Why discourse affects speakers' choice of referring expressions

Naho Orita
Graduate School of Information Sciences
Tohoku University
naho@ecei.tohoku.ac.jp

Eliana Vornov
Computer Science and Linguistics
University of Maryland
evornov@umd.edu

Naomi H. Feldman
Linguistics and UMIACS
University of Maryland
nhf@umd.edu

Hal Daumé III
Computer Science and UMIACS
University of Maryland
hal@umiacs.umd.edu

Abstract

We propose a language production model that uses dynamic discourse information to account for speakers' choices of referring expressions. Our model extends previous rational speech act models (Frank and Goodman, 2012) to more naturally distributed linguistic data, instead of assuming a controlled experimental setting. Simulations show a close match between speakers' utterances and model predictions, indicating that speakers' behavior can be modeled in a principled way by considering the probabilities of referents in the discourse and the information conveyed by each word.

1 Introduction

Discourse information plays an important role in various aspects of linguistic processing, such as predictions about upcoming words (Nieuwland and Van Berkum, 2006) and scalar implicature processing (Breheny et al., 2006). The relationship between discourse information and speakers' choices of referring expression is one of the most studied problems. Speakers' choices of referring expressions have long been thought to depend on the salience of entities in the discourse (Givón, 1983). For example, speakers normally do not choose a pronoun to refer to a new entity in the discourse, but are more likely to use pronouns for referents that have been referred to earlier in the discourse. A number of grammatical, semantic, and distributional factors related to salience have been found to influence choices of referring expressions (Arnold, 2008). While the relationship between discourse salience and speakers' choices of referring expressions is well known, there is not yet a formal account of why this relationship exists.

In recent years, a number of formal models have been proposed to capture inferences between speakers and listeners in the context of Gricean pragmatics (Grice, 1975; Frank and Goodman, 2012). These models take a game theoretic approach in which speakers optimize productions to convey information for listeners, and listeners infer meaning based on speakers' likely productions. These models have been argued to account for human communication (Jager, 2007; Frank and Goodman, 2012; Bergen et al., 2012a; Smith et al., 2013), and studies report that they robustly predict various linguistic phenomena in experimental settings (Goodman and Stuhlmüller, 2013; Degen et al., 2013; Kao et al., 2014; Nordmeyer and Frank, 2014). However, these models have not yet been applied to language produced outside of the laboratory, nor have they incorporated measures of discourse salience that can be computed over corpora.

In this paper, we propose a probabilistic model to explain speakers' choices of referring expressions based on discourse salience. Our model extends the rational speech act model from Frank and Goodman (2012) to incorporate updates to listeners' beliefs as discourse proceeds. The model predicts that a speaker's choice of referring expressions should depend directly on the amount of information that each word carries in the discourse. Simulations probe the contribution of each model component and show that the model can predict

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1639–1649,
Beijing, China, July 26-31, 2015.

speakers' pronominalization in a corpus. These results suggest that this model formalizes underlying principles that account for speakers' choices of referring expressions.

The paper is organized as follows. Section 2 reviews relevant studies on choices of referring expressions. Section 3 describes the details of our model. Section 4 describes the data, preprocessing and annotation procedure. Section 5 presents simulation results. Section 6 summarizes this study and discusses implications and future directions.

2 Relevant Work

2.1 Discourse salience

Speakers' choices of referring expressions have long been an object of study. Pronominalization has been examined particularly often in both theoretical and experimental studies. Discourse theories predict that speakers use pronouns when they think that a referent is salient in the discourse (Givón, 1983; Ariel, 1990; Gundel et al., 1993; Grosz et al., 1995), where salience of the referent is influenced by various factors such as grammatical position (Brennan, 1995), recency (Chafe, 1994), topicality (Arnold, 1998), competitors (Fukumura et al., 2011), visual salience (Vogels et al., 2013b), and so on.

Discourse theories have characterized the link between referring expressions and discourse salience by stipulating constructs such as a scale of topicality (Givón, 1983), accessibility hierarchy (Ariel, 1990), or implicational hierarchy (Gundel et al., 1993). All of these assume fixed form-salience correspondences in that a certain referring expression encodes a certain degree of salience. However, it is not clear how this form-salience mapping holds nor why it should be.

There is also a rich body of research that points to the importance of production cost (Rohde et al., 2012; Bergen et al., 2012b; Degen et al., 2013) and listener models (Bard et al., 2004; Van der Wege, 2009; Galati and Brennan, 2010; Fukumura and van Gompel, 2012) in language production. These studies suggest that only considering discourse salience of the referent may not precisely capture speakers' choices of referring expressions, and it is necessary to examine discourse salience in relation to these other factors.

