Accepted for publication in *Applied Psycholinguistics* (9/28/2020)

Title: **Limits on expectation-based processing: Use of grammatical aspect for co-reference in L2**

Short title: Limits on expectation-based processing

Authors: Theres Grüter[a] (theres@hawaii.edu; corresponding author)

Hannah Rohde[b] (hannah.rohde@ed.ac.uk)

Affiliations: [a] University of Hawai‘i at Mānoa

[b] University of Edinburgh

Keywords: prediction, expectations, reference, aspect, eye-tracking

Abstract

This study examines the use of discourse-level information to create expectations about reference in real-time processing, testing whether patterns previously observed among native speakers of English generalize to non-native speakers. Findings from a visual world eye-tracking experiment show that native (L1, *N*=53) but not non-native (L2, *N*=52) listeners' proactive coreference expectations are modulated by grammatical aspect in transfer-of-possession events. Results from an offline judgment task show these L2 participants did not differ from L1 speakers in their interpretation of aspect marking on transfer-of-possession predicates in English, indicating it is not lack of linguistic knowledge but utilization of this knowledge in real-time processing that distinguishes the groups. English proficiency, although varying substantially within the L2 group, did not modulate L2 listeners' use of grammatical aspect for reference processing. These findings contribute to the broader endeavor of delineating the role of prediction in human language processing in general, and in the processing of discourse-level information among L2 users in particular.

Trying to anticipate what will happen next is a behavior that we seem unable not to engage in. This has led to the characterization of human brains as prediction "engines" or "machines" (Clark, 2013; Van Berkum, 2010). The recognition that prediction plays an important, perhaps a fundamental, role in human cognition in general, and in language processing in particular, has transformed the field of psycholinguistics over the past decades, shifting the focus from comprehension as a result of the incremental integration of bottom-up information, to language processing as a proactive mechanism driven in large part by top-down expectations (Ferreira & Chantavarin, 2018). What remains under discussion and debate is just how ubiquitous these proactive mechanisms are, and how they might vary depending on the specific circumstances of language use (e.g., Huettig & Guerra, 2019; Huettig & Mani, 2015; Kuperberg & Jaeger, 2016).

One approach is to ask about the extent to which prediction generalizes beyond traditional groups in psycholinguistic experiments, that is, healthy, native-speaking, young adults pursuing a college education. As Huettig (2015, p. 130) concluded in a critical review of the role of prediction in language processing, "[t]he study of prediction... has a particularly strong need for more diverse participant populations" (p. 130; see also Arnett, 2008; Henrich et al., 2010). Indeed, a number of recent studies have shown substantial variability in the extent to which different types of native language users, such children, non-student and older adults, engage in predictive processing (see Huettig, 2015; Pickering & Gambi, 2018). At the same time, a flourishing but largely separate literature has investigated predictive processing among second language users (for review see Kaan, 2014). On the assumption that use of a non-native language (henceforth: L2) is normal human behavior, characteristic of the majority of language

users worldwide, it strikes us as surprising that these literatures have not been more strongly interconnected. In the belief that L2 processing constitutes one of many variations of human language processing, we situate the present study on expectation-based processing of reference among adult L2 users within the wider context of exploring variability between diverse participant populations.

More specifically, this paper aims to contribute to this broader endeavor by investigating the extent to which a particular effect attributed to expectation-based processing previously reported among college-educated native speakers of English generalizes to an equal-sized sample of college-educated L2 speakers of English. We focus on anticipatory processing involving discourse-level expectations, a domain that relative to prediction at the level of lexical and morphosyntactic processing remains underexplored, especially in the context of L2 processing. The remainder of this paper is organized as follows. After presenting a brief overview of the literature on prediction in native language (L1) processing among language users other than college-aged native speakers, we review relevant previous work on expectation-based (L1 and L2) processing at a discourse level and lay out our research questions. We then report findings from a visual world eye-tracking experiment that was designed to capture proactive coreference expectations. We probe comprehenders' expectations prior to the encounter of a disambiguating referential expression, thereby allowing for empirical evidence of prediction in the strict sense of prediction defined by Pickering and Gambi (2018, p. 1005), namely that the effect is observed *prior to* disambiguating information in the input. Findings from L1 speakers of English were previously reported in Grüter et al. (2018), and showed significant effects indicative of proactive coreference expectations

(discussed in more detail below). Here we present results from an equal-sized sample of L2 speakers of English ($N = 52$), with results showing (i) that the same factor that drove proactive coreference expectations in the L1 group does not appear to do so in the L2 group and (ii) that L2 proficiency does not modulate the pattern of effects. We discuss the findings in light of the broader question of the generalizability of expectation-based processing to a wider population of language users, and to L2 users in particular.

Prediction among language users other than native-speaking college students

The evidence for proactive language processing in healthy adult L1 speakers includes examples of prediction at just about all levels of linguistic representation—phonological form, morpho-syntax, structural dependencies, lexical semantics, up to the discourse level. For recent reviews, including critical discussion of proposed mechanisms and limitations of prediction in L1 processing, the reader is referred to Huettig (2015) and Pickering and Gambi (2018). Prediction of upcoming words based on lexical-semantic and morphosyntactic cues has also been shown in children as young as two years old (e.g., Lew-Williams & Fernald, 2007; Mani & Huettig, 2012; but see Gambi et al., 2018, for later emergence of form-related prediction), with evidence for increasing engagement in predictive processing with increasing age and vocabulary knowledge (Fernald et al., 2008; Borovsky et al., 2012). This evidence suggests that prediction is part of human language behavior from early on in development, and that 'language prediction skills' are honed in tandem with language development more generally. Potential causal relations in this tandem development during childhood meanwhile remain a matter of on-going debate (e.g., Huettig & Mani, 2015).

Predictive processing has also been investigated in adult L1 speakers who are not college students. Federmeier and colleagues have presented evidence from a series of studies suggesting that healthy adults in their 60s and above engage less in prediction than younger adults (Federmeier et al., 2010; Wlotko & Federmeier, 2012). More recently, Huettig and Janse (2016; see also Huettig & Pickering, 2019) have argued for the opposite, namely that older adults may rely *more* on prediction, based on findings that age correlates positively with predictive behavior once potentially confounding factors, such as decreased working memory and processing speed, are controlled for. Another relevant factor appears to be L1 literacy and reading experience. Reduced effects of prediction, compared to the typical college student baseline, have been reported in both adults with low literacy (Mishra et al., 2012) and adults with dyslexia (Huettig & Brouwer, 2015).

The body of research reviewed so far has shown that even among L1 users, there is substantial variability in the extent to which prediction contributes to language processing. Factors such as age, literacy, and education appear to matter and are typically framed in terms of *individual differences* among members of the wider population of speakers of the language. Research that has looked at the role of prediction in L2 processing, on the other hand, has typically framed the difference between L1 and L2 speakers as a categorical contrast between two different populations, rather than treating 'L1 vs L2 user status' as an individual difference variable among the population of users of the language. Within this general framing, the dominant research question has been whether or not L2 users are able to perform like L1 users, and what individual difference factors *among L2 users* lead to 'more native-like' performance. We return to this issue in

the Discussion, where we suggest an alternative perspective on the conceptualization of L1-L2 comparisons. For detailed reviews of prediction in L2 processing, the reader is referred to Kaan (2014), and for a more recent perspective, Schlenter (2019). In sum, findings from this literature have been variable, with some studies reporting effects of prediction in L2 similar to those observed in L1 control groups (e.g., Dijkgraaf et al., 2017; Foucart et al., 2014), while others report no or reduced prediction effects among L2 speakers (e.g., Grüter et al., 2012; Kaan et al., 2010; Martin et al., 2013; Mitsugi & MacWhinney, 2016).

Individual difference factors that might modulate L2 learners' engagement in prediction have been a key focus in recent work on L2 processing (see Kaan, 2014, for a programmatic review), with overall L2 proficiency perhaps the most frequently mentioned. While a few earlier studies reported significant modulation of a predictive effect by overall L2 proficiency (e.g., Chambers & Cooke, 2009), more recent work specifically investigating this factor has produced remarkably little evidence in support of proficiency modulating predictive processing. Looking at prediction based on lexical-semantic cues, neither Dijkgraaf et al. (2017) nor Ito et al. (2018) observed significant effects of proficiency. Similarly, Kim (2018) observed no effects of proficiency, measured through three independent tasks, on L2 learners' use of implicit causality to preactivate upcoming referents. In the domain of morphosyntax, neither Hopp (2015) nor Mitsugi (2018), focusing on case marking in L2 German and numeral classifiers in L2 Japanese respectively, found modulation of predictive effects by proficiency. Notably, both of these studies reported significant effects of proficiency on later processes of information integration, indicating that the proficiency measures employed were able to

capture relevant variability between participants; yet this variability did not modulate engagement in prediction. In order to further examine the extent to which overall L2 proficiency contributes to expectation-based processing at a discourse level, the present study included independently measured L2 proficiency as a potential predictor in the analysis.

