ABSTRACT

Title of dissertation: COMPUTATIONAL MODELING OF THE ROLE OF DISCOURSE INFORMATION IN LANGUAGE PRODUCTION AND ACQUISITION

Naho Orita, Doctor of Philosophy, 2015

Dissertation directed by: Professor Naomi Feldman
Department of Linguistics

This dissertation explores the role of discourse information in language production and language acquisition. Discourse information plays an important role in various aspects of linguistic processes and learning. However, characterizing what it is and how it is used has been challenging. Previous studies on discourse tend to focus on the correlations between certain discourse factors and speaker/comprehender's behavior, rather than looking at how the discourse information is used in the system of language and why. This dissertation aims to provide novel insights into the role of discourse information by formalizing how it is represented and how it is used. First, I formalize the latent semantic information in humans' discourse representations by examining speakers' choices of referring expressions. Simulation results suggest that topic models can capture aspects of discourse representations that are relevant to the choices of referring expressions, beyond simple referent frequency. Second, I propose a language production model that extends the rational speech act model from M. Frank and Goodman (2012) to incorporate updates to listeners' beliefs as discourse proceeds. Simulations suggest 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. Third, I examine the role of discourse information in language acquisition, focusing on the learning of grammatical categories of pronouns. I show that a Bayesian model with prior discourse knowledge can accurately recover grammatical categories of pronouns, but simply having strong syntactic prior knowledge is not sufficient. This suggests that discourse information can help learners acquire grammatical categories of pronouns. Throughout this dissertation, I propose frameworks for modeling speakers and learners using techniques from Bayesian modeling. These models provide ways to flexibly investigate the effects of various sources of information, including discourse salience, expectations about referents and grammatical knowledge.

COMPUTATIONAL MODELING OF THE ROLE OF DISCOURSE INFORMATION IN LANGUAGE PRODUCTION AND ACQUISITION

by

Naho Orita

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Naomi Feldman, Chair
Professor Jeffrey Lidz
Professor Jordan Boyd-Graber
Professor Ellen Lau
Professor Hal Daumé III

Acknowledgments

But for all the help that was generously bestowed upon me, this dissertation would not have existed as it does now, let alone my PhD. I would, if possible, like to extend my words of gratitude here to everyone who helped and supported me, but there are so many, as is often the case with writing acknowledgements, that I cannot do so here unfortunately; nevertheless, those unmentioned in what follows still deserve my deepest appreciation.

First and foremost, I would like to wholeheartedly thank my major advisor, Naomi Feldman, for being my advisor, my best mentor ever, my role model, and my Asado (Argentinian BBQ) teacher. In the course of the PhD program, She gave me a lot of help, support and encouragement, which cannot be overestimated in any way. I really enjoyed and learned a lot from the weekly meetings with her; had she not been my advisor, I couldn't have been done with my PhD. I'm also grateful to Naomi for giving a lot of advice, each piece of which was timely and hit the nail on the head; she read my papers and got them back to me with tons of great and valuable comments even when the deadlines almost caught me. She was also helpful, spending a lot of time with me, when I practiced my presentations. Speaking of her mentorship, Naomi was the person who inspired me to do what would truly engross me. Her guidance not only contributed to my improvement of techniques in computational modeling but also helped me become a full-fledged researcher.

I also thank Jeffrey Lidz, who was my language acquisition advisor and always thoughtful. I was so fortunate that I had several opportunities to do experiments with him, though they are not discussed in this dissertation. I learned, by talking with him and attending his classes, how intriguing it could be to investigate the language acquisition. I wish I had talked more with him to exchange views.

Jordan Boyd-Graber was my computational linguistics teacher and mentor, and what's more, he was friendly, for every aspect of which I'm really grateful to him. Jordan let me know how exciting it was to study computational linguistics, and I still remember the day I went to his office during his office hours to ask some questions about an assignment of implementing IBM 1 and 2; then, while I was writing a pseudocode for it, Jordan was beside me, teaching carefully how I should do it. Jordan's students were also great, from whom I learned a lot.

I would like to credit the other thesis committee members, Ellen Lau and Hal Daumé III, who gave me a lot of keen comments on my dissertation, which led me to think more about it. Having discussion with them in my defense was an invaluable as well as enjoyable experience.

Colin Phillips also deserves some words of appreciation here, since he trained the following two wonderful psycholinguists, Wingyee Chow and Sol Lago, who were both my big sisters at University of Maryland. Colin also suggested to me, when I was at a loss about my future and research, that I be a translator from linguistics to computer science, which still remains as the guiding principle to my career as a researcher.

I had great chances to become friends with a lot of nice people here; especially, many thanks go to Wingyee Chow and Sol Lago, who were my roommates and my bigger sisters (though they are younger than me!). They showed me how to survive as an international student here, and they were also always serious and enthusiastic in doing their research, which had a great influence on my attitude toward my own, though our specialties are different from each other. We were sort of workaholic, but I truly enjoyed the time with Wingyee and Sol, hanging out together in DC for dinner, movies, and etc. Having made friends with such fabulous people also justifies my decision to study in the U.S.

I was also lucky to have Dustin Chacon, Sayaka Goto, Kate Harrigan and Aaron Steven White as my 2015 cohorts and wonderful classmates for 5 years. Thanks!

Thank you to the members of UMD acquisition lab, especially, the lab manager, Tara Mease, who taught me how to do experiments with children in the lab and helped me a lot. I would also like to extend thanks to the Team Factivity members: Rachel Dudley, Morgan Moyer and Lambchop (the puppet); it was a lot of fun for me to do experiments together with them, and what's more, with very interesting results.

Expressing many thanks to my big brothers, Akira Omaki, Hajime Ono, Masaya Yoshida and Tomo Fujii, is now in order. I thank Akira for having given me a lot of advice since I first came here; I'm grateful to Hajime for helping me when I did experiments in Japan; I enjoyed drinking with Masaya, Tomo was also so kind that he helped me when I did experiments in Japan.

I am extremely grateful for many students, faculty, and staff in the Maryland Language Science community. Thank you to Valentine Hacquard, Tonia Bleam, Peggy Antonisse, Andrea Zukowski, Kathi Faulkingham, Kim Kwok and Csilla Kajtar for their kind support. Thank you to Angela Xiaoxue He, who shared a room in a hotel with me every time we went to BUCLD, which was really fun. Thank you to Shevaun Lewis, Shannon Barrios, and Annie Gagliardi for support and encouragement. Thank you to Maki Kishida, Kenshi Funakoshi, Shota Momma, Yuki Ito, Atsushi Nagai, and Kotaro Hara for their help. Thank you to Mike Fetters. It was fun to hang out with him in Osaka. Thank you to Eliana Vornov, who was my undergraduate research assistant. I enjoyed working with her. Thank you to Carolina Petersen, Julie Gerard, Rachael Richardson, Dongwoo Park, Allyson Ettinger, Lara Ehrenhofer, Anton Malko, Kasia Hitczenko, Nick Huang, and Gesoel Mendes. I really liked chatting with you guys at the department. I also thank my Computer Science friends; He He, Anupam Guha, Viet-An Nguyen, Ke Zhai, Yuening Hu, and Alvin Grissom II. They were so nice and friendly as to answer my questions, although we belonged to different departments.

I would like to say thank you to my linguistics teachers in Japan: Takao Gunji, Taisuke Nishigauchi, Kenjiro Matsuda, Michinao Matsui, Philip Spaelti and Nobuko Yoneda. They were very helpful and supportive of me when I was in Japan, and allowed me to catch a glimpse of how fascinating studying linguistics would be. It was due to their very guidance that I decided to study abroad!

I would also like to thank my linguist friends. Thank you to Koji Shimamura, my favorite drinking buddy, who saved me in the later stages of this dissertation writing by being willing to help me make sure I wasn't saying anything too stupid about syntax and semantics. Thank you to Maki Yamane, who is one of my role models, for the breaks for delicious dinner. I really enjoyed chatting with her. Thank you to Yukinori Takubo, who is my karaoke friend, for giving me a lot of valuable advice when I was writing a pos-doc fellowship application and applying for jobs.

I would also like to acknowledge The Rotary Foundation. I was supported by Rotary International Scholarship during my first year. I would like to thank my Rotary counselor Akio Hizume for his support and encouragement.

Last but not least, many thanks are extended to my family, Kazuko, Yasue, Chiho, Miho, and our cats, who are always supportive of me.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Overview

In the vast majority of contexts, language comprehension and production require the construction/updating of a discourse representation of entities under discussion (e.g., Grosz & Sidner, 1986; Kamp & Reyle, 1993), information about these entities (e.g., salience: Arnold, 2010, for comprehensive review), and goals and intentions of the speaker and listener (Clark & Marshall, 1981; Pickering & Garrod, 2004, among many). This discourse representation influences the processing of other levels of linguistic representations such as resolving lexical ambiguities (Duffy, Morris, & Rayner, 1988), predicting upcoming words (Nieuwland & Van Berkum, 2006), processing of scalar implicature (Breheny, Katsos, & Williams, 2006), processing of relative clauses (Roland, Mauner, O'Meara, & Yun, 2012), choosing referring expressions (Givón, 1983; Ariel, 1990; Gundel, Hedberg, & Zacharski, 1993), and choosing syntactic structure (Prat-Sala & Branigan, 2000; Arnold, Losongco, Wasow, & Ginstrom, 2000; Birner & Ward, 2009).

This dissertation explores how information in this discourse representation plays a role in language production and language acquisition by examining speakers' choices of referring expressions and learning of grammatical categories of pronouns.

Throughout this dissertation, **discourse information** denotes *linguistic* information beyond the sentence level, such as (i) salience: what is accessible/prominent in discourse (Givón, 1983; Ariel, 1990; Gundel et al., 1993), (ii) information status: whether an entity or event is old or new in discourse (Haliday, 1967; Prince, 1981; W. Chafe, 1994), (iii) topic: what discourse or a set of sentences is about (Grosz & Sidner, 1986; Asher, 2004; Kehler, 2004), (iv) coherence/relation: how sentences or discourse units relate (Hobbs, 1985; Mann & Thompson, 1988; Kehler, 2002; Asher & Lascarides, 2003), and so on. The role of non-linguistic discourse information such as visual information will not be addressed in this dissertation. **Discourse knowledge** refers to knowledge about discourse information. For example, knowledge about discourse salience includes knowing that an entity in a subject position tends to be salient, knowing that an entity occurred recently in a discourse tends to be salient, and so on.

Despite the importance and prevalence of discourse information, characterizing what it is and how it is used has been challenging. For example, researchers have suggested various factors to account for speakers' choices of referring expressions, such that a referent in a subject position tends to be salient, hence pronominalized (Givón, 1983; Ariel, 1990; Grosz, Weinstein, & Joshi, 1995, among many). However, previous theories and experimental studies have been focused on surface-level factors, such as recency and grammatical position, and it has not yet been well investigated how underlying semantic information in discourse such as topicality could influence speakers' choices of referring expressions (Arnold, 1998).

In addition to the difficulty of characterizing the discourse information, a question of how this information is used has been rarely explored. Previous studies on discourse have focused on the correlation between certain discourse factors and speaker/comprehender's behavior, rather than looking at how the discourse information is used in the system of language and why. For example, researchers have examined whether a referent recently occurred tends to be pronominalized (P. M. Clancy, 1980), whether a referent in a subject position tends to be pronominalized (Stevenson, Crawley, & Kleinman, 1994; Brennan, 1995), and so on. However, it has not explicitly formalized how this kind of salience factor interacts with other crucial information and processes in language production, such as production cost and inference about listeners. Thus, it remains unclear to what extent individual factors would account for the observed behavior unless we have an explicit framework to model language users. In order to look at the role of discourse information such as discourse salience, we need to represent different sources of information and how those interact each other in the framework that models speakers, listeners and/or learners.

This dissertation has three major contributions: (i) providing a computational/objective measure of topicality in speakers' choices of referring expressions, (ii) building an explanatory model that formalizes the intuition that speakers take a rational approach to choose the referring expression that balances the tradeoff between speakers' own effort (e.g., speech cost) and speakers' inference about listeners' inference about the referents based on information such as discourse salience, and (iii) formalizing the role of discourse information (information about who the referent is that is recovered from discourse) in the pronoun category learning. These case studies would be illustrative of more general problems such as how one could formalize the latent semantic information in humans' discourse representations and how one could formalize the role of discourse information in language production and language acquisition.

This dissertation also provides novel frameworks by using Bayesian models. Bayesian models allow us to explicitly formalize how the information could be used in different levels of linguistic representations/processes and flexibly test what kind of information and knowledge are necessary and/or sufficient for language production and language acquisition. Bayesian modeling is intended to describe computational level problems with respect to Marr's three levels of information processing system (Marr, 1982): the *computational* level focuses on defining the goal of the computation, the *algorithmic* level focuses on procedure of the computation, and the *hardware implementation* level focuses on the physical realization of the algorithmic level problems. This dissertation addresses the computational level. Bayesian models in this dissertation define what problems are and how these problems can be solved given certain kinds of information and interaction of them.

To formalize the above questions on the role of discourse information with computational approaches, I use techniques and resources in natural language processing. I use and extend state-of-the-art probabilistic models for the latent topic identification in NLP, topic modeling (Blei, Ng, & Jordan, 2003). I use resources that are frequently used for many NLP tasks, such as a corpus with a variety of linguistic annotations. Though the advantages of using computational tools and data beyond the laboratory have been suggested (Griffiths, 2015), there is still not much work being done by making use of them, presumably because of a gap between engineering and cognitive science community. Conversely, scientists have been developed computational models to get insights into how humans understand, use, and acquire language, but it has rarely been explored how these cognitive models could be applied to broader problems outside of cognitive science. The studies in this dissertation are in a position to bridge this gap.

1.2 Outline

Chapter 2 reviews related work on discourse salience and speakers' choices of referring expressions and defines problems that this dissertation addresses.

Chapter 3 shows a study that measures the latent semantic information –topic– in discourse representations by examining speakers' choices of referring expressions. I suggest that speakers may use semantic information recovered by the topic modeling when they choose referring expressions, which has an independent influence from the information computed by referent frequencies.

In chapter 4, I propose a language production model that uses dynamic discourse information to account for speakers' choices of referring expressions. This model extends previous rational speech act models (M. Frank & 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. This indicates that speaker's behavior can be modeled in a principled way by considering the probabilities of referents in the discourse and the information conveyed by each word.

In chapter 5, I examine the role of discourse information in language acquisition by modeling ideal learner's pronoun category learning (reflexive or non-reflexive). Little is known how children use discourse information when they learn language. Previous studies suggest that learners could use discourse information when they learn words (M. Frank, Tenenbaum, & Fernald, 2013; Horowitz & Frank, 2015), but it is not known how this information could be used to learn other kinds of linguistic knowledge such as grammatical knowledge. I propose a Bayesian model that allows us to represent discourse information and relevant grammatical knowledge and manipulate each information to examine to what extent each information helps learning grammatical categories of pronouns. I show that discourse information is sufficient to learn the grammatical categories of pronouns and discuss the necessity of prior syntactic knowledge based on an analysis of a child-directed speech corpus.

In chapter 6, I summarize results and findings in the previous chapters and discuss how these address the questions laid out in the introduction. I then present several future directions.

Chapter 2 Background

2.1 Discourse salience

Discourse theories (Givón, 1983; Ariel, 1990; Gundel et al., 1993; W. Chafe, 1994; Grosz et al., 1995) suggest that speakers use more attenuated referring expressions such as pronouns when they think that a referent is salient (or accessible/topical) in the preceding discourse, where salience of the referent is often associated with its accessibility/activation in memory (W. L. Chafe, 1974; Sanford & Garrod, 1981; Bock & Warren, 1985; Almor, 1999; Kibrik, 2000; Foraker & McElree, 2007; Rij, Rijn, & Hendriks, 2013).

2.1.1 Factors of discourse salience

It has been suggested that multiple factors/sources of information interact and have different strengths of influence on referent salience, hence speakers' choices of referring expressions. These factors are: grammatical position (Stevenson et al., 1994; Brennan, 1995; Arnold, 2001; Fukumura & Van Gompel, 2010), recency (P. M. Clancy, 1980; Fletcher, 1984; W. Chafe, 1994), topicality (Givón, 1983; Ariel, 1990; Grosz & Sidner, 1986; Arnold, 1998), competitors (Arnold & Griffin, 2007; Fukumura, Van Gompel, Harley, & Pickering, 2011; Fukumura, Hyönä, &

Scholfield, 2013), giveness (W. L. Chafe & Li, 1976; Gundel et al., 1993), order of mention (Järvikivi, Gompel, Hyönä, & Bertram, 2005; Kaiser & Trueswell, 2008), syntactic focus and syntactic topic (Cowles, Walenski, & Kluender, 2007; Foraker & McElree, 2007; Walker, Cote, & Iida, 1994), parallelism (Chambers & Smyth, 1998; Arnold, 1998), animacy (Fukumura & Gompel, 2011; Vogels, Krahmer, & Maes, 2013a), implicit causality (Stevenson et al., 1994; Arnold, 2001; Rohde, Kehler, & Elman, 2007; Fukumura & Van Gompel, 2010; Rohde & Kehler, 2014), cognitive load (Rij et al., 2013; Vogels, Krahmer, & Maes, 2014), visual salience (Fukumura, Gompel, & Pickering, 2010; Vogels, Krahmer, & Maes, 2013b), and so on.

Note that referent salience is not the only one factor that determines speakers' choices of referring expressions. The choices of referring expressions may be determined by syntactic constraints when the referent and the referring expression are in a same sentence/clause (Reinhart, 1976; Chomsky, 1981). The length of a noun phrase has an influence on the choices of referring expressions in that speakers tend to use pronouns to refer to longer antecedents (Karimi, Fukumura, Ferreira, & Pickering, 2014). The choice between strong and weak pronouns in Estonian is sensitive to the presence of contrast with other entities in discourse (Kaiser, 2010). The choice between demonstrative and personal pronouns in Finnish shows sensitivity to syntactic role and word order (Kaiser & Trueswell, 2008). These suggest that there are other factors or different levels of representations influencing on the referring choices.

2.1.2 Semantic aspect of discourse salience

In contrast to surface-level factors such as grammatical position and recency, it is not straightforward to characterize the representations of semantic aspects of discourse salience such as topicality. It has been observed that there is a correlation between a linguistic category *topic* and referent salience. Researchers have suggested that topical referents are more likely to be pronominalized (Givón, 1983; Ariel, 1990). However, as pointed out in Arnold (1998, 2010), examining the relation between topicality and speakers' choices of referring expressions is difficult for two reasons.

First, identifying the topic is known to be hard. For example in Arnold (2010), it is hard to determine what the topic is even in a simple sentence like *Andy brews beer* in that it is not clear whether the topic of this sentence is *Andy*, *beer*, or *brewing*. Second, researchers have defined the notion of "topic" differently: (i) what the sentence is about (Reinhart, 1982), (ii) prominent characters such as the protagonist (Francik, 1985), (iii) old information (Gundel et al., 1993), (iv) subjects (W. L. Chafe & Li, 1976; Grosz et al., 1995; Cowles, 2003), (v) repeated mentions (Kameyama, 1994), (vi) a referent that has already been mentioned in the preceding discourse as a pronoun or the topic of a cleft (Arnold, 1999), and (vii) a subject of a passive voice clause (Rohde & Kehler, 2014). Centering theory (Grosz et al., 1995; Brennan, 1995) formalizes the topic as a backward-looking center that is a single entity mentioned in the last sentence and in the most salient grammatical position (the grammatical subject is the most salient, and followed by the object and oblique object). Moreover, Givón (1983) suggests that all discourse entities are topical but that topicality is defined by a gradient/continuous property. Givón (1983) shows that three measures of topicality –*recency* (the distance between the referent and the referring expression), *persistence* (how long the referent would remain in the subsequent discourse), and *potential interference* (how many other potential referents of the referring expression there are in the preceding discourse) – correlate with the types of reference expressions.

The variation in the literature seems to derive from three fundamental properties. First, there is variation in the linguistic unit that bears the topic (Arnold, 1998, 2010). For example, Reinhart (1982) defines each *sentence* as having a single topic, whereas Givón (1983) defines each *entity* as having a single topic. Second, there is a variation in type of variable. For example, Givón (1983) defines topicality as a continuous property, whereas Centering seems to treat topicality as categorical based on the grammatical position of the referent. Third, many studies define 'topic' by using surface factors such as grammatical position and recency. This seems to be circular in that these surface factors are used to define referent salience but reference salience is also defined by topicality. When topicality is defined in terms of meaning, as in Reinhart (1982), we face difficulty in identifying what the topic is.

2.1.3 Summary

None of the existing definitions seem to provide a measure to capture latent semantic topic representations, and this makes it challenging to investigate their role in discourse representations. Chapter 3 formalizes this idea of latent topic representations, providing an objective measure of topicality in speakers' choices of referring expressions.

The above review also shows that previous linguistic studies have focused on identifying factors that might influence choices of referring expressions. However, it is not clear from these previous work how and why salience factors result in the observed patterns of referring expressions. The following section reviews previous proposals for the link between discourse salience and referring expressions and poses questions that this dissertation addresses.

2.2 Relation between referent salience and speakers' choices of referring expressions

Speakers do not randomly choose referring expressions such as pronouns, definite descriptions and proper names. For example, they 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 already been referred to in the discourse.

One might wonder if referring expressions have salience in the meaning itself: speakers use pronoun when they want to express the meaning 'salient discourse referent', they use definite description when they want to express the meaning 'non-salient discourse referent' and so on. This kind of idea has been elaborated in the previous discourse theories as follows.

2.2.1 Form-salience mapping

Givón (1983) suggests a single scale of topicality as in Table 2.1: The most topical referent is referred to by a zero anaphor, the least topical referent is referred to most continuous/accessible topic

zero anaphora
unstressed/bound pronouns or grammatical agreement
stressed/independent pronouns
R-dislocated definite noun phrases
neutral-ordered definite noun phrases
L-dislocated definite noun phrases
Y-moved NPs ('contrastive topicalization')
Cleft/focus construction
Referential indefinite NPs

most discontinuous/inaccessible topic

Table 2.1: Topicality scale by Givón (1983, p17)

by an indefinite NP (e.g., *a girl*), and so on. Ariel (1990) also suggests a scale of accessibility where different referring expressions are used to express the referent's accessibility as in Table 2.2: the most accessible referent is referred to by a zero anaphor, the least accessible referent is referred to by a modified full name, and so on. Gundel et al. (1993) suggest an implicational hierarchy that is associated with the referent's cognitive status as in Table 2.3: the referent in *focal attention* is referred to by zero or unstressed pronouns, the referent that is *uniquely identifiable* is referred to by a definite noun phrase, and so on.

One important question is regarding the relation between salience and referring expressions represented by these scales or hierarchies. The form-salience correspondences seem to be relative to other referring expressions in a language in that a certain referring expression encodes a certain degree of salience (or accessibility/topicality). Ariel (1990) argues that these form-salience relations are not arbitrary and suggests predictions that a form with (i) more lexical information

High accessibility

extremely high accessibility markers:
gaps, wh traces, reflexives, and agreement
cliticized pronoun
unstressed pronouns
stressed pronoun
stressed pronoun + gesture
proximal demonstrative
distal demonstrative
proximal demonstrative + modifier
distal demonstrative + modifier
first name
last name
short definite description
long definite description
full name
full name + modifier

Low accessibility

Table 2.2: Accessibility scale by Ariel (1990, p73)

cognitive status	in focus >	activated >	familiar >	uniquely > identifiable	referential >	type identifiable
referring expression	*it*	*that*, *this*, *this* N	*that* N	*the* N	indefinite *this* N	*a* N

Table 2.3: Giveness hierarchy by Gundel, Hedberg, and Zacharski (1993, p275)

(informativity), (ii) more unambiguous (rigidity), and (iii) less attenuated (attenuation) would code the lower accessibility and vice versa. However, it is not clear how this form-salience mapping holds nor why it should be. Informativity and rigidity would change depending on other entities in a context, so this mapping has to be stipulated to some extent. It is also not clear how these three measures can be represented, how they define each from-salience mapping, and why they are relevant. Thus, the form-salience relations suggested in the literature seem to be stipulated without clear explanations.

