Bilingualism: Language and Cognition 22 (3), 2019, 456–475 © Cambridge University Press 2018 doi:10.1017/S136672891800055X

A Bayesian approach to establishing coreference in second language discourse: Evidence from implicit causality and consequentiality verbs*

WEI CHENG
College of Foreign Studies, Jinan University
AMIT ALMOR
Department of Psychology, Linguistics Program
Institute for Mind and Brain, University of South Carolina

(Received: November 24, 2016; final revision received: April 02, 2018; accepted: April 04, 2018; first published online 16 May 2018)

This study investigated Chinese-speaking English learners' use of implicit causality and consequentiality biases in establishing coreference under a Bayesian view of reference interpretation, which distinguishes between context-based priors about which entity will be re-mentioned and new evidence provided by the referential expression form. In two sentence-completion experiments, participants wrote continuations to sentence fragments with either implicit causality (Experiment 1) or consequentiality (Experiment 2) biases that ended either with or without a pronoun. In both experiments, L2 speakers showed native-like re-mention biases following no-pronoun fragments, indicating native-like predictions about the next-mentioned referent. Following pronoun fragments in NP2-biasing contexts, L2 speakers produced more NP1 continuations than native speakers. We show that this difference lies in different beliefs about pronoun use in the two populations. Specifically, L2 speakers showed a stronger association between pronouns and NP1 referents than native speakers following NP2-biasing verbs.

Keywords: implicit causality, implicit consequentiality, pronoun, Chinese-speaking learners of English

1. Introduction

To comprehend a text, readers need to understand not only individual words, but also the connection between linguistic units, which often span clauses or sentences. One such connection is coreference, a mechanism in which a referring expression (e.g., a pronoun) and another element (e.g., an antecedent that is previously mentioned) refer to the same entity (Halliday & Hasan, 1976). Establishing coreference is influenced by many factors, most notably, the antecedent's structural and linear position in the sentence, with the entity in the subject position or the first-mentioned entity being the preferred referent of a subsequent pronoun in certain discourse contexts (e.g., Ariel, 1990; Arnold, 1998; Crawley, Stevenson & Kleinman, 1990; Frederiksen, 1981; Givón, 1992, 1995; Grosz, Joshi & Weinstein, 1995; Järvikivi, Van Gompel, Hyönä & Bertramet, 2005). In addition, verb meaning also affects coreference. For instance, when presented with a sentence fragment containing the verb *fear*, the connective *because*, and an ambiguous pronoun as in (1), people usually continue the sentence with the pronoun referring to the object *Sara*. By contrast, when the verb is changed to *frighten* as in (2), people tend to refer the pronoun to the subject *Mary*. This phenomenon is known as implicit causality (Garvey, Caramazza & Yates, 1976).

(1) Mary feared Sara because she ...

(2) Mary frightened Sara because she ...

The effect of verb meaning on coreference is modulated by discourse coherence relations (Kehler, Kertz, Rohde & Elman, 2008; Koornneef & Sanders, 2013; Stevenson, Knott, Oberlander & McDonald, 2000). When the connective in (1) and (2) is changed to *so*, thereby creating a different coherence relation, the biases of pronominal reference change accordingly: *Mary* in the case of *fear* and *Sara* in the case of *frighten*. This phenomenon is known as implicit consequentiality (IR hereafter with R standing for result) (Crinean & Garnham, 2006; Stewart, Pickering & Sanford, 1998).

* We thank Xiaojing Yang for helping us recruiting participants and Peter Nelson and Cameron Smith for helpful comments on an earlier version of this paper. This work was partially supported by the NSF Grant BCS0822617 (A. A.), the National Social Science Fund Grant 17CYY017 (W. C.), and the Fundamental Research Funds for National Universities (15JNYH007) – Constructing a Semantic Knowledge Network of English Words: A Multimodal Perspective (W. C.)

Address for correspondence:
Amit Almor, Department of Psychology and Linguistics Program, University of South Carolina, 29208
almor@sc.edu

The interaction between these factors in interpreting coreference during comprehension has been recently described in a Bayesian framework (Kehler et al., 2008; Kehler & Rohde, 2013). In this framework, as part of discourse processing, comprehenders make probabilistic predictions about which referent is likely to be re-mentioned in the following discourse on the basis of the semantic content of prior discourse. Upon encountering an anaphor, comprehenders then update their prediction of which is the referent by integrating their initial predictions (priors) with the referential bias (evidence) provided by the form of the anaphor: Pronouns indicate a strong subject/first-mention bias and fuller references signal biases towards non-subject antecedents. While this model has been shown to be useful in explaining many empirical findings concerning monolingual speakers, it has not been fully evaluated in the context of second language (L2) speakers' resolution of coreference. This represents an important gap in the literature, because many recent theories of L2 processing (e.g., Grüter, Rohde & Schafer, 2014, 2017; Kaan, 2014) have highlighted the role of prediction, which is an essential component of the Bayesian framework. In this paper, we aim to further our understanding of this issue by investigating advanced Chinese-speaking English learners' coreference resolution in the contexts of implicit causality and consequentiality. In the remainder of the introduction, we first explain the phenomena of IC/IR and the Bayesian model of coreference resolution in more detail. We then review previous research on the establishment of coreference by non-native speakers in their L2.

1.1. Implicit causality and implicit consequentiality

IC and IR biases appear under different discourse coherence relations. In particular, IC biases are closely related to the 'Explanation' coherence relation (Kehler et al., 2008), in which the second clause provides an explanation for the event described in the first clause. IR biases arise in the 'Result' coherence relation (Kehler et al., 2008), in which the second clause is a consequence of the event described in the first clause.

Depending on the discourse coherence relation, some verbs – usually interpersonal verbs including psychological verbs as well as action verbs – show certain IC or IR biases. When the discourse coherence relation is Explanation, some verbs show an implicit direction of causality attributing the cause of the event described by the verb to one of its two arguments (e.g., Caramazza, Grober & Garvey, 1977; Koornneef & Van Berkum, 2006). As illustrated in (1) above, *frighten* attributes the cause to the first noun phrase (NP1) or the subject, whereas *fear* attributes the cause to the second noun phrase (NP2) or the object. When the discourse coherence relation is Result, some verbs show an implicit direction of consequentiality such that one of its arguments is usually considered as bearing the consequence of the event described by the verb (e.g., Au, 1986; Stewart et al., 1998). For instance, *frighten* has an IR bias towards NP2 while *fear* has an IR bias towards NP1. The contrast between *frighten* and *fear* demonstrates that different types of verbs have distinct IC or IR biases. Some argue that the difference is due to verbs' semantic structures (e.g., Brown & Fish, 1983; Crinean & Garnham, 2006; Hartshorne & Snedeker, 2013), but others simply regard it as a reflection of world knowledge associated with different verbs (e.g., Pickering & Majid, 2007). Despite this controversy regarding the factors underlying these effects, it is uncontroversial that whether a verb has an NP1 or NP2 bias is dependent on its meaning.

IC biases are not only found in English but also in other languages, particularly for transitive psychological verbs (Hartshorne, Sudo & Uruwashi, 2013). Studies on Chinese also confirmed that IC biases are robust among many Chinese verbs (e.g., Cheng & Almor, 2015; Jiao & Zhang, 2005; Miao, 1996; Miao & Song, 1995; Sun, Shu, Zhou & Zheng, 2001). In addition, robust IR biases were also found among Chinese verbs (Cheng & Almor, 2015). Thus, IC and IR biases are arguably universal biases.

1.2. A Bayesian approach to coreference resolution

IC and IR biases influence coreference resolution during comprehension. As shown in examples (1) and (2) above, the continuations people typically produce following an ambiguous pronoun that can potentially refer to either one of two arguments of a fragment with an IC or IR verb indicate that these biases affect their resolution of ambiguous pronouns (e.g., Ehrlich, 1980; Garvey & Caramazza, 1974; Hartshorne & Snedeker, 2013). In addition, these biases also exert an influence on re-mention biases in language production (e.g., Au, 1986; Kehler et al., 2008). For example, when presented with sentence fragments such as *Mary feared Sara because*, participants usually re-mention *Sara*, the referent consistent with the verb's IC bias, in their continuation to the sentence fragment.

In addition to semantic and discourse factors such as IC and IR biases, coreference production and comprehension are also affected by syntactic and linear order factors. Numerous studies have shown that the referent in the subject position or the first-mentioned referent of the previous clause is likely to be referred to by a reduced expression such as a pronoun rather than a fuller expression such as a name (e.g., Almor & Nair, 2007; Ariel, 1990; Garrod & Sanford, 1982; Givón, 1987; Gordon, Grosz & Gilliom, 1993; Gundel, Hedberg & Zacharski, 1993). Thus, the presence of a pronoun during

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

comprehension as opposed to other fuller forms of reference provides a strong referential cue in favor of the subject/first-mention referent.

Kehler and colleagues (Kehler et al., 2008; Kehler & Rohde, 2013) synthesized the above-mentioned factors that influence coreference resolution using Bayes rule as shown in (3).

$$(3)\quad p(\mathit{referent}|\mathit{pronoun}) = \frac{p(\mathit{referent}) \times p(\mathit{pronoun} \mid \mathit{referent})}{p(\mathit{pronoun})}$$

p(referent | pronoun) represents the probability that a pronoun just encountered by the comprehender is coreferential with a particular antecedent. According to the formula, it is determined by two factors. The first is *p(referent)*, the prior probability the comprehender assigns to a referent to be re-mentioned in subsequent discourse just before encountering the pronoun. This represents a predictive process, in which language comprehenders use contextual cues to generate a prediction about the next-mentioned referent before encountering the pronoun. As the input unfolds, listeners and readers make a probabilistic evaluation of the coherence relation between clauses or sentences and then form a prediction about the next-mentioned referent consistent with the coherence relation. Since IC and IR biases are associated with the Explanation and Result coherence relations, respectively, it is in this process that comprehenders make predictions that prefer an IC or IR bias-consistent antecedent as the most probable entity to be re-mentioned.

