Coreference and Focus in Reading Times

Evan Jaffe
Department of Linguistics
The Ohio State University
jaffe.59@osu.edu

Cory Shain
Department of Linguistics
The Ohio State University
shain.3@osu.edu

William Schuler
Department of Linguistics
The Ohio State University
schuler.77@osu.edu

Abstract

This paper presents evidence of a linguistic focus effect on coreference resolution in broad-coverage human sentence processing. While previous work has explored the role of prominence in coreference resolution (Almor, 1999; Foraker and McElree, 2007), 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 this effect on coreference resolution in a broad-coverage analysis. 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 (Futrell, Gibson, Tily, Vishnevetsky, Piantadosi, and Fedorenko, in prep), a collection of "constructed-natural" narratives read by a large number of subjects. Results show a strong facilitation effect for one of these predictors on exploratory data and confirm that it generalizes to held-out data. 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.

1 Introduction

Coreference resolution has often been assumed to incur processing costs due to some form of memory retrieval or search through accessible antecedents, similar to the binding problem for syntactic dependency attachment (Felser, Phillips, and Wagers 2017). This search has been shown to be facilitated by linguistic focus (or prominence or salience) arising from syntactic, pragmatic, semantic, lexical, information structural and other factors. Previous work has investigated the role of linguistic focus in coreference resolution using constructed stimuli (Perfetti and Goldman, 1974; Greene et al., 1992; Almor, 1999; Foraker and McElree, 2007). However, as discussed in Shain et al. (2016), effects found using constructed stimuli often fail to generalize to broad-coverage sentence processing. It is possible that results obtained using constructed stimuli are due in part to (1) information-theoretic factors that such studies rarely control for (e.g. surprisal), (2) limited syntactic coverage, and/or (3) properties of the stimuli themselves that are atypical of naturalistic sentence processing (e.g. overrepresentation of rare constructions, odd semantics, or lack of context).

While previous work (Almor, 1999; Foraker and McElree, 2007) has operationalized prominence or linguistic focus using cleft constructions, such constructions are very rare (Roland et al., 2007) and therefore cannot be relied upon to predict online processing in the broad-coverage setting.

The current work addresses these concerns by deploying novel broad-coverage implementations of focus as predictors of reading times in a large corpus of naturalistic self-paced reading (SPR) by many subjects (Futrell, Gibson, Tily, Vishnevetsky, Piantadosi, and Fedorenko, in prep). Following Shain et al. (2016), the current work evaluates these predictors against a baseline including both n-gram and probabilistic context-free grammar (PCFG) estimates of incremental surprisal. Using this procedure, results show a significant facilitatory effect of predictors relating to linguistic focus on reading time latencies, supporting the hypothesis that focus effects for coreference observed using constructed stimuli do indeed gener-

Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), pages 1–9,
Salt Lake City, Utah, USA, January 7, 2018.

alize to broad-coverage sentence processing.

2 Related Work

The current study draws on two broad areas of investigation in the psycholinguistic literature: (1) the role of linguistic focus in coreference resolution and (2) the use of broad-coverage methods to test models of human sentence processing.

2.1 Linguistic focus and coreference resolution

Linguistic focus directs subjects' attention toward particularly salient or important discourse referents during sentence processing. Studies such as Perfetti and Goldman (1974), Greene et al. (1992), Almor (1999), Foraker and McElree (2007) and Sauermann et al. (2013) have explored the effects of linguistic focus on subjects' processing of coreference.

Greene et al. (1992) offer a model of pronoun resolution within a rich discourse representation that recognizes syntactic, semantic, and pragmatic factors for referent focus. Syntactic factors that can increase focus include clefting (e.g., *It was the **bird** that ate the fruit*), subject vs. object position (e.g., *The **bird** ate the fruit*), predicative vs. prenominal modification (e.g., *the red house is **beautiful***), and the status of nouns introduced as verbal complements vs. nominal compounds (e.g., *The **boat** is located in the boathouse*). Semantic and pragmatic factors include the causal role of a referent, where the perceived causal agent of a verb could be more focused than the verb's other arguments. Additionally, referents more closely related to the topic can increase focus for those referents. The Greene et al. model matches features of each anaphor automatically and in parallel to the features of all the entities in the discourse. If the match of one entity is sufficiently high, the entity is chosen, otherwise resolution is delayed or additional inference might occur.