2.2.2 Other information in the choices of referring expressions: production cost and listener model

Moreover, many studies show that speakers consider other kinds of information when they choose words: (i) speakers' own production cost relative to other possible words and (ii) inference about what listeners would infer what the word refers to given the context.

2.2.2.1 Production cost

Recent studies look at different types of production cost in a referential communication game setting. They show that speakers and listeners take production cost into account when they produce or interpret words. Rohde, Seyfarth, Clark, Jäger, and Kaufmann (2012) show that speakers tend to use ambiguous words (e.g., *the blue thing*) to refer to entities with costly unambiguous words (e.g., *the triangle-and-square thing*) if other referents can be identified with low-cost unambiguous words (e.g, *the blue circle*). Bergen, Goodman, and Levy (2012b) test the prediction by Horn's principle: expressions that are costlier (e.g., longer or less frequent) are associated with less typical/probable meanings. They show that speakers are more likely to use costly utterances (where the cost is represented as an explicit dollar value) when the target referent occurs infrequently (i.e., less probable meaning). There is also a classic study that shows speakers are slower to name infrequent objects than high frequent objects (Oldfield & Wingfield, 1965). Degen, Franke, and Jäger (2013) examine to what extent listeners estimate speakers' production cost that is represented as a word length. They show that listeners take the estimate of production cost into account when they interpret utterances. Baumann, Clark, and Kaufmann (2014) examine another aspect of production cost, speakers' pragmatic reasoning. They show that speakers use costlier forms (overspecification in their experiment) to avoid effortful pragmatic inference. These studies suggest the importance of various production costs. The remaining questions are (i) how these costs affect speakers choices of referring expressions in a natural language setting, and (ii) how they interact with other information such as discourse salience in speakers' choices of referring expressions.

2.2.2.2 Listener model

The question of to what extent speakers use their listener model is under debate. Researchers have examined this question in various experimental settings and found contrasting evidence. Since the Gricean maxims of conversational implicature (Grice, 1975), researchers have argued whether and to what extent speakers tailor their utterances to their listeners. Clark and Marshall (1981); Clark and Murphy (1982); Clark (1996) suggest that speakers form utterances based on mutually shared information with listeners (audience design). On the other hand, Barr and Keysar (2006) argue that audience design does not occur regularly, but optionally influences on language production processes, such as adjusting errors of their utterances by consulting listeners' perspective (the monitoring and adjustment hypothesis). It is also suggested that speakers might use their own knowledge as a proxy for listeners (Pickering & Garrod, 2004).

As for reference, discourse theories generally assume that speakers choose referring expressions by estimating how the referent is salient/accessible for listeners (Givón, 1983; Ariel, 1990; Gundel et al., 1993). There is a variety of psycholinguistic experiments that examine to what extent speakers take listeners' perspective into account when choosing words. These studies have mainly focused on speakers' word choice when the word ambiguously/unambiguously refers to the referent in a given context. The results seem to diverge depending on experimental situations. Speakers avoid ambiguous expressions that could lead to difficulties for listeners to interpret (Haywood, Pickering, & Branigan, 2005; Brennan & Hanna, 2009), but this is not always the case in other situations (Keysar, Barr, & Horton, 1998; Ferreira & Dell, 2000; Bard et al., 2000).

As for the choices of referring expressions, there are several findings in the previous experiments. Bard, Aylett, Trueswell, and Tanenhaus (2004) found that speakers do not adjust phonetic properties such as articulation and length to their listeners, but they do so for the referring forms, suggesting that there might be two separate production processes. Arnold and Griffin (2007) showed that speakers are more likely to use pronouns when there was no other entity in the context, but they are more likely to use a name when there was another entity that has a different gender (e.g., use *John* instead of *he* when there is another entity *Mary*). In other words, speakers use over-specific forms even when the competing referent would not cause ambiguity. Wege (2009) and Galati and Brennan (2010) found that speakers tailor specificity of the description depending on their knowledge about listeners. In contrast, Rosa and Arnold (2011) found that speakers adjust the complexity of referring forms according to their own ability of attention, but not to listeners' attention. Finally, Fukumura and Gompel (2012) tested whether speakers choose referring expressions based on listeners' discourse model of referent salience. They controlled whether the speaker and the listener share prior linguistic context (whether the listener can hear the referent mentioned in the previous sentence) and found that this manipulation does not affect speakers' choices of referring expressions. They conclude that speakers use their own discourse model when choosing referring expressions. Overall, results in previous studies seem to diverge, and it seems to be informative to explore this problem by building a formal model that allows us to explicitly examine different kinds of situations and listeners in speakers' minds.

2.2.3 Interim summary

I have argued that there seem to be no clear explanations why there should be a relation between each type of referring expression and the degree of salience. In addition, findings on the importance of production cost and listener model suggest that only considering discourse salience of the referent would not precisely capture speakers' choices of referring expressions. It is necessary to examine discourse salience in relation to those important factors, within a single framework of a speaker model.

Instead of assuming form-salience scale/hierarchies as in 2.2.1, I take an assumption that words have a meaning that is the set of entities they could refer to. [1] For example, "Alice" refers only to *Alice* whereas "she" refers to all female singular entities. This kind of meaning representations has been used in recent rational speech act models (e.g., M. Frank & Goodman, 2012). I assume that speakers choose the referring expressions based on those meanings, together considering other information such as salience of the referent (in a speaker's listener model) and production cost of the word. In Chapter 4, I propose a speaker model that formalizes the relation between discourse salience and speakers' choices of referring expressions, with considering production cost and speakers' inference about listeners. The following reviews previous formal/computational models relevant to speakers' choices of referring expressions and shows that there is a gap between questions that previous models have addressed and the questions that I have raised above.

[1] This assumption cannot be ad hoc and should be motivated in the following way: in theories such as Distributed Morphology (Halle & Marantz, 1993), intrinsic features are encoded as the lexical properties of roots in the form of bundle of features. For the words like "she" or "woman", features like [+female] and [+singular] must at least be listed as such (c.f. Harley and Ritter (2002) for a more detailed morphosyntactic cross-linguistic discussion on this point regarding pronouns). Since these features are interpretable, they must be subjected to LF and hence semantic construal. Semantic denotations can be defined in terms of the set theory (Heim & Kratzer, 1998), so that [+female] can be translated into $\{x : x \text{ is female}\}$, which is a set of female entities. Thus, "she" and "woman" are situated in the intersection of those which are female and those which are singular, which is compatible with my assumption.

2.2.4 Formalization of speakers' choices of referring expressions

There are formal/computational models relevant to speakers' choices of referring expressions, such as cognitive models, Centering, and Referring Expression Generation models. This section shows that these models are built for different reasons and none of them sufficiently explains why there is a relation between discourse salience and speakers' choices of referring expressions.

2.2.4.1 Cognitive models

There is some cognitive modeling work relevant to speakers' choices of referring expressions, but these models are data-driven rather than being explanatory. Kibrik (2000); Grüning and Kibrik (2005) and Khudyakova, Dobrov, Kibrik, and Loukachevitch (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, it is not clear why and how these factors result in the observed referring choices.

Formal models that go beyond identifying factors focus on only pronouns rather than accounting for speakers' word choices per se. Kehler, Kertz, Rohde, and Elman (2008) formalize a discrepancy between pronoun comprehension and production observed in semantically biased sentences such as (1).

(1) An example from Stevenson et al. (1994): Transfer-of-possession context (with-

pronoun condition)

a. John seized the comic from Bill. He ...

b. John passed the comic from Bill. He ...

Stevenson et al. (1994) report that people are more likely to interpret an ambiguous pronoun *he* in (1a) to refer to the subject/Goal (*John*) than the non-subject with 84.6% subject bias. On the other hand, there is no such bias (51%) in a sentence (1b). In a different condition where a subject pronoun is not given, they also found a Goal bias that matches the results in the condition with a pronoun. Across different test sentences, they also found that people tend to use a pronoun when a referent is in a subject position of the previous sentence and use a name when a referent is in a non-subject position. This shows a contrast between production and comprehension in that there is no subject bias in a sentence like (1b) where a pronoun *is* provided in the continuing sentence. Similar discrepancy has been reported with implicit causality verbs (Fukumura & Van Gompel, 2010; Rohde & Kehler, 2014). These studies suggest that pronoun production (how people refer) is insensitive to semantic biases, but a grammatical role has influence on the pronominalization. Kehler et al. (2008) formalize this discrepancy using Bayes' rule as in (2).

(2)

$$P(referent|pronoun) = \frac{P(pronoun|referent)P(referent)}{\Sigma_{referent}P(pronoun|referent)P(referent)}$$

The term $P(referent|pronoun)$ is a probability of a referent that is referred to by a pronoun, representing the interpretation bias. The term $P(pronoun|referent)$ is a probability that a speaker would use a pronoun to refer to the referent, representing the production bias. Note that they define this probability based on a grammatical position of the referent where a pronoun in a subject position would have a higher probability than non-subject positions. The term $P(referent)$ is a prior (next-mention bias) that a particular referent will be referred to, no matter what the form is. They argue that the discrepancy between production and comprehension is predicted from this formalization in that the interpretation bias is a product of the production bias and the prior.

Although this formalization can account for the experimental data, it is not clear how this could capture the form choice from multiple referring expressions. The production bias term is defined based on the grammatical position of the pronoun, but it is not clear what this probability could be for other types of forms. The prior term is defined in terms of the semantic bias, but this prior, which referent will be referred to, is known to rely on various factors depending on a context (Tily and Piantadosi (2009) conducted an empirical analysis on these factors).

Rij et al. (2013) use ACT-R (Anderson, 2007) to examine the effects of working memory load in pronoun interpretation. In a preceding experiment, Hendriks, Koster, and Hoeks (2014) examined pronoun interpretation of Dutch adults and children (age 4-6). They presented participants stories as in (3) and (4) with the last sentence starting with a potentially ambiguous pronoun. An example story in (3) has a topic shift indicated by changing a subject, and a story in (4) does not.

(3) 1. Eric/gaat/voetballen/in de sporthal.

"Eric is going to play soccer in the sports hall."

2. Philip/vraagt/Eric/om mee te rijden/naar de training.

   "Philip asks Eric to carpool to the training."

3. Philip/haalt/Eric/na het eten/met de auto op.

   "Philip picks up Eric after dinner by car."

4. Hij/voetbalt/al twintig jaar.

   "He has played soccer for twenty years."

(4) 1. Eric/gaat/voetballen/in de sporthal.

   "Eric is going to play soccer in the sports hall."

2. Eric/vraagt/Philip/om mee te rijden/naar de training.

   "Eric asks Philip to carpool to the training."

3. Eric/haalt/Philip/na het eten/met de auto op.

   "Eric picks up Philip after dinner by car."

4. Hij/voetbalt/al twintig jaar.

   "He has played soccer for twenty years."

They found that adult speakers prefer to interpret the ambiguous pronoun as referring to the subject in the previous sentence, whereas children prefer to interpret it as the first mention of the story, showing an insensitivity to the topic shift. They also found that children with a higher working memory performed more like adult speakers.

Based on the experiment in Hendriks et al. (2014), Rij et al. (2013) examine whether a topic shift signaled by a grammatical role (subject) in a preceding sentence is only available with sufficient working memory capacity. Their model simulations suggest that listeners without sufficient working memory capacity rely more on the base-level activation that depends on frequency and recency of discourse elements, whereas listeners with sufficient working memory capacity use information about the subject of the previous sentence that is realized as spreading activation that boost all discourse elements associated with it (they manipulate the amount of spreading activation as reflecting individual differences in working memory capacity). This prediction was tested in an experiment with adults, and the results confirmed the prediction.

Their computational model represents salience of discourse elements as the activation of elements in listeners' memory and suggests that how different sources of information interact in memory representations and result in listeners' interpretation of pronouns. However, as they stated, this is a listener model focusing on pronoun interpretation, and it is not clear how this model could be extended to account for speakers' choices of referring expressions.

In sum, previous cognitive models show how certain factors/information influence on pronoun production and interpretation, but it is not clear how these models would predict and account for speakers' choices of referring expressions.

2.2.4.2 Centering theory

Centering theory is a model of local discourse coherence and salience. It considers discourse as transitions across adjacent utterances and characterizes discourse coherence based on links of entities. It also predicts which entity is the most salient at each utterance. In this theory, certain discourse entities are more centered (salient/topical) than other entities, and this is assumed to constrain speakers' choices of referring expressions (pronominalization) to form a coherent discourse. For example, the third sentence in (5) seems to be odd because referent *John*, an entity that is more centered than other entities, is referred to by a name, but not by a pronoun (Grosz et al., 1995, p.215,216).

(5) a. He has been acting quite odd. (*he* = John)

b. He called up Mike yesterday.

c. John wanted to meet him quite urgently.

There is a variety of instantiations of Centering, such as algorithms for pronoun interpretation (Brennan, Friedman, & Pollard, 1987; Walker et al., 1994). Here I briefly review a primary version of the theory, Grosz et al. (1995).

There are two main states for each utterance: Backward looking centers of n-th utterance $C_b(U_n)$ and forward looking centers of n-th utterance $C_f(U_n)$. The backward looking center $C_b(U_n)$ represents the most topical/salient entity after utterance U_n is interpreted. The forward looking center is an ordered list that contains all entities in U_n. All entities in the forward looking center could be the backward looking center in the following utterance, and these are ordered according to grammatical role, such as subject > object > other, in many instantiations. In other words, the most highly ranked entity in the forward looking center that is mentioned in the following utterance is the backward looking center in the following utterance.

This theory has a deterministic rule for pronominalization that is stipulated for discourse to be coherent, known as Rule 1 as in (6).

(6) Rule 1 (Grosz et al., 1995): If any element of $C_f(U_n)$ is realized by a pronoun in utterance U_{n+1}, then $C_b(U_{n+1})$ must be realized as a pronoun also.

There are different versions of Rule 1. Grosz, Joshi, and Weinstein (1983) suggest that the backward looking center should be pronominalized if it is the same as the backward looking center in the previous utterance. Gordon, Grosz, and Gilliom (1993) suggest that the backward looking center should always be pronominalized.

Grosz et al. (1995) claim that Rule 1 represents some aspects of speakers' choices of referring expressions in that the use of a pronoun is assumed to affect "inference load placed upon the hearer" (p.208) and "signals the hearer that the speaker is continuing to talk about the same thing" (p.214). On one hand, some experiments show that this rule predicts speakers' and comprehenders' behavior to some extent. For example, speakers are more likely to use pronouns to refer to the backward looking center (Brennan, 1995), and reading times slow down when the backward looking center is not realized as a pronoun (Gordon et al., 1993). On the other hand, this rule allows some unnatural patterns. For example, it allows all other entities to be pronominalized if the backward looking center is pronominalized. It also allows speakers to repeat more explicit forms such as proper names if no pronouns are used. Crucially, Centering does not account for speakers' choices of referring expressions from multiple options. It only proposes when a pronoun is preferred for local discourse to be coherent.

object	type	clothing	position
d_1	man	wearing suit	left
d_2	woman	wearing t-shirt	middle
d_3	man	wearing t-shirt	right

Table 2.4: An example of knowledge representations for REG (Krahmer and Deemter 2012, p177)

2.2.4.3 Referring Expression Generation models

Referring Expression Generation (REG) models (Krahmer & Van Deemter, 2012; Van Deemter, Gatt, Gompel, & Krahmer, 2012, for comprehensive review) choose a description of an entity that helps listeners to identify that entity in a given context (Reiter, Dale, & Feng, 2000). This program started as computers needed to identify objects to humans (Winograd, 1972), and it has been applied to many tasks where computers/robots need to refer to things, such as news summaries (Siddharthan, Nenkova, & McKeown, 2011), weather forecasts (Turner, Sripada, Reiter, & Davy, 2008), air travel systems (White, Clark, & Moore, 2010), and a robot dialogue system (Giuliani et al., 2010).

To choose a referring expression, a typical REG algorithm first chooses a type of referring expression. If the algorithm chooses to generate a description, it decides which set of properties can distinguishes the target entity and how the set of selected properties is realized using language (Reiter et al., 2000). A domain of discourse, entities, their properties, and the target referent are usually given for this kind of task. Table 2.4 shows an example of knowledge representations in Krahmer and Van Deemter (2012).

As for the choice of referring expression, the REG algorithms have a deterministic constraint about when a pronoun is preferred. For example, Reiter et al. (2000) suggest a constraint to use a pronoun (i) if the target referent was mentioned in the previous utterance, and (ii) if that utterance does not contain any other entity that has the same agreement (e.g., gender).

Researchers have also suggested algorithms for pronoun generation. McCoy and Strube (1999) suggest important factors such as sentence boundaries, distance, and discourse structure in a pronoun generation algorithm based on Centering theory. Kibble and Power (2004) suggest a system using Centering for text generation, focusing on ideas of salience, cohesion, and continuity. As a variant, Callaway and Lester (2002) suggest an pronominalization algorithm based on parsed discourse trees and simple rules.

Although some researchers have proposed to link this domain of research with computational psycholinguistic studies of referring (Van Deemter et al., 2012; Gatt, Krahmer, Gompel, & Deemter, 2013; Gatt, Gompel, Deemter, & Kramer, 2013), most REG models have been built to do practical tasks, but not to explain how people do it.

2.2.4.4 Interim summary

There are various formal/computational models relevant to speakers' choices of referring expressions. However, these models focus on either the influence of specific factors, pronoun interpretation/production (but not the choice from multiple expressions), or deterministic constraints for pronominalization with less motivation for explaining speakers' behavior. They have not been built to explain why there is a relation between discourse salience and speakers' choices of referring expressions.

2.2.5 Uniform Information Density Hypothesis

One potential formal explanation for the relation between discourse salience and speakers' choices of referring expressions is the Uniform Information Density hypothesis (UID) (Levy & Jaeger, 2007; Tily & 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 could be considered as a refinement of Grice's maxim of quantity: speakers should provide message as informative as possible, but not more informative than is required (Grice, 1975). 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 in writers' choices of referring expressions, in that pronouns are used when a referent is predictable. This predictability of the referent was estimated by a web experiment and represented in a form of surprisal that is computed based on participants' accuracy of guessing the correct referent given the preceding discourse.

While these results appear to support UID, there is a crucial difference between previous UID studies and UID with respect to speakers' choices of referring expressions. Previous UID studies have shown the link between *form reduction* and information density at various linguistic representations. Words that convey high information content are likely to be pronounced with more duration (Bell et al., 2003; Aylett & Turk, 2006) and more articulatory and phonological details (Van Son & Van Santen, 2005) to avoid a peak in information density. Speakers are less likely to use morphosyntactic contractions such as *I'm* and shorter words such as *chimp* (*chimpanzee*) when the form conveys high information (A. Frank & Jaeger, 2008; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). Speakers are also more likely to use the optional function word *that* when the following relative clause or complement would convey high information so that the full form spreads information to avoid a peak in information density (Levy & Jaeger, 2007; Jaeger, 2010). Crucially, the target form is assumed to be already chosen by speakers in these studies. A common question across these studies is whether speakers would reduce/delete the target form for the information distribution to be uniform. On the other hand, the problem with respect to the choices of referring expression is fundamentally different in that it is not about the form reduction but *form choice*, where different candidate forms that could refer to a target referent (e.g., *she* and *Alice*) would convey different amount and content of information depending on a context. For example, *she* refers to any singular and female entity and *Alice* refers to a particular person. Therefore, the relation between discourse salience and speakers' choices of referring expression is not so obvious as UID has predicted. Chapter 4 shows that we can derive predictions about the choices of referring expressions directly from a model of language production.

2.3 Summary

I showed that previous studies have focused on individual factors of discourse salience in speakers' choices of referring expressions, but factors relevant to discourse semantics such as topicality have not been well formalized and examined. Chapter 3 suggests an objective measure of discourse topicality in speakers' choices of referring expressions.

I also showed that both theoretical/experimental and computational/formal work do not seem to provide explicit and formal explanations of why the relation between discourse salience and speakers' choices of referring expressions should exist. In addition, findings on the importance of production cost and listener model suggest that only considering discourse salience of the referent would not be sufficient to account for speakers' choices of referring expressions. To address these questions, Chapter 4 proposes a speaker model that formalizes why discourse salience affects speakers' choices of referring expressions in relation to production cost and speakers' inference about listeners.

Chapter 3 Quantifying latent topic information in speakers' discourse representations

3.1 Introduction

Among the numerous factors influencing discourse salience of a referent, this study focuses on *topicality*. Many researchers have suggested that topical referents are more likely to be salient, and thus more likely to be pronominalized (Givón, 1983; Ariel, 1990, among many). Chapter 2 showed that it has been challenging to investigate the role of topicality in speakers' discourse representations in that the representation of topicality is latent and requires inference, in contrast to surface-level factors such as grammatical position and recency. None of the existing definitions/measures seem to provide a way to capture latent topic representations, and this makes it challenging to investigate their role in discourse representations. It is this idea of latent topic representations that we aim to formalize.

This study uses topic modeling to verify the prevailing hypothesis that topical referents are more likely to be pronominalized than lexical nouns. Examining the relationship between topicality and referring expressions using topic modeling provides an opportunity to test how well the representation recovered by topic models corresponds to the cognitive representation of entities in a discourse. If we can recover the observation that topical referents are more likely to be pronominalized than more specified forms, this could indicate that topic models can capture not only aspects of human semantic cognition (Griffiths, Steyvers, & Tenenbaum, 2007), but also aspects of a higher level of linguistic representation, discourse.

Topic modeling (Blei et al., 2003; Griffiths et al., 2007) uses a probabilistic model that recovers a latent topic representation from observed words in a document. The model assumes that words appearing in documents have been generated from a mixture of latent topics. These latent topics have been argued to provide a coarse semantic representation of documents and to be in close correspondence with many aspects of human semantic cognition (Griffiths et al., 2007). This previous work has focused on semantic relationships among words and documents. While it is often assumed that the topics extracted by topic models correspond to the gist of a document, and although topic models have been used to capture discourse-level properties in some settings (Nguyen et al., 2013), the ability of topic models to capture cognitive aspects of speakers' discourse representations has not yet been tested.

This study uses topic modeling to formalize the idea of salience in the discourse, focusing on the idea of topicality as a factor of salience (Ariel, 1990; Arnold, 1998) and asks whether the latent topics that are recovered by topic models can predict speakers' choices of referring expressions. Simulations show that the referents of pronouns belong, on average, to higher probability topics than the referents of full noun phrases, indicating that topical referents are more likely to be pronominalized.

This suggests that the information recovered by topic models is relevant to speakers' choices of referring expressions and that topic models can provide a useful tool for quantifying speakers' representations of entities in the discourse.

Because of their structured representations, consisting of a set of topics as well as information about which words belong to those topics, topic models are able to capture topicality by means of semantic associations. For example, observing a word *Clinton* increases the topicality of other words associated with the topic that *Clinton* belongs to, e.g., *president*, *Washington* and so on. In other words, topic models can capture not only the salience of referents within a document, but also the salience of referents via the structured topic representation learned from multiple texts. Note that the representations recovered by the topic model is not just lexical associations. These are lexical associations derived from a set of probabilistic topics of document (i.e., the gist of document), hence representing discourse-level information.

3.2 Model

3.2.1 Recovering latent topics

This study formalizes topicality of referents using topic modeling. Topic modeling uses a probabilistic model that recovers a latent topic representation from observed words in a document. The topic model assumes that words appearing in a document have been generated from a mixture of latent topics. These latent topics have been argued to provide a gist of a document. Each document is represented as a probability distribution over topics, and each topic is represented as a probability distribution over words. In this study, each topic is represented as a probability distribution over possible referents in the corpus.

In training the topic model, all lexical nouns in the discourse are assumed to be potential referents. The topic model is trained only on lexical nouns, excluding all other words. This ensures that the latent topics capture information about which referents typically occur together in documents. Excluding pronouns from the training set introduces a confound, because it artificially lowers the probability of the topics corresponding to those pronouns. However, in this study our predicted effect goes in the opposite direction: we predict that topics corresponding to the referents of pronouns will have higher probability than those corresponding to the referents of lexical nouns. Excluding pronouns thus makes us less likely to find support for our hypothesis.

Rather than pre-specifying a number of latent topics, this study uses the hierarchical Dirichlet process (Teh, Jordan, Beal, & Blei, 2006), which learns a number of topics to flexibly represent input data. The summary of the generative process is as follows.