The other factor that affects pronoun resolution in the Bayesian model is the likelihood *p(pronoun | referent)*, which is the probability that a particular referent is referred to by a pronoun as opposed to other forms of reference. When comprehenders encounter a pronoun, their interpretation of the pronoun will reflect the product of the prior probability of each possible referent to be mentioned next and the relative probabilities that each of these referents will be referred to by a pronoun. Thus, the pronoun itself provides evidence that is integrated with the priors, resulting in the posterior probabilities of the different possible referents as antecedents of the pronoun. The referent chosen as the antecedent is the one with the highest posterior probability. Given that the referent in the subject position or the first-mentioned referent is usually referred to by a pronoun in the following clause instead of other referring expressions, pronouns typically contribute a strong subjecthood/first-mention cue (meaning that the probability that the antecedent is the subject or the first-mentioned entity of the previous clause is higher than with the priors alone).

To sum up, according to the Bayesian model, to successfully resolve an anaphoric expression amounts to calculating the posterior probabilities for all possible antecedents and picking the most probable one as the referent. This process relies on two sources of information in terms of Bayes formula: (1) *p(referent)*, i.e., the priors, which are the probabilities that each referent will be re-mentioned and which are based on the comprehension of prior contextual semantic information such as IC and IR biases, as well as discourse coherence relations; (2) *p(pronoun / referent)*, i.e., the likelihood that a given antecedent would be referred to by a pronoun as opposed to other forms of reference, which is based on prior knowledge about language, for example, that pronouns are typically used for subject or first-mentioned referents. Thus, in this view, pronoun resolution is a process that involves the integration of comprehenders' prediction of likely referents based on context, as well as multiple sources of probabilistic information about the general circumstances in which pronouns are used in the language, and finally choosing the referent with the highest posterior probability as the pronoun's antecedent.

It is important to note that much of the work in this area is based on two related assumptions that are often left implicit. The first assumption is that constrained production tasks, in which participants produce continuations for previous contexts provided to them, can yield important information about their comprehension of the preceding context. Indeed, much of the scientific understanding of the effects of IC and IR on language comprehension comes from language production sentence continuation tasks. Although the reliance on production tasks for the understanding of comprehension processes may seem problematic, it is in fact a common practice in psycholinguistics, where various production tasks such as cross-modal naming have been frequently used as means to examine the comprehension of preceding material.

The second related assumption is that similar patterns occur in both language production and comprehension, albeit possibly for different underlying reasons. In particular, IC and IR biases are assumed to occur in both language production and language comprehension, although their origin may be different in the two modalities. For example, while the choice of reference form may reflect production constraints, such as using a minimal form for referring to the most salient entity so as to minimize interference (Almor & Nair, 2007), comprehenders are sensitive to the patterns in language and can use the form of a referential expression as a source of information about the likely referent (e.g., MacDonald, 2013). Most relevant here is that, under this assumption, participants' choices in production (for example, whether they produce a continuation describing a specific referent) can be used as a measure of the probabilistic knowledge that guides their comprehension (for example, how they interpret a pronoun that was provided to them in the context fragment).

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

1.3. Establishing coreference in the L2

Compared to the large number of studies on L1 coreference resolution, only a few have looked at how L2 speakers establish coreference in discourse. Among these, to the best of our knowledge, only two studies (Cheng & Almor, 2017; Grüter et al., 2014, 2017) have adopted the Bayesian approach of coreference resolution and investigated L2 speakers' sensitivity to contextual information in resolving ambiguous reference (other studies have focused on other aspects such as native language influence (e.g., Roberts, Gullberg & Indefrey, 2008), the role of different anaphor types (e.g., Sorace & Filiaci, 2006), or the use of gender cues in online pronoun interpretation (e.g., Liu & Nicol, 2010)).

In an offline sentence-completion study, Grüter et al. (2014, 2017) investigated L2 learners' sensitivity to event structures in resolving ambiguous reference. They manipulated event structures by contrasting perfective and imperfective aspect marked on Source-Goal verbs (e.g., *hand*). The results showed that native English speakers continued the sentence with more references to the Source referent (e.g., *John*) following sentences in the imperfective aspect (e.g., *John was handing a book to Bob*) than following sentences in the perfective aspect (e.g., *John handed a book to Bob*). By contrast, despite having acquired the knowledge of English aspect as shown in an independent grammaticality test, Japanese and Korean-speaking learners of English showed a referential bias towards the Goal referent (i.e., *Bob*) following both structures. Interestingly, when presented with prompts that ended with a pronoun (e.g., *John handed/was handing a book to Bob. He . . .*), L2 speakers still did not show any difference between the aspect conditions, but, like native speakers, produced more continuations with references to the subject /first-mentioned antecedent than when no pronoun was present. Based on the Bayesian model of coreference resolution, these results indicate that although L2 speakers are sensitive to the subjecthood/first-mention cue in resolving pronouns, they are not sensitive to the aspect information in their prior prediction about which referent is likely to be re-mentioned.

On the basis of these findings, Grüter et al. (2014, 2017) proposed the RAGE hypothesis (Reduced Ability to Generate Expectations), arguing that L2 speakers are not able to engage in native-like predictions. This is a timely proposal that ties to recent trends in research on monolinguals, which has established that L1 processing is characterized by prediction (e.g., Kamide, 2008; Kuperberg & Jaeger, 2016). However, as argued by Kaan (2014), there may be no qualitative difference between L1 and L2 speakers in terms of prediction and that any differences in performance may simply reflect external factors that influence predictive processing in general, related to L2 speakers' native language influence and their proficiency in the L2. Therefore, it is not clear whether L2 participants' failure to generate native-like anticipation about which referent to be re-mentioned in Grüter et al.'s study reflects a specific difficulty in generating predictions in L2.

Cheng and Almor (2017) is another study that employed a Bayesian approach to examine L2 pronoun resolution. In two sentence-completion experiments, they investigated advanced Chinese-speaking L2 English learners' sensitivity to IC and IR biases in resolving ambiguous pronouns. They used Experiencer-Stimulus (ES) verbs such as *fear* and Stimulus-Experiencer (SE) verbs such as *frighten*, two typical types of psychological verbs that have different IC or IR biases as introduced above. Participants wrote continuations to sentence fragments ending with a pronoun prompt such as *Mary frightened/feared Sara because/so she* _. The results showed that although L2 participants resolved the pronoun in accordance with different IC or IR biases between ES and SE verbs, they could not apply this type of information as robustly as native speakers. Specifically, when the discourse-biased referent was NP2, L2 participants produced significantly more references to NP1 than native speakers.

According to the Bayesian model of coreference resolution, there are three possible explanations for Cheng and Almor's (2017) results. First, the difference between the native and the L2 speakers could be due to the latter's reduced ability to use IC and IR biases in their prediction about the referent to be re-mentioned, in line with the RAGE hypothesis (Grüter et al., 2014, 2017). A second alternative is that the L2 speakers in Cheng and Almor's study may have encountered no specific difficulty in prediction, but had problems integrating these predictions that were based on the IC and IR biases with the strong subjecthood/first-mention cue provided by the pronoun. Since all the materials in Cheng and Almor included pronouns at the end of the prompt, these two explanations cannot be teased apart. Finally, a third alternative is that, in line with Kaan (2014), Cheng and Almor's results could reflect differences between Chinese and English. By this explanation, the L2 speakers in their study behaved differently than native English speakers due to differences between Chinese and English. In particular, compared with English, which has a large number of SE verbs, Chinese has a limited set of SE verbs, as causation for SE predicates is mainly being expressed in periphrastic causatives in Chinese (Liu, 2016; Zhang, 2003). Thus, it may be that the difference between L1 and L2 speakers found by Cheng and Almor resulted from L2 participants' difficulty in understanding SE verbs, especially those without counterparts in their native language. Therefore, overall, it is unclear why L2 speakers cannot use IC and IR biases as robustly as native speakers when establishing coreference.

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

1.4. The present study

The current study aimed to address the open questions discussed above and thus further our understanding of the similarities and differences in establishing coreference between native and non-native speakers. To do so, we investigated advanced L1-Chinese L2-English speakers' use of IC and IR biases in coreference resolution in two sentence-completion experiments, one on IC (Experiment 1) and the other on IR (Experiment 2). In both experiments, participants were instructed to write up natural continuations to sentence fragments that contained two same-gender names and either an NP1-biasing or NP2-biasing verb. Each fragment ended with either a free prompt or a pronoun prompt. Materials with free prompts were used in previous studies to probe comprehenders' predictions of the next-mentioned referent (e.g., Kehler et al., 2008; Grüter et al., 2014, 2017) and thus allowed us to test if L2 speakers are able to use IC and IR biases effectively to generate expectations about the next-mentioned referent. Comparing free prompt and pronoun prompt conditions enabled us to find out whether and, if so, to what extent, L2 speakers' coreference is influenced by the subjecthood/first-mention cue provided by pronouns. While the task used here involved language production in that participants were required to generate continuations, the experimental manipulation concerns the context that they need to comprehend prior to producing the continuation. Thus, in line with most previous research in this area, we employed a task involving language production to study the comprehension that must have occurred before production was initiated.

The current study extends the Cheng and Almor (2017) study in two important ways. First, unlike Cheng and Almor, which exclusively used pronoun prompts, we used materials that contained both pronoun and free prompts, allowing us to determine whether L2 speakers are able to use discourse information to engage in both native-like predictions about the referent to be re- mentioned and native-like integration of the evidence provided by a pronoun. The second difference is that, instead of using exclusively ES and SE verbs, we included in this study a wide variety of verbs that have equivalents and exhibit similar IC or IR biases in both Chinese and English. By using a set of diverse verbs that are equivalent in terms of biases in both languages, we could exclude as much as possible the potential cross-linguistic influence from learners' native language lexicon. We next report the results from the two experiments.

