UNC Language Processing Lab, Technical Report #2, 2018

Reference frequency: what do speakers tend to talk about?

Jennifer E. Arnold, University of Chapel Hill North Carolina

Iris Strangmann, CUNY Graduate Center

Heeju Hwang, The University of Hong Kong

Sandy Zerkle, University of Chapel Hill North Carolina

Address all correspondence to: Jennifer E. Arnold, UNC Chapel Hill, Dept. of Psychology and Neuroscience, Davie Hall #337B, CB #3270, Chapel Hill, NC 27599-3270, Email: jarnold@unc.edu

Abstract

This paper presents a re-analysis of a text analysis originally conducted by Arnold (1998, chapter 2). Five written childrens' stories were analyzed to assess the frequency with which speakers tend to re-mention entities that occur in subject vs. nonsubject positions. The analysis is restricted to clauses with at least one other referent besides the subject. Results show that subjects are more likely to be rementioned than nonsubjects.

Keywords: referential likelihood, remention probability, discourse structure

Introduction

Current models of language comprehension suggest that people are guided by their experience with language, and in particular with knowledge about which structures are more probable (Farmer, Monaghan, Misyak, & Christiansen, 2011; Farmer, Fine, Misyak, & Christiansen, 2015 MacDonald, 2013; Montag & MacDonald, 2008; Wells, Christiansen, Race, Acheson, & MacDonald, 2009; Staub, 2011). In this vein, models of reference comprehension also suggest that some references are more probable, given a particular discourse context, and that this probability guides reference comprehension (Arnold, 1998, 2001; Kehler, Kertz, Rohde, & Elman, 2008; Kehler & Rohde, 2013; Frank & Goodman, 2012). For reference, models suggest that comprehenders are sensitive to the likelihood of the reference itself, independently of the probability of the words per se.

In this paper, we ask whether referential probability is associated with grammatical functions, i.e. subjects vs. objects/obliques. It is well established that the subject is a topical position (Chafe, 1976, 19984; Givon, 1983), and topicality has been linked with predictability (Givon, 1983; Prince, 1981). Does this mean that references in subject position are associated with a higher probability that the referent will be rementioned?

Arnold (1998, 2001) provided evidence in favor of this view. She analyzed written and spoken texts, and examined the frequency with which referents were mentioned in the following clause. She found that speakers and writers tend to frequently re-mention entities that occur in subject position. For example, Arnold (1998, chapter 2) analyzed the texts from written children's stories, and counted the number of references that coreferred with either the subject or an object/oblique from the previous clause. Out of 271 references to something in the previous

clause, 201 coreferred with the subject, and 70 with an object or oblique referent. This suggests that there is a striking bias for speakers to re-mention entities that occur in subject position.

This pattern is not surprising, given proposals about how speakers decide on the subject for a sentence. The subject is considered the "starting point" (Chafe, 1994). Brennan (1995) suggests that speakers puts a referent in subject position, they signal that their attention is on that referent. Thus, the subject position is reserved for entities that are in the speaker's focus of attention. If speakers tend to continue talking about things that they consider topical or worthy of attention, they should tend to re-mention subjects.

Note that this explanation does not suggest that the grammatical subject position is a causal factor in what people talk about. It seems unlikely that a speaker says "Sue ate the pizza" and then thinks "Oh look, I just mentioned Sue in subject position, therefore I better keep talking about her." Instead, the speaker's choices are driven by their communicative intentions -- i.e., what they want to communicate. These intentions drive the construction of nonlinguistic messages (Levelt, 1989), which are then turned into linguistic expressions. Thus, the tendency to re-mention subjects is a function of the tendency for speakers to stick with a topic for multiple utterances, which results in multiple mentions of a single entity.

However, there is an alternate explanation for why Arnold (1998) observed a tendency for subjects to be rementioned (although note that this explanation does not hold for the finding from Arnold, 2001). This analysis did not control for the number of entities mentioned in an utterance. Some utterances had multiple entities (e.g., "Dorothy lived in the midst of the great Kansas prairies, with Uncle Henry, who was a farmer, and Aunt Em, who was the farmer's wife,"). In this case, the statistical chance that the subject might be re-mentioned was low. On the other hand, many utterances only had one entity, which was always in subject position

("Their house was small"). If speakers are equally likely to mention all entities in an utterance, the fact there are some sentences with only subjects might result in a higher rate of subject rementioning overall. Thus, it might not be surprising that the subject would be rementioned, because the subject is the only obligatory argument in an utterance.