The observation of generally more variable findings regarding prediction among L2 speakers has led to the proposal that L2 speakers may have "reduced ability to generate expectations" during language processing, also known as the RAGE hypothesis (Grüter et al., 2014, 2017; see also Kaan et al., 2010). This hypothesis has been clearly disconfirmed in its strongest form by studies showing prediction effects of comparable magnitude in L2 and L1 groups (e.g., Djikgraaf et al., 2017). At the same time, differences between L1 and L2 users with regard to engagement in prediction have now been observed in numerous experiments, and to the best of our knowledge these differences have, without exception, been in the direction of weaker and/or delayed effects of prediction in L2 compared to L1 groups. The consistent direction of these effects, when present, strikes us as notable, and in need of explanation. We return to this point in the Discussion, where we revisit the RAGE hypothesis in light of the findings from the present study.

Proactive processing at the discourse level

The majority of research on prediction in language processing has focused on comprehenders' uptake of contextual cues to anticipate a specific upcoming word. In most cases, the cues in question immediately precede the target word or phrase, as in cues

resulting from verb semantics and argument structure restrictions (*eat… the cake*) or morphosyntactic marking such as gender marking on a determiner (*la… pelota*, 'the$_{FEM}$ ball'). Some studies have also targeted contexts where predictions about an upcoming word cannot be attributed to a single preceding cue within the same clause, but where expectations arise as a result of the wider discourse. For example, in a set of experiments with native Dutch-speaking adults (Van Berkum et al., 2005), participants were presented with contexts such as "*The burglar had no trouble locating the secret family safe. Of course, it was situated behind a…*" (original in Dutch), which were highly predictive of a specific noun, (*painting*). Taking advantage of gender-marking on adjectives preceding indefinite nouns in Dutch, participants then read continuations consisting of either an adjective with gender marking consistent with the predicted noun (*…een groot schilderij*, 'a big$_{NEU}$ painting$_{NEU}$'), or an adjective and noun with different gender (*…een grote boekenkast*, 'a big$_{COM}$ bookcase$_{COM}$). Results showed a reliable positive deflection in the ERP waveform 50–250 ms after the onset of adjective inflection on adjectives in the discourse-inconsistent condition in an ERP experiment (but see Fleur et al., 2019, for partially inconsistent results in subsequent studies), and slow-downs in the same condition in a self-paced reading study. Findings such as these suggest that prediction is not limited to immediately adjacent cues and targets, and that expectations may not always arise from a single cue, but from a more complex discourse-level representation of an event.

Just as the source of prediction cannot always be tied to a single cue, the target of an expectation is not always a specific lexical item, although this has been the paradigm case in the experimental literature. This is evident, for example, in the case of

expectations about next-mention of previously introduced discourse referents. It is a well-known phenomenon that certain verbs induce biases to remention either their subject or object in a causal dependent clause: People are more likely to continue the sentence *Helen feared Sue because…* with a mention of Sue (either with the pronoun *she* or the name *Sue*) than they are for the sentence *Helen frightened Sue because…* In these cases of 'implicit causality' (IC; Garvey & Caramazza, 1974; Hartshorne, 2014), the prediction or expectation is neither for a specific lexical item, nor is it deterministic. A number of studies have specifically investigated *when* differential bias for reactivating subject vs object referents of IC verbs emerge during real-time comprehension. Findings from adult native speakers of various languages have not been entirely consistent. Most studies have reported effects of referential bias associated with the implicit causality of the preceding verb during or shortly after the pronoun in the continuation sentence (…*because she…* see Koornneef et al., 2016, for review). Such effects are consistent with expectations built up prior to the encounter of the pronoun, and they have been interpreted explicitly as effects of prediction (e.g., Van Berkum et al., 2007). They are also consistent, however, with explanations based on integration difficulty once the pronoun is encountered, i.e., accounts not invoking prediction or expectations. Of note, three recent studies have extended the investigation of implicit causality in real-time comprehension to bilingual and L2 speakers (Contemori & Dussias, 2018; Kim, 2018; Schlenter, 2019). All three employed the visual world paradigm, and framed the investigation within a comparison between L1 and L2 groups. All three reported delay and/or reduction of IC biases in the L2 vs L1 group.

A related finding on coreference processing emerged from an ERP study that examined comprehenders' well-attested bias to remention Goal (versus Source) referents following transfer-of-possession events (Ferretti et al., 2009). As observed in a number of earlier offline continuation studies, following a transfer-of-possession event (e.g., *Bob passed the salt to Bill*), participants preferentially write continuations starting with the Goal (*Bill*) rather than the Source (*Bob*) of the event (Stevenson et al., 1994; Arnold, 2001). This bias has been shown to be modulated by whether the event is presented as completed (*passed*, perfective aspect) or ongoing (*was passing*, imperfective; Kehler et al., 2008). This is consistent with processing accounts that emphasize the status of referents in comprehenders' mental models of events (Madden & Ferretti, 2009; Magliano & Schleich, 2000; Ferretti et al., 2007): For a completed transfer-of-possession event that is part of a narrative, the Goal or recipient of the transfer act is in greater focus than in an ongoing event, since that referent is now in possession of the transferred theme. Consistent with this, Ferretti et al. (2009) presented evidence showing that participants experienced disruption, indicated by an enhanced P600 component, when a pronoun referred to the Source – the less expected referent – of a transfer event, but only when that event was portrayed as completed (*Sue handed the timecard to Fred. She…* vs. *Sue was handing the timecard to Fred. She…*). This effect observed at the pronoun is consistent with an explanation in terms of proactive expectations for reference based on comprehenders' situation models. Yet like the results of the studies on implicit causality discussed above, which found effects at or shortly after the pronoun in the dependent clause, this effect cannot be interpreted unambiguously as an effect of prediction, as it

does not satisfy the more stringent criterion for prediction as an effect measurable prior to disambiguating information in the input (Pickering & Gambi, 2018).

Identifying unambiguous effects of prediction satisfying Pickering and Gambi's criterion has been a challenge in research on reference processing. In an attempt to probe more directly for predictive effects in the context of the Goal-bias with transfer-of-possession events, Grüter et al. (2018) used the visual world paradigm to measure L1 English-speaking listeners' eye gaze to event participants during a pause between a sentence about a transfer-of-possession event and a continuation sentence. Results showed a greater preference for looking at the Goal after completed than after ongoing transfer events. This preference was temporally dispersed but sustained, and importantly, emerged well before the onset of the continuation sentence. This suggests that listeners' expectations about who would be mentioned in the continuation sentence were influenced not just by an overall Goal-bias in transfer-of-possession events, but that this bias was further modulated by whether or not the event was described as completed, as marked by grammatical aspect. This finding aligns with that of Ferretti et al. (2009), and provides unambiguous evidence that a differential bias is present prior to the encounter of a referential expression, and hence cannot be the unique result of postlexical integration when a pronoun is encountered. We interpret this bias as a reflection of comprehenders' continuously updated situation models and the relative salience of event participants therein, which in turn informs their expectations about how a discourse will continue.

This interpretation aligns well with findings from the story continuation studies mentioned above, which showed that L1 English speakers' reference choices are modulated by the event structure, indicated by grammatical aspect, of the preceding

transfer-of-possession event. This effect has been replicated with L1 speakers of other languages that mark aspect grammatically, including Japanese (Ueno & Kehler, 2016) and Korean (Kim et al., 2013). Notably, it was not replicated when the same paradigm was used with Japanese- and Korean-speaking L2 learners of English (Grüter et al., 2017), suggesting L2 users rely less on proactive expectations during reference processing, even when event structure is grammatically marked in a similar way in the speakers' L1.

Yet although Grüter et al. (2017) found no significant effects of aspect on L2 participants' choice of Source or Goal referents in their written continuations, it is important to note that aspect did significantly modulate their choice of coherence relation between the context sentence and the continuations they wrote. Specifically, participants were more likely to provide a continuation focusing on the end state of the previous event (Occasion, Result; see Kehler, 2002, for a typology of coherence relations) following a perfective-marked sentence. This effect was not different from that observed among L1 participants, and demonstrates that the L2 participants in that study were sensitive to some discourse-level implications of grammatical aspect. Grüter et al. (2017) attributed the differential effect of aspect on L2 users' reference versus coherence relation choices to the temporal sequence of decisions they had to make when actively constructing a continuation, postulating that L2 users considered aspect retroactively, at a later point in the production planning process of the continuation. However, offline measures do not allow for direct evaluation of *when* aspect-driven effects arise. The present study was conceived to address this question more directly by using a more fine-grained probe of

participants' moment-to-moment referential processing through the visual world paradigm. We thus pose the following research questions:

RQ1: Does the proactive use of discourse-level information in reference processing documented among L1 speakers generalize to L2 speakers of English drawn from the same wider population of college educated young adults?