1. Draw a global topic distribution

   $G_0 \sim \mathrm{DP}(\gamma, H)$ (where γ is a hyperparameter and H is a base distribution).

2. For each document $d \in \{1, \ldots, D\}$ (where D denotes the number of documents in the corpus),

   (a) draw a document-topic distribution

   $G_d \sim \mathrm{DP}(\alpha_0, G_0)$ (where α_0 is a hyperparameter).

(b) For each referent $r \in \{1, \ldots, N_d\}$ (where N_d denotes the number of referents in document d),

   i. draw a topic parameter $\phi_{d,r} \sim G_d$.

   ii. draw a word $x_{d,r} \sim \text{Mult}(\phi_{d,r})$.

This process generates a distribution over topics for each document, a distribution over referents for each topic, and a topic assignment for each referent. The distribution over topics for each document represents what the topics of the document are. The distribution over referents for each topic represents what the topic is about. An illustration of this representation is in Table 3.1. Topics and words that appear in the second and third columns are ordered from highest to lowest. Topicality of the referents can be represented using this probabilistic latent topic representation, measuring which topics have high probability and assuming that referents associated with high probability topics are likely to be topical in the discourse.

Word	Top 3 topic IDs	Associated words in the 1st topic
Clinton	5, 26, 61	president, meeting, peace, Washington, talks
FBI	148, 73, 67	Leung, charges, Katrina, documents, indictment
oil	91, 145, 140	Burmah, Iraq, SHV, coda, pipeline

Table 3.1: Illustration of the topic distribution

Given this generative process, we can use Bayesian inference to recover the latent topic distribution. We use the Gibbs sampling algorithm in Teh et al. (2006) to estimate the conditional distribution of the latent structure, the distributions over topics associated with each document, and the distributions over words associated with each topic. The state space consists of latent variables for topic assignments, which we refer to as $\mathbf{z} = \{z_{d,r}\}$. In each iteration we compute the conditional distribution $p(z_{d,r}|\mathbf{x}, \mathbf{z}_{-d,r}, *)$, where the subscript $-d, r$ denotes counts without considering $z_{d,r}$ and $*$ denotes all hyperparameters. Recovering these latent variables allows us to determine what the topic of the referent is and how likely that topic is in a particular document. We use the latent topic and its probability to represent topicality.

3.2.2 A measure of topicality

Discourse theories predict that topical referents are more likely to be pronominalized than more specified expressions.[1] We can quantify the effect of topicality on choices of referring expressions by comparing the topicality of the referents of two types of referring expressions, pronouns and lexical nouns. If topical words are more likely to be pronominalized, then the topicality of the referents of pronouns should be higher than the topicality of the referents of lexical nouns.

Annotated coreference chains in the corpus, described below, are used to determine the referent of each referring expression. We look at the topic assigned to each referent r in document d by the topic model, $z_{d,r}$. We take the log probability of this topic within the document, $\log p(z_{d,r}|G_d)$, as a measure of the topicality of the referent. We take the expectation over a uniform distribution of referents, where the uniform distributions are denoted $u(lex)$ and $u(pro)$, to obtain an estimate of

[1] Although theories make more fine-grained predictions on the choices of referring expressions with respect to saliency, e.g., a full name is used to refer to less salient entity compared to a definite description (c.f. accessibility marking scale in Ariel 1990), we focus here on the coarse contrast between pronouns and lexical nouns.

the average topicality of the referents of lexical nouns, $\mathbb{E}_{u(lex)}\left[\log p(z_{d,r}|G_d)\right]$, and the average topicality of the referents of pronouns, $\mathbb{E}_{u(pro)}\left[\log p(z_{d,r}|G_d)\right]$, within each document. The expectation for the referents of the pronouns in a document is computed as

$$\mathbb{E}_{u(pro)}\left[\log p(z_{d,r}|G_d)\right] = \frac{\sum_{r=1}^{N_{d,pro}} \log p(z_{d,r}|G_d)}{N_{d,pro}} \tag{3.1}$$

where $N_{d,pro}$ denotes the number of pronouns in a document d. Replacing $N_{d,pro}$ with $N_{d,lex}$ (the number of lexical nouns in a document d) gives us the expectation for the referents of lexical nouns.

To obtain a single measure for each document of the extent to which our measure of topicality predicts speakers' choices of referring expressions, we subtract the average topicality for the referents of lexical nouns from the average topicality for the referents of pronouns within the document to obtain a log likelihood ratio q_d,

$$q_d = \mathbb{E}_{u(pro)}\left[\log p(z_{d,r}|G_d)\right] - \mathbb{E}_{u(lex)}\left[\log p(z_{d,r}|G_d)\right] \tag{3.2}$$

A value of q_d greater than zero indicates that the referents of pronouns are more likely to be topical than the referents of lexical nouns.

3.3 Annotated coreference data

Our simulations use a training set of the Ontonotes corpus (Recasens, Marquez, Sapena, Martí, & Taulé, 2011, SemEval-2010 Task 1 subset of OntoNotes), which consists of news texts. We use these data because each entity in the corpus has a coreference annotation. We use the coreference annotations in our evaluation, described above. The training set in the corpus consists of 229 documents, which contain 3,648 sentences and 79,060 word tokens. We extract only lexical nouns (23,084 tokens) and pronouns (2,867 tokens) from the corpus as input to the model.[2]

Some preprocessing is necessary before using these data as input to a topic model. This necessity arises because some entities in the corpus are represented as phrases, such as in (1a) and (1b) below, where numbers following each expression represent the entity ID that is assigned to this expression in the annotated corpus. However, topic models use bag-of-words representations and therefore assign latent topic structure only to individual words, and not to phrases. We preprocessed these entities as in (2). This enabled us to attribute entity IDs to individual words (roughly heads of the noun phrases), rather than entire phrases, allowing us to establish a correspondence between these ID numbers and the latent topics recovered by our model for the same words.

1. Before preprocessing

    (a) a tradition in Betsy's family: 352

    (b) Betsy's family: 348

    (c) Betsy: 184

2. After preprocessing

    (a) tradition: 352

    (b) family: 348

[2] In particular, we extracted words that are tagged as NN, NNS, NNP, NNPS, and for pronouns as PRP, PRP$.

(c) Betsy: 184

Annotated coreference chains in the corpus were used to determine the referent of each pronoun and lexical noun. The annotations group all referring expressions in a document that refer to the same entity together into one coreference chain, with the order of expressions in the chain corresponding to the order in which they appear in the document. We assume that the referent for each pronoun and lexical noun appears in its coreference chain. We further assume that the referent needs to be a lexical noun, and thus exclude all pronouns from consideration as referents. If a lexical noun does not have any other words before it in the coreference chain, i.e., that noun is the first or the only word in that coreference chain, we assume that this noun refers to itself (the noun itself is the referent). Otherwise, if a coreference chain has multiple referents, we take its referent to be the lexical noun that is before and closest to the target word.

3.4 Results

To recover the latent topic distribution, we ran 5 independent Gibbs sampling chains for 1000 iterations.[3] Hyperparameters γ, α_0, and η were fixed at 1.0, 1.0, and 0.01, respectively.[4] The model recovered an average of 161 topics (range: 160 – 163 topics).

We computed the log likelihood ratio q_d (Equation 3.2) for each document

[3] We used a Python version of the hierarchical Dirichlet process implemented by Ke Zhai (`http://github.com/kzhai/PyNPB/tree/master/src/hdp`).

[4] Parameter γ controls how likely a new topic is to be created in the corpus. If the value of γ is high, more topics are discovered in the corpus. Parameter α_0 controls the sparseness of the distribution over topics in a document, and parameter η controls the sparseness of the distribution over words in a topic.

and took the average of this value across documents for each chain. The formula to compute this average is as follows.

```
For each chain $g$,
    1. get the final sample $s$ in $g$.
    2. For each document $d$ in the corpus,
        i. compute $q_d$ based on $s$.
    3. Compute the average of all $q_d$ in the corpus.
```

The average log likelihood ratio in each chain consistently shows values greater than zero across the 5 chains. The average log likelihood ratio across chains is 0.1359 with standard deviation 0.0104. As an example, in one chain, the average of the expected values for the referents of pronouns across documents is -2.0226 with standard deviation 0.5624. In the same chain, the average of the expected values for the referents of lexical nouns across documents is -2.1630 with standard deviation 0.4749.

We used the median test[5] to evaluate whether the two groups of the referents are different with respect to the expected values of the log probabilities of topics. The test shows a significant difference between two groups ($p = 0.024$).

We also computed the probability density $p(q)$ from the log likelihood ratio q_d for each document using the final samples from each chain. Graph 3.1 shows the probability density $p(q)$ from each chain. The peak after zero confirms the observed effect.

[5] The median test compares medians to test group differences (Siegel, 1956).

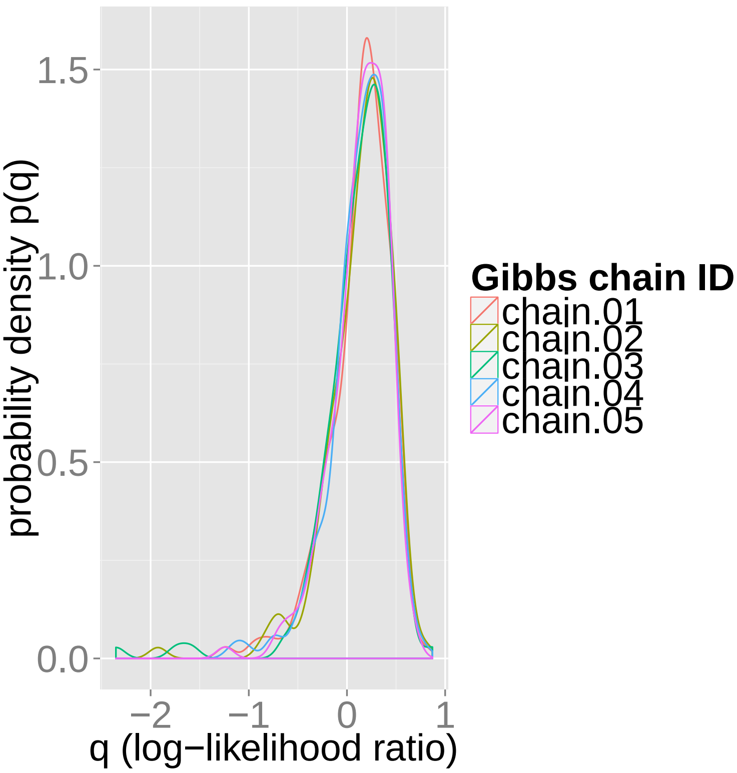


Figure 3.1: The probability density of $p(q)$ (topic probabilities)

Table 3.2 shows examples of target pronouns and lexical nouns, their referents, and the topic assigned to each referent from a document. Table 3.3 shows the distribution over topics in the document obtained from one chain. Topics in Table 3.3 are ordered from highest to lowest. Only four topics were present in this document. The list of referents associated with each topic in Table 3.3 is recovered from the topic distribution over referents. This list shows what the topic is about.

The topics associated with the pronouns *his*, *he* and *its* have the highest probability in the document-topic distribution, as shown in Table 3.3. In contrast, although the topic associated with the word *Kosovo* has the highest probability in the document-topic distribution, the topics associated with nouns *Goran* and *Albanians*

Target	Referent	Referent's Topic ID
his	Spilanovic	1
he	Spilanovic	1
its	Belgrade	1
Goran	Minister	4
Albanians	Albanians	2
Kosovo	Kosovo	1

Table 3.2: Target words, their corresponding referents, and the assigned topics of the referents

Topic ID	Assciated words	Probability
1	Milosevic, Kostunica, Slobodan, president, Belgrade, ...	0.64
2	president, Clinton, meeting, peace, Washington, ...	0.16
3	people, years, U.S., president, time, government, ...	0.16
4	government, minister, party, Barak, today, prime, ...	0.04

Table 3.3: The document-topic distribution

do not have high probability in the document-topic distribution. This is an example from one document, but this tendency is observed in most of the documents in the corpus.

These results indicate that the referents of pronouns are more topical than the referents of lexical nouns using our measure of topicality derived from the topic model. This suggests that our measure of topicality captures aspects of salience that influence choices of referring expressions.

However, there is a possibility that the effect we observed is simply derived from referent frequencies and that topic modeling structure does not play a role beyond this. Tily and Piantadosi (2009) found that the frequency of referents has a significant effect on predicting the upcoming referent. Although their finding is about comprehender's ability to predict the upcoming referent (not the type of referring expression), we conducted an additional analysis to rule out the possibility that referent frequencies alone were driving our results.

In order to quantify the effect of referent frequency on choices of referring expressions, we computed the same log likelihood ratio q_d with referent probabilities. The probability of a referent in a document was computed as follows:

$$p(r_i|doc_d) = \frac{C_{d,r_i}}{C_{d,\cdot}} \tag{3.3}$$

where C_{d,r_i} denotes the number of mentions that refer to referent r_i in document d and $C_{d,\cdot}$ denotes the total number of mentions in document d. We can directly compute this value by using the annotated coreference chains in the corpus.

The log likelihood ratio for this measure is 0.5703. The average of the expected values for the referents of pronouns across documents is -3.1110 with standard deviation 1.0444. The average of the expected values for the referents of lexical nouns across documents is -3.6813 with standard deviation 0.9459. The median test shows a significant difference between two groups. ($p < 0.0001$). We also computed the probability density $p(q)$ from the log likelihood ratio q_d. Graph 3.2 shows the probability density $p(q)$. The peak after zero confirms the observed effect. These results indicate that the frequency of a referent captures aspects of its salience that influence choices of referring expressions, raising the question of whether our latent topic representations capture something that simple referent frequencies do not.

In order to examine to what extent the relationship between topicality and referring expressions captures information that goes beyond simple referent fre-

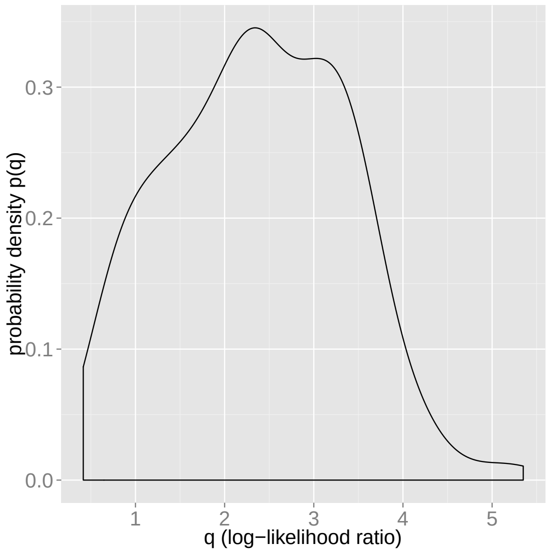


Figure 3.2: The probability density of $p(q)$ (referent frequency)

quencies, we compare two logistic regression models.[6] Both models are built to predict whether a referent will be a full noun phrase or a pronoun. The first model incorporates only the log probability of the referent as a predictor, whereas the second includes both the log probability of the referent and our topicality measure as predictors.[7]

The null hypothesis is that removing our topicality measure from the second model makes no difference for predicting the types of referring expressions. Under this null hypothesis, twice the difference in the log likelihoods between the two models should follow a $\chi^2(1)$ distribution. We find a significant difference in likelihood between these two models ($\chi^2(1) = 118.38, p < 0.0001$), indicating that the latent measure of topicality derived from the topic model predicts aspects of listen-

[6] Models were fit using `glm` in R. For the log-likelihood ratio test, `lrtest` in R package `epicalc` was used.

[7] We also ran a version of this comparison in which frequency of mention was included as a predictor in both models, and obtained similar results.

ers' choices of referring expressions that are not predicted by the probabilities of individual referents.

3.5 Discussion

In this study we formalized the correlation between topicality and choices of referring expressions using a latent topic representation obtained through topic modeling. Both quantitative and qualitative results showed that according to this latent topic representation, the referents of pronouns are more likely to be topical than the referents of lexical nouns. This suggests that topic models can capture aspects of discourse representations that are relevant to the selection of referring expressions. We also showed that this latent topic representation has an independent contribution beyond simple referent frequency.

One might wonder about the possibility that the difference between pronouns and lexical nouns appears because pronouns are referring to entities that occurred more than once in the discourse, and speakers are more likely to repeatedly reference entities that are topical. Since the topic modeling is a probabilistic model that depends on the word frequency, we cannot separate the effect of topicality from the frequency. The prediction along with this topic representation would be that the difference between pronouns and lexical nouns will be smaller if leaving out all the lexical nouns that occur as first mention. If we could confirm this prediction, it would suggest the causal relation between topicality and frequency in that the certain referents are topical because they are mentioned more frequently. However, it is not entirely clear what this kind of result show more than the results of the referent frequency model. There are also other topicality factors that do not depend on frequency, such as grammatical position, topical markers (e.g., Japanese case marker -*wa*) and syntactic constructions (e.g., clefting, passivization). It would be interesting to see to what extent these linguistic expressions could capture the difference between pronouns and lexical nouns.

This study examined only two factors: topic probabilities from the topic model and referent frequency. However, discourse studies suggest that the salience of a referent is determined by various sources of information and multiple discourse factors with different strengths of influence. For example, topical entities might be more likely to be in subject position than less topical entities. Our framework could eventually form part of a more complex model that explicitly formalizes the interaction of information sources. Having a formal model would help by allowing us to test different hypotheses and develop a firm theory regarding cognitive representations of entities in the discourse.

As summarized in Chapter 2, it has been challenging to quantify the influence of latent semantic factors such as topicality because it requires inference about the hidden meanings. The simulations in this study represent only a first step toward capturing these challenging factors. The simulations nevertheless provide an example of how formal models can help us validate theories of the relationship between speakers' discourse representations and the language they produce.

Chapter 4 Formalizing the relation between speakers' choices of referring expressions and discourse salience

4.1 Introduction

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. Speakers' choices of referring expressions have long been thought to depend on the salience of entities in the discourse, and a number of grammatical, semantic, and distributional factors related to salience have been found to influence choices of referring expressions as reviewed in Chapter 2.

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. Chapter 2 showed that previous linguistic studies have focused on identifying factors that might influence choices of referring expressions and pointed out that it is not clear from this previous work how and why these factors result in the observed patterns of referring expressions. Chapter 2 also showed that production cost and a listener model have particularly been unexplored in the formalization of this aspect of language production.

In recent years, a number of formal models have been proposed to capture inferences between speakers and listeners in a context of Gricean pragmatics (Grice, 1975; M. Frank & 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; M. Frank & Goodman, 2012; Bergen, Goodman, & Levy, 2012a; Smith, Goodman, & Frank, 2013), and studies report that they robustly predict various linguistic phenomena in experimental settings (Goodman & Stuhlmüller, 2013; Degen et al., 2013; Kao, Wu, Bergen, & Goodman, 2014; Nordmeyer & 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.

This study proposes a probabilistic model to explain speakers' choices of referring expressions based on discourse salience. Our model extends the rational speech act model from M. 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 speakers' pronominalization in a corpus. These results suggest that this model formalizes underlying principles that account for speakers' choices of referring expressions.

4.2 Speaker model

4.2.1 Rational speaker-listener model

We adopt the rational speaker-listener model from M. 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 O (e.g., visual access to object referents) as in (4.1), assuming that a speaker has chosen the word informatively.

$$P(r_s|w,O) = \frac{P_S(w|r_s,O)P(r_s)}{\Sigma_{r' \in O} P(w|r',O)P(r')} \tag{4.1}$$

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 exponentiated utility function as in (4.2).

$$P_S(w|r_s,O) \propto e^{\alpha U(w;r_s,O)} \tag{4.2}$$

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 (4.3) where $|w|$ denotes the number of referents that the word w can be used to refer to.

$$P_S(w|r_s, O) \propto \frac{1}{|w|} \tag{4.3}$$

Thus, the speaker model chooses the 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.

4.2.2 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 (4.3) 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]

$$P_S(w|r) \propto P_L(r|w) \cdot \frac{1}{C_w} \tag{4.4}$$

[1] Our speaker model corresponds to Frank and Goodman's exponentiated utility function (4.2), with α equal to one and with their cost $D(w)$ being the log of our cost C_w.

The term $P_L(r|w)$ in (4.4) represents informativeness of word w: the speaker chooses w that most helps a listener L to infer referent r. The term C_w in (4.4) 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 (4.5).

$$P_L(r|w) = \frac{P(w|r)P(r)}{\Sigma_{r'}P(w|r')P(r')} \tag{4.5}$$

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 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 (4.5) 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, M. Frank and Goodman (2012)'s speaker model in (4.3) is a special case of our speaker model in (4.4) 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 (4.4) and (4.5), yielding

$$P_S(w|r) \propto \frac{P(w|r)P(r)}{\sum_{r'} P(w|r')P(r')} \cdot \frac{1}{C_w} \tag{4.6}$$

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. [2] In this scenario, the speaker's probability of choosing a word is

$$P_S(w|r) \propto \frac{1}{\sum_{r'} P(r')} \cdot \frac{1}{C_w} \tag{4.7}$$

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 (4.7), $-\log \sum_{r'} P(r')$. This means that in deciding which word to use, the highest cost a speaker should be willing to pay for a word should depend

[2] For example, we could assume that all referents have only two referring expressions that can be used to refer to them (e.g., a pronoun and a proper name) and both expressions are equally likely to be used.

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 come from considering pronouns and nouns 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.

4.2.3 Predicting behavior from corpora

The model described in Section 4.2.2 is fully general, applying to arbitrary word choices, discourse probabilities, and cost functions. As an initial step, our simulations focus on the choice between pronouns and proper names. Our work tests the speaker model from (4.4) directly, asking whether it can predict the referring expressions from corpora of written 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 (4.8):

$$P(w|r) = \frac{1}{V} \tag{4.8}$$

where the count V is a 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 (4.5) 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 (Chapter 3) 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 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 $\frac{1}{U.}$ (count $U.$ denotes a total number of unseen entities that is estimated from a named entity list (Bergsma & 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 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 correspond to priors in the Chinese Restaurant Process (Teh et al., 2006) and the distance-dependent Chinese Restaurant Process (Blei & Frazier, 2011).

The denominator in (4.5) represents the sum of potential referents that could be referred to by word w. We assume that a pronoun can refer to a potentially infinite 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 (4.10) below.

In our simulations, the speaker's cost function C_w is estimated based on word length as in (4.9). We assume that longer words are costly to produce.

$$C_w = \text{length}(w) \tag{4.9}$$

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:

$$\begin{aligned} & P_{\text{S}}(\text{`he'}|\textit{Obama}) \\ & \propto P_{\text{L}}(\textit{Obama}|\text{`he'}) \cdot \frac{1}{C_{\text{`he'}}} \\ & = \frac{P(\text{`he'}|\textit{Obama})P(\textit{Obama})}{\Sigma_{r'} P(\text{`he'}|r')P(r')} \cdot \frac{1}{C_{\text{`he'}}} \\ & = \frac{\frac{1}{V} \cdot N_{\text{O}}}{(\frac{1}{V} \cdot N_{\text{O}}) + (\frac{1}{V} \cdot N_{\text{B}}) + (\frac{1}{V} \cdot \alpha \cdot \frac{U_{\text{sing\&masc}}}{U.})} \cdot \frac{1}{C_{\text{`he'}}} \end{aligned} \tag{4.10}$$

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 $(\frac{1}{V} \cdot \alpha \cdot \frac{U_{\text{sing\&masc}}}{U.})$ 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.

4.3 Data

4.3.1 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, Perlmann, and Greif (1984) subset of CHILDES (MacWhinney, 2000a) 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[3] 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

[3] POS tags we used are: “SUBJ”, “OBJ”, and “PMOD” in OntoNotes and ’SBJ’ and ’OBJ’ in Gleason CHILDES.

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 predicts.[4]

OntoNotes includes information about coreference chains, part of speech, and grammatical dependencies. Gleason CHILDES has parsed part of speech and grammatical dependencies (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2010), though 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 CHILDES Gleason corpus, 9 out of 70 scripts.

4.3.2 Annotation

4.3.2.1 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.[5]

[4] The part of speech used to extract the target NPs were "PRP", "NNP", and "NNPS" from OntoNotes and "pro", and "n:prop" from CHILDES.

