Is there a predictability hierarchy in reference resolution?

Weijie Xu (weijiexu@uchicago.edu)
Division of the Social Sciences, MACSS, 1155 E. 60th Street
Chicago, IL 60637 USA

Ming Xiang (mxiang@uchicago.edu)
Department of Linguistics, 1115 E. 58th Street
Chicago, IL 60637 USA

Abstract

The concept of *accessibility* has often been evoked to explain reference resolution. According to the Givenness Hierarchy theory, a referent's accessibility in the mental state of a comprehender is encoded in the form of the referent as part of its lexical semantic representation. However, the current literature has not reached a consensus on what accessibility exactly means and how to best quantify it. The factors that modulate accessibility show a great extent of overlap with another independently motivated concept of *predictability*, raising the possibility of a "Predictability Hierarchy" that mirrors Givenness Hierarchy. In a self-paced reading study, the current study examines whether there is such "Predictability Hierarchy" by manipulating the predictability and the form of a referent presented to the participants. Our results indicate that although there is no strong evidence for approximating the Givenness Hierarchy with a "Predictability Hierarchy," there is some preliminary evidence for a partial correlation between the form and the predictability of a referent.

Keywords: accessibility; predictability; reference resolution; Givenness Hierarchy; self-paced reading; psycholinguistics

Introduction

Referring is one of the central functions performed by human language. When referring to an entity, it is often the case that language users have a variety of different reference forms they can choose from. Taking the reference to the city of Chicago as an example: in English, the linguistic forms one can use range from a simplex pronoun "it," a proper name "Chicago," a noun phrase with the definite article "the city," to a demonstrative expression such as "this/that city." In light of such diversity that all languages provide to their users, the central question is how a speaker/listener negotiates the relationship between the form of a referring expression and the intended referent.

Accessibility and predictability

It is widely recognized that the discourse structure of a referent in the mental state of language users affects the processing of referring expressions. Many studies of reference resolution evoke a concept of referent *accessibility* (also referred to as *prominence* or *salience*). More specifically, a commonly held view is that there is a correlation between the accessibility of a referent and the linguistic form in which the referent is realized. Fundamentally, referents that are more accessible tend to be realized in more reduced forms. According to the Givenness Hierarchy (GH) Theory (Gundel, Hedberg, & Zacharski, 1993; Gundel, Bassene, Gordon, Humnick, & Khalfaoui, 2010) and the Accessibility Theory (Ariel, 1991, 2001), reference form should serve as a cue for comprehenders to narrow down the space of possible referents. Both theories arrive at a similar hierarchy that orders different types of reference forms. Taking Gundel et al. (1993)'s Givenness Hierarchy as an example, as presented in Table 1, reference forms are mapped onto six cognitive statuses from *in focus* to *type identifiable*, in descending order of accessibility. Each reference form encodes a certain level of referent accessibility as part of its lexical semantic representation. For example, in English, the pronoun encodes the highest level of accessibility in the sense that it is expected to refer to an entity that is highly accessible in the mental state of interlocutors.

Although the concept of accessibility seems to make intuitive sense, the current literature in psycholinguistics has not reached a consensus on a clear definition for referent accessibility. Intuitively, this concept is meant to index how easy it is to access or retrieve a referent in memory (Arnold, 2010). However, the questions regarding what this ease of information retrieval exactly means and how it can be best quantified have remained contested. Ariel (2001) claims that accessibility is a comprehensive concept and should not be measured using a single metric. Similarly, various studies show that accessibility can hardly be measured using a single factor and should reflect the interaction of multiple constraints instead (Badecker & Straub, 2002; Järvikivi, Van Gompel, Hyönä, & Bertram, 2005; Lappin & Leass, 1994). Making the issue even more complicated, Kaiser and Trueswell (2008) find that the sensitivity to different accessibility constraints may vary across reference forms, and they thus propose a form-specific multiple-constraints approach for accessibility. On the other hand, as pointed out by Arnold (2010), the lack of definition on how to exactly measure this concept leads to a practical problem with respect to logical circularity: the use of a reference form is originally theorized to be determined by referent accessibility, but the ultimate measurement of accessibility turns out to be the reference form itself.