Almor (1999) argues for a discourse focus effect in a self-paced reading paradigm. For example, Almor uses *it*-clefts to focus the subject: *It was the robin that ate the fruit. The bird seemed quite satisfied;* and *wh*-clefts to focus the object: *What the robin ate was the fruit. The bird seemed quite satisfied.* In a self-paced reading (SPR) experiment, subsequent mentions of focused referents are read more quickly.

Foraker and McElree (2007) use a speed-accuracy tradeoff (SAT) paradigm (Wicklegren, 1977) to explore the relationship between prominence and processing cost. Referents are made more prominent using constructed *it*-cleft stimuli, as in Almor (1999). They find improved accuracy for retrieval of prominent referents but — contrary to Almor (1999) — no effect on access speed.

Sturt and Lombardo (2005) explore the time course of coreference resolution, showing evidence that syntactic structure is available before the end of the utterance, and therefore that coreference decisions are plausibly occurring in an online and incremental way. They find that eye-tracking data for sentences like *The pilot embarrassed Mary and put himself/herself/him/her in a very awkward situation,* show distinct patterns between the reflexive and simple pronoun conditions, indicating that syntactic structure is available and influencing processing even before the end of the sentence. Findings like these motivate our use of SPR as a measure of incremental processing difficulty in coreference resolution.

While the present study relies on the aforementioned approaches in operationalizing focus, it extends earlier work by using coreference-based focus predictors in broad-coverage naturalistic reading and in so doing explores implementations of focus that are better adapted to broad-coverage analysis.

2.2 Broad-coverage investigation of human sentence processing

As discussed in Section 1, naturalistic stimuli have an advantage over task-specific constructed stimuli in terms of ecological validity. Several previous studies have investigated sentence processing using naturalistic stimuli. This work typically uses linear mixed-effects modeling (LME) to regress variables of interest as predictors of some measure of processing difficulty (e.g. reading fixation times). Demberg and Keller (2008) examine syntactic dependency length as a predictor of eye-tracking fixation durations during reading of the newspaper texts contained in the Dundee corpus (Kennedy et al., 2003). They do not replicate the locality effects found in constructed experiments (Gibson, 2000; Grodner and Gibson, 2005) except when the analysis is restricted to certain parts of speech. Frank and Bod (2011) use echo state networks to compare the fit of linear vs. hierarchi-

cal probabilistic language models to eye-tracking fixation durations, finding no significant contribution of hierarchy to model fit. Van Schijndel et al. (2013) implement a measure of memory retrieval cost built on a left-corner parsing strategy and find a significant *facilitation* effect for retrieval cost on the Dundee corpus, such that tokens predicted to require more costly retrieval operations were integrated more quickly during reading.

In all of the aforementioned studies, effects obtained using constructed stimuli do not generalize to naturalistic sentence comprehension. Exceptions exist, however. For example, Shain et al. (2016) show the predicted inhibitory effect of dependency length on reading times in the Natural Stories corpus (also used in the current experiments), and Brennan et al. (2016) and Lopopolo et al. (2017) find increased neural response in certain brain regions[1] to various types of probabilistic language models. To our knowledge, the current work is the first to extend these broad-coverage methods to the study of coreference resolution.

3 Data

The experiments described in this paper use the Natural Stories corpus (Futrell, Gibson, Tily, Vishnevetsky, Piantadosi, and Fedorenko, in prep), which consists of 10 stories with reading times from 181 subjects using a self-paced reading (SPR) paradigm. These stories occupy an intermediary position between isolated constructed examples on the one hand and naturally-occurring text on the other. They are written in order to sound fluent while containing an unusually high proportion of low-frequency words and syntactic constructions which are intended to test the effects of different kinds of memory usage. The corpus contains 485 sentences with 768,023 total events, where an event is one subject reading one word. Reading times exceeding two standard deviations from the subject mean, shorter than 100ms, or longer than 3000ms are excluded as outliers.