2. Experiment 1

Experiment 1 aimed to compare native English speakers' and Chinese-speaking English learners' referent choices following English sentence fragments with NP1-biasing and NP2- biasing IC verbs that ended without or with a pronoun in a causal discourse context. All verbs had Chinese equivalents with similar biases, thus minimizing the concern that different performance of L2 speakers reflects influences of their L1. This allows us to focus in this experiment on testing the two alternative Bayesian hypotheses: If L2 learners have difficulty making predictions in English, their performance should diverge from that of native English speakers in all conditions. However, if their difficulty is related to the integration of the prior predictions with the evidence provided by the pronoun, their performance should diverge from native speakers' performance only following the pronoun fragments but not following the fragments without the pronouns.

2.1. Method

Participants

Forty-three native English speakers (L1 group) were recruited from the University of South Carolina. One participant was eliminated from analysis because she had been raised in a bilingual family. The data from the remaining 42 native English participants (31 women, $M_{age} = 19.6$ years, age range: 18–39 years) were analyzed.

Forty-four Chinese-speaking English learners (L2 group) were recruited from the Guangdong University of Technology in China and received extra credit for participation. All were native speakers of Standard Mandarin, which is the lingua franca in China and the medium of instruction at all levels of schools. All participants were undergraduate students majoring in English in their sixth semester in a four-year BA program. Many of them lived in the Guangdong area and also spoke other dialects such as Cantonese, Teochew, Hakka, etc. L2 participants were required to finish two tasks: a sentence-completion task and a translation task (see details in Procedure). Only the data of those who finished both tasks were included in the analysis. In the end, 36 participants finished both tasks (28 women, $M_{age} = 21.5$ years, age range: 21–23 years). These participants started learning English as a foreign language in a school setting at an average age of 9.5 years (age range: 7–14 years) and had learned English for an average of 12 years (range: 9–15 years). At the time of testing, two of them had visited English-speaking countries for a brief period of time (10 days and 2 months, respectively), and the others had never been to English-speaking countries. The English proficiency of the L2 participants were determined by their scores on the Test for English Majors (TEM) Band 4, which classified them as advanced English learners.[1]

[1] The TEM-Band 4 is a standardized English proficiency test administered for English majors in Chinese universities in their fourth semester (Jin & Fan, 2011). The TEM-Band 4 is equivalent to the

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

In order to better understand the individual differences in their English proficiency, a C-test adopted from Schulz (2006) and composed of three short passages with 60 blanks was administered to L2 participants. The average C-test score was 35.05 out of 60 (SD = 6.46). The C-test score was used as a covariate in the analysis.

Materials and design

The experiment contained two types of verbs: 16 NP1-biasing IC verbs and 16 NP2-biasing IC verbs. To eliminate potential influence from learners' native language lexicon as much as possible, the verbs were selected from Ferstl, Garnham, and Manouilidou's (2011) norming study of English verbs' IC biases, using the following criteria: First, the English verbs must have lexical counterparts in Chinese. Second, each verb must have a strong IC bias in the same referential direction in both English and Chinese. To establish this, a norming study was conducted on Chinese verbs. The first author, an English–Chinese bilingual, translated the 300 verbs from Ferstl et al. into Chinese.[2] These verbs were then embedded in sentence fragments of the form *NP1 verb NP2 yinwei* "because", with the two NPs being common Chinese names of different genders. The 300 items were randomly divided into five lists, each consisting of 60 verbs. To counterbalance the effect of gender, five more lists were prepared by reversing the order of the two names. The norming study was conducted via paper- and-pencil surveys divided into ten booklets. 174 undergraduate students from the Guangdong University of Technology in China (different from L2 participants) filled out the surveys during class in exchange for extra credit. All were native speakers of Mandarin Chinese (106 women, M_{age} = 19.3 years, age range: 18–21 years). They were divided into ten groups almost even in size, and each group filled out one of the ten versions of the survey. Participants' continuations were coded as referring to either NP1 or NP2 by the first author and another trained native Chinese speaker. Coders were instructed to be conservative so that, as long as there was a possibility of ambiguity, the reference was coded as 'unclear'. The coding agreement rate between the two raters was 93.1%. All disagreements were resolved through discussion between coders. Disagreements that could not be resolved were coded as 'unclear'. Each verb's IC bias was determined by the percentage of NP1 references out of all NP1 and NP2 references. The Appendix shows the list of chosen verbs and their biases.

For the actual experiment, the English verbs chosen according to the above criteria were embedded in sentence fragments of the type *NP1 verb-ed NP2 because*. The two NPs were common English names of the same gender. To counterbalance the effect of gender, one half of the items had female names and the other half had male names. In the pronoun prompt condition, a pronoun of the same gender as the names in the first clause was placed after the connective *because*. In the free prompt condition, no pronoun was used. Each item appeared in both the pronoun prompt and free prompt conditions, but each participant saw each item only once in only one condition. Sample items are given in Table 1.

The experiment had a 2×2 design with the independent variables being verb bias (NP1-biasing vs. NP2-biasing verbs) and prompt type (pronoun vs. free). The dependent variable was the continuation reference to either NP1 or NP2 in the first clause. The design was counterbalanced. Every participant saw half of the items in the free prompt condition and the other half in the pronoun prompt condition. Every item was presented in the pronoun prompt condition to half of the participants and in the free prompt condition to the other half. In the end, two lists were prepared. Each list contained 32 experimental stimuli as well as 48 fillers that had the same structure as the experimental stimuli but contained non-IC verbs and other types of connectives (e.g., *and, but*, etc.). All the stimuli within a list were pseudo-randomized, with at least one filler between experimental stimuli.

Following the sentence-completion task, L2 participants were also required to finish a translation task as a measurement of their semantic knowledge of the items used in the experiment. This was a necessary step because their responses would not be meaningful if they did not know what the verb meant. The translation task was composed of the same 32 items used in the sentence-completion experiment except that participants were only presented with the first clause of the items as an independent sentence (e.g., *Mary called Sara*).

Procedure

The study was conducted via an offline paper-and-pencil survey. L1 participants took the survey in small groups of 3–7 people in a lab. L2 participants took the survey in a class. Participants were randomly and evenly assigned to one of the two lists printed on a booklet. Before the experiment started, participants were given verbal instructions on how to complete the survey. Specifically, they were asked to write down natural continuations to the sentence fragments in an intuitive way and in the

level between B1 and B2 of CEFR (Liu, 2012; Tang, Pritchard & Shi, 2012), which are roughly equivalent to the levels of intermediate high and advanced low of ACTFL, respectively (Martínez Baztán, 2008). Participants took the TEM-Band 4 one year before the present study. Given that they were English majors and the majority of their classes were taught in English, it is reasonable to assume that L2 participants in this study were advanced learners of English.

[2] Since Chinese has relatively few SE verbs, the English verbs that do not have counterparts in Chinese were translated to corresponding periphrastic causatives. For example, *disappoint* was translated as *shi shi-wang* "make disappointed". Also, the translation was guided by consulting *Oxford Advanced Learner's English-Chinese Dictionary*.

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

Table 1. *Sample Items in Experiment 1.*

	Verb Bias	
Prompt Type	NP1	NP2
Pronoun	Mary called Sara because she . . .	Jake trusted Adam because he . . .
Free	Mary called Sara because . . .	Jake trusted Adam because . . .

prescribed order. Following Goikoetxea, Pascual, and Acha (2008), participants were instructed to go over all the stimuli from the beginning to the end after the continuation phase was complete. If there was a subject pronoun in the second clause, regardless of whether it was part of the stimuli or supplied by participants themselves, they were instructed to circle the name that they intended the pronoun to refer to. Examples were given to participants to demonstrate how to do this. This step was taken to improve coding accuracy, as explained below. Participants were not constrained by time to finish the survey.

Following the fragment completion task, L2 participants were administered a translation task and an English proficiency C-test in a separate booklet. In the translation task, they needed to write down the Chinese translations of the experimental stimuli in the sentence-completion experiment (excluding fillers). In the C-test, they were asked to fill in the blanks in three short passages. The translation task was administered after the sentence-completion task to avoid potential influences of the former on sentence-completion performance. Because participants were allowed to take as much time as they needed to finish the completion task and because participants were tested in class, which did have a finite duration, participants were allowed to finish the two additional tasks in their spare time after class and turn in the answer sheet in the next class meeting one week later. They were specifically told that they were not allowed to use dictionaries if they encountered unfamiliar words.

Coding

The data in the sentence-completion experiment were coded independently by the first author and another trained native English speaker naive to the purpose of the study. Coding was done according to the following procedure: Based on participants' sentence continuations, the subject NP in the second clause was coded as referring to either the first antecedent (NP1) or the second antecedent (NP2) in the first clause. Coders were instructed to be conservative so that, as long as there was a possibility of ambiguity, the reference was coded as 'unclear'. For continuations that included a subject pronoun, coders were instructed to rely on the marking made by the participant but verify whether the entity circled by the participant made sense given the rest of the continuation. If the circled entity did not make sense given the rest of the continuation, the response was to be marked 'unclear'. Trials in which no continuation was given, or in which the continuation was nonsense, began with a plural reference or a reference to another entity, showed misunderstanding of the gender of the names, or in which the connective *because* was interpreted as part of *because of*, were also coded as 'unclear'. Table 2 illustrates different types of coded continuations.