Apart from accessibility, another independently motivated concept that contributes to referent resolution is *predictability*. It refers to the likelihood anticipated by a comprehender that a certain referent is to be mentioned next. Similar to accessibility, predictability is also comprehensive: there are multiple factors influencing the comprehender's expectancy on the "next-mention" referent. For example, Tily and Pi-

Table 1: Givenness Hierarchy of reference forms in English

in focus	>	activated	>	familiar	>	uniquely identifiable	>	referential	>	type identifiable
{*it*}		{*this* N}		{*that* N}		{*the* N}		{indefinite *this* N}		{*a* N}

antadosi (2009) demonstrate that factors such as the reference form, the frequency of mentioning, and the grammatical role, have significant effects on the prediction of the next-mention referent. Some of these factors overlap with the ones that are typically considered as affecting referent accessibility. But compared to accessibility, there is a single formalized measure of referent predictability: the probability estimated by the comprehender that a given referent is to be mentioned. The formalized notion of predictability connects the specific problem of referent resolution to the general information-theoretic approaches on parsing and comprehension (Hale, 2016; Levy, 2008). Under such models, prediction, often quantified with the measure of *surprisal*, plays an important role in modulating processing difficulty. Given that predictability is a more trackable concept than accessibility, it is theoretically important to evaluate how the two concepts are related and whether predictability can replace, or at least serve as an approximate of, accessibility. The current study makes the first attempt to address this question.

Parallels between accessibility and predictability

At an intuitive level, highly predictable information is also highly accessible, since such information is often pre-activated prior to being encountered, thus increasing the accessibility of memory retrieval and reducing processing cost. Empirically speaking, a number of parallels between predictability and accessibility have been observed.

As mentioned earlier, referents realized in pronouns are hypothesized as having the highest accessibility on the Givenness/Accessibility Hierarchy. It has been observed that referents with higher predictability also tend to have higher pronominalization rates (Arnold, 2001; Rosa & Arnold, 2017). Moreover, predictability and accessibility seem to co-vary with a similar set of factors. According to Arnold (1999), the linguistic factors that increase a referent's likelihood to be mentioned next (e.g. recency, syntactic position, focus, parallelism, and goal status) also increase its likelihood to be realized as a pronoun. A similar correlation is observed in some corpus studies, which found that particular thematic roles, such as "goal," increases both the frequency of next-mention and the rate of pronominalization at the same time (Arnold, 2001, 2003). In production, it has been observed that there is an inverse correlation between referent predictability and the length of the reference form (Jaeger & Buz, 2018; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013; Tily & Piantadosi, 2009; Arnold & Zerkle, 2019). That is, if a referent is highly predictable for comprehenders, the speaker tends to select a more reduced expression to refer to it, such as a pronoun. But we also note that this empirical generalization failed to be substantiated by a number of studies (Fukumura & Van Gompel, 2010; Modi, Titov, Demberg, Sayeed, & Pinkal, 2017; Rohde & Kehler, 2014; Modi, Titov, et al., 2017; Zhan, Levy, & Kehler, 2020). Some have suggested that referent predictability should only influence comprehension but not production (Kehler & Rohde, 2013, 2019; Rohde & Kehler, 2014).