RQ2: Among L2 speakers, does increased language proficiency lead to greater engagement in proactive use of discourse-level cues?

In addition to the visual world task reported in Grüter et al. (2018), L2 speakers also completed three additional tasks: two independent measures of English language proficiency, as well as an offline task specifically designed to test knowledge of grammatical aspect in English.

Participants

Our initial criteria for inclusion in the L2 group were (i) answer 'no' to the question 'Do you consider yourself a native speaker of English?', and (ii) first exposure to English at age 6 or older (for final $N = 52$: $M = 10.6$, $SD = 2.9$, *range* 6-21). No restrictions on participants' L1 were imposed since our interest was in generalization to L2 users in general, and since previous work had indicated that L2 speakers do not appear to benefit from facilitative L1 transfer with regard to the role of aspect marking on reference processing (Grüter et al., 2017). A total of 67 participants who fit these criteria took part

in the study. Two additional participants who answered 'yes' to the question, but indicated an age of first exposure to English of 6 or above, and did not include English as a language spoken in their childhood homes, were also included. Of these 69 participants, data from 11 was excluded prior to analysis due to non-normal vision or hearing ($N = 1$), distraction during the experiment (2), equipment malfunction (1), or poor calibration (7), and data from 6 was excluded after data inspection, due to insufficient valid data in the eye gaze record (see Results for more detail), leaving 52 participants (33 females) in the final L2 group.

L2 participants were recruited from the University of Hawai'i student community. They came from a variety of L1 backgrounds, including Bahasa Indonesia ($N = 2$), Burmese (2), Chinese (24), Croatian (1), French (1), German (2), Italian (1), Japanese (6), Korean (5), Magahi (1), Spanish (2), Swedish (1), Tagalog (1), Thai (1), Urdu (1) and Vietnamese (1). L2 participants completed three independent measures of English proficiency: the Versant English Test (Pearson, 2011), the LexTALE English Test (Lemhöfer & Broersma, 2012), and self-rating of their English language abilities on a scale of 0-10 (in all four subskills separately, as well as overall). Descriptive statistics are provided in Table 1. Pearson correlations between the three proficiency measures were moderate to strong (Versant~LexTALE: $r(48) = .65$, Versant~Self-rating: $r(48) = .68$, LexTALE~Self-rating: $r(50) = .51$, all $p < .001$). Given these consistent correlations and the fact that the Versant Test was the most comprehensive measure in terms of including direct assessment of all four subskills, overall scores from the Versant Test were used as indices of proficiency in models examining the contribution of proficiency on experimental outcomes. Conversions for Versant test scores to general level descriptors

of the Council of Europe framework (Pearson, 2011) indicate that the majority of L2 participants fall into the mid-level category of Independent User (B1: 13, B2: 16), with a few lower-level advanced Basic Users (A2: 8) and some highest-level Proficient Users (C1: 8, C2: 5).

Results from the L2 group will be compared with those from the L1 group in Grüter et al. (2018; $N = 53$, 28 females). Criteria for inclusion in the L1 group were: (i) answer 'yes' to the question 'Do you consider yourself a native speaker of English?', (ii) answer 'English' to 'What language do you feel most comfortable speaking in casual conversation now?', (iii) first exposure to English at age 5 or younger ($M = .4$, range 0-5), and (iv) self-ratings for English speaking and listening ability of 8 or higher (out of 10).

Table 1

Participant information (means and ranges)

	Age (in years)	Self-rated overall English proficiency (out of 10)	Versant English Test[1] (overall score, range 20-80)	LexTALE English[2] score (max = 100)
L2 English (n = 52)	27 (18-46)	6.7 (2-10)	60.24 (39-80)	71.73 (49-100)
L1 English (n = 53)	23 (18-49)	9.5 (7-10)	NA	NA

[1] Pearson (2011); values from 2 participants missing

[2] Lemhöfer & Broersma (2012)

Knowledge-of-aspect task

Materials and procedure

A key assumption underlying the main experiment is that listeners associate perfective aspect with completed events, and imperfective aspect with ongoing events. An offline knowledge-of-aspect task was included to verify that this assumption is justified for the L2 participants in this study. Participants read brief narratives describing events that were either ongoing or completed, and were then asked to judge the truth of a statement about a particular time point in the story. An example is provided in (1). The narrative remained on the screen until the participant made a judgment by pressing 'true', 'false', or 'not

sure'. We expect test sentences with imperfective aspect, as in (1), to be judged 'true' in the ongoing and 'false' in the completed condition.

(1) *Story beginning:*

Brenda is at the hospital visiting Anne. [picture of soup]

This is the bowl of soup that Brenda will feed to Anne.

At 11:00, Brenda is ready with the soup and a spoon.

Ongoing condition: Story end + test sentence

At 11:05, Brenda puts the first spoonful of soup into Anne's mouth.

In the afternoon, Pikachu says:

"At 11:05, Brenda was feeding the bowl of soup to Anne."

Completed condition: Story end + test sentence

At 12:00, the bowl is empty and Anne wipes her mouth.

In the afternoon, Pikachu says:

"At 12:00, Brenda was feeding the bowl of soup to Anne."

The task consisted of a total of 22 items of the type illustrated in (1), including 10 items with an imperfective-marked transfer-of-possession verb following a story in which the transfer was portrayed as ongoing (k=5; true) or completed (k=5, expected judgment: false). The transfer-of-possession verbs were the same as those used in the visual world experiment. Given the existence of crosslinguistic differences in the marking of

continuous aspect and the interaction of aspect with verb event classes (e.g., Gabriele, 2009), we considered these two conditions the most critical for determining whether L2 learners understand the function of aspect-marking in English. An additional 12 items were included to further assess learners' interpretations of other aspect, verb class, and event type combinations, including perfective-marked achievement verbs following a completed (k=4; true) or ongoing (k=4; false) event, and imperfective-marked accomplishment verbs following a completed event (k=4; false). Participants were assigned to one of four semi-randomly ordered lists counterbalanced for the presentation of verbs in the two critical conditions.

L2 participants completed this task immediately after the visual world experiment. L1 participants did not complete this task, as we had evidence from a prior L1 group (N = 61, drawn from the same population) on native speaker performance. When presenting the results from the L2 group on this task, we refer to these previous L1 data as the L1 reference group.

Results

Two L2 participants did not complete this task; we report results from the remaining 50, together with those from the L1 reference group. 'Not sure' responses were rare overall (4.5%, cf. 4.8% in L1 reference group). The mean proportion of expected judgements across all five conditions was .81 (SD = .12) in the L2 and .84 (SD = .10) in the L1 group. A logistic mixed-effect regression model (glmer(*IsExpected ~ group + (1|subject) + (1|item)*) indicated no significant difference between the two groups (B = -.21, z = -1.4, p = .16). Critically, both L1 and L2 speakers judged statements with imperfective aspect

predominantly true with ongoing transfer-of-possession events (L1: *M*=.89, *SD*=.15; L2: *M*=.85, *SD*=.19) and predominantly false with completed transfer-of-possession (L1: *M*=.86, *SD*=.17; L2: *M*=.88, *SD*=.22) and other accomplishment events (L1: *M*=.86, *SD*=.19; L2: *M*=.80, *SD*=.27). The reverse was the case for statements with perfective aspect, which were predominantly judged true with completed events (L1: *M*=.74, *SD*=.26; L2: *M*=.82, *SD*=.20) and false with ongoing events (L1: *M*=.80, *SD*=.19; L2: *M*=.86, *SD*=.17). Separate models were run on the data from each condition separately. Only one – imperfective-marked accomplishments with completed events – showed an L1-L2 difference that was significant at an unadjusted alpha = .05 level (B = -.80, z = -2.2, p = .03; all other $B < |.47|$ and $p > .14$).