The coding agreement rate between the two raters was 93.2%. All disagreements were resolved through discussion between the first author and a third independent native English- speaking coder. Disagreements which could not be resolved were coded as 'unclear'. Overall, there were 3.9% unclear responses in the L1 group ($n = 53$) and 8.9% unclear responses in the L2 group ($n = 103$).

The first author who is a Chinese–English bilingual coded L2 participants' translation data as either 'correct' or 'incorrect' by matching their translation with the intended meanings of the items. Items with missing translations were counted as 'incorrect' as well. Overall, there were 6.4% incorrect translations ($n = 74$, $M = 2$, $SD = 1.61$, range: 0–6).

2.2. Results

All data coded as 'unclear' were excluded from analysis. For the L2 group, the data whose counterparts in the translation task were coded as 'incorrect' were also excluded from analysis. This affected 3.9% of the dataset of the L1 group and 15% of the L2 group. Table 3 presents the mean proportions of NP1 references out of all NP1 and NP2 references from the remaining trials.

We used logit mixed-effects regressions to analyze the data. Logit mixed-effects models are more suitable for analyzing categorical and unbalanced data than ANOVA (Jaeger, 2008). All categorical factors were initially sum-coded to obtain main effects and interactions. Stepwise model comparison was used to estimate the significance of each term, starting with a maximal model containing all individual factors and their interactions. The interaction term was first eliminated. If the elimination did not lead to a significant loss of model fit, each of the individual factors was then removed (Baayen, 2008). If the interaction was significant, the interaction term and all embedded lower

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

Table 2. *Sample Coded Continuations in Experiment 1.*

Codes	Sample continuations
NP1	Joe frightened Luke because *Joe liked to frighten people.*
	Joe frightened Luke because he *was really muscular.*
NP2	Joe frightened Luke because *Luke thought Joe would rat him out.*
	Joe frightened Luke because *he wasn't expecting to see him around the corner*.
Ambiguous	Joe frightened Luke because he *was not paying attention*.
Nonsense	Diana respected Rebecca because she *never know*.
Other/plural entity	Ben telephoned James because *there was a family emergency*.
Gender misunderstanding	Charles cheated George because *George cheated her first*.
Because of	Barbara harmed Tiffany because *of carelessness*.

Note: Participants' continuations were in italics.

Table 3. *Mean Proportions and Standard Deviations of NP1 References by Verb Bias and Prompt Type in L1 and L2 Groups in Experiment 1.*

	L1 Verb Bias		L2 Verb Bias	
Prompt Type	NP1	NP2	NP1	NP2
Free	.78 (.17)	.05 (.08)	.73 (.19)	.04 (.08)
Pronoun	.86 (.17)	.08 (.14)	.70 (.20)	.15 (.20)

Note: Standard deviations are presented in parentheses.

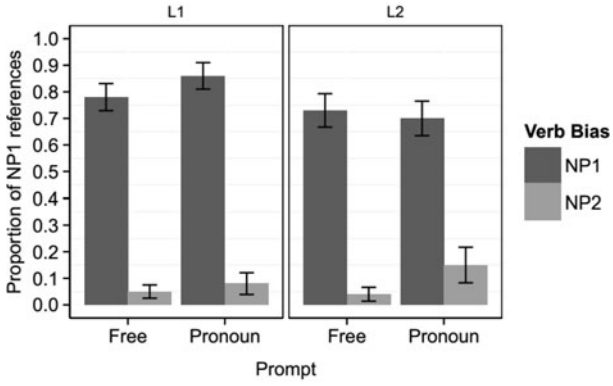

Figure 1. Proportions of NP1 references in Experiment 1. Error bars represent 95% confidence intervals.

level interactions and main effects were kept in the model. Following Barr, Levy, Scheepers, and Tily (2013), all the models contained the random effects of participants and items as well as maximal slopes when appropriate and allowed by the data. The analysis was implemented in R 3.1.0 (R Core team, 2014) using the lme4 package 1.1-7 (Bates, Maechler, Bolker & Walker, 2014), and an alpha level of .05 was used for all statistical tests. The R package lmerTest 2.0-25 (Kuznetsova, Brockhoff & Christensen, 2017) was used to estimate coefficients' p values using the Satterthwaite approximation. For pairwise comparisons, we used the R package LSmeans 2.18 (Lenth, 2016) which estimates *p* values of individual contrasts within the fitted model, using Bonferroni correction.

We performed an analysis on both the L1 and L2 data. A maximal model was fitted with group (L1 vs. L2), verb bias (NP1 vs. NP2-biasing verbs), and prompt type (free vs. pronoun), and all interactions between the three factors as the fixed effects, as well as participants and items as random effects with slopes of verb bias and prompt type for the former and slopes of prompt type and group for the latter. Removing the three-way interaction resulted in a significant loss of model fit, $\chi^2(1) = 8.28$, $p = .004$. The parameter estimates of the full model are reported in Table 4.

An examination of the model's parameters shows three important things: First, there was a main effect of prompt type with more NP1 references following the pronoun prompt than the free prompt, but no two-way interaction between group and prompt type, suggesting that L1 and L2 participants showed similar patterns of coreference in response to different types of prompts. Second, although there was a main effect of verb with more NP1 references following NP1-biasing verbs than NP2- biasing verbs, there was a two-way interaction between group and verb bias, demonstrating that L1 and L2 participants resolved reference differently in continuations following NP1 and NP2- biasing verbs. Third, there was a three-way interaction between group, verb bias, and prompt type, indicating that the effect of group on NP1 reference were modulated by the factors of verb bias and prompt type. These patterns are also illustrated in Figure 1.

Because the presence of the three-way interaction can make the interpretation of the lower order coefficients in the model problematic, we conducted further analyses to better understand the three-way interaction using the simple slope method. To this end we refitted the full model

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

Table 4. *Summary of the Logistic Regression Analysis for Variables Predicting NP1 Reference in L1 and L2 Participants' Continuations in Experiment 1.*

Predictors	B	SE B	z	p
(Intercept)	−.78	.17	−4.54	<.001*
Group (L2)	−.08	.12	−.63	.53
Verb bias (NP2)	−2.31	.18	−13.07	<.001*
Prompt type (pronoun)	.29	.10	2.94	.003*
Group × Verb bias	.26	.13	2.04	.04*
Group × Prompt type	.06	.09	.68	.50
Verb bias × Prompt type	.08	.09	.86	.39
Group × Verb bias × Prompt type	.23	.08	2.84	.004*

Note: All factors were sum-coded to obtain main effects and interactions. The L1 group, NP1 verb, and free prompt were used as the reference levels (value = −1) for the factors of group, verb bias, and prompt type, respectively. Significant effects at a $p \leq .05$ level are marked with a*.

with dummy coding using different reference levels for the factors of verb bias and prompt type. The results were adjusted using a Bonferroni p value correction. In the free prompt conditions, there was no two-way interaction between group and verb bias or an effect of group, demonstrating that L1 and L2 participants had similar re-mention biases in continuations following free prompts, that is, NP1 after NP1- biasing verbs and NP2 after NP2-biasing verbs. However, when the prompt was a pronoun, there was a two-way interaction between group and verb bias, $B = -1.85$, $SE\ B = .56$, $z = -3.31$, $p = .002$. Specifically, although L2 participants were able to distinguish IC biases of the two types of verbs, they produced fewer NP1 references than L1 participants in continuations following NP1 verbs, $B = -.94$, $SE\ B = .35$, $z = -2.70$, $p = .01$, but more NP1 references than L1 speakers in continuations following NP2 verbs, $B = .91$, $SE\ B = .42$, $z = 2.16$, $p = .06$. The latter difference also led to a two-way interaction between group and prompt type when the verbs were NP2-biasing verbs, $B = -1.18$, $SE\ B = .53$, $z = -2.25$, $p = .05$.

In order to determine whether the variance in L2 participants' English proficiency had an effect on the results, we included their C-test scores (centered) in a maximal model regressed to the L2 data only. Model comparisons showed that the C-test score did not contribute significantly to model fit, indicating that the variance in L2 participants' English proficiency did not influence their referential choice.

2.3. Discussion

This experiment investigated how L1 and L2 speakers establish coreference by using the IC information from the context. Results showed that L1 participants made reference choices following IC biases: NP1 after NP1-biasing verbs and NP2 after NP2-biasing verbs. However, the results were also affected by whether the pronoun was present or not. When L1 participants saw a pronoun prompt in the sentence fragment, they continued the sentence with significantly more references to NP1 than when they saw a free prompt. The findings are thus consistent with the Bayesian model (Kehler et al., 2008; Kehler & Rohde, 2013) on the effect of IC bias on native pronoun resolution.

Like L1 speakers, L2 participants applied the differences in verbs' IC biases to the choice of subsequent coreference. Overall, the prompt also influenced their coreference resolution with more NP1 references following the pronoun prompt than the free prompt, indicating that L2 participants were aware of the special relationship between pronouns and subject/first-mention antecedents.

Despite the general similarity in the performance of the two groups, we also observed a three-way interaction among group, prompt type, and verb bias, indicating that the extent to which L2 participants' performance resembled native speakers' performance depends on the types of prompt and verb. When there was a free prompt, L2 participants showed the same extent of re-mention biases as L1 participants, demonstrating that L2 participants had no problems using the IC information to predict the next-mentioned referent. However, when the prompt was a pronoun, L2 participants produced more NP1 references in continuations after NP2-biasing verbs than L1 participants. The discrepancy between the free and pronoun prompt conditions indicates that L2 participants resolved pronouns in different ways from L1 speakers. When the context had an NP2 IC bias, they were more likely to interpret the pronoun as referring to NP1 than L1 speakers. This 'NP1 bias' shown in the L2 data was also found in Cheng and Almor (2017). Interestingly, such 'NP1 bias' was not observed following NP1-biasing verbs in the pronoun condition. Instead, L1

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

speakers produced significantly more NP1 references than L2 speakers. This contrast will be further explored in the General Discussion.