2.2 Formal models

Computational models relevant to speakers' choices of referring expressions have been proposed, but there is a gap between questions that previous models have addressed and the questions that we have raised above.

Grüning and Kibrik (2005) and Khudyakova et al. (2011) examine the significance of various factors that might influence choices of referring expressions by using machine learning models such as neural networks, logistic regression and decision trees. Although these models qualitatively show some significant factors, they are data-driven rather than being explanatory, and have not focused on why and how these factors result in the observed referring choices.

Formal models that go beyond identifying superficial factors focus on only pronouns rather than accounting for speakers' word choices per se. For example, Kehler et al. (2008) formalize a relationship between pronoun comprehension and production using Bayes' rule to account for comprehender's semantic bias in experimental data. Rij et al. (2013) use ACT-R (Anderson, 2007) to examine the effects of working memory load in pronoun interpretation. These models show how certain factors influence pronoun production/interpretation, but it is not clear how these models would predict speakers' choices of referring expressions.

Relevant formal models in computational linguistics include Centering theory (Grosz et al., 1995; Poesio et al., 2004) and Referring Expression Generation (Krahmer and Van Deemter, 2012). These models propose deterministic constraints governing when pronouns are preferred in local discourse, but it is not clear how these would account for speakers' choices of referring expressions, nor it is clear why there should be such deterministic constraints.

2.3 Uniform Information Density

One potential formal explanation for the relation between discourse salience and speakers' choices of referring expressions is the Uniform Information Density hypothesis (UID) (Levy and Jaeger, 2007; Tily and Piantadosi, 2009; Jaeger, 2010). UID states that speakers prefer to smooth the information density distribution of their utterances over time to achieve optimal communication. This theory predicts that speakers should use pronouns instead of longer forms (e.g., *the president*) when a

referent is predictable in the context, whereas they should use longer forms for unpredictable referents that carry more information (Jaeger, 2010).

Tily and Piantadosi (2009) empirically examined the relationship between predictability of a referent and choice of referring expressions. They found that predictability is a significant predictor of writers' choices of referring expressions, in that pronouns are used when a referent is predictable.

While these results appear to support UID, there are several inconsistencies with previous UID accounts. Information content of words has been estimated using an n-gram language model (Levy and Jaeger, 2007), a verb's subcategorization frequency (Jaeger, 2010), and so on, whereas here the information content is that of referents with respect to discourse salience. In addition, selecting between a pronoun and a more specified referring expression involves deciding how much information to convey, whereas previous applications of UID (Levy and Jaeger, 2007) have been concerned with deciding between different ways of expressing the same information content. We show in the next section that we can derive predictions about referring expressions directly from a model of language production.

2.4 Summary

Previous linguistic studies have focused on identifying factors that might influence choices of referring expressions. However, it is not clear from this previous work how and why these factors result in the observed patterns of referring expressions. Where formal models relevant to this topic do exist, they have not been built to explain why there is a relation between discourse salience and speakers' choices of referring expressions. Even UID, which relates predictability to word length, is not set up to account for the choice between words that vary in their information content.

In the next section, we propose a speaker model that formalizes the relation between discourse salience and speakers' choices of referring expressions, considering production cost and speakers' inference about listeners in a principled and explanatory way.

3 Speaker model

3.1 Rational speaker-listener model

We adopt the rational speaker-listener model from Frank and Goodman (2012) and extend this model to predict speakers' choices of referring expressions using discourse information.

The main idea of Frank and Goodman's model is that a rational pragmatic listener uses Bayesian inference to infer the speaker's intended referent r_s given the word w, their vocabulary (e.g., 'blue', 'circle'), and shared context that consists of a set of objects O (e.g., visual access to object referents) as in (1), assuming that a speaker has chosen the word informatively.

$$P(r_s|w,O) = \frac{P_S(w|r_s,O)P(r_s)}{\Sigma_{r' \in O} P(w|r',O)P(r')} \quad (1)$$

While our work does not make use of this pragmatic listener, it does build on the speaker model assumed by the pragmatic listener. This speaker model (the likelihood term in the listener model) is defined using an exponentiated utility function as in (2).

$$P_S(w|r_s,O) \propto e^{\alpha U(w;r_s,O)} \quad (2)$$

The utility $U(w;r_s,O)$ is defined as $I(w;r_s,O) - D(w)$, where $I(w;r_s,O)$ represents informativeness of word w (quantified as surprisal) and $D(w)$ represents its speech cost. If a listener interprets word w literally and cost $D(w)$ is constant, the exponentiated utility function can be reduced to (3) where $|w|$ denotes the number of referents that the word w can be used to refer to.

$$P_S(w|r_s,O) \propto \frac{1}{|w|} \quad (3)$$

Thus, the speaker model chooses a word based on its specificity. We show in the next section that this corresponds to a speaker who is optimizing informativeness for a listener with uniform beliefs about what will be referred to in the discourse. The assumption of uniform discourse salience works well in a simple language game where there are a limited number of referents that have roughly equal salience, but we show that a model that lacks a sophisticated notion of discourse falls short in more realistic settings.