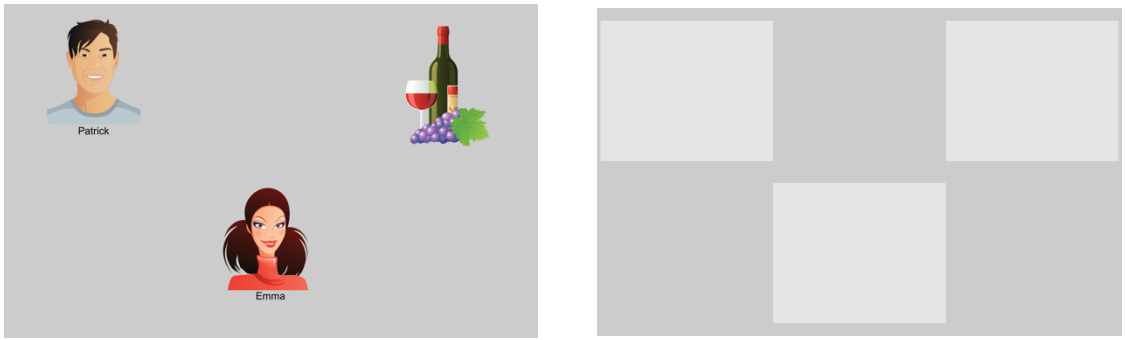
We take these results to provide overall support for the assumption that the L2 participants in this study did not differ significantly from L1 speakers in their interpretation of aspect in sentences such as those in the visual world experiment. Any between-group differences we might observe in the visual world experiment are thus unlikely to be attributable to incomplete knowledge of aspect among L2 participants.

Visual world experiment

In this task, participants listened to two-sentence mini discourses, as in (2), while looking at visual scenes on a screen. Each trial began with a 2,000 ms silent preview of the visual scene (Figure 1, panel A), followed by a context sentence (mean duration = 2,856 ms, *SD* = 277), a 2,500-ms intersentential pause, a continuation sentence (mean duration = 2,755ms, *SD* = 331), a 1,750-ms pause, and a question. 250 ms into the second pause, the visual scene changed such that the individual images were replaced with grey boxes

(Figure 1, panel B). Participants were instructed to click on the box corresponding to the position of the event participant that best answered the question. This memory component was added to engage participants and disguise the objective of the experiment. Participants' final responses were not of interest here. Of critical interest in view of the research questions under investigation are listeners' looks to event participants during the intersentential pause. In particular, in the critical transfer-of-possession events, do we see biases to differentially fixate Goals versus Sources, and most critically, are these biases modulated by Aspect in the context sentence? As previously reported in Grüter et al. (2018), such modulation, suggestive of proactive expectations, was observed among L1 speakers of English in this experimental paradigm. Here we ask to what extent the same holds for L2 listeners.

(2) Patrick {gave/was giving} Emma a bottle of nice wine. [context sentence]
{He/She} obviously knew about fancy food and drink. [continuation]
Who knew about fancy food and drink?



(a) (b)

Figure 1. Example of visual displays.

Materials

The experiment consisted of a total of 60 two-sentence items of the type illustrated in (2), comprised of 20 experimental and 40 filler items. Experimental items contained one of 10 transfer-of-possession verbs, each used in two items, in a double-object construction with a human subject/Source, a human indirect object/Goal, and an inanimate direct object/Theme. Source and Goal always differed in gender, such that the pronominal subject of the continuation sentence (She/He) disambiguated reference at that point. Four versions of each sentence were created by manipulating *Aspect* in the context sentence (perfective/imperfective) and *Reference* of the subject pronoun in the continuation (Source/Goal of context sentence, disambiguated through gender). The 4 versions were distributed across 4 experimental lists in a 2×2 Latin Square design, such that each participant encountered only one version, for a total of 5 items in each of 4 Aspect/Reference conditions.

The 40 fillers items, as well as 6 practice items presented at the beginning of the experiment, were analogous in their overall structure (context-continuation-question) but described non-transfer-of-possession events, including transitive and intransitive verbs from various verb classes. Half of the context sentences in fillers had perfective, half imperfective aspect. Continuations started with pronouns referring to human or non-human participants in the context sentence (She/He/They/It); gender disambiguated reference in some but not all filler items. Questions in fillers asked about various aspects of the context or continuation sentences. A complete list of all items is provided in Grüter

et al. (2018, Supplementary Materials). All sentences were recorded by a female native speaker of American English using a clear speaking style.

All visual scenes contained 3 areas of interest (AOIs; Figure 1). In experimental items, the 3 AOIs always represented Source, Goal and Theme of the transfer-of-possession event. Location of the three AOI types was counterbalanced between items. In filler items, visual scenes contained one or two human event participants, and one or two non-human referents from the context sentence, including objects and locations.

Data measurement, treatment and analysis approach

The experiment was conducted using an SMI RED250 eye-tracker sampling at 250 Hz. Eye gaze was recorded and exported through SMI *Experiment Suite* software, and classified as fixations, saccades and blinks using the software's default settings. For further analysis, data were binned into 20-ms samples. Next, we calculated, for each trial, the proportion of sample points containing fixations out of all sample points. Trials with values more than 2 *SD*s below the mean were excluded. This accounted for 7.1% of the L2 data. As noted earlier, we excluded participants with insufficient valid data, in this case those with fewer than 15 (out of 20) experimental trials remaining after this procedure ($N = 6$).

Following the same procedures as in Grüter et al. (2018), we calculated a 'GoalAdvantage' score as a measure of bias to fixate Goals vs Sources during two time windows in each trial. GoalAdvantage scores were calculated by subtracting the number of 20-ms bins with looks to Source from those with looks to Goal during each of two time windows. The first window ('Silence') extends from 500 ms after the offset of the

context sentence until 200 ms after the onset of the continuation (following standard procedures to offset analyses of fixations by 200 ms, Matin, Shao, & Boff, 1993). The second window ('Continuation') extends from 200 to 1500 ms after the onset of the continuation. The Silence window constitutes the critical temporal region of interest for our analysis of *anticipatory* looking behavior. The Continuation window was included in the original analysis to explore potential effects of the disambiguating pronoun and interactions with Aspect within a single omnibus analysis avoiding multiple comparisons. No such effects were found in the L1 group (Grüter et al., 2018). We likewise include this window here in order to allow for the most direct comparisons possible.

GoalAdvantage scores over relatively long temporal windows were chosen as the dependent measure for two reasons. First, unlike other effects typically investigated in visual world experiments, the hypothesized effect of Aspect cannot be tied to a unique event in the speech signal. Instead, the effect likely results from a complex combination of multiple cues in the construction of a mental event model—a critical component of discourse processing. We thus had no specific predictions about change in fixations over time during the intersentential pause, and for this reason, opted for the most conservative approach to aggregate data over the entire Silence region.[1] Second, GoalAdvantage scores were more normally distributed than empirical logits (Barr, 2008) calculated from the same trial-level data, and thus aligned better with the underlying assumptions of the linear mixed models adopted in our analysis.[2]

Results

All analyses were conducted in R (3.6.0), using the lmerTest package (Kuznetsova et al., 2017). Figure 2 presents an overview of fixations in both the L2 and the L1 groups over the course of an entire trial, collapsing across all experimental manipulations. What is perhaps most notable on visual inspection is how similar the overall pattering and timing of fixations are in the two groups; in particular, we see no evidence of overall slower or delayed processing in the L2 group. The most apparent difference is in participants' looks to the Theme during the intersentential pause: L2 listeners appeared more likely than L1 listeners to redirect their gaze away from the (last mentioned) Theme to one of the two human referents. For L1 listeners, the proportion of looks to Theme reduces below that of either human referent by the onset of the continuation sentence (0ms) whereas for L2 listeners, this happens much earlier (around -1,600 ms).This difference was not expected, and we have no ready explanation for it. However, as our critical focus is on differential looks to Source and Goal, it has no direct impact on our critical analysis.[3]