3. Experiment 2

Experiment 2 aimed to test the same hypotheses as in Experiment 1 using IR verbs embedded in resultative discourse contexts. Except for the differences in the items, the design and methods were the same as Experiment 1.

3.1. Method

Participants

New L1 and L2 participants were recruited from the same populations as in Experiment 1. Forty-nine native English speakers participated in the experiment for extra credit. Three participants were eliminated from analysis because one was an early bilingual, one had an old age (78 years), and the third one's responses were not relevant to the task. In the end, the data from 46 native English participants (38 women, M_{age} = 20.3 years, age range: 18–48 years) were analyzed.

Forty-seven Chinese-speaking English learners took part in this experiment for extra credit. Only the data of those who finished both the sentence-completion and the translation tasks were included in the analysis. In the end, 35 participants finished both tasks (34 women, M_{age} = 21.5 years, age range: 20–23 years). These participants started learning English as a foreign language in a school setting at an average age of 9.80 years (age range: 7–14 years) and had learned English for an average of 11.68 years (range: 8–14 years). At the time of testing, none of them had visited English-speaking countries. All of them took the same C-test as in Experiment 1 with an average score of 33.74 (SD = 8.17). Independent samples t-Test showed that there was no significant difference in participants' C-test scores between Experiments 1 and 2, suggesting that L2 participants in the two experiments were at comparable English proficiency levels.

Materials and design

The verbs used in Experiment 2 were IR verbs selected from a norming experiment on the 300 verbs tested in Ferstl et al.'s (2011) IC study. The verbs were embedded in sentence fragments of the form *NP1 verb-ed NP2 and as a result* with the two NPs being English names of different genders. The 300 items were randomly divided into three lists, each consisting of 100 items. To counterbalance the effect of gender, three more lists were prepared by reversing the order of the two names. Another group of native English speakers recruited from the same population (N = 115, 80 women, M_{age} = 20.7 years, age range: 18–34) took part in the norming study on the survey website Qualtrics. They were randomly assigned to each of the six lists and typed continuations to the sentence fragments. Following the same procedure as in Experiment 1, participants' continuations were coded as referring to either NP1 or NP2 by the first author and another trained native English speaker with an inter-rater agreement rate of 95.1%. In order to ensure that the verbs in the current experiment had similar IR biases in learners' native language, a norming study on Chinese verbs was administered to 180 different native Chinese speakers (153 women, M_{age} = 21.3 years, age range: 20–23) in the same way as the one conducted in Experiment 1 except that, in this study, verbs were embedded in sentence fragments of the form *NP1 verb NP2 yinci* "because of that", eliciting a Result coherence relation. Following the same procedure as in Experiment 1, continuations were coded as referring to either NP1 or NP2 by the first author and another trained native Chinese speaker with an agreement rate of 93.6%. In the end, 16 NP1-biasing and 16 NP2-biasing verbs were selected following the same procedure and criteria as in Experiment 1 (see Appendix).

The stimuli were prepared in the same way as in Experiment 1 except that the verbs were embedded in sentence fragments of the type *NP1 verb-ed NP2 and as a result*. We did not use the connective *so as* in the Cheng and Almor (2017) study because the connective *so* may denote other meanings than result (Stevenson, Crawley & Kleinman, 1994). The phrase *as a result*, by contrast, specifically indicates that the coherence relation is Result. The materials for the translation task were prepared in the same way following Experiment 1.

The design was identical to that of Experiment 1.

Procedure

The procedure was identical to that of Experiment 1.

Coding

The data were coded by the first author and another trained coder following the same procedure as in Experiment 1 with a coding agreement rate of 96.2%. Overall, there were 10% unclear responses in the L1 group (n = 143) and 17% unclear responses in the L2 group (n = 187). Furthermore, there were 6% incorrect translations in the L2 group (n = 58, M = 1.66, SD = 1.55, range: 0–6).

3.2. Results

The responses coded as 'unclear' were excluded from analysis. For the L2 group, the data whose counterparts in the translation task were coded as 'incorrect' were also excluded. Data trimming affected 10% of the dataset in the L1 group and 20% in the L2 group. Table 5 presents the mean proportions and standard deviations of NP1 references out of all NP1 and NP2 references.

The data were analyzed in the same manner as in Experiment 1. A maximal model was fitted with group,

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

Table 5. *Mean Proportions and Standard Deviations of NP1 References by Verb Bias and Prompt Type in L1 and L2 Groups in Experiment 2.*

	L1 Verb Bias		L2 Verb Bias	
Prompt Type	NP1	NP2	NP1	NP2
Free	.83 (.19)	.14 (.16)	.85 (.18)	.10 (.13)
Pronoun	.94 (.09)	.34 (.28)	.97 (.07)	.68 (.29)

Note: Standard deviations are presented in parentheses.

Table 6. *Summary of the Logistic Regression Analysis for Variables Predicting NP1 Reference in L1 and L2 Participants' Continuations in Experiment 2.*

Predictors	B	SE B	z	p
(Intercept)	.92	.17	5.54	<.001*
Group (L2)	.35	.14	2.50	.01*
Verb bias (NP2)	−2.00	.15	−13.25	<.001*
Prompt type (pronoun)	1.03	.12	8.41	<.001*
Group × Verb bias	.09	.12	.72	.47
Group × Prompt type	.39	.11	3.48	<.001*
Verb bias × Prompt type	.18	.10	1.69	.09
Group × Verb bias × Prompt type	.22	.70	2.54	.01*

Note: All factors were sum-coded to obtain main effects and interactions. The L1 group, NP1 verb, and free prompt were used as the reference levels (value = −1) for the factors of group, verb bias, and prompt type, respectively. Significant effects at a $p \leq .05$ level are marked with a*.

verb bias, and prompt type, and all interactions between the three factors as the fixed effects, as well as participants and items as random effects with slopes of verb bias and prompt type for the former and slopes of prompt type and group for the latter. Removing the three-way interaction resulted in a significant loss of model fit, $\chi^2(1) = 6.13, p = .01$. The parameter estimates of the full model are reported in Table 6. An examination of the model's parameters reveals that, in addition to the three-way interaction, there was a two-way interaction between group and prompt type. To better understand this two-way interaction, we conducted pairwise comparisons, using a Bonferroni *p*-value adjustment. For both L1 and L2 groups, there were significantly more NP1 references following the pronoun prompt than the free prompt: (L1) $B = 1.29$, $SE\ B = .29$, $z = 4.63, p < .001$; (L2) $B = 2.83$, $SE\ B = .38$, $z = 7.54, p < .001$. However, while there was no significant difference between the two groups in the free prompt condition, L2 participants produced more NP1 references than L1 participants in the pronoun prompt condition, $B = 1.47$, $SE\ B = .41$, $z = 3.57$, $p = .002$.

As in Experiment 1, we carried out additional simple slope analyses to better understand the three-way interaction as illustrated in Figure 2. Similar to Experiment 1, in the free prompt condition, there was no two-way interaction between group and verb bias, or an

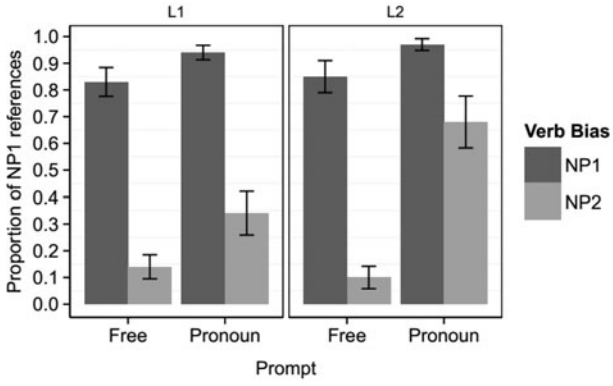

Figure 2. Proportions of NP1 references in Experiment 2. Error bars represent 95% confidence intervals.

effect of group, demonstrating that L1 and L2 participants had similar re-mention biases in continuations following free prompts, that is, NP1 after NP1-biasing verbs and NP2 after NP2-biasing verbs. When the prompt was a pronoun, there was no two-way interaction between group and verb bias. There was no two-way interaction between group and prompt type in the condition of NP1-biasing verbs. However, there was a two-way interaction between group and prompt type when the verbs were NP2-biasing

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

verbs, $B = -2.44$, $SE\ B = .48$, $z = -5.10$, $p < .001$. This interaction reflected that when verbs had an NP2 bias, L2 participants produced significantly more NP1 references than L1 participants in continuations following a pronoun prompt, $B = 2.09$, $SE\ B = .44$, $z = 4.77$, $p < .001$. In these conditions, while the L1 group showed a clear NP2 bias (NP1 percentage = 34%), the L2 group showed an opposite bias towards NP1 (NP1 percentage = 68%).

Finally, we included L2 participants' C-test scores (centered) in a maximal model regressed to the L2 data. Model comparisons showed that, as in Experiment 1, the C-test score did not contribute significantly to model fit, indicating that the variance in L2 participants' English proficiency did not influence their coreference resolution in this experiment.

3.3. Discussion

Experiment 2 investigated the extent to which L2 speakers use the information of IR in coreference resolution. Consistent with previous IR studies (e.g., Au, 1987; Stevenson et al., 1994; Stewart et al., 1998), the results showed that L1 speakers followed IR biases to establish coreference. Like in Experiment 1, the strength of the bias was affected by the referring form with significantly more NP1 references in continuations following a pronoun prompt than a free prompt. The consistent findings of the two experiments indicate that the presence of pronouns has an independent effect on coreference regardless of contextual biases.