For this work, the data is divided into 1/3 development or exploratory and 2/3 test or confirmatory partitions. All main effects are evaluated first on exploratory data, and the optimal main effect (in terms of improvement to model fit over the baseline) is then selected for evaluation on confirmatory data. This data split allows for the optimization of model predictors and parameters on the exploratory set, and eliminates the need for multiple trials correction since only one model is applied to the confirmatory partition.

[1] As measured by fMRI blood oxygen level dependent contrast imaging (BOLD)

3.1 Coreference Annotation

The current work marks all mentions that are coreferential, in contrast to many previous studies of coreference that are restricted to pronominal coreference. This allows the model to be run on all instances of coreference as well as a pronoun-only subset of the data. Due to model convergence issues for the pronoun-only subset, however, reported results are for the larger dataset of all anaphoric expressions, including pronouns and full referring forms.

All words referring to the same entity or subsets of previously mentioned sets of entities are annotated with the sentence and word index of the most recent previous mention of that entity. See Fig. 1 for example annotations. Annotation guidelines largely follow those from the OntoNotes 5.0 corpus (Weischedel et al., 2013) for identity coreference, except that (1) possessive pronouns are included in annotations, and (2) referents are associated with referring words rather than constituent spans. For example, where the OntoNotes guidelines link *a good suggestion* to *it* in the sentence, *She had a good suggestion and it was unanimously accepted*, the current annotation links the referring word, *suggestion* to the anaphor *it*.[2]

The current annotation also adds possessive determiners like *his, her, its*, which are not included in the OntoNotes identity coreference guidelines. For this study, it is assumed that such determiners require some kind of coreference resolution similar to that required for identity coreference. It is possible that a range of coreference types from strict identity coreference to more weakly related bridging anaphora, for example, would involve different processing strategies, but annotations of these distinctions is substantially more complex and left for future work.

[2] Because the reading time data is measured by word, mention spans that include multiple words would be difficult to use. That is, there is no clear procedure for assigning credit for observed latencies to the various predictors that are involved in the span. Essentially, because both the predictors and observed reading times are defined in terms of words, so must be the coreference annotation. Therefore, for multi-word mentions, the referring word is chosen.

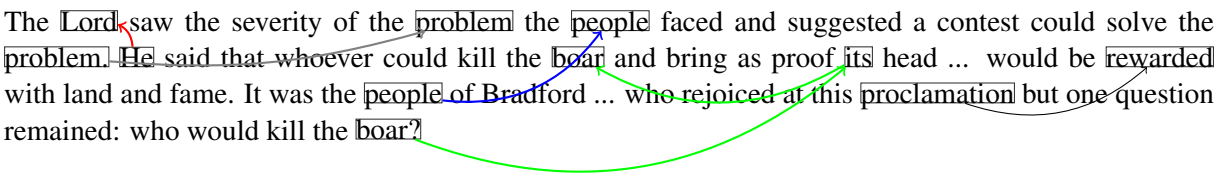

Figure 1: Example coreference annotation. Words in rectangles are linked to the most recent previous mention.

	The	Lord$_i$	saw	the	severity	of	the	problem$_j$	the	people	faced	and	suggested	a	contest	could	solve	the	problem$_j$.
MentionCount	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
WordDistance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
ReferentDistance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**5**
	He$_i$	said	that	whoever	could	kill	the	boar$_k$	and	bring	as	proof	its$_k$	head	would	be	rewarded	with	land	and	fame.
MentionCount	**1**	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
WordDistance	**18**	0	0	0	0	0	0	0	0	0	0	0	**4**	0	0	0	0	0	0	0	0
ReferentDistance	**9**	0	0	0	0	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0	0	0

Table 1: Example predictor values. MentionCount is the number of previous mentions to the same referent. The two other predictors measure distance in words or referents, respectively, back to the antecedent. Words sharing a subscript value are coreferential.

3.2 Baseline Predictors

In order to isolate new effects, it is necessary to statistically control for known effects. These experiments use word length, *n*-gram surprisal, syntactic surprisal, and story position.