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

4.3.2.2 Agreement annotation

Many mentions (46246 out of 56575 mentions in OntoNotes and 10141 out of 10530 mentions in CHILDES Gleason) were automatically annotated using agreement information from the named entity list in Bergsma and Lin (2006), leaving 10329 to be manually annotated from OntoNotes (about 18 %) and 389 from CHILDES (about 4%).[6]

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.

4.3.2.3 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 coreference guidelines (BBN Technologies, 2007).[7]

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

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

4.4 Experiments

Our experiments are designed to quantify the contributions of the various components of the complete model described in Section 4.2.2 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 a probability $\frac{1}{V} \cdot \alpha \cdot \frac{1}{C_{\cdot}}$ to unseen referents in the denominator of (4.5), 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 these two types of corpora to examine to what extent each model captures speakers' referring expressions. We select 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 with these criteria, 553 pronouns and 1332 proper names (total 1885 items) in the OntoNotes corpus, and 165 pronouns and 149 proper names (total 314 items) in the CHILDES Gleason corpus

Corpus	Model	Discourse	Total acc.	Pronoun acc.	Proper name acc.	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 4.1: Accuracies and model log-likelihood

remained for use in the analysis.

Each model chooses referring expressions given information extracted from each corpus as described in Section 4.3.1. 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.

4.4.1 Results

Table 4.1 summarizes the results of each model with OntoNotes and CHILDES dataset. The new referent hyperparameter α and the decay parameter for discourse recency salience were fixed at 0.1 and 3.0 respectively.[8]

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

4.4.1.1 News

Overall, the pronoun accuracies are low. This would be due to the poor estimates of discourse salience. We currently have either recency or frequency, but there are many other factors that influence on the pronominalization such as grammatical position and topicality as reviewed in Chapter 2.

Across different models, 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 as in Table 4.4.1.1 that breaks up the sum of the log posterior into pronouns and proper names in the complete model. 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.

salience	pronoun	proper name
recency	-1022.33	-222.76
frequency	-491.81	-467.06

Table 4.2: Model log likelihood computed over pronouns and proper names (complete model)

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

Figure 4.1 shows the information conveyed by each evaluated word as defined in (4.7). At the beginning of the discourse, words that have high information content are used, and the information content of words gradually decreases as discourse proceeds, showing that the words are getting more predictable with more discourse information.

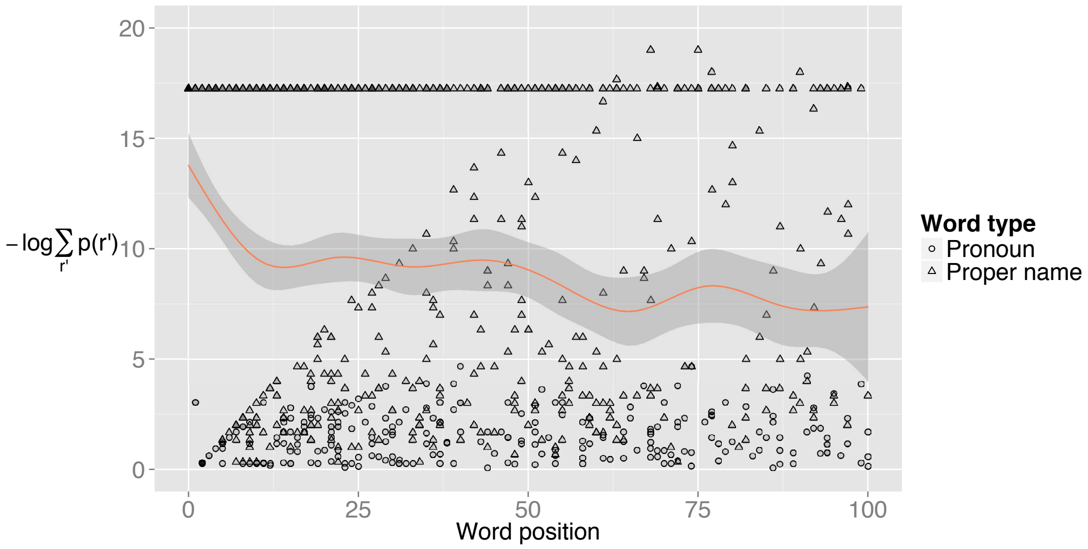


Figure 4.1: Information content of pronouns and proper names (aggregated across documents (discourse salience: recency)): The red line is a smoothed trend line with confidence interval around it.

4.4.1.2 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 the topic (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 the 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 perforamnce. In contrast to the news corpus, speakers in the child-directed speech corpus often use pronouns for the first mention. This causes a difference in pronoun accuracies between the complete model and the model without unseen referents. The better accuracy of the model without unseen referents is because of this data distribution.

4.4.2 Summary

Experiments with the adult-directed news corpus show a close match between speakers' utterances and model predictions. On the other hand, experiments with the 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.

4.5 Discussion

This study proposes a language production model that extends the rational speech act model from M. 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 a formal explanation 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 & Gompel, 2012; Bard, Hill, Foster, & Arai, 2014, among many). 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 (4.4) might differ between contexts and could also be extended to sum over possible listener identities q in mixed contexts as in Equation 4.11.

$$P_L(r|w) = \Sigma_q P(r|w,q)P(q) \tag{4.11}$$

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

One of important differences between our model and Frank and Goodman model is the assumption about the potential referents. Our model assumes that pronouns can refer to anything that matches agreement information (gender and number). However, we could also assume that pronouns refer to entities in a particular discourse/context, as in Frank and Goodman model. This assumption raises a question of how we could estimate possible referents in the particular context. It is easy to get objects in O in Frank and Goodman model because it is a simple language game setting where interlocutors know all possible referents, but it is not straightforward in the case of natural language setting in that listeners usually do not know what speakers are going to refer to till they get enough information about what speakers are talking about. It would be interesting to estimate possible referents in a particular discourse, but this seems to be outside of the scope of the speaker model in that it requires world knowledge, common sense reasoning, and so on.

Although the simulations in this study 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 (Chapter 3), affects speakers' choices of referring expressions. Future work will test to what extent this latent discourse information could affect the model predictions.

We also hope to extend this work to look at a broader range of referring expressions, such as definite descriptions. Dealing with definite descriptions raises a challenging problem in that they tend to be less informative but longer than proper names (e.g., *Alice* vs. *the girl*, *the girl sitting at the bank*, and so on). The current model would be more likely to prefer proper names in most of the cases because many definite descriptions would be less informative and more costly than proper names. Another problem is identifying the range of referents that can be referred to by the definite description. The size of competitors would change depending on the description (it would be much less predictable than pronouns that can be identified by gender and number in English), such as *the woman* vs. *the young* *woman*. The model would need to have rich knowledge about entities in the world and corresponding linguistic expressions so that it can identify which referent in the discourse matches the target description (e.g., knowledge that *the young girl* can be a description of *Alice* in 'Alice in wonderland', but cannot be a description of *Red Queen*). The discourse salience term in the current model also poses a problem with respect to the definite descriptions in that speakers could use the definite descriptions to refer to something in common knowledge such as *the president of the U.S.*, but not already referred to in the preceding discourse. The current measure of discourse salience based solely on the linguistic information would not be able to capture these cases. These potential problems suggest that the current model would have to be able to access rich knowledge about the world and corresponding linguistic expressions.

This study focuses on the role of discourse salience, but there are other factors that affect the choice of referring expressions, but do not affect $p(r)$, the probability of a referent to be mentioned. 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 the grammatical position does affect *how* you refer. Kehler et al. (2008) also use grammatical position to define the word probability $p(w|r)$ and it captures the experimental data. Similarly, the syntactic constraints (such as Binding principles) do have influence on the form choices, and we assume that this kind of knowledge may also be reflected in the word probability, $p(w|r)$. We hope to address how we could have a better representation of the word probability in the future.

Despite these limitations, this work suggests that the observed speakers' behavior seem to reflect the subtle probabilistic interactions with other information such as production cost and the inference about listeners. Our work provides a framework of examining how discourse information interacts with other information that plays an important role in language production. We hope this framework will help exploring how discourse information could be used in various processes of language production.

Chapter 5 Learning grammatical categories of pronouns using discourse information

5.1 Introduction

The work in the previous chapter as well as studies reviewed in Chapter 2 suggest that speakers choose the referring expressions by considering listeners who share discourse information. In turn, the choices of referring expressions provide listeners information about what speakers want to refer to based on the shared discourse information. Listeners are able to recover what speakers want to refer to because they have knowledge about the referring expressions (e.g., *she* refers to a singular and female entity) and discourse (e.g., salience). This raises some questions. What if listeners do not know properties of referring expressions? How do they infer the speakers' intended referent without knowledge about those words? Would discourse information help them infer the referent of the unknown words? This kind of situation exactly parallels to children who are acquiring language. In this chapter, I ask to what extent discourse information could help learners learning syntactic aspects of some referring expressions. This chapter particularly examines to what extent discourse information (information about who the referent is based on discourse) helps learning of grammatical categories of pronouns: reflexive and non-reflexive pronouns.

An interpretation of a sentence that contains a pronoun depends on which entity the pronoun refers to and syntactic restrictions on the relation between the pronoun and the reference. Though both *herself* and *her* in (7a) and (7b) refer to an entity that is singular and female, English speakers know that the sentence in (7a) means that Alice saw Alice in the mirror and the sentence in (7b) means that Alice saw someone else in the mirror.

(7) a. Alice saw herself in the mirror.

b. Alice saw her in the mirror.

This difference reflects English speakers' syntactic knowledge of pronouns. The syntactic distribution of English pronouns is governed by two syntactic properties: c-command and locality. In short, pronouns like *herself* must have c-commanding and local antecedents, and pronouns like *her* occur elsewhere. This means that the grammatical relation between pronouns and antecedents, as characterized by locality and c-command, defines the distribution of grammatical categories of pronouns in English: reflexive pronouns, e.g., *myself* and *herself*, a group of pronouns that have c-commanding and local antecedents, and non-reflexive pronouns, e.g., *me* and *her*, a group of pronouns that have either non-c-commanding, non-local, or non-c-commanding and non-local antecedents.

How can learners acquire these grammatical categories of pronouns? In other words, how can learners identify syntactic distributional categories of pronouns and their word distributions? This is a potentially difficult problem for children acquiring language because identifying the entity that the pronoun refers to and identifying the grammatical categories of pronouns depend on each other. In order to learn that *herself* is reflexive, learners need to interpret the sentence in (7a) as 'Alice saw Alice', recognizing that *Alice* and *herself* co-refer to the same entity. However, in order to interpret the meaning of the sentence (i.e., identifying the entity that the pronoun refers to), they might need to use the knowledge that *herself* is a reflexive pronoun that takes the local and c-commanding antecedent, whereas *her* is a non-reflexive pronoun that must not be c-commanded by its antecedent in the local domain. It seems to pose a chicken-or-the-egg problem. How can learners solve this problem? This study examines one way in which learners can overcome this circularity problem.

Recent developmental studies suggest that children as young as two years old can use discourse information such as discourse continuity and discourse relation to learn new words (Horowitz & Frank, 2015; Sullivan & Barner, 2015). This study provides another aspect of how discourse information could help word learning by showing that discourse information can help learners identify syntactic categories of pronouns and their word distributions. Discourse information in this study especially refers to information about who the referent is that is estimated from discourse. If learners could predict that the pronoun *herself* in (7a) is likely to refer to *Alice* and the pronoun *her* in (7b) is likely to refer to someone else based on the discourse, this provides information that can help them categorize these pronouns into different classes based on the syntactic position of the entity that they think the pronoun refers to. This study aims to measure the degree to which the discourse information, that could be noisy or ambiguous, can help learning grammatical categories of pronouns by using a computational model.

Though the data used for the computational modeling were taken from child-directed speech, the distributions of verbs and pronouns were balanced for the purpose of the experiment, and they do not reflect the actual input children receive. To get further insights on the simulation results, the distributions of input children receive with respect to reflexive and non-reflexive pronouns will be investigated.

This chapter is organized as follows. Section 5.2 introduces the relevant basic syntactic properties of English pronouns (Chomsky, 1973, 1981; Reinhart, 1976). Section5.3 defines what the learning problem in this study is. Section 5.5 describes a behavioral experiment that measures the discourse information available to listeners. Section 5.6 presents Bayesian modeling that shows discourse information can help to learn the grammatical categories of pronouns. Section 5.7 investigates the distributions of reflexive and non-reflexive pronouns in the input children actually receive and compare them with the model input data. The last section discusses open questions.

5.2 Grammatical distribution of pronouns

This section illustrates the distribution of the pronoun-antecedent relations by introducing two syntactic properties: c-command and locality and defines the learning problem in this study.

English pronouns are subject to the following grammatical constraints (Chomsky, 1981).

(8) a. Reflexive pronouns must be bound in a local domain.

b. Non-reflexive pronouns must be free in a local domain.

A binding relation in (8) is defined in (9).

(9) α binds β if α c-commands β and α, β are coindexed.

The first property is c-command. Reflexive pronouns must be c-commanded by their antecedent (Reinhart, 1976). C-command is defined in (10).

(10) C-command: α c-commands β if α does not dominate β and every γ that dominates α dominates β.

In the sentence (11), English speakers know that the antecedent of *herself* is not *Alice*, but *Alice's sister*.

(11) Alice's sister saw herself in the mirror.

That is, when the constituent structure of the sentence is represented as a tree in Figure 5.1, the reflexive *herself* is contained in the sister node of its antecedent *Alice's sister*.

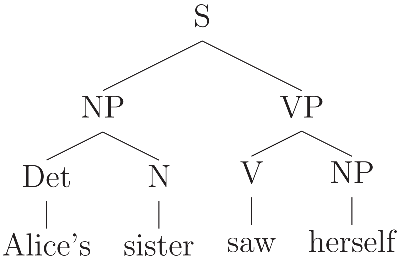


Figure 5.1: Syntactic tree showing a c-command relationship between the antecedent *Alice's sister* and the pronoun *herself*.

The second property is locality. As in the sentence (12), even though *Alice's sister* c-commands *herself*, the antecedent of *herself* is not *Alice's sister*, but *Alice.*

(12) Alice's sister thought that Alice saw herself in the mirror.

Locality refers to the domain of the syntactic relation between the pronoun and its antecedent. Reflexive pronouns must have their antecedents in the local domain, corresponding to the minimal clause containing the reflexive (Chomsky, 1973, 1981). In (12), although both the NP *Alice* and the NP *Alice's sister* c-command the reflexive, only the former can be the antecedent, illustrating the relevance of locality. There are a number of ways that generalize the relevant local domain in the theory, but for the purpose of this study, it is sufficient to take the local domain in English to be the minimal clause containing the pronoun and its antecedent.

Reflexive and non-reflexive pronouns occur in nearly complementary distribution: non-reflexive pronouns appear in contexts in which the antecedent is either non-local (13a), not in a c-commanding position (13b), or both (13c) (Chomsky, 1973, 1981).

(13) a. The Red Queen$_i$ said that Alice likes her$_i$ (*herself$_i$).

b. Alice$_i$'s sister likes her$_i$ (*herself$_i$).

c. While the Red Queen$_i$ is sleeping, Alice kicked her$_i$ (*herself$_i$).

In summary, the relationships between the grammatical positions of antecedents and pronouns, as characterized by locality and c-command, define the distribution of grammatical categories of pronouns in English. Note that there are cases where the complementarity between reflexives and non-reflexives breaks down as in (14).

(14) a. John$_i$ knew that the reports about himself$_i$ were fabricated. (Pollard & Sag, 1992)

b. I know what Mary, Sue, and Bill have in common. Mary likes him, Sue likes him, and Bill$_i$ likes him$_i$ too. (Conroy, Takahashi, Lidz, & Phillips, 2009)

The reflexive pronoun *himself* in (14a) and the non-reflexive pronoun *him* in (14b) are in noncomplementary positions where both reflexive and non-reflexive pronouns could be used. In (14a), *himself* does not have a local binder. In (14b), *him* has a local binder. This study does not address these cases, but they are important features of the distribution of pronouns and ultimately must be explained.

Cross-linguistically, languages differ in many aspects with respect to the pronoun-antecedent relations. Some languages have different reflexive pronoun forms on the basis of having an antecedent in subject vs. non-subject position (c.f. Koster & Reuland, 1991), such as *seg selv* (its antecedent is in a subject position) and *ham selv* (its antecedent is in a non-subject position) in Norwegian.

Languages such as Spanish distinguish between reflexive and non-reflexive pronouns only in third person pronouns. First and second person pronouns can be used as both non-reflexives (15a) and reflexives (15b). On the other hand, third person pronouns cannot have local antecedents (15d), but a reflexive form *se* must be used as in (15e).

(15) a. Juan me vió

Juan me saw
‘Juan saw me’

b. (Yo) me ví

I me saw
'I saw myself.'

c. (Yo) lo ví

I him saw
'I saw him.'

d. * Juan$_i$ lo$_i$ vió

Juan him saw
'Juan saw him'

e. Juan$_i$ se$_i$ vió

Juan self saw
'Juan saw himself'

The local domain differs across languages. In Icelandic (Hyams & Sigurjónsdóttir, 1990), the monomorphemic reflexive *sig* can be bound across a subjunctive (16b) or infinitival (16c) clause boundary, but not a finite clause (16d).

(16) a. Jón$_i$ rakađi sig$_{i,*j}$

'John shaves himself.'

b. Jón$_i$ segir ađ Pétur$_j$ raki sig$_{i,j}$

'John says that Peter shaves (SUBJ) himself.'

c. Jón$_i$ skipađi Pétur$_j$ ađraka sig$_{i,j}$

'John ordered Peter to shave (INF) himself.'

d. Jón$_i$ veit ađ Pétur$_j$ rakar sig$_{*i,j}$

'John knows that Peter shaves (IND) himself.'

Languages like Kannada have reflexives that can be bound across indicative clauses as in (17) (Amritavalli, 2000).

(17) Hari$_i$ Rashmi tann$_i$-annu hoDe-d-aLu anta heeL-id-a

Hari Rashmi self-ACC hit-PST-3SF that say-PST-3SM
'Hari said that Rashmi hit him.'

In sum, distributional constraints on pronouns substantially vary within a language and across languages.

5.3 Learning problem

From the perspective of language acquisition, the above cross-linguistic variability suggests that learners need to determine which syntactic features are relevant for learning pronoun categories. In order for learners to determine the relevant grammatical features, they must be able to identify a potential antecedent of a pronoun because the relevant features are about the relation between the pronoun and its antecedent. They also need to map each type of pronoun in their language onto its distribution defined by the relevant grammatical features.

These pose a potentially difficult problem for children acquiring language for several reasons: First, distributional constraints on pronouns vary within a language. For example, languages that have the long-distance reflexives also have non-reflexives and only-local reflexives. These different pronouns may have different local domains and the distributions of these pronouns may overlap. Second, distributional constraints on pronouns vary across languages. As illustrated above, the local domain differs across languages and the morphological paradigm of reflexives differs across languages. This variability suggests that learners need to have a hypothesis space that is flexible enough. Third, learning requires combining semantic and syntactic features. Learning pronouns is not only about the assignment of referent which is known to be hard (Quine, 1960), but also depends on grammatical restrictions that vary within and across languages as shown above.

Despite these problems, people acquire this knowledge at some point. One of potential information that helps learners to learn the grammatical categories of pronouns is a guess about meaning of a sentence containing a pronoun. Presumably they guess the meaning of the sentence from discourse information by using their discourse knowledge. If they can use discourse knowledge to infer the meaning of the sentence such that (18a) means *Alice saw Alice in the mirror* and (18b) means *Alice saw someone else in the mirror*, this could provide information that can help them categorize these pronouns into different syntactic classes.

(18) a. Alice saw herself in the mirror.

b. Alice saw her in the mirror.

In other words, the expectations about referents of the pronouns would help learners identifying the syntactic distributional profiles of the pronoun categories such that pronouns like *herself* have c-commanding and local antecedents. Learners might be able to categorize pronouns based on these expectations about the referents along with the relevant grammatical features.

This study aims to build a computational model on the learning of grammatical categories of pronouns in English with minimal specification of a hypothesis space defined by c-command and locality. The goal is to measure to what extent discourse information (the expectations about the referents of pronouns), along with the prior syntactic knowledge help learners to categorize pronouns in English. The key to this learning problem is if/how learners could use discourse information. The following section shows evidence that children have discourse knowledge to some extent and can use it to learn new words.

5.4 Children's discourse knowledge

Developmental studies suggest that children have discourse knowledge to some extent and can use it in language comprehension and production. In addition, researchers have started examining how discourse information help learners learning new words. The following reviews studies on what children know about discourse and how discourse information could help them learning new words.

5.4.1 Children's discourse knowledge

Some studies have investigated how a certain salience factor, the first-mention bias, influences on children's pronoun resolution when the pronoun is ambiguous. For example in (19), most adult English speakers prefer to interpret *she* as referring to *Jane Austen*.

(19) An example from Hartshorne, Nappa, and Snedeker (2014):

Jane Austen was born long before Agatha Christie. She wrote many books.

Song and Fisher (2005, 2007) examined whether children's pronoun resolution is affected by the first-mention bias. Like adults, three year old children in their experiments prefer to choose the first-mention in the preceding context as a referent of the pronoun. On the other hand, Arnold, Brown-Schmidt, and Trueswell (2007) reported that three to five year old children only use the gender information but not the first-mention information in determining the referent of the pronoun. Arnold et al. (2007) also points out that children's first-mention effect in Song and Fisher's eye-tracking experiment was much later than the gender effect in their experiment, suggesting that children's on-line processing of discourse information is still developing. Hartshorne et al. (2014) raise a possibility that the divergence in previous experiments could be due to methodological differences and/or U-shaped development (a developmental trajectory that shows a systematic drop in performance typically due to children's over-generalization). They show that children can use first-mention information, but do so too slowly to have been caught in the previous experiments.

A number of studies across languages show that children's use of referring expressions is similar to adults' in that they appear to select the expressions according to referents' salience. That is, they use attenuated forms to refer to salient referents and use more informative forms to refer to less salient referents (P. Clancy, 1993; Guerriero, Cooper, Oshima-Takane, & Kuriyama, 2001; Allen & Schröder, 2003; Serratrice, 2005; Skarabela, 2007; Serratrice, 2008; Salomo, Lieven, & Tomasello, 2010; Serratrice, 2013) Some studies also show children's sensitivity to visual and/or social cues in the choices of referring expressions, such as visual presence of the referent, joint attention, and shared information (Campbell, Brooks, & Tomasello, 2000; Matthews, Lieven, Theakston, & Tomasello, 2006; Serratrice, 2008; Skarabela, Allen, & Scott-Phillips, 2013; Hughes & Allen, 2013, 2014).

Though the range of phenomena is limited, these studies suggest that children have adult-like discourse knowledge to some extent (first-mention bias and salience in particular) and may be able to use this knowledge to understand and produce language.

5.4.2 The role of discourse information in word learning

Recent studies have started investigating how children's discourse knowledge could help using discourse information for word learning. M. Frank et al. (2013) suggest a role of discourse information in word learning by examining video-recorded child-caregiver interaction (6-18 months old). They annotated eye-gaze, hand position, and discourse continuity and found that individual social cues (eye-gaze and hand position) were noisy, but combining the discourse continuity cue with these cues would help estimating the speakers' intended referent. A follow-up study (Rohde & Frank, 2014) found that many cues (social, lexical, syntactic) that are used to signal discourse topicality in adult speech are also available in child-directed speech, suggesting that children may be able to use discourse topicality information even if they cannot interpret individual sentences. Based on these two studies, Horowitz and Frank (2015) conducted experiments to investigate adults' and children's sensitivity to discourse continuity to infer speakers' intended referents, without social cues such as pointing and eye-gaze. They found that adults and children (3-6 years old, but not 2 years old) can use discourse information (how utterances are relate within a discourse) to infer speakers' intended referents. They suggest that this kind of information may help learners smooth noisy social cues and provide clearer discourse structure, hence helping word learning. Similarly, Sullivan and Barner (2015) show that children as young as 2 years old can learn new words based on the inference of the relation between the new words and surrounding discourse context, but not based on individual words within the discourse.

These studies suggest that discourse information may be a key to learn new words, particularly in a situation where other learning strategies such as social cues (Baldwin, 1993; Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998) and syntactic cues (L. Gleitman, 1990) are not available (Pinker, 1979; Sullivan & Barner, 2015). Our work provides additional evidence that discourse information, that is expectations about the referents of the pronouns estimated from discourse, could help ideal learners to learn grammatical categories of pronouns. The next section illustrates how we estimate the expectations about the referents of the pronouns from discourse.