3.2 Incorporating discourse salience

To extend Frank and Goodman's model to a natural linguistic situation, we assume that the speaker estimates the listener's interpretation of a word (or referring expression) w based on discourse information. We extend the speaker model from (3) by assuming that a speaker S chooses w to optimize a listener's belief in speaker's intended referent r relative to the speaker's own speech cost C_w. This cost

is another factor in the speaker model, roughly corresponding to utterance complexity such as word length.[1]

$$P_S(w|r) \propto P_L(r|w) \cdot \frac{1}{C_w} \quad (4)$$

The term $P_L(r|w)$ in (4) represents informativeness of word w: the speaker chooses w that most helps a listener L to infer referent r. The term C_w in (4) is a cost function: the speaker chooses w that is least costly to speak.

The speaker's listener model, $P_L(r|w)$, infers referent r that is referred to by word w according to Bayes' rule as in (5).

$$P_L(r|w) = \frac{P(w|r)P(r)}{\Sigma_{r'} P(w|r')P(r')} \quad (5)$$

The first term in the numerator, $P(w|r)$, is a word probability: the listener in the speaker's mind guesses how likely the speaker would be to use w to refer to r. The second term in the numerator, $P(r)$, is the discourse salience (or predictability) of referent r. The denominator $\Sigma_{r'} P(w|r')P(r')$ is a sum of potential referents r' that could be referred to by word w. The terms in this sum are non-zero only for referents that are compatible with the meaning of the word. If there are many potential referents that could be referred to by word w, that word would be more ambiguous thus less informative. The whole of the right side in Equation (5) represents the speaker's assumption about the listener: given word w the listener would infer referent r that is salient in a discourse and less ambiguously referred to by word w.

If $P(r)$ is uniform over referents and $P(w|r)$ is constant across words and referents, this listener model reduces to $\frac{1}{|w|}$. Thus, Frank and Goodman (2012)'s speaker model in (3) is a special case of our speaker model in (4) that assumes uniform discourse salience and constant cost.

Our model predicts that the speaker's probability of choosing a word for a given referent should depend on its cost relative to its information content. To see this, we combine (4) and (5), yielding

$$P_S(w|r) \propto \frac{P(w|r)P(r)}{\sum_{r'} P(w|r')P(r')} \cdot \frac{1}{C_w} \quad (6)$$

Because the speaker is deciding what word to use for an intended referent, and the term $P(r)$ denotes the probability of this referent, $P(r)$ is constant in the speaker model and does not affect the relative probability of a speaker producing different words. We further assume for simplicity that $P(w|r)$ is constant across words and referents. This means that all referents have about the same number of words that can be used to refer to them, and that all words for a given referent are equally probable for a naive listener. In this scenario, the speaker's probability of choosing a word is

$$P_S(w|r) \propto \frac{1}{\sum_{r'} P(r')} \cdot \frac{1}{C_w} \quad (7)$$

where the sum denotes the total discourse probability of the referents referred to by that word.

The information content of an event is defined as the negative log probability of that event. In this scenario, the information conveyed by a word is the logarithm of the first term in (7), $-\log \sum_{r'} P(r')$. This means that in deciding which word to use, the highest cost a speaker should be willing to pay for a word should depend directly on that word's information content.

This relationship between cost and information content allows us to derive the prediction tested by Tily and Piantadosi (2009) that the use of referring expressions should depend on the predictability of a referent. For referents that are highly predictable from the discourse, different referring expressions (e.g., pronouns and proper names) will have roughly equal information content, and speakers should choose the referring expression that has the lowest cost. In contrast, for less predictable referents, proper names will carry substantially more information than pronouns, leading speakers to pay a higher cost for the proper names. These are the same predictions that have been discussed in the context of UID, but here the predictions are derived from a principled model of speakers who are trying to provide information to listeners. The extent to which our model can also capture other cases that have been put forward as evidence for the UID hypothesis remains a question for future research.