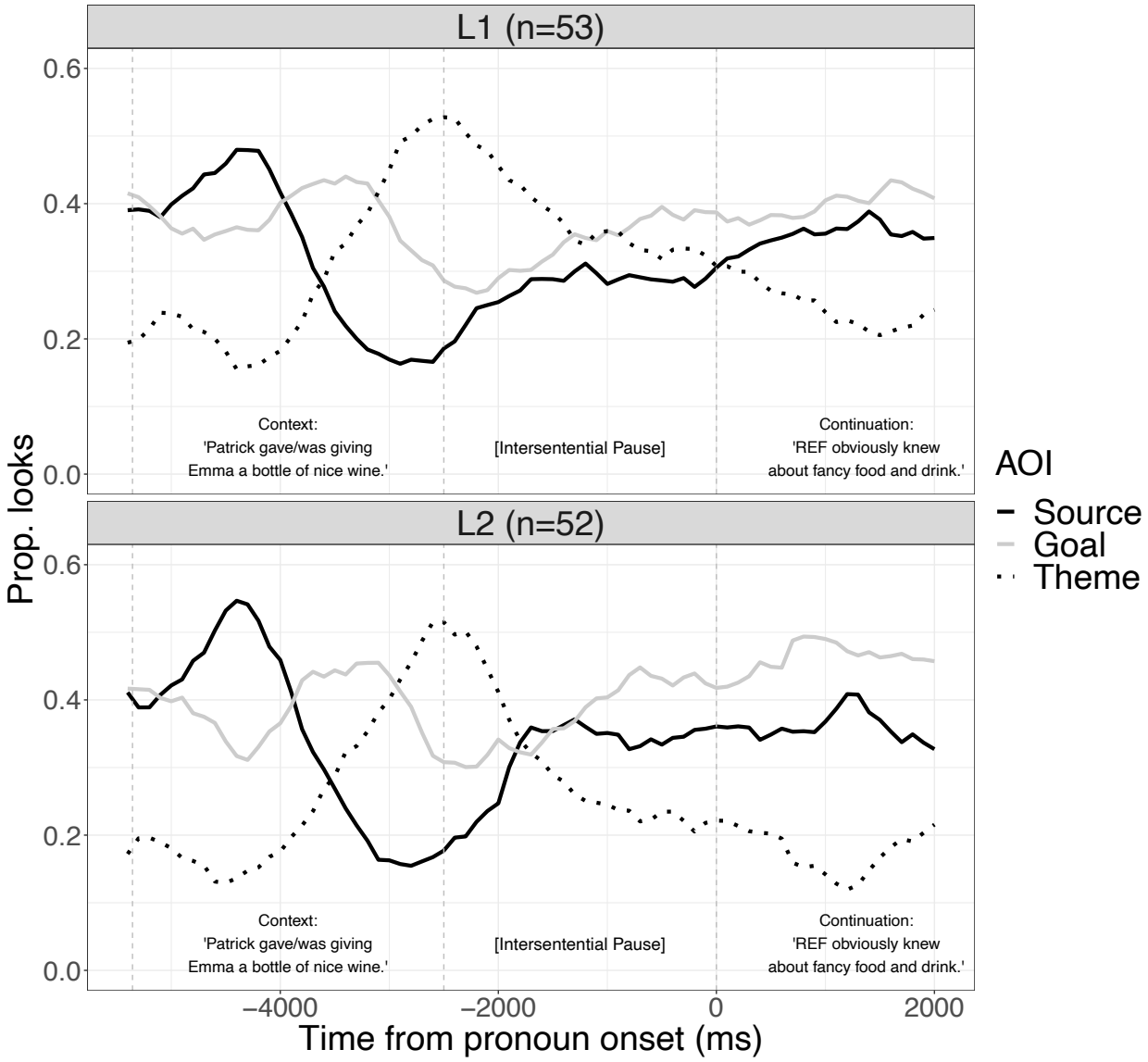


Figure 2. Time course of fixations by Group, collapsed over Aspect and Reference. Proportions to each AOI are calculated out of all fixations to an AOI. Trials were aligned by the onset of the Continuation. (In half the items the subject of the Continuation refers to the Source, in the other half to the Goal. Hence when collapsing over Reference, roughly equal fixations to Source and Goal are expected in the Continuation.)

Also apparent in Figure 2 is an overall bias, emerging in both groups during the second half of the intersentential pause, to look more at the Goal than the Source. This is consistent with the well-documented preference to remention Goals after transfer-of-possession events (Arnold, 2001; Stevenson et al., 1994), and suggests that this preference originates in proactive expectations prior to the encounter of any referential expressions in the continuation. We further examine this Goal bias, and critically its potential modulation by Aspect in the context sentence, through mixed-effect modeling following the same procedures as in Grüter et al. (2018). We begin by modelling the L2 data alone to address our two primary research questions: Do the effects previously observed among L1 speakers of English generalize to L2 speakers drawn from the same wider population (RQ1), and to what extent does proficiency modulate L2 speakers' engagement in proactive reference processing (RQ2)? We will then combine the L2 data with the L1 data from Grüter et al. (2018) to further probe whether the overall effect of Aspect remains robust in this larger yet more diverse sample, and whether native-speaker status defined categorically (L1, L2) constitutes a potentially modulating factor.

Analysis of L2 data

We began by fitting the final model for the L1 data (GoalAdv ~ Aspect * Reference * Window + (1 + Aspect * Reference | Subject) + (1 + Aspect * Reference | Item); Grüter et al., 2018, Appendix B) to the L2 data, with Aspect (contrast-coded; perfective: -0.5, imperfective: 0.5), Reference (Source: -0.5, Goal: 0.5), and Window (Silence/Continuation; treatment-coded, reference-level: Silence) constituting the fixed effects. (Note that when a factor is treatment-coded, all effects not involving this factor

are calculated not as *main* effects but as *simple* effects at the reference level of this factor. The treatment-coding of Window here means that the effects of Aspect and Reference in the model output reflect the effects of these two factors in the Silence window only, i.e., in the time period of critical interest.) This model generated a singular-fit warning and a convergence code of zero, indicating that the model is likely too complex for these data. We thus removed slopes from the random effect terms, and reran the model with only intercepts for random effects. Fixed effects are summarized in Table 2. The only effect that reached significance was the interaction between Reference and Window, indicating that, as expected, looks to Goal and Source were affected by the disambiguating pronoun at the onset of the Continuation. Critically, no significant effect of Aspect was observed (b = .25, p = .91). We also note the positive intercept term (b = 3.67, p = .096), indicating a marginal trend towards an overall bias to look at Goals during the (reference-level) Silence window.[4]

Table 2

L2 data (N = 52): Model statement and summary of fixed effects in linear mixed effects model of GoalAdvantage. Reference level for Window = Silence.

*Formula: GoalAdv ~ Aspect * Reference * Window + (1 | Subject) + (1 | Item)*

Fixed effects:	Estimate	Std. Error	t	p
(Intercept)	3.674	2.140	1.717	0.096 .
Aspect	0.252	2.339	0.108	0.914
Reference	-2.327	2.338	-0.995	0.320
WindowContinuation	0.785	1.653	0.475	0.635
Aspect:Reference	0.356	4.678	0.076	0.939
Aspect:WindowContinuation	2.578	3.306	0.780	0.436
Reference:WindowContinuation	18.832	3.306	5.697	1.4e-08 ***
Aspect:Reference:WindowContinuation	5.012	6.611	0.758	0.448

Table 3

L1 data (N = 53): Model statement and summary of fixed effects in linear mixed effects model of GoalAdvantage. Reference level for Window = Silence.

*Formula: GoalAdv ~ Aspect * Reference * Window + (1 | Subject) + (1 | Item)*

Fixed effects:	Estimate	Std. Error	t	p
(Intercept)	3.817	1.494	2.555	0.015 *
Aspect	-5.411	2.118	-2.555	0.011 *
Reference	4.020	2.118	1.898	0.058 .
WindowContinuation	-2.344	1.497	-1.566	0.118
Aspect:Reference	4.035	4.236	0.953	0.341
Aspect:WindowContinuation	3.378	2.994	1.128	0.259
Reference:WindowContinuation	8.634	2.994	2.884	0.004 **
Aspect:Reference:WindowContinuation	-4.230	5.988	-0.706	0.480

In order to verify that the previously reported effects in the L1 group remain when the L1 data are submitted to the simplified model fitted to the L2 data, we fit the same model to the previous L1 data (Table 3). The pattern of results reported in Grüter et al. (2018) remained unchanged: In addition to the expected Reference-by-Window interaction (b = 8.63, p = .004), the only other fixed effect that reached significance was Aspect (b = -5.41, p = .011). The intercept term also remained significant (b = 3.82, p = .015). We further probed whether the simpler model presents a significant decrease in model fit compared to the fuller model previously reported. Model comparison using the

anova() function indicated that this was not the case (Chisq = 9.23, df = 18, p = .95). We thus base all further analyses on models with random intercepts only.

Returning to the L2 data, we next probed for potentially modulating effects of English proficiency by adding participants' overall scores from the Versant English test (centered) to the model (Table 4).[5] Model comparison with the previous model indicated a marginal increase in model fit (ChiSq = 14.9, df = 8, p = .07). Unexpectedly, the simple effect of Proficiency was significant (b = -2.76, p = .03), showing a decrease in the overall bias to look at Goals with increasing Proficiency during the (reference-level) Silence window. More importantly, Proficiency did not interact with Aspect (b = .47, p = .87). The only other significant effect, beyond the previously observed and expected Reference-by-Window interaction, was an unexpected 3-way interaction between Aspect, Reference and Proficiency (b = 11.3, p = .02). To explore this interaction and better understand the contribution of Proficiency, we divided the data from the Silence region by median split of Versant scores, and inspected the data separately for higher (Versant Score >= 60, N=26) and lower proficiency participants (N=24). In neither of these subgroups did we find effects of Aspect, Reference, or an interaction between the two. Given that the factor Reference should not be observable during the Silence region (i.e., the listener does not yet know whether the pronoun in the Continuation will refer to the Source or Goal), and that we found no further support for an Aspect-by-Reference interaction in the higher and lower proficiency subgroups, we believe the observed 3-way interaction may be spurious and not further informative with regard to our research questions. The simple effect of Proficiency in the overall model was reflected in the subgroup models through a significant intercept term in the lower-proficiency (b = 7.4, p

= .02) but not in the higher-proficiency subgroup (b = 1.0, p = .7). Interestingly, it is the pattern in the lower-proficiency subgroup that is more similar to what was observed in the L1 data, where a similar overall preference for Goals was found.

