreference systems such as CherryPicker (Rahman and Ng, 2009). The NLP tools and techniques discussed above can be applied to cross-document coreference resolution as well (see Bagga and Baldwin, 1998, for discussion of a meta document), although training the systems for narratives like ours would involve much more manual annotation and supervision, particularly because different authors usually assign different names to a given character. In order to limit the amount of manual annotation, unsupervised methods for coreference resolution (Ng, 2008; Poon and Domingos, 2008; Haghighi and Klein, 2007) could be used. This, however, would require a larger number of picture books and human-produced narratives.

Coreference is far from a simple phenomenon, both for theory and application. Nevertheless, ultimately it would be desirable to improve the automatic coreference resolution systems in ways that reflect corpus-linguistic and psycholinguistic findings – e.g., referential distance effects (Givón, 1992), and the privileged status in memory of discourse entities in the immediately preceding clause (Clark and Sengul, 1979). The goal would be to represent as many of the interacting factors in referential choice as possible, with a weighting scheme or a ranking algorithm sensitive to these multiple factors.

References

Jennifer E. Arnold and Zenzi M. Griffin. 2007. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, 56: 521-536.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of LREC Workshop on Linguistic Coreference*, pages 563-566.

Susan Brennan. 1995. Centering attention in discourse. *Language and Cognitive Processes*, 10: 137-167.

Patricia M. Clancy. 1980. Referential choice in English and Japanese narrative discourse. In Wallace L. Chafe, editor, *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Ablex, Norwood, NJ.

Herbert H. Clark and C. J. Sengul. 1979. In search of referents for nouns and pronouns. *Memory and Cognition*, 7(1): 35-41.

Thomas Givón. 1992. The grammar of referential coherence as mental processing instructions. *Linguistics*, 30:5-55.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of ACL 2007*, pages 848–855.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of EMNLP 2009*, pages 1152–1161.

Choonkyu Lee. 2012. Situation model and salience. The LSA 2012 Special Session on Information Structure and Discourse: In Memory of Ellen F. Prince. Portland, Oregon.

Choonkyu Lee and Karin Stromswold. submitted. Situation model and accessibility: Referring expressions in narrative production.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 Shared Task. In *Proceedings of the CoNLL-2011 Shared Task*, pages 28-34.

Rainer Malaka and Alexander Zipf. 2000. Deep Map: Challenging IT research in the framework of a tourist information system. In Daniel R. Fesenmaier, Stefan Klein, and Dimitrios Buhalis, editors, *Information and Communication Technologies in Tourism 2000: Proceedings of the International Conference in Barcelona, Spain*, pages 15-27. Springer, Wien.

Mercer Mayer. 1969. *Frog, Where Are You?* Penguin Books, New York.

Mercer Mayer. 1974. *Frog Goes to Dinner.* Penguin Books, New York.

Mercer Mayer and Marianna Mayer. 1975. *One Frog Too Many.* Penguin Books, New York.

Kathleen F. McCoy and Michael Strube. 1999. Taking time to structure discourse: Pronoun generation beyond accessibility. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*, pages 378-383. Lawrence Erlbaum Associates, Mahwah, NJ.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, pages 197-214. Peter Lang, Frankfurt.

Vincent Ng. 2009. Unsupervised models for coreference resolution. In *Proceedings of EMNLP 2008*, pages 640-649.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of ACL 2010*, pages 1396-1411.

Brendan O'Connor and Michael Heilman. 2011. ARKref is a Noun Phrase Coreference System. Website at http://www.ark.cs.cmu.edu/ARKref/

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov logic. In *Proceedings of EMNLP 2008*, pages 650-659.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of CoNLL 2011*.

Ellen Prince. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223-256. Academic Press, New York.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of EMNLP 2009*, pages 968-977.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*, pages 45-52.

Wietske Vonk, Lettica G. M. M. Hustinx, and Wim H. G. Simons. 1992. The use of referential expressions in structuring discourse. *Language and Cognitive Processes*, 7(3/4): 301-333.

Rolf A. Zwaan. 1999. Situation models: The mental leap into imagined worlds. *Current Directions in Psychological Science*, 8(1):15-18.

Rolf A. Zwaan and Gabriel A. Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162-185.