5.5 Measuring discourse information

If learners can use discourse information to infer the referent of the pronoun, this could provide information that can help them categorize pronouns into different grammatical classes. This section describes an experiment in Orita, McKeown, Feldman, Lidz, and Boyd-Graber (2013) that measures to what extent discourse context is informative for adults to predict which entities are likely to be referred to by pronouns. The results of this experiment provides the upper bound[1] of the information in discourse that learners can potentially use to learn the grammatical

[1] This experiment provides *upper* bound because it is estimated from adult speakers who have the target knowledge children would eventually acquire.

categories of pronouns. Note that this experiment measures discourse information as empirically opposed to the last two studies in this dissertation where I examined what factors contribute to discourse information.

Orita et al. (2013) used a variant of the human simulation paradigm (Gillette, Gleitman, Gleitman, & Lederer, 1999) to determine whether learners could guess the identity of an unknown word that was originally either a reflexive pronoun, non-reflexive pronoun, or lexical noun phrase, using only language contexts.

The human simulation paradigm has previously been used to investigate learning of lexical nouns and verbs (L. R. Gleitman, Kimberly, Nappa, Papafragou, & Trueswell, 2005). Past experiments using the human simulation paradigm have examined whether adults (Gillette et al., 1999; Snedeker & Gleitman, 2004; Kako, 2005) or older children (Piccin & Waxman, 2007) can guess the identity of common nouns and/or verbs from contextual information. In this paradigm, adult participants are usually provided partial clues such as a sentence with a word replaced with a nonsense word and asked to guess the identity of that word, based on various linguistic and/or information from the scene. This paradigm simulates what can be inferred about the meaning of a word by a language learner who hears a word but does not know its meaning.

In Orita et al. (2013), adult participants were shown text excerpts of real conversations between adults and children from CHILDES ENG-USA section (MacWhinney, 2000b). From a list of choices provided, they had to correctly guess the identity of a phrase that had been blanked out, which was originally either a reflexive pronoun, non-reflexive pronoun, or lexical noun phrase. The goal of this experiment is to see whether the provided contextual information is sufficient for adults to guess the meaning of the missing word. If adult participants can guess the identity of the missing word, this would be evidence that language learners might possibly determine the referent of unknown pronouns from conversational context alone.

Adult participants (n=20) saw 75 discourse excerpts from CHILDES. Each discourse excerpt contains one bolded sentence where a word had been blanked out as in Table 5.1. This target sentence always came from an adult utterance. There were 12 lines of dialogue before the target sentence and six lines afterwards. In order to factor out any possible contribution of verb knowledge to determine which pronoun was intended, the target sentence contained one of five verbs: *see*, *cover*, *dry*, *hurt*, *help*. These verbs were chosen based on the frequency and fraction of the time they were used with reflexive object in adult utterances in the US English section of the CHILDES database.

Participants' task was to identify the missing noun phrase from 15 possible choices as in Table 5.1. The choices always included the same five reflexive pronouns (*yourself*, *myself*, *ourselves*, *himself*, *themselves*), non-reflexive pronouns (*you*, *me*, *us*, *him*, *them*), and five lexical noun phrases which would have been prominent in each conversation: e.g., the names of the participants (including Mommy or Daddy) and prominent people or objects mentioned in the conversational excerpts.

The deleted noun phrases belonged to one of three categories: 25 were reflexive pronouns (4 tokens of *myself*, 1 token of *ourselves*, 7 tokens of *himself*, 10 tokens of *yourself*, and 3 tokens of *themselves*), 25 were non-reflexive pronouns (4 tokens of *me*, 1 token of *us*, 7 tokens of *him*, 10 tokens of *you*, 3 tokens of *them*), and 25 were

Participants:	Meghan (age three years), Mother
Mother:	Get out. Didn't you want to get out?
Meghan:	yeah.
Mother:	Where's your...
Meghan:	Ouch! [mumbles] hard.
Mother:	Dry your back. Here.
Meghan:	I can't.
Mother:	**You can dry ______.**
Meghan:	I can't.
Mother:	Just turn around here. Pick up the feet. You've got to dry it.
Meghan:	Uhhuh.
Mother:	Go ahead.

From the following options, circle the one which you think goes in the blank.

him	himself	me	Meghan	myself	ourselves
the bathmat	the feet	the tub	them	themselves	us
you	your back	yourself			

Table 5.1: Example stimulus item in Orita et al. (2013)

lexical NPs. This led to a total of 75 test items.

Overall, participants in their experiment were reasonably accurate at guessing the correct word from a list of 15 choices. The first row in Table 5.2 breaks up guesses of the correct word by syntactic category of the NP (reflexive pronouns, non-reflexive pronouns, or lexical NPs). Individual participants chose the correct NP out of 15 choices an average of 63.8% of the time. This ranged from 32.4% for the least accurate participant to 84.2% for the most accurate participant, with a standard deviation of 10.6%, and was significantly better than chance ($t(39) = 34.19$, $p < 0.0001$). These results show that adults can usually guess the identity of a missing NP given only a small amount of linguistic context.

However, these results underestimate participants' ability to guess what is

	Lexical NP	Non-reflexive	Reflexive
% correct word	61.75	70.25	64.25
% plausibly correct word	66.75	81.25	68

Table 5.2: Percentage of correct answers and answers with a plausibly correct referent in Orita et al. (2013)

	Lexical NP	Non-reflexive	Reflexive
% Lexical NP guesses	**71.8**	23.4	15.8
% Non-reflexive guesses	23.2	**73.4**	16
% Reflexive guesses	5	3.2	**68.2**

Table 5.3: Confusion matrix obtained in Orita et al. (2013)

being referred to. The second row in Table 5.2 shows guesses of a plausibly correct word, a word that plausibly had the same intended referent as the correct word (for instance, a pronoun with the same gender/number features as the name that had actually been used, or vice versa). These results show that adults are good at guessing which entity is referred to given a context.

Table 5.3 breaks up the results by syntactic category of the NP. Participants' guesses were usually of the same category that the actual word had been. Importantly, adults usually guessed correctly whether the missing word had been a reflexive pronoun—when the word actually had been reflexive, participants guessed a reflexive 68.2% of the time. When the word had been a lexical NP or a non-reflexive pronoun, they almost never guessed that it had been a reflexive.

This task parallels that of a child identifying an unfamiliar word. Of course, the parallel is not complete. In some ways, adult participants were provided with less information than the children they were meant to simulate: they only received a small excerpt of the conversation and did not receive any visual information. In other ways, the participants had more data: they already knew the meanings of all of the other words in the conversation, they were limited to 15 choices of possible meaning, and they had full syntactic and discourse knowledge where children might only have partial knowledge. Furthermore, choosing an answer in their experiment was not subject to any time pressures, whereas in actual acquisition processing speed could potentially impact the learner's ability to use the discourse context as an information source.

However, to the extent that the adult simulation reflects the prior information presented in the discourse, it provides an estimate of the upper bound of the information that children might have access to. Where adults (who already know the distribution of reflexives) can guess that a missing word is reflexive, a child might be able to guess that a missing word co-refers with a specific NP. Together with syntactic knowledge of locality and c-command, this should provide learners with useful information for acquiring grammatical categories of pronouns (here reflexive and non-reflexive). To explore this possibility, we formalize a Bayesian model that learns to categorize pronouns.

5.6 Bayesian Model

Learning grammatical categories of pronouns (i.e., learning the distributional profiles of the pronoun categories and their word distributions) poses a circularity problem. In order to learn the grammatical category of the pronoun, learners need to know the reference of the pronoun, but in order to infer the reference, they might need to know the grammatical category of the pronoun. The hypothesis is that if learners can guess which entity the pronoun is likely to refer to using discourse knowledge, this guess can help them categorize pronouns into different grammatical classes.

In this section, we develop a Bayesian model that integrates the discourse knowledge estimated in Orita et al. (2013) to investigate the degree to which discourse knowledge and syntactic knowledge help an ideal learner find the correct grammatical categories of English pronouns. The model discovers:

(20) a. how many pronoun categories there are in a language

b. the distribution of pronouns in each category

c. which syntactic position of an antecedent is associated with each pronoun category

This ideal learner is assumed to have the following prior knowledge.

(21) a. discourse knowledge that helps define the distribution of the potential antecedents

b. syntactic knowledge relevant to pronoun categories (details follow)

c. lexical knowledge that is sufficient for distinguishing pronouns from lexical noun phrases

Though the prior discourse knowledge is estimated from adults' responses in the experiment and the overall performance was good as seen in the previous section, the responses sometimes wrong or ambiguous depending on the given dialogue.

Other linguistic information relevant to pronouns, such as gender and number, are not represented in our model.

Regarding (b) above, this ideal learner is assumed to already know locality and c-command before learning pronoun categories, and is further assumed to know that these are relevant for categorizing pronouns. Thus, the learner is able to identify the syntactic position of each potential antecedent. The model distinguishes four syntactic positions based on the knowledge of locality and c-command; [+local, +c-command], [+local, -c-command], [-local, +c-command], and [-local, -c-command]. In English, if an antecedent is in a syntactic position described by [+local, +c-command] as in (22), an unknown pronoun *blick* must be a reflexive pronoun.

(22) Alice$_i$ likes *blick*$_i$. (*Alice* is in [+local, +c-command])

If the potential antecedent is elsewhere as in (23), an unknown pronoun *splink* must be a non-reflexive pronoun.

(23) a. Alice$_i$'s sister likes *splink*$_i$. (*Alice* is in [+local, -c-command])

b. Alice$_i$ said that Red Queen likes *splink*$_i$. (*Alice* is in [-local, +c-command])

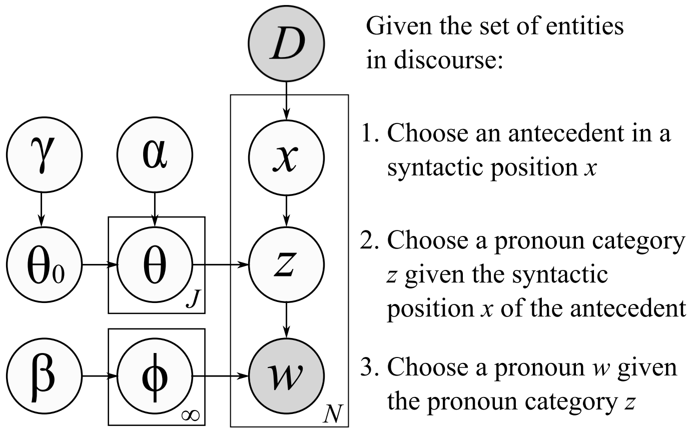
c. While the Red Queen$_i$ is sleeping, Alice kicked her$_i$. (*Red Queen* is in [-local, -c-command])

However, this learner does not know in advance how many pronoun categories there are and which syntactic position is associated with which pronoun category, and needs to recover them from the input.

This model is a computational-level model (Marr, 1982): The model investigates whether the information in discourse could be sufficient to learn the grammatical categories of English pronouns given the prior knowledge in (21) in principle. We do not assume any specific algorithm (such as Gibbs sampling) to compute probabilities that might be used by learners. This model provides a measure of whether there is sufficient information to allow a model with a particular structure to acquire the appropriate categories. In other words, the model tells us whether a certain way of thinking about the problem (i.e., the structure of the model) combined with a certain kind of information (i.e., the input as it is represented by the model) could lead to successful learning.

5.6.1 Generative Model

We use a Bayesian network to illustrate how the observations (i.e., pronouns a learner observed) could have been generated (see Appendix A). Our model assumes the following generative process. For each pronoun, an antecedent in one of the four syntactic positions described above is chosen given prior discourse knowledge ($\mathcal{D}$). Then a pronoun category is chosen based on the syntactic position of the antecedent (more specifically, the antecedent is in a syntactic position which bears certain features relative to the pronoun), and a pronoun is generated from the chosen pronoun category. Figure 5.2 illustrates this process with a graphical model. This model is a nonparametric extension to the author-topic model (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004) that allows for an infinite number of categories (called topics in their model).



D
Given the set of entities in discourse:
γ
α
x
1. Choose an antecedent in a syntactic position *x*
θ0
θ
J
z
2. Choose a pronoun category *z* given the syntactic position *x* of the antecedent
β
ϕ
∞
w
N
3. Choose a pronoun *w* given the pronoun category *z*

Figure 5.2: Graphical Model

Each antecedent-category distribution θ_j is a random variable that encodes the distribution over pronoun categories favored by an antecedent in syntactic position j. For example, if the model works correctly, then the category *reflexive* would have high probability in the distribution $\theta_{[+\text{local},+\text{c-command}]}$. (Here we use the category name *reflexive* for exposition, but the model does not associate any labels with the pronoun categories it recovers.) Each category-word distribution ϕ_k is a distribution over words that encodes the probability distribution over pronouns in pronoun category k. For example, if the model works correctly, pronouns such as *herself* and *myself* would have high probabilities in the distribution $\phi_{\text{reflexive}}$. In addition to learning this distribution, our model learns the number of pronoun categories needed to describe the data. For each pronoun in the corpus, an antecedent in a syntactic position x is assumed to be sampled from a distribution we refer to as discourse knowledge $\mathcal{D}$ (see the section 5.6.2 for the details). A pronoun category z is then sampled from the multinomial distribution with parameter θ associated with the syntactic position x of the antecedent and a pronoun w is sampled from a multinomial distribution with parameter ϕ associated with pronoun category z.

To learn the number of pronoun categories based on the observed data, we use Chinese Restaurant Franchise in the hierarchical Dirichlet process set up (Teh et al., 2006). This allows potentially infinite number of categories but has a bias towards fewer categories, and it allows categories to appear across multiple grammatical contexts (see Appendix A for more details on Chinese Restaurant Franchise). The distribution θ_0 is the distribution of global pronoun categories shared across different grammatical contexts, and the distribution θ_j is the distribution of each grammatical context over pronoun categories. This structure ensures that the model can share pronoun categories across and within different grammatical contexts θ_j. In the Chinese Restaurant Franchise metaphor, the distributions θ_j correspond to the lower-level restaurants, and the distribution θ_0 corresponds to the franchise restaurant. A table in the lower-level restaurant θ_j is a metaphor of the index that associates pronoun tokens with a pronoun category, and a dish in the franchise restaurant θ_0 is a metaphor of the index of a pronoun category. The summary of the generative process follows.

1. Draw a distribution over pronoun categories $\theta_0 \sim \mathrm{GEM}(\gamma)$, where GEM is the Griffiths, Engen, McCloskey distribution (Pitman, 2002).
2. For each antecedent syntactic position $j = 1 \ldots 4$, draw a pronoun category distribution $\theta_j \sim \mathrm{DP}(\alpha, \theta_0)$.
3. For each pronoun category $k = 1 \ldots \infty$, draw a distribution over tokens $\phi_k \sim \mathrm{Dir}(\beta)$.

4. For each pronoun in the corpus $n = 1 \ldots N$
   (a) Draw an antecedent syntactic position from the discourse knowledge $x_n \sim \mathcal{D}$
   (b) Draw a pronoun category $z_n \sim \mathrm{Mult}(\theta_{x_n})$.
   (c) Draw a word $w_n \sim \mathrm{Mult}(\phi_{z_n})$.

Intuitively, the GEM distribution is an infinite version of the Dirichlet prior distribution. Where a Dirichlet distribution is a prior over finite-dimensional multinomial distributions, a GEM distribution is a prior over infinite-dimensional multinomial distributions.

5.6.2 Prior Discourse Knowledge

The observed discourse knowledge distribution $\mathcal{D}$ defines a prior distribution over potential antecedents in the discourse. Recall that our ideal learner maps each antecedent in the discourse deterministically to its syntactic position (defined in terms of locality and c-command), and in this way $\mathcal{D}$ defines a distribution over syntactic positions x for each pronoun's antecedent.

Rather than specify a parametric form for this prior distribution, we estimate it directly from participants' responses from Orita et al. (2013). In one experimental item, for example, participants guessed the identity of the missing word in the sentence "You drying __ off?". Nine out of 20 participants guessed that the missing word is *yourself*, six out of 20 guessed *him*, three out of 20 guessed *me*, and two out of 20 guessed *Seth*. Under the assumption that experimental participants have sampled their responses from a shared prior distribution over entities in the discourse, these guesses provide an estimate of participants' beliefs about how likely each entity is to be referred to in the discourse.

Where participants chose *yourself*, the antecedent of this pronoun is *you*, which is a local and c-commanding antecedent. Where participants chose *him* and *me*, the antecedents could be in any of the remaining three syntactic positions, but in this particular dialogue the only potential antecedents for non-reflexives are neither local nor c-commanding. In cases of non-reflexive guesses where potential antecedents appeared in multiple syntactic positions, we assumed the prior probability for each syntactic position to be proportional to the number of potential antecedents in that position. We ignored responses in which participants chose lexical NPs (here *Seth*) based on the assumption that learners distinguish pronouns from lexical NPs. We then normalized each count by the total number of pronoun guesses. The resulting prior distribution over syntactic positions for antecedents in this example is $p(x_{[+local,+c\text{-}command]}|\mathcal{D}) = 0.5$ and $p(x_{[-local,-c\text{-}command]}|\mathcal{D}) = 0.5$ where $\mathcal{D}$ stands for discourse knowledge. In this way the results from Experiment 1 provide us with an informative prior distribution regarding which entities are likely to be referred to in the discourse, and through simulations we can test whether this prior knowledge helps an ideal learner acquire pronoun categories. Note that this prior distribution differs from the distribution seen in Tables 5.2 and 5.3 because it is based on individual experimental items rather than on aggregated data.

5.6.3 Inference

Given this generative process, we can use Bayesian inference to recover the learner's beliefs about pronoun categories. We want to estimate four sets of latent variables: the antecedent-category parameter θ, the category-word parameter ϕ, the antecedent's syntactic position x, and the pronoun category z. We use the Gibbs sampling algorithm from Rosen-Zvi et al. (2004) to estimate these unknown parameters. Instead of directly estimating parameters θ and ϕ, we compute the posterior on x and z (parameters θ and ϕ are integrated out), and use the results to estimate the antecedent-category parameter θ and the category-word parameter ϕ. A reader who is satisfied with this idea can safely skip the technical details presented in this section.

Gibbs sampling (Geman & Geman, 1984) is a method to approximate the posterior distribution of observing data that best-fits the model we define. This sampling method is often used in Bayesian models when we cannot compute the posterior distribution analytically. Gibbs sampling works as follows: Suppose that we want to compute probabilities of latent categories $p(Z) = p(z_1, ..., z_N)$ In the pronoun category learning problem, Z corresponds to a vector of the pronoun category assignments, e.g., [category 1, category 1, category 3, category 2, ..., category 2]. It is hard to compute $p(Z)$ directly, but we can sample z_i from $p(z_i|Z_{-i})$ where term Z_{-i} indicates Z removed z_i In the pronoun category learning problem, $p(z_i|Z_{-i})$ corresponds to the probability of i_{th} pronoun category assignment given all pronoun category assignments removed i_{th} category assignment. Gibbs sampling first initializes Z randomly, computes the conditional probability of $p(z_i|Z_{-i})$ and decides new z_i, computes $p(z_{i+1}|Z_{-i+1})$ given updated Z, and repeats this process for all z $(z_{1,\ldots,N})$. When this process is sufficiently repeated, $Z_{\text{final state}}$ is known to be derived from the true distribution $p(Z)$.

In this model, the assignments of x and z for a particular token are sampled as a block, conditioned on everything else, so that in each iteration we compute the conditional distribution $p(x_i, z_i|w_i, \mathbf{x}_{-i}, \mathbf{z}_{-i})$ where $\mathbf{x}_{-i}$ and $\mathbf{z}_{-i}$ denote all syntactic position and category assignments not including the ith pronoun. This is proportional to

$$p(w_i|x_i, z_i, \mathbf{x}_{-i}, \mathbf{z}_{-i}) \cdot p(z_i|x_i, \mathbf{x}_{-i}, \mathbf{z}_{-i}) \cdot p(x_i|\mathcal{D}) \tag{5.1}$$

where the first term is the likelihood, the second term is defined by the Chinese Restaurant Franchise as described in Appendix A, and the third term is estimated directly from participants' responses in Orita et al. (2013).

The likelihood in the equation (5.1) is the probability of w_i assigned to $x_i = j,\ z_i = k$. The parameter ϕ is integrated out, yielding

$$p(w_i = m|x_i = j, z_i = k, \mathbf{z_{-i}}, \mathbf{x_{-i}}) = \frac{\beta + V_{k,m}}{\Sigma_i^V \beta_i + V_{k,i}} \tag{5.2}$$

where $w_i = m$ denotes the observation that the ith pronoun is the mth pronoun in the pronoun lexicon, $z_i = k$ and $x_i = j$ denote the assignments of the ith pronoun to category k and grammatical context of the antecedent j respectively. Term $V_{k,m}$

denotes the number of times pronoun m is used in pronoun category k, not including the current instance. Note that the likelihood for new categories is the same as for old categories with all $V_{k,m} = 0$, and can be written as $\frac{1}{|V|}$.

The following is the second term in the equation (5.1) defined by the Chinese Restaurant Franchise.

$$p(z_i = k | x_i = j, \mathbf{x_{-i}}, \mathbf{z_{-i}}) = \begin{cases} \frac{N_{j,k}}{\alpha + N_{j,\cdot} - 1} \quad N_{j,k} > 0 \\ \text{(existing category in grammatical context } j\text{)} \\ \\ \frac{\alpha}{\alpha + N_{j,\cdot} - 1} \cdot \frac{M_k}{\gamma + M_{\cdot} - 1} \quad M_k > 0 \\ \text{(new category in grammatical context } j\text{,} \\ \text{existing category across all grammatical contexts)} \\ \\ \frac{\alpha}{\alpha + N_{j,\cdot} - 1} \cdot \frac{\gamma}{\gamma + M_{\cdot} - 1} \\ \text{(new category in grammatical context } j\text{,} \\ \text{new category across all grammatical contexts)} \end{cases} \tag{5.3}$$

Term $N_{j,k}$ denotes the number of times pronoun category k is used in grammatical context j, not including the current instance, and term M_k denotes the number of times pronoun category k is associated with a group of pronouns across all grammatical contexts.

By plugging-in the equations in (5.2) and (5.3) to the equation (5.1), we have

$$p(x_i = j, z_i = k | w_i = m, \mathbf{z_{-i}}, \mathbf{x_{-i}}) \propto \begin{cases} \frac{\beta + V_{k,m}}{\Sigma_i^V \beta_i + V_{k,i}} \cdot \frac{N_{j,k}}{\alpha + N_{j,\cdot}} \cdot p(x_i = j | \mathcal{D}) \\ \text{(existing category in grammatical context } j\text{)} \\ \\ \frac{\beta + V_{k,m}}{\Sigma_i^V \beta_i + V_{k,i}} \cdot \frac{\alpha}{\alpha + N_{j,\cdot}} \cdot \frac{M_k}{\gamma + M.} \cdot p(x_i = j | \mathcal{D}) \\ \text{(new category in grammatical context } j\text{,} \\ \text{existing category across all grammatical contexts)} \\ \\ \frac{1}{|V|} \cdot \frac{\alpha}{\alpha + N_{j,\cdot}} \cdot \frac{\gamma}{\gamma + M.} \cdot p(x_i = j | \mathcal{D}) \\ \text{(new category in grammatical context } j\text{,} \\ \text{new category across all grammatical contexts)} \end{cases} \tag{5.4}$$

We use the equation in (5.4) to do Gibbs sampling. Note that we do not have -1 in the denominators unlike the equation (5.3) because when sampling we decrement the current instance before computing (5.4).

In sum, we want to recover the ideal learner's belief about pronoun categories, i.e., the distributions of pronoun categories associated with certain syntactic environment (distributions parameterized by θ) and the distributions of pronouns in each category (distributions parameterized by ϕ). The problem is that it is hard to directly estimate these distributions. Instead, we sample category and syntactic position assignments together given the prior distribution about the possible antecedents. This nicely breaks up into three computable terms as in Equation

(5.1) by using Bayes rule, and by using resulting sampling chain, we can infer the distributions we are interested in.