As far as the L2 group is concerned, the results showed that L2 participants also produced more continuations with NP1 references following a pronoun prompt than a free prompt. The findings are similar to those in Experiment 1, indicating that, in both the IC and IR contexts, the referring form affected L2 speakers' establishment of coreference such that the presence of a pronoun increased the likelihood of NP1 reference.

As in Experiment 1, we once again observed a three-way interaction between group, verb bias, and prompt type in this experiment. In the free prompt condition, L2 participants showed the same extent of re-mention biases as native speakers, indicating that they were able to derive IR biases from the context and use them to predict which referent would be re-mentioned in the following discourse. However, when the pronoun was present, L2 participants were more likely than L1 participants to resolve the pronoun towards NP1, even though the context had an NP2 bias. Thus, as found in Experiment 1, L2 participants demonstrated a 'NP1 bias' for pronoun interpretation. The 'NP1 bias' in this experiment was strong enough to flip the reference bias choice in the L2 group to an NP1 bias even after contexts with NP2-biasing verbs. This finding is consistent with Cheng and Almor (2017) and helps pinpoint the reason for differences in L2 reference resolution to the stage of the Bayesian integration of priors with the evidence provided by reference form rather than their ability to make predictions.

4. General discussion

This study investigated the extent to which advanced Chinese-speaking L2 English learners rely on IC and IR biases when establishing coreference. According to the Bayesian model (Kehler et al., 2008; Kehler & Rohde, 2013), when establishing coreference between a pronoun and its antecedent, comprehenders rely on 1) contextual information such as IC and IR biases as well as discourse relations to predict which referent will be re-mentioned, and 2) integrating these prior predictions with the probabilistic information provided by the subjecthood/first-mention cue derived from the presence of a pronoun. We tested L2 speakers' performance in these two aspects of processing by manipulating verb bias (NP1-biasing vs. NP2-biasing verbs) and prompt type (free prompt vs. pronoun prompt) in two sentence-completion experiments that focused on IC and IR, respectively. The experiments yielded converging results in line with the hypothesis that L2 speakers are able to form predictions from the context preceding a pronoun that are comparable to those of native speakers but then tend to behave differently when integrating the information provided by the pronoun with their prior predictions about which referent is most likely to be mentioned next.

With respect to forming predictions from the context preceding the pronoun, our experiments found that, in the free prompt conditions, there were no significant differences between the L1 and L2 groups. Like L1 participants, in both the IC (Experiment 1) and IR (Experiment 2) contexts, L2 participants produced more NP1 re-mentions following NP1-biasing verbs and more NP2 re-mentions following NP2-biasing verbs. The results, therefore, indicate that L2 speakers are able to derive IC and IR biases from the context and use them in a native-like way to generate predictions about the referent to be re-mentioned.

These results contradict those of Grüter et al. (2014, 2017) who found that advanced L1- Japanese and L1-Korean L2 English learners were not able to use discourse information to generate native-like re-mention biases. One possible reason for the difference between our study and theirs is related to the different linguistic phenomena tested in both studies. To create distinct kinds of discourse contexts, Grüter et al. manipulated verb aspect whereas we manipulated verb meaning. Aspect is a well-known area of difficulty for L2 learners (Bardovi-Harlig, 2000). It is also linguistically encoded in different ways in Grüter et al.'s L2 participants' native languages (Korean and Japanese) and the target language of English (Shirai, 1998). By contrast, IC and IR biases are argued to be universal

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

biases involving simple interpersonal verbs (Hartshorne et al., 2013), which may be comparatively easier to master, especially when learners' L1 and L2 have equivalent lexical items, as was the case in our study. Therefore, it is likely that L2 speakers may find it more difficult to use aspect than IC and IR biases to establish coreference in discourse. Although factors related to participants' L2 proficiency could also account for the different findings, the fact that, in both Grüter et al.'s and our studies, participants demonstrated native-like knowledge of the phenomena tested, speaks against such explanations.

Overall, the free prompt results in the present study provide important insights into the nature of predictive processing in L2. Specifically, L2 speakers' native-like performance in the free prompt condition indicates that they are able to use verb-bias and discourse-coherence information to generate native-like coreference expectations. This is incompatible with the RAGE hypothesis (Grüter et al., 2014, 2017), which assumes a general reduced ability for L2 speakers to generate expectations.

In the pronoun prompt condition, L2 participants patterned with L1 participants in producing more continuations with NP1 references than in the free prompt condition. This finding is consistent with Grüter et al.'s (2014, 2017) study, which also found that the presence of a pronoun increased L2 speakers' references to NP1 despite their difficulty in using the information of aspect to establish coreference. Given that pronouns are the preferred referential form for subject/first-mention antecedents (e.g., Almor & Nair, 2007; Ariel, 1990; Garrod & Sanford, 1982; Givón, 1987; Gordon et al., 1993; Gundel et al., 1993), the current study provides further evidence that L2 speakers are sensitive to the subjecthood/first-mention cue provided by the presence of a pronoun.

However, when the context had an NP2 bias, L2 participants were more likely to interpret the pronoun as referring to NP1 than L1 participants. These findings replicate Cheng and Almor (2017), who also observed an 'NP1 bias' in L2 speakers' pronoun interpretation in IC and IR contexts. In that study, however, SE verbs were used, which might have posed a general problem for the L1-Chinese participants because SE verbs are rare in Chinese and difficult to acquire for Chinese-speaking English learners (Juffs, 1996; Zhang, 2003). In the present study, we used a more diverse set of verbs that are shared in both Chinese and English. The fact that we replicated what appears like an enhanced L2 'NP1 bias' in this study indicates that it is not due to the specific verbs chosen, but rather reflects a general stronger-than-native tendency to resolve the pronoun to the subject or first-mentioned referent by L2 speakers in the NP2-biasing context.

The remaining questions are why there is a stronger 'NP1 bias' in L2 speakers' pronoun interpretation following NP2-biasing verbs, and whether this bias can be explained in the Bayesian framework of Kehler et al. (2008) and Kehler and Rohde (2013). As the results of our free prompt conditions show, L2 speakers understand IC and IR biases and apply them to the formation of re-mention predictions. Therefore, according to the Bayesian framework of coreference resolution, the source of L2 speakers' 'NP1 bias' in the NP2-biasing verb conditions must reflect the Bayesian updating of context-dependent prior predictions about the next-mentioned referent with the evidence provided by the pronoun to derive the posterior probabilities for the potential antecedents to be the actual antecedent of the pronoun. L2 speakers' stronger posterior preference for NP1 in the NP2-biasing context could, therefore, reflect (1) their assigning weaker prior context-based predictions to each of the antecedents than L1 speakers, or (2) their assigning a higher probability than L1 speakers to a pronoun to be used as the referential form for subject or first-mentioned antecedents. A third alternative is that the basic Bayesian framework cannot account for our results, perhaps because L2 speakers assign a greater weight to the subjecthood/first-mention cue provided by the pronoun than L1 speakers when integrating it with prior context-based expectations. While such weighing can be added to the Bayesian framework, it will require modifying this account. As each of these (not mutually exclusive) alternatives has important implications for theories of language processing in L2, we consider each of them in turn.

According to the first explanation, L2 speakers may form weaker prior probabilities for which referent will be re-mentioned, thus allowing the evidence provided by the pronoun to have a stronger effect on their posterior probabilities of which is the most likely antecedent. This explanation is not likely given the native-like performance of L2 speakers in the absence of a pronoun.

According to the second explanation, L2 speakers assign a stronger-than-native likelihood to pronouns to be used as the referential form for subject or first-mentioned antecedents following NP2-biasing verbs. Under the assumption that participants' comprehension and production preferences resemble each other, this means that L2 participants may show a stronger production bias to use a pronoun (as opposed to other forms of reference) for subject or first-mentioned referents in the NP2-biasing contexts but not in the NP1-biasing contexts. To determine if this is the case, we examined participants' choice of referring expressions in the free prompt condition, in which they were free to choose any form (e.g., pronoun, repeated name) to refer to a referent. We coded their binary choice between pronouns and names for the subject of the second clause referring to either NP1 or NP2 in the context clause and calculated the percentage of pronominalization with regard to verb bias

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

Table 7. *Differences in Proportions of Subject Pronouns in Continuations in the Free Prompt Condition and Differences in Proportions of NP1 References in Experiments 1 and 2.*

			Antecedent Pronominalizing Rate			Proportion of NP1 References			Estimated Posterior
Experiment	Verb Bias	Group	NP1	NP2	NP1−NP2	Free Prompt	Pronoun Prompt	Pronoun−Free	*p(NP1 | pronoun)*
1	NP1	L1	.89 (216/242)	.39 (26/67)	.50	.78	.86	.08	.89
		L2	.61 (99/161)	.30 (18/60)	.31	.73	.70	−.03	.84
	NP2	L1	.80 (12/15)	.57 (173/300)	.23	.05	.08	.03	.06
		L2	.80 (8/10)	.26 (66/251)	.54	.04	.15	.11	.11
2	NP1	L1	.75 (116/155)	.05 (2/40)	.70	.83	.94	.11	1.00
		L2	.51 (82/161)	.04 (1/25)	.47	.85	.97	.12	.97
	NP2	L1	.43 (12/28)	.16 (42/257)	.27	.14	.34	.20	.32
		L2	.62 (13/21)	.03 (5/198)	.59	.10	.68	.58	.75

Note: The formula within parentheses represent the frequency of pronouns divided by the frequency of pronouns and names. The numbers in the columns of Proportion of NP1 References are copied from Tables 3 and 5.

and referent for both L1 and L2 participants. The results are presented in Table 7. [3]

As noted above, we assume that the rate at which our participants produce pronouns is related to how they estimate someone else would use a pronoun. Under the Bayesian approach, this assumption entails that participants' own pronoun production rates would be related to their pronoun interpretation performance in the pronoun prompt condition. To assess how strong a signal the pronoun prompt provides to our participants for deciding between the two possible referents, we focus on the differences between pronominalization rates of NP1 and NP2 references in each condition. For example, if participants pronominalize the same proportion of their NP1 and NP2 continuations, encountering a pronoun in comprehension would not be informative for choosing the referent, but if instead participants pronominalize 89% of their NP1 responses but only 39% of their NP2 responses, encountering a pronoun in comprehension should provide a strong cue in favor of NP1 being the antecedent. Table 7 shows these pronominalization-rate differences.