Word length is a baseline predictor measured as the number of characters in each word. Longer words are predictive of longer reading times.

Surprisal (Hale, 2001) is the log of the inverse frequency, which increases as the frequency decreases. The log transform makes surprisal a more linear measure of exponential changes in stimulus. The linearity of surprisal is desirable not only because it allows LMER fitting, but because it corresponds with the Weber-Fechner law (Fechner, 1966), which maintains that perception of stimuli increase additively as stimulus strength increases multiplicatively. Stevens' power law (Stevens, 1957) expresses a similar relationship. For word frequencies, which exhibit a Zipfian curve, the log of the probability essentially converts the frequencies to a linear perception curve, allowing easier differentiation of the relative rarity of words that occur exponentially more or less frequently.

Ngram Surprisal controls for conditional word frequency, given preceding words as context, and is a commonly used baseline effect (Monsalve et al., 2012; van Schijndel and Schuler, 2015). 5-gram probability is calculated as the linear combination of most likely *n*-grams up to 5 words long, including the target word. Because longer *n*-grams are often infrequent and thus have poor or non-existent frequency estimates, Kneser-Ney smoothing allows the full sequence to be estimated as an interpolation of shorter *n*-grams. Following Shain et al. (2016), this work uses 5-gram probabilities from the Gigaword 4.0 corpus (Graff and Cieri, 2003) using the KenLM toolkit (Heafield et al., 2013):

$$S(w_i) = -\ log\ P(w_i|w_{i-n}...w_{i-1}) \quad (1)$$

To control for the effect of surprisal due to syntactic context, the current work estimates the probability of syntactic tree structure at each given word (Shain et al., 2016; van Schijndel and Schuler, 2015). Syntactic context is defined as the linear combination of all previous syntactic rule productions up to the current word.

Probabilistic Context-Free Grammar (PCFG) Surprisal follows that used by van Schijndel and Schuler (2015) and comes from an incremental parser (van Schijndel et al., 2013) using the Generalized Categorial Grammar (GCG) framework of Nguyen et al. (2012). Specifically, PCFG surprisal is defined as the sum of negative log probabilities of words given possible trees that span from the first word to the current word. This is analogous to *n*-gram surprisal, but uses hierarchic tree con-

text rather than linear context:

$$S(w_i) = -\ log\ P(T_i = w_i | T_1 ... T_{i-1} = w_1 ... w_{i-1}) \quad (2)$$

where T is a random variable over all trees and $T_1 ... T_i$ are its first i leaf nodes.

Story Position is a measure of progress through the story, where each value is computed as the current sentence index divided by the total number of sentences in the story. For example, the 50th sentence in a 100 sentence story would have a story position of 0.5 for each word in that sentence. This predictor could be interpreted as a percent completion measure that is intended to model order effects due to fatigue, practice or environmental factors, and generally control for a base rate of reading as the story progresses. There is potential for discourse predictability to also be captured with the baseline predictor, analogous to sentence position but generalized to the discourse level, where the space of possible continuations decreases as more information becomes available.

Sentence Position was originally included in the baseline, but was removed as the weakest predictor in order to overcome model convergence issues.

3.3 Broad-coverage implementations of focus

Because naturalistic stimuli in English rarely contain the kinds of constructions used to control linguistic focus in constructed stimulus experiments (Roland et al., 2007), it is necessary to implement focus in some other way. This work explores two types of implementations: frequency-based and recency-based.

The frequency-based implementation, *MentionCount*, is calculated as the running count of mentions in a coreference chain. The first mention has count 0, the subsequent mention count 1, and so on. This measure is closely related to the notion of *thematization* used in Perfetti and Goldman (1974), who also use repetition as an index of focus. As a predictor, *MentionCount* is meant to test the hypothesis that more frequent referents are faster to access. Incidentally, *MentionCount* is quite similar to the measures of *topicality* proposed by Givón (1983), suggesting a potential connection between the discourse notion of topicality and the attendant psychological effects that is left for future research.