In the current paper, we re-analyze the text data from Arnold (1998), limiting our analysis to utterances that have 2 or more entities, and taking a closer look at other potential confounding factors. Using the same dataset as Arnold (1998), we examined the text from five children's stories, counting the frequency with which the subject was re-mentioed, compared to object/oblique arguments. For example, one sentence was "Now one night Mr. and Mrs. Whittaker went to a church sociable in the village." This contains four referents (Mr. and Mrs. Whittaker (subject); church sociable; village; one night). We examined the next sentence to see which of these referents was mentioned. In this case only the subject was re-mentioned: "And just as usual, they didn't even bother [to lock the door"].

From a theoretical standpoint, we care about the frequency of reference continuation because it is one measure of referential probability. Several models suggest that language users are sensitive to frequency patterns in their language (Arnold, 1998; MacDonald, 2013; Tily & Piantadosi, 2011; Kehler & Rohde, 2013). Thus, people might learn from exposure to language that some types of entities are more likely to be mentioned than others. For example, subject referents may be more likely to be mentioned than referents in other positions.

However, it is unclear how comprehenders calculate probability. One possibility is that people track estimates at a very fine grain. E.g., following our example sentence above, people might estimate the specific probability that each referent might be mentioned in subject position of the following clause. Such a prediction might also be influenced by factors like parallelism,

which would lead to a stronger expectation for the subject to be mentioned. On the other hand, people might make a very general calculation, such as "the Whittakers were the topic of the last sentence, so they are likely to be mentioned again sometime soon at some point in the story."

At this point, we do not know how peole calculate referential frequency, if indeed they do. Therefore we decided to use the following clause as a sample of referential continuation. If the same referent is mentioned in the following clause, it provides good evidence of referential probability. We examined all references in the following clause, and not just the subject, to provide the broadest estimate of reference frequency.

It is also likely that language users are sensitive to the fact that discourse entities can be either directly mentioned, or indirectly evoked. For example, once a building is mentioned, its door (for example) becomes inferable (Prince, 1992), and mention of a part of the whole still functions to connect to the earlier discourse. We therefore consider both direct references and indirect ones.

2. Analysis of children's books

2.1 Methods

Using the corpus from Arnold (1998), we analyzed the full texts from the five children's books shown in Table 1. We first identified independent clauses as the focus our analysis. An independent clause is a tensed clause, not including relative clauses, sentential complements, or other untensed descriptive clauses (e.g., "fixing the tire"). We included all main clauses (e.g., *He lives in the Whittaker house*), and tensed subordinate clauses that began with while, although, if, because etc, for example *because it was even older than the rocking chairs and more antique*.

We excluded quotes, standalone noun-phrase expressions, interjections, questions, and the last utterance in each text.

We then limited our analysis to only clauses with two or more references, where a reference is defined as a noun phrase with a referent. For example, *But the robbers went so fast,* mentions only the subject, so would not be included in the analysis. We included clauses that had a dummy "there" or "it", but excluded these from our counts for the subject. The analysis included a total of 345 clauses. Of these, the majority (258) had only one other reference besides the subject, but 87 had 2 or 3 nonsubject NPs.

Table 1. Texts used in the analysis of children's books

Title	Author	# clauses	# analyzed
Georgie and the Robbers	Robert Bright	113	60
The Hodag	Caroline Arnold	250	76
The Wonderful Wizard of Oz, chapter 1	L. Frank Baum	119	76
Peter Pan chapter 2	J. M. Barrie	132	55
Sylvester and the Magic Pebble	William Steig	173	78

For simplicity, we also limited our analysis to only arguments in the independent clause itself, and not any sentential arguments or complements. For example, in *And just as usual, they didn't even bother [to lock the door],* the object of "bother" is a sentential complement. This was excluded from our analysis, so this clause only had a single reference "they", and thus was not coded. In theory, people might also track reference probability for embedded clauses, but the current dataset was not large enough to distinguish the effects of main vs. embedded subjects, so all embedded clauses were excluded. However, for the purpose of deciding whether a referent

was mentioned in the next clause, we included the entire next independent clause and all of its embedded phrases.

Thus, for each clause we counted each noun phrase referent, and tallied whether it was mentioned in the next clause or not. If the clause was followed by a quote, we ignored the quote and considered the following clause to judge next mentions. It was possible for a referent to be mentioned both directly and indirectly. A direct reference was either a repeat of the name or description, a pronoun, or a different expression to refer to the same entity. We also included indirect references, as shown in Table 2.