To establish whether these differences in pronoun production biases can account for our results, we also calculated the difference between the proportions of NP1 references in the pronoun and free prompt conditions for each verb bias and group combination in the two experiments. These differences, which are also shown in Table 7, reflect the impact of encountering a pronoun on participants' choice of referent. Our aim is to compare L1 and L2 groups' pronoun production biases and see if the conditions where one group exhibits a stronger bias for pronominalizing NP1 referents than the other group are the same conditions in which that group shows a bigger difference in the proportion of NP1 references between the free and pronoun prompt conditions. In other words, we now compare the patterns of Columns 6 and 9 in Table 7 to determine whether the two types of differences follow similar patterns in the two groups.

In Experiment 1, for NP1-biasing verbs, L1 speakers' pronoun production patterns show that they had a stronger pronoun-NP1 connection in this condition than L2 speakers (.50 vs. .31). This is mirrored by the coreference patterns in how the two groups interpreted pronouns following these verbs (.08 vs. -.03). For the NP2-biasing verbs in Experiment 1, it is the L2 speakers' production that shows a greater pronoun-NP1 connection than the L1 speakers' production (.54 vs. .23). Here too, this is mirrored by the coreference patterns: The L2 group was more strongly affected by the pronoun cue than the L1 group (.11 vs. .03) in their interpretation of pronouns in the NP2-biasing context. Indeed, the three-way interaction reflects this: In the free prompt, there was no group-by-verb-bias interaction (meaning L1 and L2 speakers responded to verb bias similarly to generate their priors), but in the pronoun prompt, there were differences by group and verb bias.

In Experiment 2, for NP1-biasing verbs, similar to Experiment 1, L1 speakers' pronoun production patterns show that they had a stronger pronoun-NP1 connection in this condition than L2 speakers (.70 vs. .47). However, in this case, this is not mirrored by the coreference patterns, which are comparable in the two groups (.11 vs. .12). Note, yet, that this likely reflects the fact that, in this condition, NP1 coreference choices are almost at ceiling for both groups in the pronoun prompt condition (.94 vs. .97), which could obscure any potential difference between the two groups. For NP2-biasing verbs, again, similar to Experiment 1, L1 speakers showed a weaker pronoun-NP1 connection than L2 speakers (.27 vs. 59). As in Experiment 1, the coreference patterns mirror this pattern in that L1 speakers were less affected by the pronoun cue than L2 speakers (.20 vs. .58). In this experiment too, the three-way interaction reflects this, although, in the pronoun prompt, the group differences only emerge for NP2-biasing verbs and not for NP1-biasing verbs (due to ceiling effects).

Thus, we can conclude that both L1 and L2 participants' interpretations of pronouns, as reflected in their continuations following the pronoun prompt, conform to the Bayesian principles in that they follow the same patterns shown by pronoun productions following the free prompt. In this sense, our results are in line with the predictions of the Bayesian framework of Kehler et al. (2008) and Kehler and Rohde (2013) and are consistent with the second explanation mentioned earlier.

As a final test of the Bayesian framework, we calculated for each experiment the posterior probabilities of NP1 references in the pronoun prompt condition (reflecting the final outcome of interpreting the pronouns). The proportion of NP1 references in the free prompt condition was used as an estimate of the prior *p(referent=NP1)*, the pronominalization rate of NP1 responses in the free prompt condition was used as an estimate of *p(pronoun | NP1)*, and the pronominalization rate in the free prompt condition for both NP1 and NP2 responses was used as an estimate of *p(pronoun)*. Formula (4) shows the calculation, and the rightmost column in Table 7 shows the results of this estimation for each verb bias and group combination. Formula (5) illustrates the calculation for the first row in Table 7 (Experiment 1, NP1 biasing verbs, L1 speakers).

$$(4)\quad p(NP1 \mid Pronoun) = \frac{p(NP1) \times p(pronoun \mid NP1)}{p(pronoun)}$$

$$(5)\quad p(NP1 \mid pronoun) = \frac{.78 \times .89}{(216 + 26)/(242 + 67)} = .89$$

As can be seen in Table 7, the posterior probabilities estimated on the basis of the free prompt condition

[3] We are grateful to an anonymous reviewer for suggesting the table and related analysis and discussion.

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

match quite well the actual proportion of NP1 references observed in the pronoun prompt condition. This reinforces the validity of the Bayesian framework to describe the performance of both our L1 and L2 participants, and furthermore allows us to attribute the differences in their pronoun interpretation performance to differences in their beliefs about pronoun use, as reflected in their own pronoun production data, rather than to the predictions they generate prior to encountering the pronoun, as reflected in their re-mention biases in the free prompt condition, or the integration of these predictions with the evidence provided by the pronouns.

Another issue concerning the stronger 'NP1 bias' in L2 speakers' pronoun interpretation following NP2-biasing verbs is whether such bias reflects a subject preference or a first-mention preference for pronouns, given that, in our sentence fragments, the referents in the subject position were also the first-mentioned entities in the sentence. Although previous studies on L1 pronoun resolution have identified both grammatical role and order of mention as important factors in pronoun resolution (e.g., Järvikivi et al., 2005), the respective role of each factor is still a matter of debate. Some studies (e.g., Gordon & Hendrick, 1998; Gordon et al., 1993; Gordon, Ledoux & Yang, 1999) claim that referents in the subject position are more accessible than referents in other syntactic positions and thus the preferred antecedents for subsequent pronouns. By contrast, others (e.g., Gernsbacher, 1990; Gernsbacher & Hargreaves, 1988) hold that order of mention is more important than syntactic structure in coreference processing, arguing that first-mentioned referents are retrieved more easily than later-mentioned referents and are thus more likely to be interpreted as antecedents of subsequent pronouns. One problem in most of these studies is that, similar to our study, grammatical roles and order of mention are confounded, as the subject is usually the first-mentioned entity in English. A recent study in Finnish, a language that allows both SVO and OVS structures and therefore makes it possible to disentangle these two factors, showed an effect of first-mention but not of subjecthood on participants' interpretation of pronouns in IC contexts (Järvikivi, Van Gompel & Hyönä, 2017). However, another study in Chinese, a language that also allows a relatively free word order, found that order of mention had no effect on pronoun resolution in Chinese (Xu, 2015). Given these contradictory findings in the L1 literature, and the fact that, in our study, grammatical subjects were also the first mentioned antecedents, we cannot say whether our L2 speakers' 'NP1 bias' reflects a preference for the subject or the first-mention entity to be the referent of the pronoun. We thus leave the resolution of this issue for future research.

In summary, by investigating advanced Chinese-speaking English learners' sensitivity to IC and IR biases in making re-mention decisions and resolving pronominal reference, this study furthers our understanding of how L2 speakers establish coreference in discourse. In both the IC and IR contexts, L2 participants showed native-like re-mention biases. This indicates that L2 speakers are able to generate native-like predictions about the next-mentioned referent based on discourse-level information. However, unlike native speakers, L2 participants exhibited an 'NP1 bias' in pronoun resolution by producing more NP1 references following the NP2-biasing context than native speakers. A close inspection of pronominalization rates in the free prompt conditions under the Bayesian framework suggests that this reflects differences between the groups in their beliefs about the likelihood of pronoun use in different conditions. Specifically, L2 participants show a weaker association between pronouns and NP1 referents than L1 speakers following NP1 biasing verbs, and a stronger association between pronouns and NP1 referents than L1 speakers following NP2 biasing verbs. Future research will have to explore the reasons for this difference and establish whether it may reflect influences of L1 or other factors. More generally, this work highlights the utility of using a Bayesian approach in L2 research as a means for capturing and explaining what might otherwise be complex findings. This is helpful for identifying specific factors that have a probabilistic effect on L2 processing.

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

Appendix. *Mean Proportions of NP1 References for Verbs Used in Experiments 1&2*

		L1		L2	
Verb bias	Verbs	Free	Pronoun	Free	Pronoun
		Experiment 1			
NP1	apologize to	.95	.95	.94	.94
	telephone	.75	.90	1.00	.94
	lie to	.84	.00	.75	1.00
	betray	.72	.71	.67	1.00
	confess to	.00	.81	.88	.94
	cheat	.86	.90	.70	.71
	harass	.35	.63	.60	.40
	harm	.52	.76	.80	.53
	annoy	.90	.86	.50	.41
	anger	.86	1.00	.35	.41
	attract	.83	1.00	.94	.94
	bother	.90	.79	.72	.81
	frighten	.89	.90	.69	.63
	please	.71	.90	.80	.59
	astonish	.79	.86	.71	.89
	stimulate	.68	.70	.15	.22
NP2	congratulate	.00	.00	.00	.18
	thank	.05	.10	.14	.20
	praise	.05	.10	.00	.12
	punish	.00	.00	.00	.12
	reward	.00	.14	.00	.05
	condemn	.05	.05	.06	.18
	correct	.15	.30	.13	.32
	scold	.10	.10	.06	.21
	envy	.05	.00	.05	.13
	admire	.00	.00	.00	.06
	respect	.05	.00	.00	.20
	fear	.10	.00	.06	.19
	pity	.06	.14	.12	.05
	distrust	.05	.19	.06	.19
	worry about	.00	.10	.00	.11
	dislike	.05	.00	.12	.20
		Experiment 2			
NP1	kill	1.00	1.00	.93	1.00
	dislike	.83	.96	1.00	1.00
	envy	.95	.91	1.00	1.00
	admire	.84	.96	.80	1.00
	fear	.91	1.00	.79	1.00
	pity	.90	.91	.50	.94
	trust	.76	.91	.50	.90