The recency-based implementations follow e.g. McElree (2001) in assuming that more recently mentioned entities are more prominent and thus more likely to be remembered better. Specifically, these experiments use two measures of the distance between the current word and the most recent mention of its referent: number of intervening words, and number of intervening discourse referents. Following Gibson (2000) discourse referents are operationalized in the latter option as nouns or verbs, here including pronouns and non-finite verbs. Experiments also evaluate log-transformed versions of each of these distance measures, modeling the possibility of non-linear decay over time in likelihood that linguistic focus for mentioned entities results in processing facilitation.

Table 1 shows example values for the MentionCount and word- and referent-based recency predictors. Log transformed versions of the recency predictors are not shown in this figure. For the first sentence, *problem* is mentioned twice. The first mention has zero previous mentions, while the second has one. Distance in words is 10 between the two mentions, and distance in referents (nouns and verbs) is 5.

4 Statistical evaluation

Each main effect predictor is evaluated on the exploratory data via likelihood ratio test (LRT) of two fitted linear mixed effects (LME) models, one including the main effect as a fixed effect and one excluding it. Both models also contain a set of baseline fixed effects: word length, 5-gram forward surprisal, incremental PCFG surprisal, and story position. All models include all baseline fixed effects. Models also include by-subject random slopes for the main effect and every baseline effect, with the exception of syntactic surprisal, whose by-subject random slopes were removed as the weakest predictor in order to overcome lack of convergence.

Experiments evaluate each main effect over all instances of coreference, as the smaller pronoun-only subset did not converge reliably.

Delays in the time course of processing effects can be modeled by spillover (Erlich and Rayner, 1983), where the effect of an independent variable is predicted to be observed *n* words later. Using standard linear regression on the exploratory dataset, we found the best-fit spillover position of the baseline predictors to be zero (*in situ*) with the exception of PCFG surprisal, which is optimally

	Effect Size (ms)	
Effect	Predictor units	SD
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 2: Effect sizes for main and baseline predictors on confirmatory partition of data. The main effect, spilled over MentionCount, is highly significant ($p = 7.05e-5$). Negative effect direction indicates a speed-up in reading times. SD shows β-effect in milliseconds per unit of standard deviation. Predictor Units are the effect size 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.

spilled over by 1 position. In addition to optimizing the baseline predictors, we consider both *in situ* and spillover-1 variants of each of our main effects.[3]

The reading time measures are transformed following Box and Cox (1964) to match assumptions of normality by the likelihood ratio test. These experiments use a coefficient of λ = -0.63.[4] All predictors are also centered and z-transformed prior to regression.

[3]The reason for choosing a single optimal spillover position for each variable rather than considering multiple spillover positions simultaneously (as in Smith and Levy, 2013, for example) is that our data are too sparse to support such highly parameterized models given that we are controlling for heterogeneity in the population via by-subject random slopes for each independent variable. Since there are 181 subjects in the dataset, each additional independent variable (including each additional modeled spillover position for a given independent variable) contributes 181 additional slopes to estimate.

[4]The effect estimates given in Table 2 are presented in milliseconds for expository purposes. However, this is in fact a back-transformation of β into milliseconds using the equation $\beta\text{-ms} = (\lambda\bar{y}' + \lambda\beta + 1)^{1/\lambda} - (\lambda\bar{y}' + 1)^{1/\lambda}$, where $\bar{y}'$ is the mean of the transformed reading times (1.55 in our data). Because Box and Cox (1964) introduces non-linearity, β-ms is only valid at the back-transformed mean, holding all other effects at their means.

5 Results

MentionCount in the spilled-over position is highly significant on exploratory data. Results for recency-based predictors in the exploratory data partition are extremely weak, and so they are not evaluated on confirmatory data.