3.3 Predicting behavior from corpora

The model described in Section 3.2 is fully general, applying to arbitrary word choices, discourse probabilities, and cost fuctions. As an initial step, our simulations focus on the choice between pronouns and proper names. Our work tests the speaker model from (4) directly, asking whether it can predict the referring expressions from corpora of writ-

[1] Our speaker model corresponds to Frank and Goodman's exponentiated utility function (2), with α equal to one and with their cost $D(w)$ being the log of our cost C_w.

ten and spoken language. Implementing the model requires computing word probabilities $P(w|r)$, discourse salience $P(r)$, and word costs C_w.

We simplify the word probability $P(w|r)$ in the speaker's listener model as in (8):

$$P(w|r) = \frac{1}{V} \tag{8}$$

where the count V is the number of words that can refer to referent r. We assume that V is constant across all referents. Our reasoning is as follows. There could be many ways to refer to a single entity. For example, to refer to entity *Barack Obama*, we could say 'he', 'The U.S. president', 'Barack', and so on. We assume that there are the same number of referring expressions for each entity and that each referring expression is equally probable under the listener's likelihood model.

In our simulations, we assume that a speaker is choosing between a proper name and a pronoun. For example, we assume that an entity *Barack Obama* has one and only one proper name 'Barack Obama', and this entity is unambiguously associated with male and singular. Although we use an example with two possible referring expressions, as long as $P(w|r)$ is constant across all referents and words, it does not make a difference to the computation in (5) how many competing words we assume for each referent.

To estimate the salience of a referent, $P(r)$, our framework employs factors such as referent frequency or recency. Although there are other important factors such as topicality of the referent (Orita et al., 2014) that are not incorporated in our simulations, this model sets up a framework to test the role and interaction of various potential factors suggested in the discourse literature.

Salience of the referent is computed differently depending on its information status: old or new. The following illustrates the speaker's assumptions about the listener's discourse model:

For each referent $r \in [1, R_d]$:

1. If $r = old$, choose r in proportion to N_r (the number of times referent r has been referred to in the preceding discourse).
2. Otherwise, $r = new$ with probability proportional to α (a hyperparameter that controls how likely the speaker is to refer to a new referent).
3. If $r = new$, sample that new referent r from the base distribution over entities with probability $\frac{1}{U_\cdot}$ (count $U_\cdot$ denotes a total number of unseen entities that is estimated from a named entity list (Bergsma and Lin, 2006)).

The above discourse model is frequency-based. We can replace the term N_r for the old referent with $f(d_{i,j}) = e^{-d_{i,j}/a}$ that captures recency, where the recency function $f(d_{i,j})$ decays exponentially with the distance between the current referent r_i and the same referent r_j that has previously been referred to. This framework for frequency and recency of new and old referents exactly corresponds to priors in the Chinese Restaurant Process (Teh et al., 2006) and the distance-dependent Chinese Restaurant Process (Blei and Frazier, 2011).

The denominator in (5) represents the sum of potential referents that could be referred to by word w. We assume that a pronoun can refer to a large number of unseen referents if gender and number match, but a proper name cannot. For example, 'he' could refer to all singular and male referents, but 'Barack Obama' can only refer to *Barack Obama*. This assumption is reflected as a probability of *unseen referents* for the pronoun as illustrated in (10) below.

In our simulations, the speaker's cost function C_w is estimated based on word length as in (9). We assume that longer words are costly to produce.

$$C_w = \text{length}(w) \tag{9}$$

Suppose that the speaker is considering using 'he' to refer to *Barack Obama*, which has been referred to N_O times in the preceding discourse, and there is another singular and male entity, *Joe Biden*, in the preceding discourse that has been referred to N_B times. In this situation, the model computes the probability that the speaker uses 'he' to refer to *Barack Obama* as follows:

$$\begin{aligned} & P_{\text{S}}(\text{'he'}|\textit{Obama}) \\ & \propto P_{\text{L}}(\textit{Obama}|\text{'he'}) \cdot \tfrac{1}{C_{\text{'he'}}} \\ & = \tfrac{P(\text{'he'}|\textit{Obama})P(\textit{Obama})}{\Sigma_{r'} P(\text{'he'}|r')P(r')} \cdot \tfrac{1}{C_{\text{'he'}}} \\ & = \tfrac{\frac{1}{V} \cdot N_{\text{O}}}{(\frac{1}{V} \cdot N_{\text{O}}) + (\frac{1}{V} \cdot N_{\text{B}}) + (\frac{1}{V} \cdot \alpha \cdot \frac{U_{\text{sing\&masc}}}{U_\cdot})} \cdot \tfrac{1}{C_{\text{'he'}}} \end{aligned} \tag{10}$$

where count $U_{\text{sing\&masc}}$ in the denominator of the last line denotes the number of unseen singular & male entities that could be referred to by 'he'. We estimate this number for each type of pronoun we

evaluate (singular-female, singular-male, singular-neuter, and plural) based on the named entity list in Bergsma and Lin (2006). The term $(\frac{1}{V} \cdot \alpha \cdot \frac{U_{\text{sing\&masc}}}{U_{\cdot}})$ is the sum of probabilities of unseen referents that could be referred to by the pronoun 'he'. The unseen referents can be interpreted as a penalty for the inexplicitness of pronouns. In the case of proper names, the denominator is always the same as the numerator, under the assumption that each entity has one unique proper name.