Table 2. Examples of indirect reference

type of indirect reference	Example
Whole-part	• The Whittaker house… the stairs • They…their teeth • The whole forest… one side
Part-whole	• The Hodag's tail…the Hodag
Possessive	• Dorothy…their house

Another open question is how to treat coordinated sentences. For example, consider "They drove a fast truck and wore maks". This sentence includes two clauses: 1) *They drove a fast truck*, and 2) *and wore masks*. The subject of the second clause is not mentioned explicitly, due to the fact that this clause is syntactically coordinated with the first clause. At one level, examples like this might provide evidence to readers that the subject is likely to be involved in

the subsequent discourse, because the writer continues discussing the actions of the subject character. On the other hand, this structure is limited to situations where the subject is continued, which means that including them would necessarily increase the rate of subjects being re-mentioned. In the results reported here, we adopt the more conservative approach, which is to exclude clauses that are followed by a coordinated clause like this one (n=37). Note that in the end this choice is irrelevant, because the same basic pattern of results obtains if we include coordinated clauses.

2.2 Results

Here we report the frequency with which subjects and nonsubjects (objects or oblique references) are rementioned in the following clause. We report the results in two ways. First, we look at the total frequency of mentioning a subject or nonsubject, out of the total number of in each category. Given that clauses can have more than one nonsubject, the total N is higher for nonsubjects than subjects in this analysis. This analysis reflects the calculation "given that I have seen a reference in subject/nonsubject position, what is the likelihood that it will be mentioned again"? The second analysis instead considers the clause as a unit, asking whether one or more nonsubjects are mentioned in the following clause. This analysis instead reflects the calculation "given that I have seen this clause, is the author more likely to re-mention the subject or one of the other entities?" In all cases we assess the significance of re-mention frequencies with a chi-square test of independence.

Analysis 1: Frequency of re-mention out of all mentions

As shown in table 3, authors were relatively more likely to mention the referents that occurred in subject position than those that were mentioned in object or oblique position. If we

compare full or indirect remention with no remention, the difference between subjects and nonsubjects is significant (χ^2= 20.0, p < .001). Likewise, if we only examine full re-mentions (categorizing indirect rementions with the no-rementions), the difference is still significant (χ^2= 25.07, p < .001).

Table 3. Frequency of re-mention for subjects vs. objects/oblique references, considering the sample of all references.

	full remention	full or indirect remention	total N
subject	37%	46%	303
object or oblique	20%	30%	400

Analysis 2: Frequency of re-mention for each clause

Table 4 reports the frequency of re-mentioning the subject vs. the frequency of re-mentioning another referent in the clause, considering the clause as a unit. The total N for subjects is lower than that for objects/obliques, because the 5 clauses with a dummy subject are excluded from this count. This analysis is the most conservative, because many clauses have more than one nonsubject reference. Here we count whether any one of the nonsubject references is rementioned, which should make it statistically more likely that the next clause would mention a nonsubject than a subject. Nevertheless, we still see that subjects are more likely to be mentioned than nonsubjects. This difference is significant (χ^2=7.4, p = .007).

Table 4. Frequency of re-mention for subjects vs. objects/oblique references, considering each clause as a unit

	full or indirect remention	total N
subject	46%	303
object or oblique	35%	308

4. General Discussion

The main finding from the text analysis presented here is that reference is systematic: writers tend to continue referring to some things more than others, and these patterns are correlated with grammatical function. If the writer mentions something in subject position, it is more likely to be mentioned again than something that is mentioned in nonsubject position.

The analysis presented here is conservative, because it considers only clauses with more than one reference, and only clauses not followed by a conjoined clause. It also specifically examines the frequency of re-mention, as opposed to analyzing references to understand the characteristics of their antecedents. In this way, it goes beyond the original analysis presented in Arnold (1998).

On the other hand, a limitation of the current text analysis is that only tests one type of language, namely written fictional stories. Nevertheless, there is reason to expect a similar pattern to occur in other genres. Arnold (2001) reports an analysis of the Aligned Hansard corpus, which is a transcript of Canadian parliament. This analysis focused on transfer verbs (e.g., *give, receive*), which included at least three arguments (the goal, them, and source). In this sample, subjects were also more likely to be rementioned than nonsubjects. Similarly, subjects are considered to have high persistence, meaning that they tend to be continued for more clauses in the following discourse than other entities (Givon, 1983).