Due to the separation of data into exploratory and confirmatory partitions, and subsequent testing on confirmatory data only once, no multiple trials correction is required. Our results are consistent with a general pattern of smaller effect estimates in naturalistic vs. constructed studies of human sentence processing (Demberg and Keller, 2008; Smith and Levy, 2013; van Schijndel and Schuler, 2015; Shain et al., 2016). It might be the case that relatively muted tendencies in naturalistic human sentence processing are exaggerated in artificial settings devoid of conversation context or the implicit intended use of language for communication. The MentionCount values range from 0 to 90, with $\mu = 2.4$ and $SD = 9.3$. The baseline predictors all have plausible effect estimates. The Word length effect is positive, as expected, indicating a slowdown as word length increases. The linear 5-gram and hierarchic syntactic surprisal effects are both positive, indicating that processing difficulty increases with unpredictability of the current token given its context. Story position effect is negative, showing a general decrease in reading times as the story progresses.

As a sanity check, a simpler linear only model (no random effects) was run with the baseline predictors but not MentionCount. Figure 2 presents the residuals mapped to the MentionCount predictor value, showing a slight negative trend that demonstrates that for high values of MentionCount, the baseline's predictions of reading times are too high. This negative correlation between MentionCount and reading times is evident in the full LMER result. Additionally, there is no obvious confound from excessive residuals being due to items at any given MentionCount value.

6 Discussion

These results complement previous work on coreference resolution in constructed stimuli by providing strong evidence of a broad-coverage discourse focus effect on coreference resolution. The implementation of linguistic focus that successfully improved model fit was based on frequency rather than recency of mention. This is a potentially im-

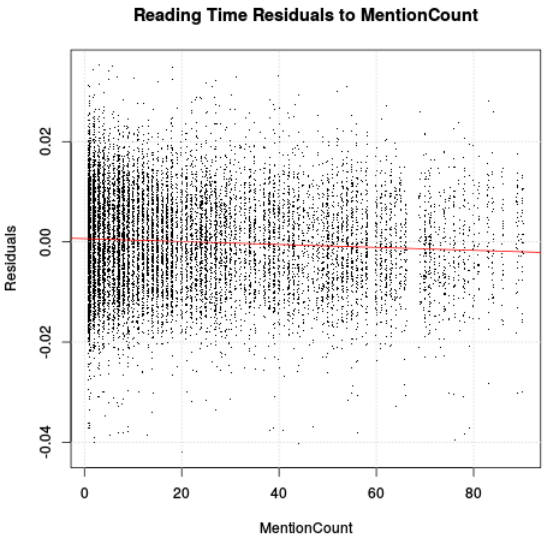

Figure 2: Scatterplot of residuals from simple linear model (no random effects) without MentionCount plotted to spilled-over MentionCount predictor. Fit line shows slight downward trend, indicating main effect of MentionCount to reduce reading times.

portant secondary finding, since recency-based effects were found in syntactic dependency locality effects (Shain et al., 2016). The current negative result for coreference-based recency effects does coincide with related lack of recency effects for syntactic dependencies from Demberg and Keller (2008) (who also used somewhat naturalistic stimuli), and could be attributable to a number of factors. It is possible that a hybrid estimator — taking into account both recency and frequency of mention — might show stronger effects than those presented here. Additionally, since proforms are unlikely to occur at great distance to their antecedents, separating recency effects by anaphor type (full-referring vs. proform) could result in better predictors. Lastly, recency effects might be weak at short to moderate distances where coreference succeeds, but could increase in strength for constructed stimuli where the pronouns are used further from antecedents than is normal, and initial coreference fails, resulting in reanalysis. Of course, these unnatural recency effects would not be detectable or applicable when analyzing naturalistic stimuli.

It is possible that what we have interpreted as a linguistic focus effect is in fact related to surprisal. If subjects are attempting to predict discourse mentions in advance, it is possible that they are reallocating probability mass to mentions of entities as a function of the number of times they have been mentioned in the past, thereby reducing surprisal and facilitating processing of mentions consistent with this prediction. Whether the effect is indeed driven by focus or is instead driven by prediction is also left to future research.

Finally, after considering that high values of MentionCount can only exist toward the end of stories, we considered a potential confound of story position, or relative completion of the story. Story position turns out to be an extremely strong predictor that we argue should be added to future baselines for this type of data. Despite this, spilled-over MentionCount is still highly significant over this more rigorous baseline.