Table 4

*L2 data, including Proficiency (N = 50): Model statement and summary of fixed effects in linear mixed effects model of GoalAdvantage. Reference level for Window = Silence. Formula: GoalAdv ~ Aspect * Reference * Window * scale(Versant) + (1 | Subject) + (1 | Item)*

Fixed effects:	Estimate	Std. Error	t	p
(Intercept)	4.089	2.062	1.983	0.056 .
Aspect	0.469	0.469	0.198	0.843
Reference	-1.294	2.362	-0.548	0.584
WindowContinuation	0.490	1.670	0.293	0.769
Versant	-2.753	1.278	-2.154	0.033 *
Aspect:Reference	1.010	4.726	0.214	0.831
Aspect:WindowContinuation	2.707	3.340	0.810	0.418
Reference:WindowContinuation	18.984	3.340	5.684	1.52e-08 ***
Aspect:Versant	-0.400	2.364	-0.169	0.866
Reference:Versant	1.283	2.371	0.541	0.588
WindowContinuation:Versant	3.223	1.670	1.929	0.054 .
Aspect:Reference:WindowContinuation	3.622	6.680	0.542	0.588
Aspect:Reference:Versant	11.340	4.729	2.398	0.017 *
Aspect:WindowsContinuation:Versant	2.838	3.341	0.849	0.396
Reference:WindowsContinuation:Versant	-3.515	3.341	-1.052	0.293
Aspect:Reference:WindowsCont:Versant	-5.501	6.681	-0.82	0.410

In sum, modelling of the L2 data following the same procedures as Grüter et al. (2018) did not reveal the effect of Aspect previously observed in an equal-sized dataset from L1 speakers drawn from the same wider college-student population. Figure 3 presents a visual illustration of (the absence of) the Aspect effect in the L2 (right panel) and the L1 sample (left panel). This visualization indicates that there is little evidence that a potential effect of Aspect is present but weaker or noisier in the L2 data. Our analyses also showed that there was no evidence for proficiency interacting with Aspect. Given that the distribution of proficiency scores in the L2 group was wide, and the group included a substantial number of participants at the higher end of the proficiency spectrum, these results suggest that it is unlikely that the effect of Aspect is one that might emerge with increasing proficiency.

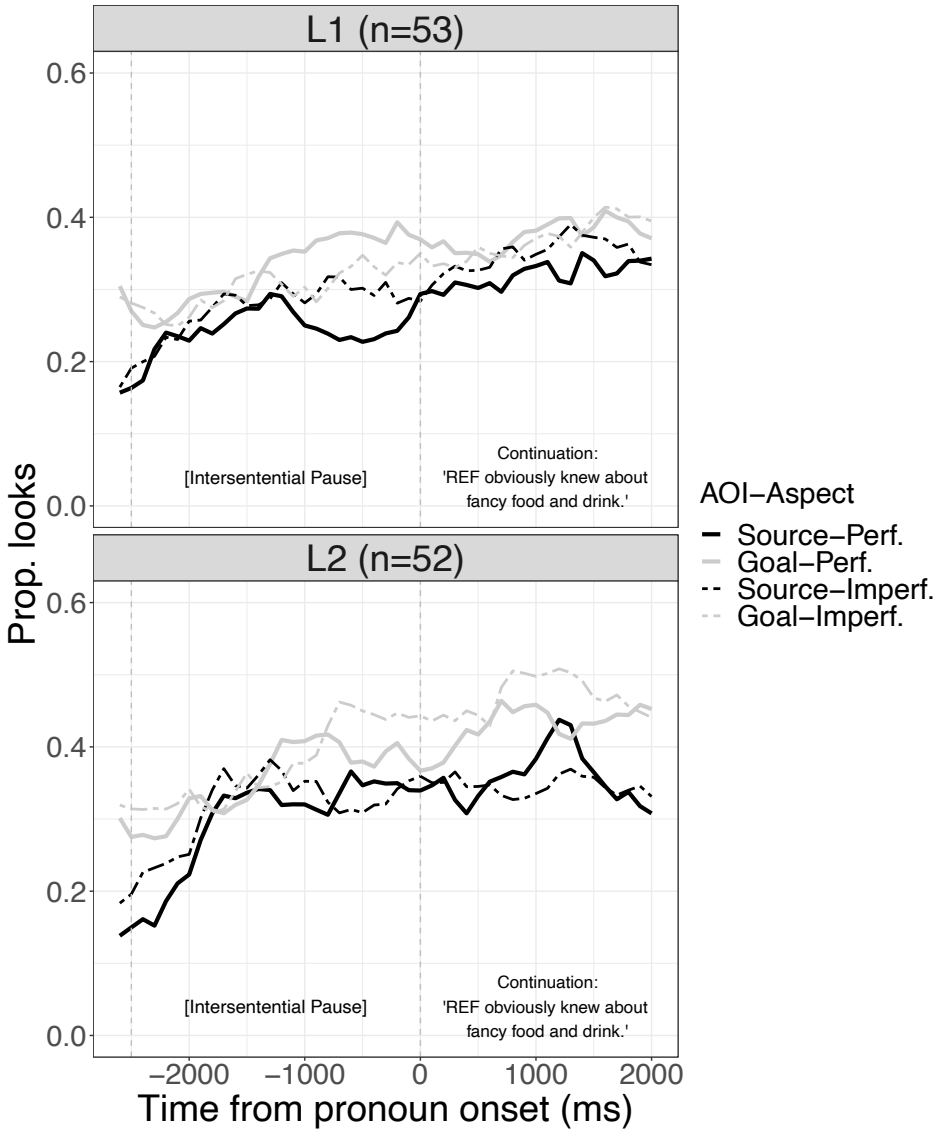


Figure 3. Time course of fixations by Group and Aspect, collapsed over Reference

Analysis of combined dataset

Combining the L1 and L2 data into a single larger dataset provides an additional opportunity for assessing whether the effect of Aspect observed among L1 speakers scales up to a more heterogeneous sample of language users. We thus submitted the

combined data (N = 105) to the same model as above (GoalAdv ~ Aspect * Reference * Window + (1 | Subject) + (1 | Item)). The effect of Aspect did not reach significance (b = -2.56, p = .105). The only significant effects were the expected Reference-by-Window interaction (b = 13.72, p < .001) and the intercept term (b = 3.74, p = .027). Next, in order to evaluate whether speaker status categorically perceived (L1, L2) was a significant predictor of participants' reference expectations, we added Group (contrast-coded and centered) to the model (Table 5). Model comparison indicated that this addition improved model fit (ChiSq = 15.7, df = 8, p = .048). The model returned a marginal interaction between Group and Aspect (b = 5.74, p = .069); the main effect of Aspect remained marginal (b = -2.58, p = .103). The model also indicated significant interactions between Group and Reference (b = -6.38, p = .044) and between Group, Reference, and Window (b = 10.20, p = .022). Further investigation of these latter effects revealed a non-significant trend in the L1 group for greater Goal-bias in the Silence region for items whose reference would later be disambiguated towards Goal, whereas in the Continuation region, the effect of Reference, highly significant in both groups, was somewhat stronger in the L2 group. We do not have an explanation for these interaction effects, but do not believe they impact conclusions with regard to the research questions under investigation. No other effects reached significance.

In sum, analysis of the combined dataset (N = 105) showed only a non-significant trend in the expected direction for the effect of Aspect. This trend was modulated by a marginally significant interaction with Group. Analysis of the L2 data alone (N = 52) had shown no effect of Aspect, whereas the effect appeared robust in the L1 group (N = 53).

Table 5

L1 and L2 data combined (N = 105): Model statement and summary of fixed effects in linear mixed effects model of GoalAdvantage. Reference level for Window = Silence.