4 Data

4.1 Corpora

Our model was run on both adult-directed speech and child-directed speech. We chose to use the SemEval-2010 Task 1 subset of OntoNotes (Recasens et al., 2011), a corpus of news text, as our corpus of adult-directed speech. The Gleason et al. (1984) subset of CHILDES (MacWhinney, 2000) was chosen as our corpus of child-directed speech.

The model requires coreference chains, agreement information, grammatical position, and part of speech. These were extracted from each corpus, either manually or automatically. The coreference chains let us easily count how many times/how recently each referent is mentioned in the discourse, which is necessary for computing discourse salience. The agreement information (gender and number of each referent) is required so that the model can identify all possible competing referents for pronouns. For instance, *Barack Obama* will be ruled out as a possible competitor for the pronoun *she*. The grammatical position that each proper name occupies[2] determines the form of the alternative pronoun that could be used there. For example, the difference between *he* and *him* is the grammatical position that each can appear in. The part of speech is used to identify the form of the referring expression (pronouns and proper names), which is what our model aims to predict.[3]

OntoNotes includes information about coreference chains, part of speech, and grammatical dependencies. Gleason CHILDES has parsed part of speech and grammatical dependencies (Sagae et al., 2010), but it does not have coreference chains. Neither corpus has agreement information. The following section describes manual annotations that we have done for this study. Due to time constraints, we annotated only a part of the CHILDES Gleason corpus, 9 out of 70 scripts.

4.2 Annotation

4.2.1 Mention annotation

We considered only maximally spanning noun phrases as mentions, ignoring nested NPs and nested coreference chains. For the sentence "Both Al Gore and George W. Bush have different ideas on how to spend that extra money" from OntoNotes, the extracted NPs are *Both Al Gore and George W. Bush* and *different ideas about how to spend that extra money*.

These maximally spanning NPs were automatically extracted from the OntoNotes data, but were manually annotated for the CHILDES data using brat (Stenetorp et al., 2012) by two annotators.[4]

4.2.2 Agreement annotation

Many mentions (46,246 out of 56,575 mentions in OntoNotes and 10,141 out of 10,530 mentions in CHILDES Gleason) were automatically annotated using agreement information from the named entity list in Bergsma and Lin (2006), leaving 10,329 to be manually annotated from OntoNotes (about 18%) and 389 from CHILDES (about 4%).[5]

The guidelines we followed for this manual agreement annotation were largely based on pronoun replacement tests. NPs that referred to a single man and could be replaced with *he* or *him* were labeled 'male singular', NPs that could be replaced by *it*, such as *the comment*, were labeled 'neuter singular', and so on. NPs that could not be replaced with a pronoun, such as *about 30 years earnings for the average peasant, who makes $145 a year*, were excluded from the analysis.

4.2.3 Coreference annotation

We used the provided coreference chains for the OntoNotes data, but for the CHILDES data, it was necessary to do this manually using brat. The guidelines we followed for determining whether mentions coreferred came from the OntoNotes corefer-

[2]Dependency tags used were 'SUBJ', 'OBJ', and 'PMOD' in OntoNotes and 'SBJ' and 'OBJ' in CHILDES.

[3]The part of speech used to extract the target NPs were 'PRP' (pronoun), 'NNP' (proper name), and 'NNPS' (plural proper name) from OntoNotes and 'pro' (pronoun) and 'n:prop' (proper name) from CHILDES.

[4]Interannotator agreement for the CHILDES mention annotation was: precision 0.97, recall 0.98, F-score 0.97 (for two scripts).

[5]Interannotator agreement for the manual annotation of agreement information was 97% (for 500 mentions).

ence guidelines (BBN Technologies, 2007).[6]

5 Experiments

Our experiments are designed to quantify the contributions of the various components of the complete model described in Section 3.2 that incorporates discourse salience, cost, and unseen referents. We contrast the complete model with three impoverished models that lack precisely one of these components. The comparison model without discourse uses a uniform discourse salience distribution. The model without cost uses constant speech cost. The model without good estimates of unseen referents always assigns probability $\frac{1}{V} \cdot \alpha \cdot \frac{1}{C.}$ to unseen referents in the denominator of (5), regardless of whether the word is a proper name or pronoun. In other words, this model does not have good estimates of unseen referents like the complete model does.

