Coreference and discourse focus in broad-coverage stimuli

Evan Jaffe, Cory Shain, William Schuler (The Ohio State University)
jaffe.59@osu.edu

Previous work argues that discourse prominence facilitates coreference resolution [1, 6]. However, these studies use constructed stimuli with specific syntactic patterns (e.g. cleft constructions) which could have idiosyncratic frequency confounds. This paper explores the generalizability of a discourse prominence effect on coreference resolution in a broad-coverage, naturalistic analysis, given that behavioral data for naturally occurring, contextualized sentences have been argued to complement experimental evidence [4]. In particular, the current work proposes several new estimators of prominence appropriate for broad-coverage sentence processing and evaluates them as predictors of reading behavior in the Natural Stories corpus [7], which includes self-paced reading (SPR) times for ten "constructed-natural" narratives from 181 subjects.

The experimental setup consists of a linear mixed-effects regression (LMER) model fitting reading times to predictors of interest. For each predictor of interest, a likelihood ratio test is performed to compare the baseline model to a new model including the baseline predictors plus the predictor of interest. Following previous work on the Natural Stories corpus, baseline predictors of syntactic surprisal [11], n-gram surprisal [9, 12], and word length are included. Similar to [2, 3], an additional baseline predictor for order effects is added, called *Story Position*, calculated as the proportional sentence position in the entire narrative, ranging from 0 to 1.

After filtering outliers and data from inattentive subjects, the data consist of 59,632 reading time events for both anaphoric proforms and anaphoric fully referring words. Data is divided into exploratory and confirmatory partitions prior to running any model fitting to optimize predictors on exploratory data and eliminate the need for multiple trials correction. Coreference annotation consists of references back to the most recent antecedent for each anaphor, largely following OntoNotes 5.0 [13] guidelines for identity coreference, but adding possessive determiners (*her*, *their*, *its*, etc.). Also, the current annotation indicates only the head word of expressions as coreferring (as opposed to a mention span) to better match up with the word-level reading time data.

This work defines several predictors of discourse prominence based on coreference, including distance to antecedent and thematization. Distance to antecedent is measured by number of intervening words or referents (defined as nouns or verbs), generating the *Coref Length Word* and *Coref Length Referent* predictors, respectively. The *Mention Count* predictor quantifies an entity's overall importance in the narrative and is measured as the entity's running count of repeated mentions. *Mention Count* is similar to thematization, which has been measured as the total number of propositions containing the entity [10], but this work generalizes the measure to be a running total in order to model incremental processing effects. *Mention Count* is also somewhat related to topic persistence [8], which is measured as number of clauses to the right that a referent continues an uninterrupted presence as a semantic argument. However, *Mention Count* is a looser definition that allows reference to the entity to be interrupted; it does not assume perfect continuity. Models including variously spilled-over versions of all predictors are also optimized on an exploratory partition of the data in order to model the possibility that processing effects may have variable time course and potentially occur at some temporal distance from the target stimulus [5].

Table 1 shows a significant ($p = .00007$) reading time facilitation for the *MentionCount* predictor spilled-over by 1 word position. Distance-based predictors are not significant predictors of reading time latencies on exploratory data, and therefore are not evaluated on the confirmatory set. These results provide broad-coverage support for the hypothesis that coreference resolution is easier when the target entity is focused by discourse properties, resulting in faster reading times.

	Effect Size (ms)	
Effect	Predictor units	Z
Word Length	2.17	4.23
Syntactic Surprisal	0.36	1.65
5-gram Surprisal	2.34	3.57
Story Position	-19.2	-6.62
MentionCount***	-0.14	-2.81

Table 1: Effect sizes for main and baseline predictors on confirmatory partition of SPR data. The main effect, spilled over MentionCount, is highly significant ($p = .00007$). No p values exist for baseline predictors, which are included in both models input to the likelihood ratio test. Negative effect direction indicates a speed-up in reading times. Z shows the fixed population β value in milliseconds per unit of standard deviation of the predictor. Predictor Units are the β value in milliseconds, rescaled to the original predictors' units. Model includes observations from spilled over anaphors, totaling 59,632 observations. Word Length is measured in characters, Surprisal is measured in bits, and Story Position is the proportion of sentences completed, scaled between 0 and 1. Note that MentionCount ranges from 1-90, so a word referring to an entity with 70 previous mentions is predicted to be read approximately 10ms faster, relative to a singleton mention.

References

[1] A. Almor. Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, 106(4):748–765, October 1999.

[2] Harald Baayen, Shravan Vasishth, Reinhold Kliegl, and Douglas Bates. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94(Supplement C):206 – 234, 2017.

[3] R. Harald Baayen, Jacolien van Rij, Cecile de Cat, and Simon Wood. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In Dirk Speelman, Kris Heylen, and Dirk Geeraerts, editors, *Mixed Effects Regression Models in Linguistics*. Springer, Berlin, 2018.

[4] V. Demberg and F. Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2), 2008.

[5] K. Erlich and K. Rayner. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning & Verbal Behavior*, 22:75–87, 1983.

[6] S. Foraker and B. McElree. The role of prominence in pronoun resolution: Active versus passive representations. *Journal of Memory and Language*, 56(3):357–383, 2007.

[7] R. Futrell, E.Gibson, H. Tily, I. Blank, A. Vishnevetsky, S. Piantadosi, and E. Fedorenko. In *in prep*.

[8] Talmy Givón. Topic continuity in discourse: An introduction. In Talmy Givón, editor, *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, pages 1–41. John Benjamins, Amsterdam, 1983.

[9] I. F. Monsalve, S. L. Frank, and G. Vigliocco. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 398–408, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[10] Charles A. Perfetti and Susan R. Goldman. Thematization and sentence retrieval. *Journal of Verbal Learning and Verbal Behavior*, 13(1):70 – 79, 1974.

[11] C. Shain, M. van Schijndel, R. Futrell, E. Gibson, and W. Schuler. Memory access during incremental sentence processing causes reading time latency. *COLING 2016, workshop on Computational Linguistics for Linguistic Complexity*, 2016.

[12] M. van Schijndel and W. Schuler. Hierarchic syntax improves reading time prediction. In *NAACL 2015*, 2015.

[13] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, El-Bachouti M., Belvin R., and A. Houston. Ontonotes release 5.0., 2013. LDC Catalog No.: LDC2013T19.