Yet some variability by verb type is also to be expected. Rohde & Kehler (2014; see also Kehler & Rohde, 2013; Kehler et al., 2008) use a sentence-completion paradigm to examine implicit causality verbs. For example, in *Liz amazed Ana because…,* people tend to think that

Liz was the cause of the event, where in *Liz admired Ana because*…., people tend to think that Ana was the cause. When participants completed these sentence fragments, they were more likely to mention the implicit cause, regardless of whether it was in subject or object position. This led Rohde & Kehler to conclude that next-mention biases are unrelated to grammatical position, at least for this verbtype. Note that their predictions are specific to situations where the following clause provides an explanation of the first clause, which is constrained by the use of the word "because". This causal coherence relation is expected to influence expectations for who will be mentioned next.

The sentence-continuation paradigm may be problematic (see Zerkle & Arnold, 2018), but this idea also has some support in a corpus analysis. Arnold & Weatherford (under revision) examined implicit-causality verbs in the Fisher corpus of spoken conversations. In contrast with Rohde & Kehler (2014), they found that there was no preference for implicit causes to be frequently mentioned again. However, neither did they find that subjects tended to be mentioned again in this sample. This analysis also included a sample of transfer verbs, and here (in contrast with Arnold, 2001), they also did not find a tendency for subjects to be continued.

The take-home story from these different studies is that verb type and coherence relation may matter. However, in many cases the subject position is correlated with a high likelihood of remention, as in the current study. This raises the possibility that these frequencies affect the way people represent and process language. While either reading or listening, a comprehender can anticipate that recently mentioned entities – both subjects and nonsubjects – are relatively likely to be mentioned again, or be indirectly related to subsequent references. On top of a general expectation for given information, comprehenders may expect remention of subjects relatively more than nonsubjects.

One thing we do not know is how fine-grained these expectations are. The text analysis presented here includes a wide variety of verbs and coherence relations. This suggests that overall, subjects are frequently re-mentioned. On the basis of this, people may calculate a general expectation for subject re-mention, and generalize it to other verbs and other situations. Atlernatively, people may track fine-grained distinctions between genre types, tenses, coherence relations, and verbs. Additional research is needed to demonstrate whether the frequencies in texts are related to the actual expectations of comprehenders, and if so, whether they are calculated based on broad patterns or fine-grained ones.

Acknowledgements

This research was funded by NSF Grant 1651000 to J. Arnold

References

Arnold, J. E. (1998). *Reference form and discourse patterns* (Doctoral Dissertation). Stanford University.

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference Continuation. *Discourse Processes*, *31*(2), 137–162.

Brennan, S. E. (1995). Centering attention in discourse. *Language And Cognitive Processes*, *10*(2), 137-167. doi:10.1080/01690969508407091

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (Ed.), *Subject and Topic*, Presented at the Symposium on Subject and Topic, Univ. of California, University of California.

Chafe, W. (1994). *Discourse, consciousness, and time*. Chicago: Chicago University Press.

Farmer, T. A., Fine, A. B., Misyak, J. B., & Christiansen, M. H. (2017). Reading span task performance, linguistic experience, and the processing of unexpected syntactic events. *The Quarterly Journal of Experimental Psychology*, *70*(3), 413-433. doi:10.1080/17470218.2015.1131310

Farmer, T. A., Monaghan, P., Misyak, J. B., & Christiansen, M. H. (2011). Phonological typicality influences sentence processing in predictive contexts: reply to Staub, Grant, Clifton, and Rayner (2009). *Journal of Experimental Psycholog: Learning, Memory, and Cognition*, *37*(5), 1318–1325.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*(6084), 998. doi: 10.1126/science.1218633

Givón, T. (1983). Topic continuity in discourse: the functional domain of switch reference. *Typological Studies in Language*, *2*, 51–82.

Kehler, A., & Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, *39*(1–2), 1–37.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*(1), 1–44.

Levelt, W. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA, US: The MIT Press.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 1–16.

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in eight and twelve year old children and adults. *Journal of Experimental Psychology: General*, *144*, 447–468.

Prince, E. (1981). *Toward a taxonomy of given-new information*. In P.Cole, (Ed.). Radical Pragmatics (pp. 223-256). NY: Academic Press.

Prince, E. F. (1992). The ZPG letter: Subjects, definiteness, and iInformation-status. *Discourse description: diverse linguistic analyses of a fund-raising text*., ed. by W. C. Mann & S. A. Thompson, 295-325. Amsterdam: John Benjamins.

Rohde, H., & Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, *29*, 912-927. doi:10.1080/01690965.2013.854918

Staub, A. (2011). The effect of lexical predictability on distributions of eye fixation durations. Psychonomic Bulletin & Review, 18, 371-376.

Tily, H., & Piantadosi, S. (2009). Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference*. Amsterdam, The Netherlands.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*(2), 250–271.