*Formula: GoalAdv ~ Aspect * Reference * Group * Window + (1 | Subject) + (1 | Item)*

Fixed effects:	Estimate	Std. Error	t	p
(Intercept)	3.733	1.605	2.325	0.028 *
Aspect	-2.578	1.579	-1.632	0.103
Reference	0.841	1.578	0.533	0.594
Group	-0.162	1.686	-0.096	0.924
WindowContinuation	-0.774	1.116	-0.693	0.488
Aspect:Reference	2.168	3.158	0.686	0.493
Aspect:Group	5.741	3.157	1.818	0.069 .
Reference:Group	-6.375	3.157	-2.020	0.044 *
Aspect: WindowContinuation	2.985	2.232	1.337	0.181
Reference: WindowContinuation	13.728	2.232	6.151	8.46e-10 ***
Group: WindowContinuation	3.170	2.232	1.420	0.156
Aspect:Reference:Group	-3.79061	6.314	-0.600	0.548
Aspect:Reference: WindowContinuation	0.387	4.464	0.087	0.931
Aspect:Group:WindowsContinuation	-0.799	4.464	-0.179	0.858
Reference:Group:WindowsContinuation	10.198	4.464	2.285	0.022 *
Aspect:Reference:Group:WindowsCont.	9.242	8.9274	1.035	0.301

Discussion and Conclusion

The goal of this study was to examine to what extent proactive use of discourse-level information in real-time reference processing, as previously observed among L1 speakers, is also characteristic of comprehenders who are not native speakers of the language. To this end, we compare the results of a visual world eye-tracking experiment with L1 speakers of English, previously reported in Grüter et al. (2018) to an equal-sized group of L2 speakers of English. L2 participants additionally completed an independent offline task to assess their understanding of the linguistic contrast hypothesized to give rise to the predictive effect in real-time processing. Results from this knowledge-of-aspect task showed that the L2 participants associated perfective aspect with completed events and imperfective aspect with ongoing events and that their judgments for transfer-of-possession events in this offline task did not differ significantly from those of L1 speakers. Differences in the use of aspect to create proactive expectations about reference in real-time processing are thus unlikely to be attributable to limitations in L2 participants' knowledge of aspect in English.

L2 participants' fixations on Source and Goal referents during an intersentential pause following sentences containing perfective- or imperfective-marked transfer-of-possession verbs were submitted to the same analyses as those from L1 speakers who completed the same experiment. Whereas these analyses had revealed a significant effect of aspect among L1 speakers, who fixated more on Goal referents following perfective-marked than following imperfective-marked transfer verbs, no such effect was found in the L2 group. Further inspection and visualization of the effect did not indicate any trends in the expected direction, suggesting that it is unlikely that a larger sample with increased

power would yield a different outcome. With regard to our first research question (RQ1), we therefore conclude that the effect observed by Grüter et al. (2018) among native English-speaking college students does not easily generalize to non-native speakers of English drawn from the same wider population of college educated young adults.

Further support for this conclusion comes from the analysis of the combined data from L1 and L2 speakers, in which the main effect of Aspect was no longer robust. The marginal interaction between Aspect and Group ($p = .069$) indicates that the categorical classification of participants as native versus non-native constitutes a potentially meaningful predictor of listeners' engagement in expectation-based processing. A prima facie explanation for this pattern lies in participants' overall English proficiency. L1 speakers' self-ratings of their overall English proficiency were significantly higher than those among L2 participants (see Table 1; $W = 118.5$, $p < .001$). However, if overall English proficiency were a significant factor for engaging in predictive processing, we would expect proficiency measured as a continuous variable to be a significant modulator of the Aspect effect within the L2 group. Yet our analyses showed no such modulation, and no evidence for any trends towards an effect of Aspect among the more highly proficient L2 participants. This is noteworthy in that our L2 sample showed substantial variability in overall proficiency, covering the range of A2 to C2 levels, and included participants at the highest (C2) level. Moreover, the measure of proficiency used in our analyses (Versant English Test overall scores) correlated moderately to strongly with two other, less comprehensive measures of proficiency commonly used in L2 processing studies (LexTALE, self-ratings), suggesting that the variance captured by this measure is representative of the construct of L2 proficiency as used in other research in the field. In

answer to our second research question (RQ2), we thus conclude that increased overall language proficiency did not lead to greater engagement in proactive use of discourse-level cues in this experiment.

The null effect of proficiency may appear surprising as it is a commonly stated assumption that proficiency modulates L2 learners' engagement in predictive processing (e.g., Kaan, 2014). As discussed above, however, empirical support for this assumption has been remarkably weak, with multiple studies targeting different linguistic phenomena, in different languages, and using a variety of proficiency measures, reporting no significant effects of proficiency on predictive processing (Djikgraaf et al., 2017; Hopp, 2015; Ito et al., 2018; Kim, 2018; Mitsugi, 2018). Proficiency also did not significantly modulate L2 learners' reference and coherence choices in Grüter et al.'s (2017) story continuation study. The null effect of proficiency in the present study is thus consistent with accumulating evidence from recent work suggesting that despite common and perhaps intuitive assumptions, overall proficiency does not appear to play a notable role in L2 speakers' engagement in expectation-based processing.

A broader goal of this study is to contribute to the understanding of variability among different types of language users with regard to engagement in proactive, expectation-driven mechanisms during language processing. We found that the effect observed among adult L1 speakers did not generalize to L2 users drawn from the same wider population of college-educated young adults. More specifically, unlike L1 users, L2 users did not appear to make use grammatical aspect as an indicator of event structure to dynamically update their situation model and the discourse expectations associated with it during the intersentential pause. This is consistent with the findings from Grüter et

al.'s (2017) offline story continuation study, and consistent with the RAGE hypothesis in its broad statement that "L2 speakers have Reduced Ability to Generate Expectations" (2017, p. 224).

What remains unknown is *why* we see differential reliance on expectations in different types of language users. The original RAGE hypothesis suggested that explanations would lie at the level of L2 users' *ability* to generate and continuously update expectations during processing, consistent with reduced-capacity models of L2 processing more generally (e.g., Hopp, 2010). We have come to realize, however, that this is not the only possibility (see also Grüter et al., 2020). There may be circumstances when relying on, or even generating, expectations would not enhance processing success, and thus would not be a processing strategy that rationally optimizes available resources. In other words, under certain circumstances, the costs always associated with prediction, that is, the potential need to revise a proactively created representation that turns out to be false, may not outweigh its benefits (see also Federmeier, 2007). For L2 users, this could apply when relevant L2 knowledge is represented differently as a result of how young (L1) versus older (L2) learners extract information from the input (see e.g., Grüter et al., 2012, following proposals by Arnon & Ramscar, 2012). In this case, a particular linguistic cue may not be reliable enough within an L2 user's system of knowledge representations to make the risks that come with launching a prediction worthwhile. Reduced engagement in predictive processing under this scenario would then not necessarily be a reflection of reduced *ability* to predict, but of reduced *utility* of prediction under these circumstances.

This perspective is consistent with Kuperberg and Jaeger's (2016) view of prediction as a utility function, whereby a resource-bound rational comprehender weighs its advantages and disadvantages in a given situation. The extent to which a language user will engage in prediction is then seen as "a function of its expected utility, which, in turn, may depend on comprehenders' goals and their estimates of the relative reliability of their prior knowledge and the bottom-up input" (p. 32). Exploring the possibility that L2 users' reduced engagement in prediction in certain domains is not due to differences in *ability* but to differences in *utility* of prediction requires a more fundamental reframing of the often implicit research question in the field of L2 processing, namely whether L2 learners can "achieve L1 processing efficiency." Instead, we may need to consider that what is most efficient for L1 users might not always be most efficient for L2 users; or, put differently, that reduced engagement in prediction (as compared to L1 speakers) is in fact the most efficient and adaptive processing strategy for L2 users under certain circumstances.

Identifying what exactly these circumstances are is an issue that future work will need to explore. As a direction for future research into modulating factors of prediction in L2 processing, we would like to suggest a reframing of the RAGE hypothesis that allows for the possibility that differential processing outcomes between L1 and L2 speakers could result not just from reduced *ability* to engage in prediction, but from reduced *utility* of prediction when weighing its potential costs and benefits in a given knowledge system and context. In other words, we suggest that when reduced engagement in prediction is observed in L2 users vis-a-vis an L1 comparison group, we consider the possibility that this is not deficient but maximally adaptive processing behavior.

To be very clear, the RAGE hypothesis, both in its original formulation and in the updated version we suggest here, does not constitute a *scientific* hypothesis in the sense that it would allow us to derive clearly falsifiable predictions (no pun intended). Yet we believe we simply do not know enough at this point to be able to formulate a more clearly testable hypothesis concerning the specific circumstances and contexts under which L2 users are predicted to show reduced engagement in or reliance on proactive processing mechanisms. At present, the RAGE hypothesis in its current (and previous) form merely offers a broad statement of the empirical observation that in some, but attestedly not all, circumstances L2 users show reduced effects of expectation-based processing (see also Grüter et al., 2017, p. 224). Further exploratory research is needed to hone our understanding of when L2 users employ predictive mechanisms. We believe the consideration of *utility*, in the sense of Kuperberg and Jaeger (2016), in addition to *ability* will be productive in designing such research. The RAGE hypothesis thus constitutes an invitation to further explore the role that predictive mechanisms play in L2 processing, seen as an instance of human language processing more generally, in the hope that such further exploration will eventually allow us to formulate more directly testable predictions about the contexts in which we will or will not see reduced engagement in expectation-based processing among L2 users.