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

Appendix. *Continued*

		L1		L2	
Verb bias	Verbs	Free	Pronoun	Free	Pronoun
	worry about	.87	1.00	.87	.95
	dream about	1.00	.83	.95	.93
	appreciate	.81	1.00	.83	1.00
	love	.75	.95	1.00	.92
	fancy	.77	.91	.45	.92
	miss	.91	1.00	.93	1.00
	like	.71	.87	.90	1.00
	worship	.43	.83	1.00	.92
	hate	.79	1.00	1.00	1.00
NP2	guide	.14	.22	.08	.65
	warn	.04	.30	.00	.53
	interrupt	.23	.64	.18	.70
	support	.14	.55	.07	.83
	punish	.05	.22	.17	.79
	criticize	.11	.30	.00	.57
	accuse	.06	.30	.11	.69
	praise	.20	.30	.00	.54
	attract	.18	.64	.29	.80
	annoy	.16	.39	.11	.85
	bother	.14	.48	.07	.72
	frighten	.00	.26	.20	.76
	astonish	.53	.35	.20	.79
	stimulate	.13	.18	.00	.64
	encourage	.00	.09	.13	.57
	inspire	.20	.22	.11	.46

References

Almor, A., & Nair, V. A. (2007). The form of referential expressions in discourse. *Language and Linguistics Compass, 1,* 84–99.

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

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

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,* 104–122.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning and use*. Oxford: Blackwell.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4* (R package version 1.1-7). http://CRAN.R-project.org/package=lme4.

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

Cheng, W., & Almor, A. (2015). The effect of lexical and periphrastic causatives on pronoun resolution: Evidence from Chinese. Poster session presented at the 28th CUNY Conference on Human Sentence Processing, Los Angeles, March, 2015.

Cheng, W., & Almor, A. (2017). The effect of implicit causality and consequentiality on nonnative pronoun resolution. *Applied Psycholinguistics*, *38,* 1–26.

Crawley, R., Stevenson, R., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research, 4,* 245–264.

Crinean, M., & Garnham, A. (2006). Implicit causality, implicit consequentiality and semantic roles. *Language and Cognitive Processes, 21,* 636–648.

Ehrlich, K. (1980). Comprehension of pronouns. *Quarterly Journal of Experimental Psychology, 32,* 247–255.

https://doi.org/10.1017/S136672891800055X Published online by Cambridge University Press

Ferstl, E. C., Garnham, A., & Manouilidou, C. (2011). Implicit causality bias in English: A corpus of 300 verbs. *Behavioral Research Methods, 43,* 124–135.

Frederiksen, J. (1981). Understanding anaphora: Rules used by readers in assigning pronominal referents. *Discourse Processes, 4,* 323–347.

Garrod, S., & Sanford, A. J. (1982). The mental representation of discourse in a focused memory system: Implications for the interpretation of anaphoric noun-phrases. *Journal of Semantics, 1,* 21–41.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry, 5,* 459–464.

Garvey, C., Caramazza, A., & Yates, J. (1976). Factors influencing assignment of pronoun antecedents. *Cognition, 3,* 227–243.

Gernsbacher, M.A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Gernsbacher, M. A., & Hargreaves, D. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language, 27,* 699–717.

Givón, T. (1987). *On understanding grammar*. New York, NY: Academic Press.

Givón, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics, 30,* 5–55

Givón, T. (1995). *Functionalism and grammar*. Philadelphia: John Benjamins.

Goikoetxea, E., Pascual, G., & Acha, J. (2008). Normative study of the implicit causality of 100 interpersonal verbs in Spanish. *Behavior Research Methods, 40,* 760–772.

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

Gordon, P. C., & Hendrick, R. (1998). The representation and processing of coreference in discourse. *Cognitive Science, 22,* 389–424.

Gordon, P.C., Hendrick, R., Ledoux, K., & Yang, C.L. (1999). Processing of reference and the structure of language: An analysis of complex noun phrases. *Language and Cognitive Processes, 14,* 353–379.

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

Grüter, T., Rohde, H., & Schafer, A. (2014). The Role of Discourse-Level Expectations in Non-Native Speakers' Referential Choices. In W. Orman & M. J. Valleau (Eds.), *BUCLD 38: Proceedings of the 38th annual Boston University Conference on Language Development* (pp. 179–191). Somerville, MA: Cascadilla Proceedings Project.

Grüter, T., Rohde, H., & Schafer, A. J. (2017). Coreference and discourse coherence in L2: The roles of grammatical aspect and referential form. *Linguistic Approaches to Bilingualism, 7*, 199–229.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69,* 274–307.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hartshorne, J. K., & Snedeker, J. (2013). Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. *Language and Cognitive Processes, 28,* 1474–1508.

Hartshorne, J. K., Sudo, Y., & Uruwashi, M. (2013). Are implicit causality pronoun resolution biases consistent across languages and cultures? *Experimental Psychology, 60,* 179–196.

Jaeger, F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–46.

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

Järvikivi, J., Van Gompel, R. P. G., & Hyönä, J. (2017). The interplay of implicit causality, structural heuristics, and anaphor type in ambiguous pronoun resolution. *Journal of Psycholinguistic Research, 46,* 525–550.

Jiao, J., & Zhang, B. (2005). The effect of the implicit causality of Chinese verbs on pronoun resolution. *Psychological Science, 28,* 1082–1085.

Jin, Y., & Fan, J. (2011). Test for English Majors (TEM) in China. *Language Testing, 28,* 589–596.

Juffs, A. (1996). Semantics-syntax correspondences in second language acquisition. *Second Language Research, 12,* 177–121.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4,* 257–282.

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Language and Linguistics Compass, 2,* 647–670.

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

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

Koornneef, A. W., & Sanders, T. J. M. (2013). Establishing coherence relations in discourse: The influence of implicit causality and connectives on pronoun resolution. *Language and Cognitive Processes, 28,* 1169–1206.

Koornneef, A. W., & Van Berkum, J. J. A. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language, 54,* 445–465.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(*13*), pp. 1–26. doi: 10.18637/jss.v082.i13.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31,* 32–59.

Lenth, R. V. (2016). Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software, 69,* 1–33. doi:10.18637/jss.v069.i01

Liu, J. (2012). Aligning TEM-4 with the CEFR (Unpublished master's thesis). Henan Normal University, China.

Liu, M. C. (2016). Emotion in lexicon and grammar: Lexical-constructional interface of Mandarin emotional predicates. *Lingua Sinica*, *2*(*4*). doi:10.1186/s40655-016-0013-0

Liu, R., & Nicol, J. (2010). Online processing of anaphora by advanced English learners. In M. T. Prior, S. K. Lee (Eds.), *Selected proceedings of the 2008 second language research forum* (pp. 150–165). Somerville, MA: Cascadilla Proceedings Project.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in psychology*, *4,* 226.

Martínez Baztán, A. (2008). *La evaluacion oral: una equivalencia entre las guidelines de ACTFL yalgunas escalas del MCER* (Unpublished doctoral dissertation). Retrieved from http://hera.ugr.es/tesisugr/17457853.pdf

Miao, X. (1996). A study of semantic and grammatical factors influencing pronoun processing. *Acta Psychological Sinica, 28,* 352–358.

Miao, X., & Song, Z. (1995). The influence of verb meaning and sentence grammar on pronoun processing. *Psychological Science, 18,* 197–200.

Pickering, M. J., & Majid, A. (2007). What are implicit causality and implicit consequentiality? *Language and Cognitive Processes, 22,* 780–788.

R Core Team. (2014). R: A language and environment for statistical computing [Software]. Available from https://www.R-project.org/

Roberts, L., Gullberg, M., & Indefrey, P. (2008). Online pronoun resolution in L2 discourse: L1 influence and general learner effects. *Studies in Second Language Acquisition, 30,* 333–357.

Schulz, B. (2006). *Wh-scope marking in English interlanguage grammars: Transfer and processing effects on the second language acquisition of complex questions* (Unpublished doctoral dissertation). University of Hawaii, Manoah.

Shirai, Y. (1998). Where the progressive and the resultative meet: Imperfective aspect in Japanese, Chinese, Korean and English. *Studies in Language, 22,* 661–692.

Sorace, A., & Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. *Second Language Research, 22,* 339–368.

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

Stevenson, R., Knott, A., Oberlander, J., & McDonald, S. (2000). Interpreting pronouns and connectives: Interactions among focusing, thematic roles and coherence relations. *Language and Cognitive Processes, 15,* 225–262.

Stewart, A. J., & Pickering, M. J., & Sanford, A. J. (1998). Implicit consequentiality. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th annual conference of the cognitive science society* (pp. 1031–1036). Mahwah, NJ: Erlbaum.

Sun, Y., Shu, H., Zhou, X., & Zheng, X. (2001). The effect of implicit verb causality on pronoun processing. *Psychological Science, 24,* 39–41.

Tang, J., Pritchard, N., & Shi, L. (2012). Calibrating English language courses with major international and national EFL tests via vocabulary range. *Chinese Journal of Applied Linguistics, 35,* 24–43.

Xu, X. (2015). The influence of information status on pronoun resolution in Mandarin Chinese: evidence from ERPs. *Frontiers in Psychology, 6,* 873. http://doi.org/10.3389/fpsyg.2015.00873

Zhang, J. (2003). *The acquisition of English psych predicates by Chinese-speaking learners* (Unpublished doctoral dissertation). Guangdong University of Foreign Studies, China.