5.6.4 Simulations

In order to test the effectiveness of discourse information for the categorization of pronouns, simulations compare four models: Adult-like discourse + Strong syntax model, Uniform discourse + Weak syntax model, Adult-like discourse + Weak syntax model, and Uniform discourse + Strong syntax model.

The Adult-like discourse + Strong syntax model has the adult-like discourse knowledge estimated in Orita et al. (2013) and built-in knowledge of the grammatical constraints on reflexive and non-reflexive pronouns in English. This model knows there are two grammatical categories of pronouns in English. Furthermore, it knows that pronouns that have local c-commanding antecedents are reflexive pronouns and that pronouns that do not have local c-commanding antecedents are non-reflexive pronouns (i.e., the antecedent-category parameter θ is observed and syntactic position of an antecedent x is deterministic). Thus, the model only needs to learn the distribution of each category over pronouns. In other words, this model knows everything relevant beforehand except the mapping of pronouns into categories. If this model is able to learn the correct grammatical categories of pronouns, then we can ask which prior knowledge was necessary for the ideal learner to succeed in this learning task.

The Uniform discourse + Weak syntax model has information about locality and c-command, but it needs to learn both how many pronoun categories there are, what the distributional profile of each category is, and how to map each pronoun into those categories. It also lacks information about which entities are likely to be referred to in the discourse. It assumes that potential antecedents are sampled uniformly, so that $p(x_i|\mathcal{D})$ is defined by counting the number of discourse entities that appear in each syntactic position.

The Adult-like discourse + Weak syntax model is identical to the Uniform discourse + Weak syntax model, but it contains the adult-like discourse knowledge estimated in Orita et al. (2013). Comparing the performance of this model to the Uniform discourse + Weak syntax model allows us to quantify the degree to which discourse information helps an ideal learner acquire pronoun categories.

The Uniform discourse + Strong syntax model is similar to the Uniform discourse + Weak syntax model in that it assumes that potential antecedents are sampled uniformly. This model is similar to the Adult-like discourse + Strong syntax model in that it incorporates built-in knowledge of the grammatical constraints on reflexive and non-reflexive pronouns in English. This model only needs to learn the distribution of each category over pronouns. Comparing this model to the Uniform discourse + Weak syntax model allows us to examine whether this type of strong prior syntactic knowledge is sufficient to help learners categorize pronouns.

Each model was trained on 50 dialogues from Orita et al. (2013), 25 with reflexive and 25 with non-reflexive pronouns. For each dialogue, the model was provided with the pronoun, a prior distribution over possible antecedents for that pronoun, and the syntactic positions of those antecedents relative to the pronoun. Through the unsupervised learning procedure described above, the models recovered a distribution over categories associated with each syntactic position (e.g., one category may be associated with a syntactic position *c-commanding and local*, which corresponds to the reflexives in English) and a distribution over pronouns for each category (e.g., pronouns such as *herself* and *myself* would have high probabilities in one category that is associated with a syntactic position *c-commanding and local*).

5.6.5 Results

For each model, we ran 10 independent Gibbs sampling chains for 2000 iterations each (we obtained similar results by running 1000 iterations).

Hyperparameters α, β, and γ were fixed at 1.0, 0.01, and 0.001, respectively. The same parameter values were used for all three models. These are the best parameter values based on multiple runs, but results were qualitatively consistent across a range of parameter values. We also tried to resample these hyperparameters given the current assignments of x and z by using slice sampling (Neal, 2003), but performance did not reach the best parameter values chosen by hand. It seemed to get trapped in local maxima due to the data size which is very small. Note the Strong syntax models needs only the parameter $\beta = 0.01$ because these models do not have to learn the number of pronoun categories and the association between syntactic positions of antecedents and pronoun categories.

We computed pairwise F-scores (see details in Appendix B) using the final samples from each chain. The pairwise F-score ranges between 0 to 1; higher scores mean better performance. This is computed by counting how pairs of pronouns are assigned to the category by the model. Using pairwise F-scores is a reasonable choice because this measure does not require adult-like gold-standard knowledge of pronoun categories which is not available to learners, but still correlates with gold-standard measures (S. Frank, Goldwater, & Keller, 2009).

The Adult-like discourse + Strong syntax model perfectly categorized English pronouns into two classes, achieving a mean pairwise F-score of 1.00 across the 10 sampling runs. Now we can ask which prior knowledge was necessary or unnecessary for the successful learning. Table 5.4 shows the distribution over pronouns belonging to each category obtained at the 2000th iteration of the sampling run with the highest likelihood. The maximum likelihood estimate $p(word|category)$ gives the proportion of times each pronoun occurs in a category, based on a single sample from the posterior distribution over z and x.

Category 1		Category 2	
Word	p(word|category)	Word	p(word|category)
myself	0.16	myself	0.0
ourselves	0.04	ourselves	0.0
yourself	0.4	yourself	0.0
himself	0.28	himself	0.0
themselves	0.12	themselves	0.0
me	0.0	me	0.16
us	0.0	us	0.04
you	0.0	you	0.4
him	0.0	him	0.28
them	0.0	them	0.12

Table 5.4: The Adult-like discourse + Strong syntax model results

The Uniform discourse + Weak syntax model consistently failed to learn the correct categories, achieving a mean pairwise F-score of 0.55 across the 10 sampling chains. In all 10 chains, the model learned 3-4 categories, where the

Category 1		Category 2	
Word	p(word|category)	Word	p(word|category)
myself	0.0	myself	0.16
ourselves	0.0	ourselves	0.0
yourself	0.0	yourself	0.4
himself	0.0	himself	0.28
themselves	0.0	themselves	0.12
me	0.29	me	0.0
us	0.0	us	0.04
you	0.0	you	0.0
him	0.5	him	0.0
them	0.21	them	0.0

Category 3	
Word	p(word|category)
myself	0.0
ourselves	0.09
yourself	0.0
himself	0.0
themselves	0.0
me	0.0
us	0.0
you	0.91
him	0.0
them	0.0

Table 5.5: Uniform discourse + Weak syntax model results

correct number of categories is two. Table 5.5 shows the distribution over pronouns belonging to each category obtained at the 2000th iteration of the sampling run with the highest likelihood.

The Adult-like discourse + Weak syntax model performed much better than the Uniform discourse + Weak syntax model, achieving a mean pairwise F-score of 0.97 across the 10 sampling runs. In seven of the 10 runs, the model perfectly categorized English pronouns into two classes. In two additional runs, the model learned two categories, but the membership was not consistent. In the final run, the model learned three categories. Table 5.6 shows the pronouns belonging to each category, obtained at the 2000th iteration of the Gibbs sampling run which had the highest likelihood. The pronouns associated with each category are reflexive pronouns and non-reflexive pronouns, respectively. This model also learned

Category 1		Category 2	
Word	p(word|category)	Word	p(word|category)
myself	0.16	myself	0.0
ourselves	0.04	ourselves	0.0
yourself	0.4	yourself	0.0
himself	0.28	himself	0.0
themselves	0.12	themselves	0.0
me	0.0	me	0.16
us	0.0	us	0.04
you	0.0	you	0.4
him	0.0	him	0.28
them	0.0	them	0.12

Table 5.6: Adult-like discourse + Weak syntax model results

that there are exactly two categories, as expected. These results indicate that discourse information can help an ideal learner categorize pronouns with knowledge of the relevance of c-command and locality to defining pronoun distributions.

Although the Uniform discourse + Weak syntax model has prior knowledge of c-command and locality, it is still possible that the low performance in this model might result from insufficient syntactic knowledge. For this reason, We compare the Uniform discourse + Strong syntax model with the Uniform discourse + Weak syntax model to see whether even stronger prior syntactic knowledge is sufficient for categorizing pronouns. The mean F-score was 0.56 for this Uniform discourse + Strong syntax model. Table 5.7 shows the pronouns in each category, obtained at the 2000th iteration of a Gibbs sampling run which had the highest likelihood. The lack of improvement of the Uniform discourse + Strong syntax model over the Uniform discourse + Weak syntax model suggests that simply having strong prior syntactic knowledge is not sufficient for acquiring grammatical categories of pronouns.

In summary, these simulation results suggest that knowing which entities are likely to be referred to in the discourse can help learners acquire grammatical cat-

Category 1		Category 2	
Word	p(word|category)	Word	p(word|category)
myself	0.12	myself	0.0
ourselves	0.0	ourselves	0.06
yourself	0.29	yourself	0.0
himself	0.21	himself	0.0
themselves	0.09	themselves	0.0
me	0.0	me	0.25
us	0.0	us	0.06
you	0.0	you	0.63
him	0.21	him	0.0
them	0.09	them	0.0

Table 5.7: Uniform discourse + Strong syntax model results

egories of pronouns. On the other hand, simply having stronger prior knowledge about the grammatical distribution of pronouns is not sufficient to support the acquisition of pronoun categories. The comparison between the Adult-like discourse + Weak syntax model and the Uniform discourse + Weak syntax model shows that the model does not need to know the number of categories and the syntactic property of each category antecedently if it has the adult-like discourse knowledge and the knowledge of the relevance of c-command and locality to defining pronoun distributions. The comparison between the Uniform discourse + Weak syntax model and the Uniform discourse + Strong syntax model shows that simply having strong prior syntactic knowledge (i.e., the knowledge of two syntactic categories that correspond to reflexive and non-reflexive) is not sufficient for acquiring grammatical categories of pronouns. In other words, these four models together show (i) that having knowledge that there are two syntactic categories that correspond to reflexive and non-reflexive is neither necessary nor sufficient and (b) that discourse knowledge is both necessary and sufficient, along with the knowledge that c-command and locality are relevant. Table 5.8 shows a summary of model comparison.

Model	Discourse knowledge	Syntactic knowledge	Result
Uniform + WeakSyntax	uniform	locality, c-command	failed
Discourse + WeakSyntax	adult-like	locality, c-command	learned
Uniform + StrongSyntax	uniform	reflexive and non-reflexive	failed
Discourse + StrongSyntax	adult-like	reflexive and non-reflexive	learned

Table 5.8: Summary of model comparison

5.6.6 Summary

Simulation results suggest that knowing which entities are likely to be referred to in the discourse can help learners acquire grammatical categories of pronouns if learners can pay attention to the relevant syntactic position of the potential antecedent. On the other hand, simply having the strong syntactic prior knowledge about the distribution of pronouns (reflexive pronouns require local and c-command antecedents) is not sufficient to identify an unknown pronoun as reflexive unless the model has a reasonably good estimate of a pronoun's reference.

Though the experimental materials in Orita et al. (2013) were taken from CHILDES corpus, they do not fully reflect actual input children receive at least for three reasons: (i) the number of reflexive pronouns and non-reflexive pronouns are balanced as equal, (ii) verbs that take pronouns are balanced, and (iii) there is no *non-c-commanding and local* antecedent for non-reflexive pronouns in the input data. The first and second were necessary from the perspective of experimental design because these factored out potential confounds of frequency and verb in the experiment. The third means that it is sufficient to categorize pronouns by only using the locality feature with the current data. However, it is not entirely clear how the distribution of pronouns in the actual input looks like. The next section explores the distribution of pronouns in the child-directed speech.

5.7 Corpus study

The data that the model was trained with does not seem to reflect the distribution of actual input. The model was trained with 25 reflexive pronouns and 25 non-reflexive pronouns. Types of verbs that take pronouns as arguments are controlled. There are only five types of verbs (*cover*, *dry*, *help*, *hurt*, *see*) in the data. These items in Orita et al. (2013) were collected by searching many different corpus in CHILDES.

In this section, we investigate the distribution of pronouns in the relation to their antecedents. The goal of this preliminary corpus study is to explore the distribution of pronouns in more realistic data and compare with the model input.

5.7.1 Data

The Brown Adam corpus (Brown, 1973) in the CHILDES database (MacWhinney, 2000b) was used. His family was middle class, educated, and speakers of Standard American English. This corpus contains spontaneous speech recorded at home. All 55 files were used for search. His age ranges from 2;3 to 5;2. Note that this corpus consists of 2 hours of recording per 2 weeks. It is a tiny part of the utterances that Adam heard, but we suppose that this small subset of the input could approximate the true distribution that Adam received over the course of development.

5.7.2 Coding procedure

The coding scheme in Table 5.9 was created to extract as much relevant information as possible. Utterances that contain pronouns were extracted by using UNIX commands. First, his parents' speech was extracted, then reflexive and non-reflexive pronouns were extracted out of the parents' speech. Pronouns were coded based on the scheme in Table 5.9. Only object pronouns were coded in order to compare reflexive pronouns and non-reflexive pronouns. Pronouns counted are: *me*, *us*, *you*, *her*, *him*, *them*, *myself*, *ourselves*, *yourself*, *herself*, *himself*, and *themselves*. Pronouns *it* and *itself* were not counted because these pronouns were not in the model input data. Consecutive repetitions and uninterpretable fragment utterances were excluded. Fragments were separately tagged and counted.

The scope of the analyses is intra-sentence level but not discourse level. Discourse-relevant factors are not included in the data coding and the data analyses. The extra-sentential referents are not identified (i.e., just coded as a referent outside of the sentence).

If non-reflexive pronouns have extra-sentential antecedents, some properties of those antecedents were not coded. For example, tags 4, 5, 7 in Table 5.9 are unspecified for *them* in an utterance *why don't you wash them off?*. Coding examples are in Appendix C.

tag #	
1	Is an utterance fragment or not?
2	What is the form of a pronoun? (e.g., *myself*, *him*)
3	Is an antecedent extra-sentential or intra-sentential?
4	What is the form of the antecedent if intra-sentential? (e.g., lexical NP, pronoun, PRO, trace, imperative subject)
5	Does the antecedent c-command the pronoun and/or is it local if intra-sentential? (+C+L, +C-L, -C+L, -C-L, intensifier)
6	What is the grammatical position of the pronoun? (e.g., object, oblique object, indirect object, adnominal intensifier, adverbial intensifier)
7	What is the grammatical position of the antecedent if intra-sentential? (e.g., subject, object, oblique object, indirect object, possessive. If an intensifier, what is the grammatical position of the focus?)
8	What is the lexical head that takes the pronoun? (e.g., *hurt*, *by*)
9	What is the category of the head? (e.g., transitive verb, ditransitive verb, complex transitive verb, intransitive verb, preposition)
10	What is the tense of the clause containing the pronoun? (e.g., finite, non-finite, imperative)
11	What kind of the clause contains the pronoun? (e.g., main clause, imperative, complement, adjunct, bare-infinitive, relative clause, cleft)

Table 5.9: Coding scheme

5.7.3 Results

The total number of utterances from parents is 114,081. The total number of utterances that contain object pronouns listed above (both reflexive and non-reflexive) is 1,366. This number includes fragments. The total number of non-fragments is 1,274. The total number of fragments is 92. Of fragments, there are 3 reflexives (2 *himself*, 1 *yourself*) and 87 non-reflexives (62 *you*, 10 *them*, 9 *me*, 6 *him*, 2 *her*). The following analyses are based on 1,274 non-fragment sentences. In the following all tables, a number in each cell indicates frequency. Percentage next to each number indicates a proportion of the particular pronoun among reflexives and non-reflexives respectively.

The total number of reflexive pronouns is 78. The total number of non-reflexive pronouns is 1180. The ratio between reflexives and non-reflexives is 6 : 94, which is considerably different from the ratio in the model training data where we have 25 reflexives and 25 non-reflexives. Table 5.10 and Table 5.11 show the breakdown of each type of pronouns. There is a parallel with the model training data[2]. The most frequent item is a second person pronoun and the least frequent item is a first person plural pronoun.

myself	ourselves	yourself	himself	herself	themselves	total
8 (10.25%)	1 (1.28%)	56 (71.79%)	12 (15.38%)	1 (1.28%)	0 (0%)	78 (100%)

Table 5.10: Reflexive object pronouns in Adam corpus

me	us	you	him	her	them	total
246 (20.85%)	16 (1.36%)	377 (31.95%)	167 (14.15%)	99 (8.39%)	275 (23.31%)	1180 (100%)

Table 5.11: Non-reflexive object pronouns in Adam corpus

Table 5.12 breaks up pronouns by their grammatical positions. The prominent differences between Adam corpus and the model training data are oblique objects and intensifiers. Though there were no such items in the model training data, the majority proportion of pronouns in object position in Adam corpus seems to be reflected in the model data. In the Adam corpus, there are 7 reflexive pronouns used as intensifiers as in (24). All of the 7 intensifiers are adverbial intensifiers.

[2] Of 25 reflexive pronouns in the model training data, there were 4 tokens of *myself*, 1 token of *ourselves*, 10 tokens of *yourself*, 7 tokens of *himself*, and 3 tokens of *themselves*. Of 25 non-reflexive pronouns in the model training data, there were 4 tokens of *me*, 1 token of *us*, 10 tokens of *you*, 7 tokens of *him*, and 3 tokens of *them*.

	object	oblique object	intensifier	total
reflexives	48 (61.54%)	23 (29.49%)	7 (8.97%)	78 (100%)
non-reflexives	907 (76.86%)	273 (23.14%)	0	1180 (100%)
total	955 (75.91%)	296 (23.53%)	7 (0.56%)	1258 (100%)

Table 5.12: Grammatical positions of pronouns

(24) a. MOTHER: he'd like to do it himself. (adam45.cha)

b. MOTHER: you have some yourself. (adam15.cha)

Table 5.13 shows the distribution of antecedents of reflexive and non-reflexive pronouns, broken up by whether the antecedent is intra-sentential or extra-sentential. All reflexive pronouns in the Adam corpus have antecedents within a sentence. There is no exempt anaphor in this data set. This distribution parallels to the model training data. As for non-reflexive pronouns, most non-reflexive pronouns have extra-sentential antecedents (98.81%). This is slightly different from the distribution of non-reflexives in the model input in that there are 12% non-reflexives (3 items) that have intra-sentential antecedents.

	intra-sentential	extra-sentential	total
reflexives	78 (100%)	0 (0%)	78 (100%)
non-reflexives	15 (1.27%)	1165 (98.73%)	1180 (100%)
total	92 (7.31%)	1166 (92.69%)	1258 (100%)

Table 5.13: Intra-sentential and extra-sentential antecedents

Table 5.14 breaks up the intra-sentential antecedents in Table 5.13 by the syntactic position of the antecedent defined by c-command and locality. The abbreviation +C means that the antecedent c-commands the pronoun, vice versa for -C. The abbreviation +L means that the antecedent and the pronoun are in the same clause, and vice versa for -L. Examples of intra-sentential non-reflexive antecedents are shown in (25). In the model training data, there are 3 non-reflexive pronouns that have intra-sentential antecedents, and all of 3 antecedents are in a c-commanding and non-local position. This seems to reflect the proportion in Adam corpus in that the majority of non-reflexives that have intra-sentential antecedents have the c-commanding and non-local antecedents (73%). The implication of the distribution of non-reflexive pronouns in Table 5.14 will be discussed later.

	+C, +L	+C, -L	-C, +L	-C, -L	intensifier	total
reflexives	71 (91.03%)	0	0	0	7 (8.97%)	78 (100%)
non-reflexives	0	11 (73.33%)	1 (6.67%)	3 (20.00%)	0	15 (100%)
total	71 (76.34%)	11 (11.83%)	1 (1.08%)	3 (3.23%)	7 (7.53%)	93 (100%)

Table 5.14: Intra-sentential antecedents broken up by syntactic positions

(25) a. +C, -L

MOTHER: **you** ask Urs(u)la to tell **you** about this. (adam26.cha)

b. -C, +L

MOTHER: why don't you take **Ursula**'s briefcase over to **her**? (adam02.cha)

c. -C, -L

MOTHER: I guess I put all the air in **Bobo** when I blew **him** up yesterday. (adam36.cha)

Table 5.15 breaks up antecedents by their categories. Only intra-sentential antecedents were counted. Category *unpronounced* includes PRO (5 items), trace (6 items) and imperative subjects (4 items). The majority of antecedents are pronouns. A similar distribution is observed in the model data: Of 3 non-reflexive pronouns that have intra-sentential antecedents, all of them have pronoun antecedents. Similarly, 72% of reflexive pronouns have pronoun antecedents in the model input data. This pattern is reasonable because the most frequent pronoun is a second person pronoun in both reflexives and non-reflexives (i.e., the antecedents of *you* and *yourself* are *you*).

	pronoun	lexical noun	unpronounced	total
reflexives	61 (78.20%)	1 (1.28%)	15 (19.23%)	78 (100%)
non-reflexives	11 (73.33%)	4 (26.67%)	0	15 (100%)
total	72 (77.42%)	4 (4.30%)	15 (16.13%)	93 (100%)

Table 5.15: Categories of antecedents

Table 5.16 breaks up clauses that contain pronouns by tense types. Category *non-finite* includes *to*-infinitives, bare-infinitives, and gerund. This distribution parallels to the model training input in that the most frequent tense is finite and non-finite next.

	finite	non-finite	imperative	total
reflexives	62 (79.49%)	14 (17.95%)	2 (2.56%)	78 (100%)
non-reflexives	829 (70.25%)	161 (13.64%)	190 (16.10%)	1180 (100%)
total	891 (70.83%)	175 (13.91%)	192 (15.26%)	1258 (100%)

Table 5.16: Tense of clauses that contain pronouns

Table 5.17 breaks up lexical heads that take pronouns by their grammatical categories. Category *complex* in Table 5.17 means a complex transitive verb such as *made* in *Alice made her happy*. Category *none* is for intensifiers which do not have lexical heads. Ratios between verb heads and preposition heads in Adam corpus are: 68 : 32 in reflexives and 77 : 23 in non-reflexives. The model input data does not reflect this aspect because there is no preposition heads.

	transitive	ditransitive	complex	intransitive	copula	preposition	none	total
reflexives	46	2	0	0	0	23	7	78
non-reflexives	718	151	33	1	2	275	0	1180
total	764	153	33	1	2	298	7	1258

Table 5.17: Categories of lexical heads

Table 5.18 shows top 20 head words that take reflexive pronouns and non-reflexive pronouns respectively. There are a few overlapping words between the model training data and the Adam corpus. In the model training data, verbs *cover*, *dry*, *help*, *hurt*, and *see* were used. In the Adam corpus, there are the verb *hurt* in both reflexives (25 tokens) and non-reflexives (15 tokens) and the verb *help* (14 tokens) in non-reflexives. It is hard to conclude anything from this small amount of data, but we speculate that certain words occur with reflexive pronouns more frequently than non-reflexive pronouns. This may suggest that the distribution of adjacent words could help categorization. Gulzow (2006) reported that children around age 3 frequently produce *by X-self* compared to reflexive *X-self*. This seems parallel with the distribution of head words in the Adam corpus in that the second most frequent head word of reflexive pronouns is a preposition *by*. However, the categories based on only this distribution could only help learners to decide which pronouns go together, but not to acquire knowledge of what the distributional profile of each category is.

5.7.4 Summary

There are similarities and differences between Adam corpus and the model training data. These data sets are similar in that (i) the most frequent pronoun is a second person, and the least frequent pronoun is a first person plural, (ii) the distribution

Reflexives		Non-reflexives	
Word	count	Word	count
hurt	25	tell	112
by	13	give	83
on	5	let	77
like	4	for	73
push	4	to	67
to	3	put	50
cut	2	show	45
give	2	want	39
at	1	ask	37
check	1	with	33
eat-up	1	of	27
find	1	see	25
for	1	on	23
hear	1	make	19
injure	1	hurt	15
lose	1	call	14
of	1	get	14
scratch	1	help	14
take	1	behind	12
trick	1	excuse	12

Table 5.18: Top 20 heads

of tense type of the clauses that contain pronouns parallels, and (iii) the majority of the antecedents of non-reflexives are extra-sentential. These data sets differ in that (i) the ratio between reflexives and non-reflexives in Adam corpus is 6 : 94 (original ratio 78 : 1180), while 50 : 50 (original ratio 25 : 25) in the model input, (ii) there are some intensifiers (9%) in Adam corpus, while there are no intensifiers in the model input, (iii) head word vocabulary is rich in Adam corpus, while there are only 5 types of head verbs in the model input (verbs *hurt* and *help* were found in both data sets), and (iv) there is a considerable number of pronouns located in the oblique object position (23%) in Adam corpus, while no oblique object pronouns in the model input.