The current study

Despite the parallels observed between accessibility and predictability, there have not been any studies that directly evaluate whether the two notions are empirically equivalent. The current study explores this question. Recall that in the Givenness Hierarchy (GH) theory, each reference form encodes a certain level of accessibility as part of their lexical semantic representation. In this regard, if predictability can serve as a reasonable approximation of accessibility, a parallel "Predictability Hierarchy" should be expected. That is to say, each reference form is expected to encode a certain level of predictability in the same order as the GH. To assess whether this is indeed the case, the current study examines the relationship between predictability and reference form in comprehension. Most of the previous studies on the relationship between predictability and reference form investigate production. These previous studies often assume that speakers make production choices based on an estimate of the predictability of a referent in a listener's mental state. As mentioned in the last section, there are both empirical and theoretical challenges to this approach. We therefore will focus on comprehension in the current study.

Two different probability measures are relevant to the comprehension of referring expressions. The first one is $P(referent)$, which is the probability of a given referent in the current discourse, and this probability represents the concept of referent predictability. The second one, $P(RF|referent)$, is the probability of a reference form (RF) given a specific referent. If each reference form encodes a certain level of predictability in the same order as the GH, higher $P(referent)$ should lead to higher probability of those reference forms that are closer to the pronoun end of the hierarchy in Table 1.

If there exists a "Predictability Hierarchy" in the same way as the Givenness Hierarchy, a penalty would arise when the comprehender encounters a referent whose actual predictability $P(referent)$ in the discourse mismatches the level of referent-predictability indicated by the form of the referent. The current study assesses this prediction with a self-paced reading experiment by manipulating the predictability of a referent and the form with which this referent is expressed. In general, stronger mismatch should lead to larger penalty

and thus result in longer reading times on the target referent.

Experiment

Participants read a context passage and then self-paced read a one-sentence continuation, as in (1). One of the NPs in the continuation sentence was the target referent, whose reference form was manipulated as shown in the curly bracket in (1).

(1) Sample experiment stimuli

> *Context Passage*: Today, in Rich's Kitchen we'll learn about the fine attributes of baking a cake. Since I am not a phenomenal baker we will be assisted by the use of Little Debbie in using one of their fine cake mixes.
>
> *Continuation*: In order to/ properly make/ {it/this cake/that cake/the cake}/ we/ will/ need/ some vegetable oil/ and/ a couple of eggs.

The materials were selected from the data set constructed by Modi, Titov, et al. (2017). All the sentences in the original data set were selected from the InScript corpus (Modi, Anikina, Ostermann, & Pinkal, 2017). In Modi, Titov, et al. (2017), the predictability of each target referent was measured using the referent cloze game first created by Tily and Piantadosi (2009). In the referent cloze game, participants were presented with a passage that was cut off right before the target NP, and were asked to guess what the upcoming NP would most likely refer to by choosing among all the existing discourse referents or by writing down a new referent. For instance, in example (1), participants were presented with a discourse from "Today" (in *context passage*) up until the verb "make" (in *continuation*), and they were asked to choose among "Rich," "Kitchen," "we," "cake," "I," "baker," "Little Debbie," and "cake mix," or to write down a new referent. For this example, the target NP from the original corpus is "this cake." The predictability of the target referent was then computed by the percentage of correct guesses that match the actual referent in the original sentence from the corpus. In the present study, we kept the discourse context preceding the target referent identical as in the Modi and colleagues' study. In this way, the predictability of the target referent measured in the original study was available to us. Since the discourse context was at least three sentences long, only the sentence containing the target referent was presented in the SPR paradigm, with the rest of the discourse context displayed as a whole (see the Procedure section).

A subset of the reference forms from Table 1 was investigated in the current study. As shown in (1), for each item, the target referent appeared in four different forms: simplex third-person pronoun ("he," "she," "it," and "they"), proximal demonstratives ("this N"), distal demonstratives ("that N"), and nominals with a definite article ("the N"). As in Table 1, these four forms represent four different cognitive statuses with descending accessibility: the simplex pronoun occupies the highest status *in focus*; the proximal "this N" encodes the second-highest status *activated*, followed by the distal "that N" as *familiar* and the definite "the N" as *uniquely identifiable*.