We use the adult- and child-directed corpora to examine to what extent each model captures speakers' referring expressions. We selected pronouns and proper names in each corpus according to several criteria. First, the referring expression had to be in a coreference chain that had at least one proper name, in order to facilitate computing the cost of the proper name alternative. Second, pronouns were only included if they were third person pronouns in subject or object position, and indexicals and reflexives were excluded. Finally, for the CHILDES corpus, children's utterances were excluded.

After filtering pronouns and proper names according to these criteria, 553 pronouns and 1,332 proper names (total 1,885 items) in the OntoNotes corpus, and 165 pronouns and 149 proper names (total 314 items) in the CHILDES Gleason corpus remained for use in the analysis.

Each model chooses referring expressions given information extracted from each corpus as described in Section 4.1. For evaluation, we computed accuracies (total, pronoun, and proper name) and model log likelihood (summing $\log P_S(w|r)$ for the words in the corpus) for each model.

5.1 Results

Table 1 summarizes the results of each model with the OntoNotes and CHILDES datasets. The new referent hyperparameter α and the decay parameter for discourse recency salience were fixed at 0.1 and 3.0, respectively.[7]

5.1.1 News

Overall, the recency salience measure provides a better fit than the frequency salience measure with respect to accuracies, suggesting that recency better captures speakers' representations of discourse salience that influence choices of referring expressions. On the other hand, the models with frequency discourse salience have higher model log likelihood than the models with recency do. This is because of the peakiness of the recency models. Model log likelihood computed over pronouns and proper names (complete model) were -1022.33 and -222.76, respectively, with recency, and -491.81 and -467.06 with frequency. The recency model tends to return a higher probability for a proper name than the frequency model does. Some pronouns receive a very low probability for this reason, and this lowers the model log likelihood.

The model without discourse and the model without cost consistently failed to predict pronouns (these models predicted all proper names). This happens because in the model without discourse, the information content of pronouns is extremely low due to the large number of consistent unseen referents. In the model without cost, pronouns are disfavored because they always convey less information than proper names. The log likelihoods of these models were also below that of the complete model. These results show that pronominalization depends on subtle interaction between discourse salience and speech cost. Neither of them is sufficient to explain the distribution of pronouns and nouns on its own.

The total accuracy of the model without good estimates of unseen referents was the worst among the four models, but this model did predict pronouns to some extent. Because the number of proper names is larger than the number of pronouns in this dataset, the difference in total accuracies between the model without good estimates of unseen referents and the models without discourse or cost reflects this asymmetry. Comparison between the complete model and the model without good estimates of unseen referents also suggests that having knowledge of unseen referents helps correctly pre-

[6] Interannotator agreement for CHILDES coreference annotation was computed using B^3 (Bagga and Baldwin, 1998): precision: 0.99, recall: 1.00 (for one script).

[7] We chose the best parameter values based on multiple runs, but results were qualitatively consistent across a range of parameter values.

Corpus	Model	Discourse	Total accuracy	Pronoun accuracy	Proper name accuracy	Log-lhood
OntoNotes	complete	recency	80.27%	59.49%	88.89%	-1245.09
		frequency	73.10%	62.74%	77.40%	-958.87
	-discourse	NA	70.66%	0.00%	100.00%	-6904.77
	-cost	recency	70.66%	0.00%	100.00%	-1537.71
		frequency	70.66%	0.00%	100.00%	-1017.38
	-unseen	recency	64.14%	68.17%	62.46%	-1567.51
		frequency	56.98%	76.67%	48.80%	-1351.58
CHILDES	complete	recency	49.68%	11.52%	91.95%	-968.64
		frequency	46.18%	10.30%	85.91%	-360.28
	-discourse	NA	47.45%	0.00%	100.00%	-2159.22
	-cost	recency	47.45%	0.00%	100.00%	-1055.54
		frequency	47.45%	0.00%	100.00%	-392.72
	-unseen	recency	50.31%	13.94%	90.60%	-961.54
		frequency	48.41%	21.21%	78.52%	-332.73

Table 1: Accuracies and model log-likelihood

dict the use of proper names in the first mention of a referent.