Acknowledgments

This work was supported by a standard grant from the National Science Foundation (BCS-1251450). We thank Amy J. Schafer, as well as A. L. Blake, Amber Camp, Catherine Gardiner, Victoria Lee, Wenyi Ling, Ivana Matson, Maho Takahashi and Aya Takeda for their various contributions to this project, and the reviewers of this journal for their very helpful feedback.

References

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist, 63*, 602–614.

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

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*, 292–305.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 457–474.

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112*, 417–436.

Chambers, C. G., & Cooke, H. (2009). Lexical competition during second-language listening: Sentence context, but not proficiency, constrains interference from the

native lexicon. *Journal of Experimental Psychology: Learning, Memory and Cognition, 35*, 1029–1040.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–204.

Contemori, C., & Dussias, E. P. (2018). Prediction at the discourse level in L2 English speakers: an eye-tracking study. In A. B. Bertolini & M. J. Kaplan (Eds.), *Proceedings of the 42nd annual Boston University conference on language development* (pp. 159–171). Cascadilla.

Dijkgraaf, A., Hartsuiker, R., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition, 20*, 917–930.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44,* 491-505.

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language, 115*, 149–161.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.), *Developmental psycholinguistics: On-line methods in children's langauge processing* (Vol. 44, pp. 97-135). John Benjamins.

Ferreira, F., & Chantavarin, S. (2018). Integration and prediction in language processing: A synthesis of old and new. *Current Directions in Psychological Science, 27*, 443-448.

Ferretti, T. R., Kutas, M., & McRae, K. (2007). Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 182–196.

Ferretti, T. R., Rohde, H., Kehler, A., & Crutchley, M. (2009). Verb aspect, event structure, and coreferential processing. *Journal of Memory and Language, 61*, 191–205.

Fleur, D. S., Flecken, M., Rommers, J., & Nieuwland, M. S. (2019). Definitely saw it coming? An ERP study on the role of article gender and definiteness in predictive processing. Preprint posted on BioRxiv, doi: https://doi.org/10.1101/563783

Foucart, A., Martin, C. D., Moreno, E. M., & Costa, A. (2014). Can Bilinguals See It Coming? Word Anticipation in L2 Sentence Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1461-1469.

Gabriele, A. (2009). Transfer and transition in the SLA of aspect: A bidirectional study of learners of English and Japanese. *Studies in Second Language Acquisition, 31*, 371-402.

Gambi, C., Gorrie, F., Pickering, M. J., & Rabagliati, H. (2018) The development of linguistic prediction: predictions of sound and meaning in 2-to-5 year olds. *Journal of Experimental Child Psychology, 173*, 351–370.

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

Grüter, T., Lau, E., & Ling, W. (2020). How classifiers facilitate predictive processing in L1 and L2 Chinese: The role of semantic and grammatical cues. *Language, Cognition and Neuroscience, 35*, 221-234.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research, 28*, 191–215.

Grüter, T., Rohde, H., & Schafer, A. J. (2014). The role of discourse-level expectations in non-native speakers' referential choices. In W. Orman & M. J. Valleau (Eds.), *Proceedings of the 38th Annual Boston University Conference on Language Development* (pp. 179-191). Cascadilla Press.

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.

Grüter, T., Takeda, A., Rohde, H., & Schafer, A. J. (2018). Intersentential coreference expectations reflect mental models of events. *Cognition, 177*, 172-176.

Hartshorne, J. K. (2014). What is implicit causality? *Language, Cognition and Neuroscience, 29*, 804–824.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–135.

Hopp, H. (2010). Ultimate attainment in L2 inflectional morphology: Performance similarities between non-native and native speakers. *Lingua, 120*, 901–931.

Hopp, H. (2015). Semantics and morphosyntax in L2 predictive sentence processing. *International Review of Applied Linguistics in Language Teaching, 53*, 277–306.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626*, 118–135.

Huettig, F., & Brouwer, S. (2015). Delayed anticipatory spoken language processing in adults with dyslexia—Evidence from eye-tracking. *Dyslexia, 21*, 97–122.

Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. *Brain Research, 1706*, 196-208.

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience, 31*, 80–93.

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience, 31*, 19–31.

Huettig, F., & Pickering, M. J. (2019). Literacy advantages beyond reading: Prediction of spoken language. *Trends in Cognitive Science, 23*, 464-475.

Ito, A., Corley, M., & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism, Language and Cognition, 21,* 251–264.

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

Kaan, E., Dallas, A., & Wijnen, F. (2010). Syntactic predictions in second-language sentence processing. In J.-W. Zwart & M. de Vries (Eds.), *Structure preserved. Festschrift in the honor of Jan Koster* (pp. 207-213). John Benjamins.

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

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

Kim, H. (2018). *Cross-linguistic activation in Korean L2 learners' processing of remention bias in English* (Doctoral dissertation, University of Hawai'i, Honolulu, U.S.A.). Retrieved from http://hdl.handle.net/10125/62788

Kim, K., Grüter, T., & Schafer, A. J. (2013) Effects of event-structure and topic/focus-marking on pronoun reference in Korean. Poster presented at the 26th annual CUNY conference on human sentence processing, Columbia, SC.

Koornneef, A., Dotlačil, J., van den Broek, P., & Sanders, T. (2016). The influence of linguistic and cognitive factors on the time course of verb-based implicit causality. *The Quarterly Journal of Experimental Psychology, 69*, 455–481.

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

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

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44*, 325–343.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science, 18*, 193 – 198.

Madden, C. J., & Ferretti, T. R. (2009). Verb aspect and the mental representation of situations. In W. Klein & P. Li (Eds.), *The expression of time* (pp. 217–240). Mouton de Gruyter.

Magliano, J. P., & Schleich, M. C. (2000). Verb aspect and situation models. *Discourse Processes, 29*, 83–112.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 843–847.

Martin, C. D., Thierry, G., Kuipers, J., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language, 69*, 574-588.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception, & Psychophysics, 53*, 372–380.

Mishra, R. K., Singh, N., Pandey, A., & Huettig, F. (2012). Spoken language-mediated anticipatory eye movements are modulated by reading ability: Evidence from Indian low and high literates. *Journal of Eye Movement Research, 5*, 1–10.

Mitsugi, S. (2018). Generating predictions based on semantic categories in a second language: A case of numeral classifiers in Japanese. *International Review of Applied Linguistics in Language Teaching.* (published online 2018-07-03). https://doi.org/10.1515/iral-2017-0118

Mitsugi, S., & MacWhinney, B. (2016). The use of case marking for predictive processing in second language Japanese. *Bilingualism: Language and Cognition, 19*, 19–35.

Pearson (2011). *Versant English Test: test description and validation summary.* Palo Alto, CA: Pearson Knowledge Technologies. (http://www.versanttest.com)

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 144*, 1002-1044.

Schlenter, J. (2019). *Predictive language processing in late bilinguals: Evidence from visual-world eye-tracking.* (Doctoral dissertation, University of Potsdam, Potsdam, Germany). Retrieved from https://doi.org/10.25932/publishup-43249

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

Ueno, M., & Kehler, A. (2016). Grammatical and Pragmatic Factors in the Interpretation of Japanese Null and Overt Pronouns. *Linguistics, 54*, 165-1222.

Van Berkum, J. J. A. (2010). The brain is a prediction machine that cares about good and bad – Any implications for neuropragmatics? *Italian Journal of Linguistics, 22*, 181-208.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 443–467.

Wlotko, E. W., & Federmeier, K. D. (2012). Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension. *Psychophysiology, 49*, 770–785.

[1] The first 500 ms after the offset of the context sentence were excluded in order to allow for sentence wrap-up effects and for fixations to the last mentioned entity, the Theme, to dissipate.

[2] This was further confirmed by inspection of the distribution of model residuals, which was also more normal for models of GoalAdvantage scores versus models based on (weighted) empirical logits.

[3] If anything, the fact that looks to Source and Theme combined constitute an overall greater proportion of all fixations in the L2 vs. the L1 data should increase the power to detect modulations of these looks by Aspect in the L2 data.

[4] Prompted by a question from a reviewer regarding potential L1 influence, we applied the same model to the data from the largest L1 subgroup in our L2 sample, i.e., L1-speakers of Chinese ($N = 24$). The pattern of results was the same as that reported in Table 2, with the exception that the intercept was fully significant, $b = 5.982$; $p = .015$; Aspect was not significant, $b = -0.774$; $p = .82$.

[5] Data from the two participants with missing Versant scores are omitted from this model, thus $N = 50$.