Meanwhile, for each target referent and each item, we also coded for several variables related to the context in which the target referent appeared, as shown in (2). These variables served as corpus-related control variables in the subsequent reading time (RT) analyses.

(2) Corpus-related control variables

- *Phi-featured Referents*: the number of referents with the same gender and number features as the target referents in the discourse so far
- *Recency*: the distance measured as the number of intervening characters from the last-mentioned antecedent
- *Frequency*: the number of mentions of the target referent in the discourse so far
- *Intervening Referents*: the number of referents between the target referent and its last-mentioned antecedent
- *Previous Referents*: the number of referents that have been mentioned so far in the discourse
- *Grammatical Role*: the grammatical role, coded as "subject" and "object," of the target referent
- *Previous Grammatical Role*: the grammatical role, coded as "subject," "object," and "others," of the last-mentioned antecedent of the target referent
- *If in SPR*: a boolean value indicating whether the last-mentioned antecedent was presented in the self-paced reading sentence or in the separate context passage

The prediction with the design in (1) is as follows. If referent predictability and accessibility are empirically equivalent, the Givenness Hierarchy should give rise to a parallel Predictability Hierarchy: each reference form thus encodes a certain level of predictability in the same order as the Givenness Hierarchy. That is, among all the four types of reference forms to be investigated, "the N" should encode the lowest level of predictability, followed by "that N," whose encoded predictability should be higher than "the N" but lower than "this N," and in the end, the pronoun encodes the highest predictability level. Given this Predictability Hierarchy, with each reference form, there will be a mismatch if the predictability encoded by that reference form is inconsistent with the actual predictability of the referent. On the reference form "the N," the mismatch effect should become stronger when the referent predictability increases, since "the N" encodes the lowest predictability level compared with the other reference forms, resulting in increasing reading times (RTs). Then, moving from "the N" to "that N," since a higher predictability level is encoded, the mismatch effect should be dampened for highly predictable referents and be enhanced for referents that are less predictable. Iteratively, if we move along the hierarchy from the lowest predictability level ("the N") all the way to the highest one ("pronoun"), when the actual referent predictability increases, the mismatch between the referent predictability and the reference form is expected

to vary gradually. In other words, we expect an interaction effect between referent predictability and reference form at every reference form level compared with the adjacent form on the hierarchy.

Method

Participants 112 native speakers of English living in the U.S. (mean age: 32.89) were recruited through Amazon Mechanical Turk.

Materials Experiment materials were selected from Modi, Titov, et al. (2017)'s dataset, which consists of narrative stories from InScript (Modi, Anikina, et al., 2017), a corpus of spoken language collected by asking participants to describe a scenario in daily life as if explaining it to a child. There were 20 experimental items and 20 filler items in the present study. Each item consisted of a discourse context followed by a one-sentence continuation. The target referent, whose reference form was manipulated, was in the continuation sentence. The materials were presented to participants in a Latin Square design: each participant only saw one of the four types of reference forms in each item. As mentioned earlier, the target referents' predictability measures were available from the original data set created by Modi, Titov, et al. (2017). In selecting items from the original set, we balanced the distribution of target referents' predictability across items by dividing the predictability scale (a 0-1 interval) into 5 equally spaced bins (with an increment of 0.2 intervals each), and each bin contained four items. Filler items were designed to contain both an acceptable and an unacceptable condition. The acceptable filler sentences were naturally produced utterances drawn from the original InScript corpus. To create the unacceptable filler sentences, we modified the acceptable fillers by introducing semantic and pragmatic anomalies. Filler sentences served as attention check items for the data analysis. We removed 18 participants from the data analysis since the mean of their acceptability ratings for the acceptable fillers was lower than the unacceptable ones.