5.1.2 Child-directed speech

Unlike the adult-directed news text, neither recency nor frequency discourse salience provides a good fit to the data. The low accuracies of pronouns and the high accuracies of proper names in all models indicate that the models are more likely to predict proper names than pronouns. There are several possible reasons for this. First, the CHILDES transcripts involve long conversations in a natural settings. Compared to the news, interlocutors are not focusing on a specific topic, but rather they often switch topics (e.g., a child interrupts her parents' conversation about her father's coworker to talk about her eggs). This topic switching makes it difficult for the model to estimate discourse salience using simple frequency or recency measures. Second, interlocutors are a family and they share a good deal of common knowledge/background (e.g., a mother said *she* as the first mention of her child's friend's mother). The current model is not able to incorporate this kind of background knowledge. Third, many referents are visually available. The current model is not able to use visual salience. In general, these problems arise due to our impoverished estimates of salience, and we would expect a more sophisticated discourse model that accurately measured salience to show better performance.

5.2 Summary

Experiments with the adult-directed news corpus show a close match between speakers' utterances and model predictions. On the other hand, experiments with child-directed speech show that the models were more likely to predict proper names where pronouns were used, suggesting that the estimates of discourse salience using simple measures were not sufficient to capture a conversation.

6 Discussion

This paper proposes a language production model that extends the rational speech act model from Frank and Goodman (2012) to incorporate updates to listeners' beliefs as discourse proceeds. We show that the predictions suggested from UID in this domain can be derived from our speaker model, providing an explanation from first principles for the relation between discourse salience and speakers' choices of referring expressions. Experiments with an adult-directed news corpus show a close match between speakers' utterances and model predictions, and experiments with child-directed speech show a qualitatively similar pattern. This suggests that speakers' behavior can be modeled in a principled way by considering the probabilities of referents in the discourse and the information conveyed by each word.

A controversial issue in language production is to what extent speakers consider a listener's discourse model (Fukumura and van Gompel, 2012). By incorporating an explicit model of listeners, our model provides a way to explore this question. For example, the speaker's listener model $P_L(r|w)$ in (4) might differ between contexts and could also be extended to sum over possible listener identity q in mixed contexts, as in (11).

$$P_L(r|w) = \Sigma_q P(r|w,q)P(q) \qquad (11)$$

This provides a way to probe speakers' sensitivity to differences in listener characteristics across situations.

Although the simulations in this paper employed simple measures for discourse salience (referent frequency and recency), the discourse models used by speakers are likely to be more complex. Studies show that semantic information that cannot be captured with these simple measures, such as topicality (Orita et al., 2014) and animacy (Vogels et al., 2013a), affects speakers' choices of referring expressions. Future work will test to what extent this latent discourse information could affect the model predictions.

Our model predicts a tight coupling between the probability of a referent being mentioned, $p(r)$, and the choice of referring expression. However, these two quantities appear to be dissociated in some cases. For example, Fukumura and Van Gompel (2010) show that semantic bias (as a measure of predictability) affects *what* to refer to (i.e., the referent), but not *how* to refer (i.e., the referring expression), while grammatical position does affect *how* you refer. One way of dissociating the probability of mention from the choice of referring expression in our model would be through the likelihood term, the word probability $p(w|r)$. While we have assumed this word probability to be constant across words and referents, Kehler et al. (2008) use grammatical position to define this probability and show that their model captures experimental data. Syntactic constraints (such as Binding principles) also influence form choices, and this kind of knowledge may also be reflected in the word probability. Examining the role of the word probability $p(w|r)$ more closely would allow us to further explore these issues.

Despite these limitations, our model provides a principled explanation for speakers' choices of referring expressions. In future work we hope to look at a broader range of referring expressions, such as null pronouns and definite descriptions, and to explore the extent to which our model can be applied to other linguistic phenomena that rely on discourse information.

Acknowledgments

We thank the UMD probabilistic modeling reading group for helpful comments and discussion.

References

John R Anderson. 2007. *How can the human mind occur in the physical universe?* Oxford University Press.

Mira Ariel. 1990. *Accessing noun-phrase antecedents*. Routledge.

Jennifer Arnold. 1998. *Reference form and discourse patterns*. Ph.D. thesis, Stanford University Stanford, CA.

Jennifer Arnold. 2008. Reference production: Production-internal and addressee-oriented processes. *Language and cognitive processes*, 23(4):495–527.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566.

Ellen Gurman Bard, Matthew P Aylett, J Trueswell, and M Tanenhaus. 2004. Referential form, word duration, and modeling the listener in spoken dialogue. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, pages 173–191.

BBN Technologies. 2007. OntoNotes English coreference guidelines version 7.0.

Leon Bergen, Noah Goodman, and Roger Levy. 2012a. That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Leon Bergen, Noah D Goodman, and Roger Levy. 2012b. That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the thirty-fourth annual conference of the cognitive science society*.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Sydney, Australia, July. Association for Computational Linguistics.