The relative frequency of reflexive pronouns could be problematic. The difference between the frequency of reflexives and non-reflexives in Adam corpus is huge. For the purpose of comparison between reflexives and non-reflexives, We only looked at object forms, but there are many more non-reflexive pronouns in other forms, i.e., subject forms (e.g., *I*) and possessive forms (e.g., *my* and *mine*). The proportion of reflexive pronouns must be very small in the actual input. It is an open question if and how this frequency difference in the input matters for the learning of a pronoun system. The current learning model might treat this small proportion of reflexive pronouns as noise in the input and thus not categorize them into one group.

There are very few non-reflexives that have intra-sentential antecedents (15 instances in Adam corpus, 1.27%). It is striking that there was just 1 non-reflexive pronoun that has a non-c-commanding and local antecedent as in (25b). There is no such instance in the model input data. This implies that there is very little evidence that shows c-command is relevant to define the grammatical distribution of pronouns. With this kind of input, only knowing locality is sufficient to distinguish the distribution of reflexive pronouns and non-reflexive pronouns.

5.8 Discussion

We showed a Bayesian model with prior discourse knowledge estimated in the experiment along with relevant prior syntactic knowledge can accurately recover grammatical categories of pronouns. This suggests the possibility that discourse information can help learners acquire grammatical categories of pronouns. We have posed a circular problem in that identifying the referent of the pronoun and identifying the grammatical categories of the pronoun depend on each other. To learn that the pronoun is reflexive (or non-reflexive), learners need to identify the referent of the pronoun. However, to identify the referent of the pronoun, they might need to use the grammatical knowledge that the pronoun is reflexive that takes the local and c-commanding antecedent. This study suggests one way in which learners can overcome this circularity problem by showing that the discourse information could be sufficient to infer the potential referents of the pronouns.

Discourse knowledge in the current model is directly estimated in the experiment. This is the maximum knowledge in a sense that it is estimated from adult English speakers. To make simulations more explicit, we hope to build a discourse model that approximates children's discourse knowledge.

The model is assumed to have prior knowledge of c-command and locality and its relevance to the pronoun acquisition. Is this a reasonable assumption? Knowledge of c-command appears to be available to children at age four or even earlier (Lidz & Musolino, 2002; Sutton, Fetters, & Lidz, 2012). We also have firm evidence that English children around five years old consistently demonstrate their knowledge of locality (Zukowski, McKeown, & Larsen, 2008). However, these findings are not sufficient to justify the assumption for two reasons. First, even if children have knowledge of c-command, it is not yet known whether they know its relevance to pronouns. Second, it is not yet known whether children have already acquired locality before learning grammatical categories. It would be valuable to investigate (i) whether we can further weaken the prior syntactic knowledge and still acquire the correct categories and (ii) whether the model is flexible enough to acquire pronoun categories in other languages that have different syntactic distributions.

The corpus study shows that there is very little evidence showing the relevance of c-command in defining the grammatical distribution of pronouns. This may suggest two learning hypotheses: (i) Learners antecedently know that c-command is relevant. If they had to decide whether c-command was relevant, they might decide it was not based on the input data (suppose the distribution in the Adam corpus approximates the input a child generally receive) and they would have learned the wrong grammar. Nonetheless English speakers acquire this knowledge at some point. This may suggest that learners must not have to decide whether c-command is relevant. (ii) Learners do not know whether c-command is relevant. Their initial hypothesis about the grammatical distribution of pronouns would be wrong if the input parallels to Adam corpus. As getting relevant input, they learn that c-command is relevant. If there is little relevant evidence available in child-directed speech, then they must get evidence from other sources such as written language. This suggests that the timing of the acquisition would be late.

To test these hypotheses, we need to investigate following two questions. First, is the input data really insufficient? Since Adam corpus is a tiny part of speech heard by Adam, there might have been a sufficient number of such non-reflexive pronouns that were not recorded. It is also possible that this distribution is accidental and there is sufficient evidence in other child-directed speech corpora. Second, do children know the relevance of c-command? It is not yet clear whether children have the knowledge that c-command is relevant to pronouns because previous developmental studies on c-command (e.g., Wexler & Chien, 1985; McKee, 1992) had methodological problems (c.f. Grimshaw & Rosen, 1990).

This study examined the potential utility of discourse information as a cue to the acquisition of pronoun categories. We showed that a Bayesian model with prior knowledge of discourse information and relevant syntax can accurately recover grammatical categories of pronouns without knowing in advance how many categories are present in a language. This supports a role for discourse information in helping learners acquire grammatical categories of pronouns and shows one way in which they can overcome the circularity problem inherent to language acquisition at the syntax-semantics interface.

Chapter 6 Conclusion

6.1 Overview

This dissertation explores the role of discourse information in language acquisition and language production. In Chapter 3, I formalized the latent semantic information in humans' discourse representations by examining speakers' choices of referring expressions. Simulation results suggest that topic models can capture aspects of discourse representations that are relevant to the choices of referring expressions. I also showed that this latent topic representation has an independent contribution beyond simple referent frequency. In Chapter 4, I proposed a language production model that extends the rational speech act model from M. Frank and Goodman (2012) to incorporate updates to listeners' beliefs as discourse proceeds. Simulations suggest 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. Chapter 5 examined the role of discourse information in language acquisition, focusing on the learning of grammatical categories of pronouns. I showed that the Bayesian model with prior discourse knowledge can accurately recover grammatical categories of pronouns, but simply having the strong syntactic prior knowledge is not sufficient. This suggests that the discourse information can help learners acquire grammatical categories of pronouns. Overall, the Bayesian models used in this dissertation allowed me to flexibly investigate the effects of various sources of information, including discourse salience, expectations about the referents and grammatical knowledge.

6.2 Implications and future directions

All models in this dissertation are about how discourse information affects our language use and learning. I tested different types of discourse information in each study: topicality in Chapter 3, recency and frequency in Chapter 4, and direct estimates of which entity is likely to be referred to from the human experiment in Chapter 5. These information sources provide probabilistic representations of discourse entities given the preceding discourse, and models in this dissertation suggest frameworks to use these representations to examine various questions. However, this information would not have an independent influence but rather interacts with each other. For example, discourse topicality estimated from the topic modeling (Chapter 3) can be incorporated into the discourse salience prior in the speaker model (Chapter 4), and this discourse salience prior could also be used as a discourse model for the word learning (Chapter 5). The current models set us up to ask various questions and further investigate the nature of our discourse knowledge. In the remainder of this chapter, I discuss implications, challenges and some directions for future work.

6.2.1 Topicality

The work in Chapter 3 suggests that speakers might use latent topic information that is recovered by the topic modeling when they choose the referring expressions. The notion of discourse topicality has been argued to be one of discourse salience factors that influences speakers' choices of referring expressions, but it has been difficult to objectively measure topicality because it is about the degree/strength of latent meanings (see Chapter 2 for a summary of this problem). I argue that the topic information recovered by topic modeling could objectively approximate topic information in speakers' discourse representations (hence an aspect of discourse salience information). An important question is what the topic information derived from the topic modeling actually represents. The latent topic information derived from the topic modeling cannot be separated out from the effect of frequency because the model recovers the topic distributions based on the frequencies of words within and across documents, along with the probabilistic structure of the model. Thus, the topic representations derived from the topic modeling do not provide *pure* topicality that does not depend on frequency. Nevertheless, I have quantitatively shown that the topic information recovered by topic modeling captures something beyond the information recovered from the simple referent frequencies. The question is what there is between these two measures. I speculate that this gap might be something derived from the structure of the topic modeling, presumably approximating semantic associations of words in the discourse representations.

Though Chapter 3 has shown that the topic model captures something beyond the simple referent frequencies, there are some aspects of discourse topicality that this work might have missed. First, the topic model that I used in this work assumes a bag-of-words representation, so it does not consider any structural information such as a sequence/flow of topics in a document. A more sophisticated extention of the standard topic modeling that can recover the sequence of topics might be able to capture topicality better than this work (for preliminary results in this direction, see Vornov (2015)). Second, topicality would not depend on only linguistic information, but it would be influenced by non-linguistic information such as visual information. The data that I used to measure topicality of the referents is from broadcast news. This kind of data would have generally been presented with visual information such as pictures, movies, captions and so on, so the topic information estimated only from linguistic information would have missed the effects of these kinds of information. A model that can incorporate visual information might be able to capture discourse topicality better than the current work.

Despite these limitations, the work in Chapter 3 suggests a formal account of topicality. This could be applied to future work on the discourse representations. It is not yet known to what extent this measure of topicality could capture other phenomena that are supposed to be influenced by discourse topicality, such as passivization (Christianson & Ferreira, 2005).

6.2.2 Discourse salience

Chapter 2 showed a variety of factors that might influence on discourse salience. An important question for future research is what the nature of discourse salience is. Is discourse salience just a reflection of a set of various factors? Is there any generic mechanism that is responsible for the observed salience patterns? One of the potential directions that worth pursuing is to explore memory representations of discourse entities (e.g., Bock & Warren, 1985; Foraker & McElree, 2007; Rij et al., 2013).

The discourse salience prior in the speaker model (Chapter 4) uses either referent frequency or recency, using the idea of Bayesian non-parametric models (Teh et al., 2006; Blei & Frazier, 2011). The denominator in the speaker's listener model represents a sum of potential referents that could be referred to by the word speaker is considering. It is possible that these terms could be more reasonably represented using generic memory representations such as ACT-R (Anderson & Milson, 1989).

ACT-R has a declarative memory module where entities are represented as chunks. These chunks can be connected with other chunks. Each chunk has an activation value that is computed based on the occurrences of that chunk in the preceding discourse and this activation value decays as time passes. If a certain chunk is retrieved, then that chunk spreads activation to other chunks associated with that chunk. This representation of an entity in memory seems to be a better alternative for the discourse salience prior and the denominator representing potential referents. It seems to be more reasonable to assume that discourse salience is a frequency of the referent in the preceding context that decays with time, rather than treating frequency and recency as independent factors. It also makes sense to assume that speakers would not memorize all entities in the current discourse. The entities in speakers' memory representations would have different activation values and these should be reflected in the denominator computation. It would be interesting to experiment with ACT-R to see to what extent it approaches actual speakers' behavior.

6.2.3 Listeners' discourse model

Studies in Chapter 3 and Chapter 4 can be considered as providing distant pictures of how discourse salience affects speakers' behavior because these models do not use any sentence-level information that is known to affect speakers' choices of referring expressions, such as grammatical roles. This could be problematic particularly when we want to use the current discourse model to examine listeners' behavior (e.g., reference resolution).

Many researchers have observed that implicit causality, a semantic bias derived from verb semantics and the connectives such as *because* and *so*, affects the choice of *referents* (Garvey & Caramazza, 1974; Brown & Fish, 1983; Au, 1986; Stevenson et al., 1994, among many). For example, from Fukumura and Van Gompel (2010), people are more likely to complete the clause in (26a) by referring to *John* such that *because John/he was very clever*. On the other hand, they tend to complete the clause in (26b) by referring to *Mary* such that *because she/Mary was very clever*.

(26) a. John impressed Mary because ...

b. John admired Mary because ...

Recent studies have observed that a semantic bias derived from verb semantics and the connectives such as *because* and *so* affects the choice of *referents* but not the choice of *referring expressions* (Stevenson et al., 1994; Kehler et al., 2008; Fukumura & Van Gompel, 2010; Rohde & Kehler, 2014). This suggests that reference production and comprehension sometimes use different information and goes against the speaker model in Chapter 4. The speaker model in Chapter 4 is a recursive model as it is based on the rational speech act model (M. Frank & Goodman, 2012), so we can use the embedded listener model to examine listeners' behavior. However, this means that both the speaker model and listener model use the same kind of discourse prior. Whether this is a reasonable assumption is exactly the question under debate in psycholinguistics, and this question would also challenge the plausibility of recent recursive pragmatic reasoning models (e.g., M. Frank & Goodman, 2012; Smith et al., 2013). As an initial step, we could compare a listener model using the same discourse prior as the speaker model with a listener model using a different discourse prior from the speaker model.

6.2.4 Information quality in word learning

I showed that the speaker's listener model in Chapter 4 corresponds to the information content conveyed by a word given preceding discourse information. I suggest that this information content of a word could be used to measure information quality of a word in word learning situations.

Developmental studies have proposed the "fast-mapping" hypothesis: learners acquire a new word heard only once or a very few times, that is, a large number of instances is not necessary (Nappa, Wessel, McEldoon, Gleitman, & Trueswell, 2009; Spiegel & Halberda, 2011; Medina, Snedeker, Trueswell, & Gleitman, 2011). This hypothesis predicts that information of a new word should be highly informative. A recent study (Cartmill et al., 2013) tested this prediction and showed that quality of input matters for word learning, where this aspect of input quality is measured as referential transparency: that is, the quality of nonverbal cues available in the immediate extralinguistic context. On the other hand, some studies show that nonverbal cues such as eye-gaze and pointing would not be strong cues for identifying the speaker's intended referent in that these cues tend to be noisy and irregularly used but the combination of these social cues along with discourse continuity information provide better information about the referent (M. Frank et al., 2013; Rohde & Frank, 2014; Horowitz & Frank, 2015; Sullivan & Barner, 2015). This finding suggests that highly informative input does not seem to solely depend on nonverbal social cues. Discourse information seem to smooth noise in these nonverbal cues.

These findings raise a question of what information exactly makes learning instances transparent. The amount and content of information of a new word may vary by the type of that word, a learning environment (e.g., visual and social cues), and linguistic context surrounding that word. To investigate this question, it would be interesting to use the listener model in Chapter 4 to explore whether and how information that is accumulated as discourse proceeds changes the quality of information conveyed by a new word. This would provide a way to characterize what makes a new word more informative, thus leading a successful word learning.

6.2.5 Discourse representations

All models in this dissertation use probability to represent discourse information. These are in line with a body of research that support the probabilistic models of language processing and learning (Chater & Manning, 2006, for overview). Understanding and using discourse representations would involve uncertain inference on noisy and complex information, such as identifying topics and reading between the lines from partial information, so it seems to be reasonable to think that these kinds of processes could be represented using probability.

The work in Chapter 3 has shown that the probabilistic model can capture information that the simple measure can not. This suggests that probability is not just another way to represent information people might use, but it does capture different information that requires inference about something latent. The speaker model in Chapter 4 uses probability to formalize speakers' choices of referring expressions. The simulation results seem to be indicative in that the model that has subtle interactions of different sources of information captured speakers' behavior the best. However, I have not compared with any other non-probabilistic models that are derived from different theories. Future work will compare this model with different models that do not use probability representations (e.g., a model with a simple salience threshold) to examine the nature of discourse representations. The learning model in Chapter 5 represents an ideal learner's estimates of who the referent is, using probability, and simulation results show that this information (along with some prior syntactic knowledge) is sufficient to learn the grammatical categories of pronouns. This suggests that it seems to be a promising direction to explore how the referents (entities) are represented in the discourse using probability. Overall, I hope that formal frameworks in this dissertation provide some ways to further explore the more general question of how humans' linguistic representations would look like.

The studies in this dissertation explore limited aspects of discourse information, namely salience represented by the notion of topicality, recency, and frequency. It is not clear how these relate to other linguistic knowledge, such as pragmatics. In Chapter 4, I have formalized the interaction between informativity of word and production cost in the choices of referring expressions. Speakers' choices of referring expressions seem to be at least relevant to some maxims of Gricean cooperative principle (Grice, 1975) such as the maxim of quantity (provide information as informative as possible, but not more informative than is required) and the maxim of manner (be perspicuous: avoid ambiguity, be brief, etc.). I argue that the speaker model in this dissertation could potentially be one way of formalizing some aspects of Gricean pragmatics in that it seems to explicitly formalize some components of the cooperative principle: informativity (discourse salience), ambiguity (the denominator in the speakers' listener model) and brevity (production cost). Similarly, researchers recently started looking at the relation between conversational implicature and inference of discourse relations (e.g., Asher, 2013; Irmer, 2013). It is an open question whether the inference of discourse relations and the conversational implicature share the same underlying mechanism. It would be interesting to use existing computational frameworks such as abductive reasoning (Hobbs, 2004) to formalize Gricean pragmatics of inference. This would allow us to integrate different theories and provide deeper insights into the underlying mechanism of discourse.

Another question that this dissertation did not address is how different types of discourse information could be represented at the same time and be used in real-time processing. Studies here have particularly focused on topicality and salience. However, this information should not be independently represented nor used. In other words, it is an open question whether our discourse knowledge is a list of different information sources/factors (such as topic, salience, etc.) or what we observe is a product of the underlying mechanism. The models in this dissertation provided objective measures to test what kind of discourse knowledge best approximate the observed speakers' and learners behavior. I hope that these models help contributing toward exploring what our discourse knowledge would look like.

6.3 Conclusion

This dissertation represents an contribution to the computational study of psycholinguistics and language acquisition. I proposed flexible frameworks for modeling speakers and learners using techniques from Bayesian models. I showed how these models formalize problems that have been unclear in the literature of discourse and language acquisition: (i) I suggested a way to represent semantic information in humans' discourse representations, (ii) I provided a principled account for how discourse salience affects speakers' choices of referring expressions, and (iii) I proposed a learning model that flexibly integrates different sources of information and allows us to test various learning hypotheses. Simulation results confirmed the significance of discourse information in language production and language acquisition and suggested various implications to other domains of research and future directions as discussed above.

Appendix A: Bayesian modeling

This appendix briefly illustrates stochastic frameworks I use in this dissertation. Each chapter also has an intuitive explanation for each model. A Reader who is satisfied with it can safely skip the technical details presented here.

A.1 Bayesian Networks

Probabilistic graphical models provide efficient framework to illustrate probability distributions. I use Bayesian Networks, which are a special case of graphical models, to express the joint distribution with random variables and their dependencies in a directed graph. In a graphical model, nodes are random variables and edges indicate dependence. For example in figure A.1, shaded w is an evidence node that is observed. Here let w be a word we observe. A variable z is a hidden node that is latent, such as unknown cause, category, hypothesis, and so on. Here let z be an unknown category. The direction of the arrow indicates that w depends on z. We use plates when a model have repetitions of nodes. The number of repetitions is given by the index variable at the lower of the right corner.

We can infer the probability of the category z given the observed word w, $P(z|w)$ (the conditional probability or the *posterior* probability), by applying Bayes'

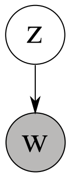


Figure A.1: Example of a graphical model

rule as in (A.1).

$$P(z|w) = \frac{P(w|z)P(z)}{P(w)} \tag{A.1}$$

$P(z)$ denotes the probability that this model believes z is true before seeing any data. This is called the *prior* probability. The term $P(w|z)$ the probability of the word given the category, called the *likelihood*. The denominator $P(w)$ is called *evidence*. We can compute this probability by summing $P(w|z)P(z)$ over all possible category Z: $\Sigma_{z' \in Z} P(w|z')P(z')$.

A.2 Hierarchical Dirichlet Process

The hierarchical Dirichlet process (Teh et al., 2006) is one of Bayesian nonparametric models that can learn its complexity according to the data observed. This framework is used in Chapter 3 to recover topic distributions in a corpus. It also allows the pronoun category learning model in Chapter 5 to learn the number of pronoun categories and which syntactic position is associated with which pronoun category according to the data observed.

The Dirichlet process is a stochastic process that generates the distribution G

from the base distribution G_0 (base measure) as follows.

$$G \sim DP(\alpha_0, G_0) \tag{A.2}$$

The concentration parameter $\alpha > 0$ is a learnable parameter that controls how similar the distribution G is to the base distribution G_0.

The hierarchical Dirichlet process has a set of random probability measure G_j and the global measure G_0. G_0 itself is a draw from a Dirichlet process with parameter γ and base measure H.

$$\begin{aligned} G_0 &\sim DP(\gamma, H) \\ G_j &\sim DP(\alpha_0, G_0) \quad \text{for each } j \end{aligned} \tag{A.3}$$

The random probability measures G_j are conditionally independent given G_0. The Dirichlet process generates the distributions G_j with base measure G_0 and concentration parameter α_0. Each G_j is infinite-dimensional. The atoms of G_0 are shared among the distributions G_j. This structure ensures that the model can share atoms across and within the different groups.

A.3 Chinese Restaurant Franchise

A stochastic process called Chinese Restaurant Franchise (Teh et al., 2006) is used to cluster the data in the hierarchical Dirichlet process setup. As for the pronoun category learning model in Chapter 5, this framework allows the model to learn the number of pronoun categories and syntactic environment associated with each category, instead of specifying them.

The following illustration uses Chinese restaurant metaphor as in the original study (Teh et al., 2006). There are Chinese restaurants $G_{j_{1:J}}$ and a restaurant franchise G_0 that serves a global menu across J restaurants. Each restaurant G_j has infinite number of tables. The franchise restaurant G_0 has infinite number of dishes. At each restaurant G_j, a new customer chooses a table at that restaurant according to the following probabilities:

$$\begin{aligned} \text{existing table with probability } &\propto \tfrac{N_{j,k}}{\alpha+N_{j,\cdot}-1} \\ \text{new table with probability } &\propto \tfrac{\alpha}{\alpha+N_{j,\cdot}-1} \end{aligned} \tag{A.4}$$

Term $N_{j,k}$ denotes the number of times dish k is used in the lower restaurant j, not including the current instance. Concentration parameter α controls how likely customers are to sit down at a new table. If the value of α is high, more customers would sit down at a new table.

If a customer chooses the new table, he needs to order a new dish from the global menu at the franchise restaurant (only one dish). This dish is shared among all other customers who sit at that table. The new dish can be either an existing dish or a brand-new dish in the global menu. The customer at the new table chooses the dish according to the following probabilities:

$$
\begin{aligned}
\text{existing dish with probability } & \propto \tfrac{M_k}{\gamma + M.-1} \\
\text{new dish with probability } & \propto \tfrac{\gamma}{\gamma + M.-1}
\end{aligned}
\tag{A.5}
$$

where M_k denotes the number of times dish k is used across all lower restaurants. The franchise restaurant keeps track of the number of tables across lower-level restaurants that serve dish k. Parameter γ controls how likely a new dish is to be created. If the value of γ is high, more dishes are created.

In the pronoun category learning model in Chapter 5, a customer corresponds to a pronoun token, a table corresponds to the index that associates pronoun tokens with a pronoun category, and a dish corresponds to the index of a pronoun category.

The same dish can be served across tables and restaurants, but for efficiency, I use the minimal path assumption (Wallach, 2008, p. 56,57) in the pronoun category learning model. The minimal path assumption assumes that a dish for a new table comes from an existing dish in the franchise restaurant and it comes from a new dish if and only if there is no table with the appropriate dish. No two internal tables will have the same dish. Figure A.2 illustrates this sampling process. It has been shown that the minimal path assumption works as well as explicitly sampling seating assignments (Nguyen et al., 2014).

The Chinese Restaurant Franchise has two important properties: *exchangeability* and *rich get richer* property. *Exchangeability* means that the probability of

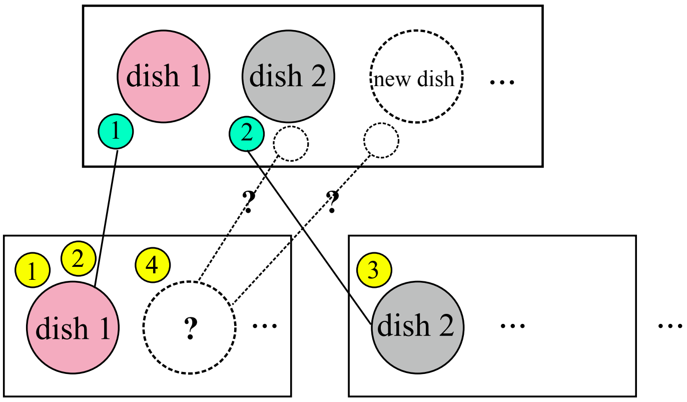


Figure A.2: Seating assignments in Chinese Restaurant Franchise under the minimal path assumption: Under the minimal path assumption, no two internal draws will have the same value. When a new table is created in a left-side lower restaurant, a dish for this table could be either the existing dish 2 or a new dish in the upper level restaurant, but not the existing dish 1.

the dish assignment $p(z)$ only depends on the number of dishes (type counts) and the size of each dish (the number of customers eating dish k). It does not depend on the seating order. This property makes inference easier. The *Rich get richer* property means that tables with many customers will get more customers because a new customer will sit at an existing table with probability proportional to the number of customers at the existing table. This property allows a model to learn a smaller number of categories but still be flexible enough to learn categories according to the data observed.