Procedure Participants recruited through Amazon Mechanical Turk were directed to the server of Ibex Farm to take part in the experiment. After submitting the consent form and completing a questionnaire of language background, participants were presented with the instructions of the experiment. For each trial, participants were asked to go through three steps. In the first step, they read a context passage that was part of a story describing a common scenario in daily life. In the second step, the context passage disappeared, and the participants were directed to another page that presented in self-paced reading paradigm a one-sentence continuation of the previous context passage. The last step was an acceptability judgment task, in which the participants were asked to rate on a 1-7 scale the naturalness of the continuation sentence they read. There were four practice trials in total to familiarize participants with the experiment procedure, two of which were explicitly assigned as part of the instructions, and another two were implicitly assigned after the participants were notified that the experiment session would start.

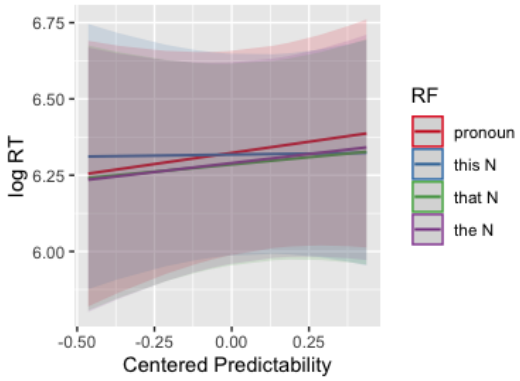

Figure 1: Critical regions: Model predicted interactions between Predictability and Reference Form.

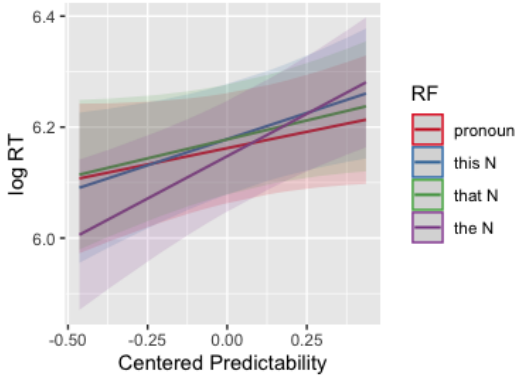

Figure 2: Spill-over regions: Model predicted interactions between Predictability and Reference Form.

Results and discussion

Linear mixed-effects models (LMEMs) over log RTs were performed for the critical referent region and the spill-over region with the maximal random effects that allow the model to converge. The critical fixed-effect predictors are the Reference Form (RF) and the Predictability of the referent. The regression model, as presented in (3), also controls for a number of corpus-related effects as in (2) as well as several experiment-control variables[1]. The model predicted RTs are presented in Figure 1 for the critical region and in Figure 2 for the spill-over region. To compare each RF with the left adjacent RF on the hierarchy in Table 1, the variable Reference

[1]Experiment-control variables include *Word Length* (the orthographic length of the chunk), *Chunk Position* (the position of the chunk in the SPR sentence), and *RT Previous* (the logRT of the previous chunk).

Table 2: Results of corpus-related control variables

Effect		Critical Region				Spill-Over Region			
		Beta	std.	*t* value	*p* value	Beta	std.	*t* value	*p* value
Recency		**-0.123**	**0.032**	**-3.887**	**0.001**	-0.015	0.009	-1.707	0.104
Frequency		0.086	0.125	0.689	0.501	**-0.117**	**0.034**	**-3.494**	**0.002**
Intervening Referents		**0.316**	**0.095**	**3.336**	**0.004**	0.035	0.026	1.385	0.182
Previous Referents		**-0.057**	**0.017**	**-3.3**	**0.004**	-0.002	0.005	-0.332	0.744
Previous Grammatical Role	Other	-0.162	0.196	-0.828	0.42	0.054	0.055	0.992	0.334
	Subject	**-0.381**	**0.169**	**-2.248**	**0.039**	-0.054	0.044	-1.23	0.233
If in SPR		**-0.703**	**0.242**	**-2.902**	**0.01**	-0.131	0.063	-2.071	0.052

Form was backward-difference coded. On neither the critical region nor the spill-over region did we find step-by-step Predictability × RF interaction moving from the pronoun to the definite "the N." However, in the spill-over region, there is an RF × Predictability interaction when comparing "the N" (Figure 2) to the pronoun ($\beta = 0.188$, $p = 0.018$) and to "that N" ($\beta = 0.169$, $p = 0.034$). That is to say, for the definite "the N," log RT in the spill-over region linearly increased when the referent predictability increased. This linear relation is dampened for other reference forms.