7 Conclusion

This work provides evidence of a linguistic focus effect based on reading time latencies from a coreference-annotated corpus of naturalistic stimuli. Experiments on naturalistic stimuli suggest that mention count is a plausible broad-coverage implementation of linguistic focus and show that more mentions of an entity are correlated with faster reading times.

Acknowledgments

Thank you to four anonymous reviewers for excellent feedback. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1343012, and NSF grant no. 1551313. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Amit Almor. 1999. Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review* 106(4):748–765.

George E. P. Box and David R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* 26(2):211–252. http://www.jstor.org/stable/2984418.

Jonathan R. Brennan, Edward P. Stabler, Sarah E. Van Wagenen, Wen-Ming Luh, and John T.

Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language* 157:81 – 94. https://doi.org/10.1016/j.bandl.2016.04.008.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2):193–210.

Kate Erlich and Keith Rayner. 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning & Verbal Behavior* 22:75–87.

Gustav Theodor Fechner. 1966. *Elements of psychophysics Elemente der Psychophysik*, volume 1. United States of America: Holt, Rinehart and Winston.

Stephani Foraker and Brian McElree. 2007. The role of prominence in pronoun resolution: Active versus passive representations. *Journal of Memory and Language* 56(3):357–383. https://doi.org/10.1016/j.jml.2006.07.004.

Stefan Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science* .

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*. MIT Press, Cambridge, MA, pages 95–126.

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

David Graff and Christopher Cieri. 2003. *English Gigaword LDC2003T05*.

Steven B. Greene, Gail McKoon, and Roger Ratcliff. 1992. Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 18:266–283.

Daniel J. Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input. *Cognitive Science* 29:261–291.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics*. Pittsburgh, PA, pages 159–166.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 690–696.

Alan Kennedy, James Pynte, and Robin Hill. 2003. The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Alessandro Lopopolo, Stefan L. Frank, Antal van den Bosch, and Roel M. Willems. 2017. Using stochastic language models (slm) to map lexical, syntactic, and phonological information processing in the brain. *PLOS ONE* 12(5):1–18. https://doi.org/10.1371/journal.pone.0177794.

Brian McElree. 2001. Working memory and focal attention. *Journal of Experimental Psychology, Learning Memory and Cognition* 27(3):817–835.

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. 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*. Association for Computational Linguistics, Stroudsburg, PA, USA, EACL '12, pages 398–408. http://dl.acm.org/citation.cfm?id=2380816.2380866.

Luan Nguyen, Marten van Schijndel, and William Schuler. 2012. Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*. Mumbai, India, pages 2125–2140.

Charles A. Perfetti and Susan R. Goldman. 1974. Thematization and sentence retrieval. *Journal of Verbal Learning and Verbal Behavior* 13(1):70 – 79. https://doi.org/http://dx.doi.org/10.1016/S0022-5371(74)80032-0.

Douglas Roland, Frederic Dick, and Jeffrey L Elman. 2007. Frequency of basic english grammatical structures: A corpus analysis. *Journal of memory and language* 57 3:348–379.

Antje Sauermann, Ruth Filik, and Kevin B. Paterson. 2013. Processing contextual and lexical cues to focus: Evidence from eye movements in reading. *Language and Cognitive Processes* 28(6):875–903. https://doi.org/10.1080/01690965.2012.668197.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. *COLING 2016, workshop on Computational Linguistics for Linguistic Complexity* .

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128:302–319.

Stanley Smith Stevens. 1957. On the psychophysical law. *Psychological Review* 64(3):153–181.

Patrick Sturt and Vincent Lombardo. 2005. Processing coordinate structures: Incrementality and connectedness. *Cognitive Science* 29:291–305.

Marten van Schijndel, Andy Exley, and William Schuler. 2013. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science* 5(3):522–540.

Marten van Schijndel and William Schuler. 2015. Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL-HLT 2015*. Association for Computational Linguistics.

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. 2013. Ontonotes release 5.0. https://catalog.ldc.upenn.edu/ldc2013t19. LDC Catalog No.: LDC2013T19.

Wayne Wicklegren. 1977. Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica* 41:67–85.