Appendix B: Pairwise F-score

The pairwise F-score is computed by counting whether two token are assigned to the same category by the model or not. For example, we count:

- *me* in same category as *him* as **hit**
- *me* in different category than *him* as **miss**
- *me* in same category as *myself* as **false alarm**
- *me* in different category than *myself* as **correct rejection**

F-score was computed by taking the harmonic mean of accuracy (a) and completeness (c) defined as follows.

$$F = \frac{2 * a * c}{a + c} \tag{B.1}$$

where accuracy (a) and completeness (c) are defined as

$$\text{accuracy} = \frac{\text{hits}}{\text{hits} + \text{false alarms}} \tag{B.2}$$

$$\text{completeness} = \frac{\text{hits}}{\text{hits} + \text{misses}} \tag{B.3}$$

Appendix C: Corpus study: coding examples

- Mother: I don't hurt myself. (adam03.cha)

    1. not a fragment
    2. myself
    3. intra-sentential antecedent
    4. pronoun
    5. +C, +L
    6. object
    7. subject
    8. hurt
    9. transitive verb
    10. finite
    11. main clause

- Mother: He'd like to do it himself. (adam45.cha)

    1. not a fragment
    2. himself
    3. intra-sentential antecedent
    4. PRO
    5. none (intensifier)
    6. adverbial intensifier
    7. subject is focused (intensifier)
    8. none (intensifier)
    9. none (intensifier)
    10. non-finte
    11. complement

- Mother: why don't you take Ursula's briefcase over to her? (adam02.cha)

1. not a fragment
2. her
3. intra-sentential antecedent
4. lexical NP
5. -C, +L
6. oblique object
7. possessive
8. to
9. preposition
10. finite
11. main clause

- Mother: remember you had a froggie who had beans in him? (adam19.cha)

1. not a fragment
2. him
3. intra-sentential
4. lexical NP
5. -L, +C
6. oblique object
7. object
8. in
9. preposition
10. finite
11. relative clause

- Mother: don't burn you! (adam16.cha)

1. not a fragment
2. you
3. extra-sentential antecedent
4. none
5. none (-C, -L)
6. object
7. none
8. burn
9. transitive verb
10. imperative
11. imperative

References

Allen, S. E., & Schröder, H. (2003). Preferred argument structure in early inuktitut spontaneous speech data. In J. W. Du Bois, L. Kumpf, & W. Ashby (Eds.), *Preferred argument structure: Grammar as architecture for function* (pp. 301–338). John Benjamins, Amsterdam.

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological review*, *106*(4), 748.

Amritavalli, R. (2000). Lexical anaphors and pronouns in kannada. In B. Lus, K. Wali, J. Gair, & K. V. Subbarao (Eds.), *Lexical anaphors and pronouns in selected south asian languages: A principled typology.* Berlin: Mouton de Gruyter.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703.

Ariel, M. (1990). *Accessing noun-phrase antecedents.* Routledge.

Arnold, J. (1998). *Reference form and discourse patterns.* Unpublished doctoral dissertation, Stanford University Stanford, CA.

Arnold, J. (1999). Marking salience: The similarity of topic and focus. *Unpublished manuscript, University of Pennsylvania.*

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

Arnold, J. (2010). How speakers refer: the role of accessibility. *Language and Linguistics Compass*, *4*(4), 187–203.

Arnold, J., Brown-Schmidt, S., & Trueswell, J. (2007). Children's use of gender and order of mention in pronoun processing. *Language and Cognitive Processes*, *22*, 527–565.

Arnold, J., & Griffin, Z. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, *56*(4), 521–536.

Arnold, J., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 28–55.

Asher, N. (2004). Discourse topic. *Theoretical Linguistics*, *30*(2-3), 163–201.

Asher, N. (2013). Implicatures and discourse structure. *Lingua*, *132*, 13–28.

Asher, N., & Lascarides, A. (2003). *Logics of conversation.* Cambridge University Press.

Au, T. K.-F. (1986). A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory and Language*, *25*(1), 104–122.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5), 3048–3058.

Bagga, A., & Baldwin, B. (1998). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference* (Vol. 1, pp. 563–566).

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of child language*, *20*(02), 395–418.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, *42*(1), 1–22.

Bard, E. G., Aylett, M. P., Trueswell, J., & Tanenhaus, M. (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*, 173–191.

Bard, E. G., Hill, R. L., Foster, M. E., & Arai, M. (2014). Tuning accessibility of referring expressions in situated dialogue. *Language, Cognition and Neuroscience*, *29*(8), 928–949.

Barr, D. J., & Keysar, B. (2006). Perspective taking and the coordination of meaning in language use.

Baumann, P., Clark, B., & Kaufmann, S. (2014). Overspecification and the cost of pragmatic reasoning about referring expressions. In *Proceedings of the 36th annual conference of the cognitive science society.*

BBN Technologies. (2007). *OntoNotes English co-reference guidelines version 7.0.*

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001–1024.

Bergen, L., Goodman, N., & Levy, R. (2012a). 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.*

Bergen, L., Goodman, N. D., & Levy, R. (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.*

Bergsma, S., & Lin, D. (2006, July). 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* (pp. 33–40). Sydney, Australia: Association for Computational Linguistics.

Birner, B. J., & Ward, G. (2009). Information structure and syntactic structure. *Language and Linguistics Compass*, *3*(4), 1167–1187.

Blei, D. M., & Frazier, P. I. (2011). Distance dependent chinese restaurant processes. *The Journal of Machine Learning Research*, *12*, 2461–2488.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, *21*(1), 47–67.

Breheny, R., Katsos, N., & Williams, J. (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.

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

Brennan, S. E., Friedman, M. W., & Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25th annual meeting on association for computational linguistics* (pp. 155–162).

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2), 274–291.

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press.

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition*, *14*(3), 237–273.

Callaway, C. B., & Lester, J. C. (2002). Pronominalization in generated discourse and dialogue. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 88–95).

Campbell, A. L., Brooks, P., & Tomasello, M. (2000). Factors affecting young children's use of pronouns as referring expressions. *Journal of Speech, Language, and Hearing Research*, *43*(6), 1337–1349.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, i–174.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283.

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

Chafe, W. L. (1974). Language and consciousness. *Language*, 111–133.

Chafe, W. L., & Li, C. N. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view in subject and topic.

Chambers, C. G., & Smyth, R. (1998). Structural parallelism and discourse coherence: A test of centering theory. *Journal of Memory and Language*, *39*(4), 593–608.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, *10*(7), 335–344.

Chomsky, N. (1973). Conditions on Transformations. In S. R. Anderson & P. Kiparsky (Eds.), *A festschrift for Morris Halle.* New York: Holt, Rinehart & Winston.

Chomsky, N. (1981). *Lectures on Government and Binding.* Dordrecht.

Christianson, K., & Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (odawa). *Cognition*, *98*(2), 105–135.

Clancy, P. (1993). Preferred argument structure in korean acquisition. In *Proceedings of the 25th annual child language research forum* (pp. 307–314).

Clancy, P. M. (1980). Referential choice in english and japanese narrative discourse. In *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production* (Vol. 3, pp. 127–201). Norwood: Ablex.

Clark, H. H. (1996). *Using language* (Vol. 1996). Cambridge university press Cambridge.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge University Press.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J. F. L. Ny & W. Kintsch (Eds.), *Language and comprehension* (pp. 287–299). Amsterdam.

Conroy, A., Takahashi, E., Lidz, J., & Phillips, C. (2009). Equal treatment for all antecedents: How children succeed with Principle B. *Linguistic Inquiry*, *40*, 446–486.

Cowles, H. W. (2003). *Processing information structure: Evidence from comprehension and production.* Unpublished doctoral dissertation, University of California, San Diego.

Cowles, H. W., Walenski, M., & Kluender, R. (2007). Linguistic and cognitive prominence in anaphor resolution: topic, contrastive focus and pronouns. *Topoi*, *26*(1), 3–18.

Degen, J., Franke, M., & Jäger, G. (2013). Cost-based pragmatic inference about referential expressions. In *Proceedings of the 35th annual conference of the cognitive science society* (pp. 376–381).

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of memory and language*, *27*(4), 429–446.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology*, *40*(4), 296–340.

Fletcher, C. R. (1984). Markedness and topic continuity in discourse processing. *Journal of Verbal Learning and Verbal Behavior*, *23*(4), 487–493.

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

Francik, E. P. (1985). *Referential choice and focus of attention in narratives.* Stanford University.

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th annual meeting of the cognitive science society* (pp. 933–938).

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frank, M., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1–24.

Frank, S., Goldwater, S., & Keller, F. (2009). Evaluating Models of Syntactic Category Acquisition without Using a Gold Standard. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Fukumura, K., & Gompel, R. P. van. (2011). The effect of animacy on the choice of referring expression. *Language and Cognitive Processes*, *26*(10), 1472–1504.

Fukumura, K., & Gompel, R. P. van. (2012). Producing pronouns and definite noun phrases: Do speakers use the addressee ! !! s discourse model? *Cognitive science*, *36*(7), 1289–1311.

Fukumura, K., Gompel, R. P. van, & Pickering, M. J. (2010). The use of visual context during the production of referring expressions. *The Quarterly Journal of Experimental Psychology*, *63*(9), 1700–1715.

Fukumura, K., Hyönä, J., & Scholfield, M. (2013). Gender affects semantic competition: The effect of gender in a non-gender-marking language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1012.

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

Fukumura, K., Van Gompel, R. P., Harley, T., & Pickering, M. J. (2011). How does similarity-based interference affect the choice of referring expression? *Journal of Memory and Language*, *65*(3), 331–344.

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

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic inquiry*, 459–464.

Gatt, A., Gompel, R. P. van, Deemter, K. van, & Kramer, E. (2013). Are we bayesian referring expression generators. In *Proceedings of the 35th annual conference of the cognitive science society, berlin, germany* (Vol. 35).

Gatt, A., Krahmer, E., Gompel, R. P. van, & Deemter, K. van. (2013). Production of referring expressions: Preference trumps discrimination. In *Proceedings of the 35th annual conference of the cognitive science society, berlin, germany.*

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6(6)*, 721–741.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–176.

Giuliani, M., Foster, M. E., Isard, A., Matheson, C., Oberlander, J., & Knoll, A. (2010). Situated reference in a hybrid human-robot interaction system. In *Proceedings of the 6th international natural language generation conference* (pp. 67–75).

Givón, T. (1983). *Topic continuity in discourse: A quantitative cross-language study* (Vol. 3). John Benjamins Publishing.

Gleason, J. B., Perlmann, R. Y., & Greif, E. B. (1984). What's the magic word: Learning language through politeness routines ! v. *Discourse Processes*, *7*(4), 493–502.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55.

Gleitman, L. R., Kimberly, C., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1(1)*, 23–64.

Goodman, N., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive science*, *17*(3), 311–347.

Grice, H. (1975). Logic and conversation. *Syntax and semantics*, *3*, 41–58.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

Grimshaw, J., & Rosen, S. T. (1990). Knowledge and obedience: The developmental status of the binding theory. *Linguistic Inquiry*, *21,2*, 187–222.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st annual meeting on association for computational linguistics* (pp. 44–50).

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, *12*(3), 175–204.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*(2), 203–225.

Grüning, A., & Kibrik, A. A. (2005). Modeling referential choice in discourse: A cognitive calculative approach and a neural network approach. In R. Mitkov (Ed.), *Anaphora processing: Linguistic, cognitive and computational modelling* (pp. 163–198). John Benjamins.

Guerriero, S., Cooper, A., Oshima-Takane, Y., & Kuriyama, Y. (2001). A discourse-pragmatic explanation for argument realization and omission in english and japanese children ! !! s speech. In *Proceedings of the 25th annual boston university conference on language development* (Vol. 1, pp. 319–330).

Gulzow, I. (2006). *Intensifiers in language acquisition: A comparative analysis of english and german.* De Gruyter.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.

Haliday, M. A. (1967). Notes on transitivity and theme in english. *Journal of linguistics*, *3*, 37–81.

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection.

Harley, H., & Ritter, E. (2002). Person and number in pronouns: A feature-geometric analysis. *Language*, *78*(3), 482–526.

Hartshorne, J. K., Nappa, R., & Snedeker, J. (2014). Development of the first-mention bias. *Journal of child language*, 1–24.

Haywood, S. L., Pickering, M. J., & Branigan, H. P. (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science*, *16*(5), 362–366.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar* (Vol. 13). Blackwell Oxford.

Hendriks, P., Koster, C., & Hoeks, J. C. (2014). Referential choice across the lifespan: why children and elderly adults produce ambiguous pronouns. *Language, Cognition and Neuroscience*, *29*(4), 391–407.

Hobbs, J. R. (1985). *On the coherence and structure of discourse.* CSLI.

Hobbs, J. R. (2004). Abduction in natural language understanding. In L. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 724–741).

Horowitz, A. C., & Frank, M. (2015). Young children ! !! s developing sensitivity to discourse continuity as a cue for inferring reference. *Journal of experimental child psychology*, *129*, 84–97.

Hughes, M. E., & Allen, S. E. (2013). The effect of individual discourse-pragmatic features on referential choice in child english. *Journal of Pragmatics*, *56*, 15–30.

Hughes, M. E., & Allen, S. E. (2014). The incremental effect of discourse-pragmatic sensitivity on referential choice in the acquisition of a first language. *Lingua.*

Hyams, N., & Sigurjónsdóttir, S. (1990). The Development of "Long-Distance Anaphora": A cross-linguistic comparison with special reference to Icelandic. *Linguistic Acquisition*, *1*, 57–93.

Irmer, M. (2013). Inferring implicatures and discourse relations from frame information. *Lingua*, *132*, 29–50.

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

Jager, G. (2007). Game dynamics connects semantics and pragmatics. In A.-V. Pietarinen (Ed.), *Game theory and linguistic meaning* (pp. 89–102). Elsevier.

Järvikivi, J., Gompel, R. P. van, Hyönä, J., & Bertram, R. (2005). Ambiguous pronoun resolution contrasting the first-mention and subject-preference accounts. *Psychological Science*, *16*(4), 260–264.

Kaiser, E. (2010). Effects of contrast on referential form: Investigating the distinction between strong and weak pronouns. *Discourse Processes*, *47*(6), 480–509.

Kaiser, E., & Trueswell, J. C. (2008). Interpreting pronouns and demonstratives in finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes*, *23*(5), 709–748.

Kako, E. (2005). Information sources for noun learning. *Cognitive Science*, *29*, 223–260.

Kameyama, M. (1994). Indefeasible semantics and defeasible pragmatics. In *Cwi report cs-r9441 and sri technical note 544.*

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory* (No. 42). Springer Science & Business Media.

Kao, J. T., Wu, J., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*.

Karimi, H., Fukumura, K., Ferreira, F., & Pickering, M. J. (2014). The effect of noun phrase length on the form of referring expressions. *Memory & cognition*, 1–17.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar.* CSLI.

Kehler, A. (2004). Discourse topics, sentence topics, and coherence. *Theoretical Linguistics*, *30*(2-3), 227–240.

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

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current directions in psychological science*, 46–50.

Khudyakova, M. V., Dobrov, G. B., Kibrik, A. A., & Loukachevitch, N. V. (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).*

Kibble, R., & Power, R. (2004). Optimizing referential coherence in text generation. *Computational Linguistics*, *30*(4), 401–416.

Kibrik, A. A. (2000). A cognitive calculative approach towards discourse anaphora. In *Proceedings of the discourse anaphora and anaphor resolution conference (daarc2000).* Lancaster: University Centre for Computer Corpus Research on Language.

Koster, J., & Reuland, E. (1991). *Long-distance anaphora.* Cambridge University Press.

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the 20th conference on neural information processing systems (nips).*

Lidz, J., & Musolino, J. (2002). Children's command of quantification. *Cognition*, *84*, 113–154.

MacWhinney, B. (2000a). *The CHILDES project: Tools for analyzing talk* (Third ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2000b). *The CHILDES Project: Tools for analyzing talk. Third Edition.* (Tech. Rep.). Mahwah, NJ: Lawrence Erlbaum Associates.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco, CA: W. H. Freeman.

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2006). The effect of perceptual availability and prior discourse on young children's use of referring expressions. *Applied Psycholinguistics*, *27*(03), 403–422.

McCoy, K., & Strube, M. (1999). Generating anaphoric expressions: Pronoun or definite description. In *Proceedings of the acl workshop on the relation of discourse/dialogue structure and reference* (pp. 63–71).

McKee, C. (1992). A comparison of pronouns and anaphors in English and Italian. *Language Acquisition*, *3*, 21–55.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Nappa, R., Wessel, A., McEldoon, K. L., Gleitman, L. R., & Trueswell, J. C. (2009). Use of speaker's gaze and syntax in verb learning. *Language Learning and Development*, *5*(4), 203–234.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, *31*, 705–767.

Nguyen, V.-A., Boyd-Graber, J., Resnik, P., Cai, D. A., Midberry, J. E., & Wang, Y. (2013). Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 1–41.

Nguyen, V.-A., Boyd-Graber, J., Resnik, P., Cai, D. A., Midberry, J. E., & Wang, Y. (2014). Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine learning*, *95*(3), 381–421.

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of cognitive neuroscience*, *18*(7), 1098–1111.

Nordmeyer, A. E., & Frank, M. (2014). A pragmatic account of the processing of negative sentences. In *Proceedings of the 36th annual conference of the cognitive science society.*

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Orita, N., McKeown, R., Feldman, N. H., Lidz, J., & Boyd-Graber, J. (2013). Discovering pronoun categories using discourse information. In *Proceedings of the 35th annual conference of the cognitive science society.*

Piccin, T. B., & Waxman, S. R. (2007). Why nouns trump verbs in word learning: New evidence from children and adults in the Human Simulation Paradigm. *Language Learning & Development*, *3(4)*, 295–323.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(02), 169–190.

Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*(3), 217–283.

Pitman, J. (2002). Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, *11*, 501–514.

Pollard, C., & Sag, I. (1992). Anaphors in English and the Scope of Binding Theory. *Linguistic Inquiry*, *23*, 261-303.

Prat-Sala, M., & Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in english and spanish.

Journal of Memory and Language, *42*(2), 168–182.

Prince, E. F. (1981). Toward a taxonomy of given-new information. *Radical pragmatics*.

Quine, W. (1960). *Word and Object*. New York: Wiley.

Recasens, M., Marquez, L., Sapena, E., Martí, M. A., & Taulé, M. (2011). *SemEval-2010 task 1 OntoNotes English: Coreference resolution in multiple languages.*

Reinhart, T. (1976). *The Syntactic Domain of Anaphora*. Unpublished doctoral dissertation, MIT.

Reinhart, T. (1982). Pragmatics and linguistics: An analysis of sentence topics. *Philosophica anc Studia Philosophica Gandensia Gent*, *27*(1), 53–94.

Reiter, E., Dale, R., & Feng, Z. (2000). *Building natural language generation systems*. MIT Press.

Rij, J., Rijn, H., & Hendriks, P. (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.

Rohde, H., & Frank, M. (2014). Markers of topical discourse in child-directed speech. *Cognitive science*, *38*(8), 1634–1661.

Rohde, H., & Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, *29*(8), 912–927.

Rohde, H., Kehler, A., & Elman, J. L. (2007). Pronoun interpretation as a side effect of discourse coherence. In *Proceedings of the 29th annual conference of the cognitive science society* (pp. 617–622).

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., & Kaufmann, S. (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.*

Roland, D., Mauner, G., O'Meara, C., & Yun, H. (2012). Discourse expectations and relative clause processing. *Journal of Memory and Language*, *66*(3), 479–508.

Rosa, E. C., & Arnold, J. E. (2011). The role of attention in choice of referring expressions. *Proceedings of PRE-Cogsci: Bridging the gap between computational, empirical and theoretical approaches to reference.*

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. In *20th Conference on Uncertainty in Artificial Intelligence.*

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, *37*(03), 705–729.

Salomo, D., Lieven, E., & Tomasello, M. (2010). Young children's sensitivity to new and given information when answering predicate-focus questions. *Applied Psycholinguistics*, *31*(01), 101–115.

Sanford, A. J., & Garrod, S. C. (1981). *Understanding written language: Explorations of comprehension beyond the sentence*. Wiley New York.

Serratrice, L. (2005). The role of discourse pragmatics in the acquisition of subjects in italian. *Applied Psycholinguistics*, *26*(03), 437–462.

Serratrice, L. (2008). The role of discourse and perceptual cues in the choice of referential expressions in english preschoolers, school-age children, and adults. *Language Learning and Development*, *4*(4), 309–332.

Serratrice, L. (2013). The role of number of referents and animacy in children's use of pronouns. *Journal of Pragmatics*, *56*, 31–42.

Siddharthan, A., Nenkova, A., & McKeown, K. (2011). Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, *37*(4), 811–842.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences.

Skarabela, B. (2007). Signs of early social cognition in children's syntax: The case of joint attention in argument realization in child inuktitut. *Lingua*, *117*(11), 1837–1857.

Skarabela, B., Allen, S. E., & Scott-Phillips, T. C. (2013). Joint attention helps explain why children omit new referents. *Journal of Pragmatics*, *56*, 5–14.

Smith, N. J., Goodman, N., & Frank, M. (2013). Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in neural information processing systems* (pp. 3039–3047).

Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts? In G. Hall & S. R. Waxman (Eds.), *Weaving a Lexicon.* Cambridge, MA: MIT Press.

Song, H.-j., & Fisher, C. (2005). Who's 'she'? discourse prominence influences preschoolers' comprehension of pronouns. *Journal of Memory and Language*(52), 29–57.

Song, H.-j., & Fisher, C. (2007). Discourse prominence effects on 2.5-year-old children's interpretation of pronouns. *Lingua*, *117*(11), 1959–1987.

Spiegel, C., & Halberda, J. (2011). Rapid fast-mapping abilities in 2-year-olds. *Journal of experimental child psychology*, *109*(1), 132–140.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the demonstrations session at eacl 2012.*

Stevenson, R. J., Crawley, R. A., & Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, *9*(4), 519–548.

Sullivan, J., & Barner, D. (2015). Discourse bootstrapping: preschoolers use linguistic discourse to learn new words. *Developmental science.*

Sutton, M., Fetters, M., & Lidz, J. (2012). Parsing for principle c at 30 months. In *Proceedings of the 36th Boston University Conference on Language Development* (pp. 581–593).

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*.

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.*

Turner, R., Sripada, S., Reiter, E., & Davy, I. P. (2008). Using spatial reference frames to generate grounded textual summaries of georeferenced data. In *Proceedings of the fifth international natural language generation conference* (pp. 16–24).

Van Deemter, K., Gatt, A., Gompel, R. P. van, & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in cognitive science*, *4*(2), 166–183.

Van Son, R. J., & Van Santen, J. P. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, *47*(1), 100–123.

Vogels, J., Krahmer, E., & Maes, A. (2013a). When a stone tries to climb up a slope: the interplay between lexical and perceptual animacy in referential choices. *Frontiers in psychology*, *4*.

Vogels, J., Krahmer, E., & Maes, A. (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.

Vogels, J., Krahmer, E., & Maes, A. (2014). How cognitive load influences speakers' choice of referring expressions. *Cognitive science*.

Vornov, E. (2015). *Using sequential lda to further quantify the role of discourse topicality in speakers' choices ofreferring expression.* Undergraduate honors thesis, University of Maryland.

Walker, M., Cote, S., & Iida, M. (1994). Japanese discourse and the process of centering. *Computational linguistics*, *20*(2), 193–232.

Wallach, H. M. (2008). *Structured topic models for language.* Ph.D. dissertation, University of Cambridge.

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

Wexler, K., & Chien, Y. C. (1985). The development of lexical anaphors and pronouns. In *Papers and reports on child language development* (Vol. 24). Stanford University, Stanford, California.

White, M., Clark, R. A., & Moore, J. D. (2010). Generating tailored, comparative descriptions with contextually appropriate intonation. *Computational Linguistics*, *36*(2), 159–201.

Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, *3*(1), 1–191.

Zukowski, A., McKeown, R., & Larsen, J. (2008). A tough test of the locality requirement for reflexives. In *Proceedings of the 32nd Boston University Conference on Language Development* (pp. 586–597).