(3) Linear mixed effects model over logRTs.

Fixed Effects
predictability × reference form + phi-featured referents + recency + frequency + intervening referents + previous referents + grammatical role + previous grammatical role + if in SPR + word length + chunk position + RT previous
Random Effects
Critical region: (predictability|subject) + (1|item)
Spill-over region: (1|subject) + (1|item)

Although the primary focus of the current study is on the interaction between predictability and referent forms, it is worth noting that we also found a significant main effect of Predictability on the spill-over region ($\beta = 0.187$, $p = 0.019$), showing slower RTs for referents with higher predictability. This may seem counterintuitive at the first sight given the overwhelming evidence in the literature that showed a robust negative correlation between the predictability of a word and the time it takes to read that word. But we note that previous findings are mostly based on content words. Words that are predictable in a context can have their lexical semantic features pre-activated, facilitating the lexical processing when the word is actually encountered. This facilitation effect of predictability does not apply to the pronoun condition in the current study in that pronouns do not have inherent lexical semantic features and their interpretation is completely dependent on the antecedent. It is also the case that the linking hypothesis in the existent literature is between a word's predictability in a context and its RT; whereas the predictability measured in the current study is the predictabiity of a referent, which can be expressed in different forms/words. There is not necessarily a linear correlation between the predictability of a referent and the RT measure. The positive effect of referent predictability observed in the current results needs to be further validated in future studies.

A number of variables presented in (2) also showed an effect on RTs as summarized in Table 2. As expected, if the target referent has been repeatedly mentioned in the discourse (Frequency), the RT on the target referent is faster (on the spill-over region); if the antecedent of the target referent was in a subject position (Previous Grammatical Role), or if the antecedent and the target form were in the same clause (If in SPR), the target referent was also read faster; and if there were many intervening referents between the target and its antecedent (Intervening Referents), it led to slower RT on the target referent. What is somewhat unexpected is that both Recency, which measures the linear distance between the target and its antecedent, and Previous Referents, which tracks how many other referents are present in the same discourse, had a significant negative effect on the RTs at the critical referent region. For Recency, it is possible that distance measured in terms of the number of characters might not be an appropriate representation of the amount of time elapsed between the target and its antecedent, especially given the self-paced reading paradigm. For Previous Referents, it is possible that the number of referents in the discourse co-varies with the richness of the context, and more contextual information facilitates reference resolution.

In summary, moving from one reference form to the next one on the GH, we did not find a step-by-step interaction between reference form and predictability. Thus, there is no strong evidence that different reference forms constitute a "Predictability Hierarchy" in the same order as the GH. Referent predictability and accessibility, based on the current evidence, should therefore be kept apart as two distinct theoretical constructs. It is worth noting, however, a weaker version of the "Predictability Hierarchy" hypothesis may still hold. The current results found differences between "the N" on the one hand and the other reference forms on the other, indicating that "the N" encodes a different degree of predictability of the referent, distinguishable from other reference forms. More specifically, the results indicate that the predictability encoded by "the N" is lower than other RFs. This is consistent

with the "Predictability Hierarchy" hypothesis, suggesting a somewhat partial correlation between referent predictability and accessibility.