David M Blei and Peter I Frazier. 2011. Distance dependent Chinese restaurant processes. *The Journal of Machine Learning Research*, 12:2461–2488.

Richard Breheny, Napoleon Katsos, and John Williams. 2006. Are generalised scalar implicatures generated by default? an on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100(3):434–463.

Susan E Brennan. 1995. Centering attention in discourse. *Language and Cognitive Processes*, 10(2):137–167.

Wallace Chafe. 1994. Discourse, consciousness, and time. *Discourse*, 2(1).

Judith Degen, Michael Franke, and Gerhard Jäger. 2013. Cost-based pragmatic inference about referential expressions. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 376–381.

Michael Frank and Noah Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Kumiko Fukumura and Roger PG Van Gompel. 2010. Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language*, 62(1):52–66.

Kumiko Fukumura and Roger PG van Gompel. 2012. Producing pronouns and definite noun phrases: Do speakers use the addressee's discourse model? *Cognitive Science*, 36(7):1289–1311.

Kumiko Fukumura, Roger PG Van Gompel, Trevor Harley, and Martin J Pickering. 2011. How does similarity-based interference affect the choice of referring expression? *Journal of Memory and Language*, 65(3):331–344.

Alexia Galati and Susan E Brennan. 2010. Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62(1):35–51.

Talmy Givón. 1983. *Topic continuity in discourse: A quantitative cross-language study*, volume 3. John Benjamins Publishing.

Jean Berko Gleason, Rivka Y Perlmann, and Esther Blank Greif. 1984. What's the magic word: Learning language through politeness routines. *Discourse Processes*, 7(4):493–502.

Noah D Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*.

H Paul Grice. 1975. Logic and conversation. *Syntax and semantics*, 3:41–58.

Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

André Grüning and Andrej A Kibrik. 2005. Modeling referential choice in discourse: A cognitive calculative approach and a neural network approach. In Ruslan Mitkov, editor, *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, pages 163–198. John Benjamins.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Florian T Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Gerhard Jager. 2007. Game dynamics connects semantics and pragmatics. In Ahti-Veikko Pietarinen, editor, *Game theory and linguistic meaning*, pages 89–102. Elsevier.

Justine T Kao, Jean Wu, Leon Bergen, and Noah D Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L Elman. 2008. Coherence and coreference revisited. *Journal of Semantics*, 25(1):1–44.

Mariya V Khudyakova, Grigory B Dobrov, Andrej A Kibrik, and Natalia V Loukachevitch. 2011. Computational modeling of referential choice: Major and minor referential options. In *Proceedings of the CogSci 2011 Workshop on the Production of Referring Expressions. Boston (July 2011)*.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Proceedings of the 20th Conference on Neural Information Processing Systems (NIPS)*.

Brian MacWhinney. 2000. The CHILDES project: Tools for analyzing talk.

Mante S Nieuwland and Jos JA Van Berkum. 2006. When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7):1098–1111.

Ann E Nordmeyer and Michael Frank. 2014. A pragmatic account of the processing of negative sentences. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Naho Orita, Eliana Vornov, Naomi H Feldman, and Jordan Boyd-Graber. 2014. Quantifying the role of discourse topicality in speakers' choices of referring expressions. In *Association for Computational Linguistics, Workshop on Cognitive Modeling and Computational Linguistics*.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

Marta Recasens, Lluis Marquez, Emili Sapena, M. Antònia Martí, and Mariona Taulé. 2011. SemEval-2010 task 1 OntoNotes English: Coreference resolution in multiple languages.

Jacolien Rij, Hedderik Rijn, and Petra Hendriks. 2013. How WM load influences linguistic processing in adults: A computational model of pronoun interpretation in discourse. *Topics in Cognitive Science*, 5(3):564–580.

Hannah Rohde, Scott Seyfarth, Brady Clark, Gerhard Jäger, and Stefan Kaufmann. 2012. Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In *The 16th Workshop on the Semantics and Pragmatics of Dialogue, Paris, September*.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(03):705–729.

Nathaniel J Smith, Noah Goodman, and Michael Frank. 2013. Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in neural information processing systems*, pages 3039–3047.

Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Junichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101.

Harry Tily and Steven Piantadosi. 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*.

Mija Van der Wege. 2009. Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, 60(4):448–463.

Jorrig Vogels, Emiel Krahmer, and Alfons Maes. 2013a. When a stone tries to climb up a slope: the interplay between lexical and perceptual animacy in referential choices. *Frontiers in psychology*, 4.

Jorrig Vogels, Emiel Krahmer, and Alfons Maes. 2013b. Who is where referred to how, and why? the influence of visual saliency on referent accessibility in spoken language production. *Language and Cognitive Processes*, 28(9):1323–1349.