The reference forms we examined in this study included demonstratives like "this/that N." It is worth noting that demonstratives are not only anaphoric but also deictic. Some previous studies have observed that demonstratives and pronouns demonstrated different sensitivities to linguistic contexts (Kaiser & Trueswell, 2008). Their divergent behaviors might stem from the fact that, as attention-managing devices, deictics are used to perform different pragmatic functions than regular anaphors, pointing to the possibility that in communication demonstratives may trigger additional cognitive processes (Ehlich, 1982). In the current study, however, neither accessibility nor predictability is designed to capture the deictic nature of demonstratives. In this regard, the current "Predictability Hierarchy" might not be able to fully reflect the processing of "this/that N" and therefore is hard to detect the difference between these two forms and the others.

There is also a potential concern with the statistical power of the current experiment. There are two critical predictors plus eight corpus-related predictors in our statistic models, but we only have a limited number of experimental items. We conducted a power analysis using the SIMR package in R (Green & MacLeod, 2016). For all the significant corpus-related effects in Table 2, with the current effect sizes, the power reached 80%. But for the Predictability $\times$ RF interaction, to reach an 80% power on the critical region would require approximately 265 participants, and an 80% power on the spill-over region would require 155 participants. The absence of the interaction effect in the current study could be the result of insufficient power. We therefore need future work that is better powered.

Conclusions

In conclusion, while there is no robust support to entirely approximate the Givenness Hierarchy with a "Predictability Hierarchy," there is some preliminary evidence for a partial correlation between the form of a referent and the predictability of a referent. More future work is needed to further understand the intricate relationship between the form of a referent and the communicative function it signals.

References

Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of pragmatics*, *16*(5), 443–463.

Ariel, M. (2001). Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, *8*, 29–87.

Arnold, J. E. (1999). Reference form and discourse patterns. *Unpublished doctoral dissertation*.

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

Arnold, J. E. (2003). *Multiple constraints on reference form: null, pronominal, and full reference in mapudungun. preferred argument structure: grammar as architecture for function, ed. by jw du bois, le kumpf and wj ashby, 225–45.* Amsterdam: John Benjamins.

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

Arnold, J. E., & Zerkle, S. A. (2019). Why do people produce pronouns? pragmatic selection vs. rational models. *Language, Cognition and Neuroscience*, *34*(9), 1152–1175.

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 748.

Ehlich, K. (1982). Anaphora and deixis: Same, similar, or different. *Speech, place and action. Studies in deixis and related topics*, *315*, 338.

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.

Green, P., & MacLeod, C. J. (2016). Simr: an r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.

Gundel, J. K., Bassene, M., Gordon, B., Humnick, L., & Khalfaoui, A. (2010). Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics*, *42*(7), 1770–1785.

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

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397–412.

Jaeger, T. F., & Buz, E. (2018). Signal reduction and linguistic encoding. In E. M. Fernández & H. S. Cairns (Eds.), *Blackwell handbooks in linguistics. the handbook of psycholinguistics* (pp. 38–81). Wiley-Blackwell.

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

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.

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

Kehler, A., & Rohde, H. (2019). Prominence and coherence in a bayesian theory of pronoun interpretation. *Journal of Pragmatics*, *154*, 63–78.

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics*, *20*(4), 535–561.

Levy, R. (2008). Expectation-based syntactic comprehen-

sion. *Cognition*, *106*(3), 1126–1177.

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.

Modi, A., Anikina, T., Ostermann, S., & Pinkal, M. (2017). Inscript: Narrative texts annotated with script information. *arXiv preprint arXiv:1703.05260*.

Modi, A., Titov, I., Demberg, V., Sayeed, A., & Pinkal, M. (2017). Modeling semantic expectation: Using script knowledge for referent prediction. *Transactions of the Association for Computational Linguistics*, *5*, 31–44.

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

Rosa, E. C., & Arnold, J. E. (2017). Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language*, *94*, 43–60.

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.*

Zhan, M., Levy, R., & Kehler, A. (2020). Pronoun interpretation in mandarin chinese follows principles of bayesian inference. *Plos one*, *15*(8), e0237012.