

Title: Is there a cost when predictions are not met? A VWP study investigating L1 and L2 speakers

Short title: Is there a cost when predictions are not met?

Leigh B. Fernandez¹, Lauren V. Hadley², Aybora Koç¹, John C.B. Gamboa¹, and Shanley E.M. Allen¹

¹ University of Kaiserslautern-Landau (RPTU), Kaiserslautern, Germany

² Hearing Sciences - Scottish Section, School of Medicine, University of Nottingham, UK

Corresponding author:

Leigh B. Fernandez

Postfach 3049

67653 Kaiserslautern Germany

leigh.fernandez@rptu.de

Phone: +49(0)631-205-4138

Fax: +49(0)631-205-5182

Research has found that both L1 and L2 speakers make predictions about upcoming linguistic information, with predictive behavior being impacted by individual differences and methodological factors. However, it is not clear if a cost is incurred when a prediction is made, but not met. L2 speakers have less experience with their L2 and parsing can be cognitively demanding, which together may lead L2 speakers to incur prediction costs differently relative to L1 speakers. In this study using the visual world paradigm, we test whether L1 and L2 speakers predict in the same way, within the same time frame, and incur the same costs if predictions are not met. We also explore the role of proficiency and speech rate. We found that both groups predict in a similar way and within a similar timeframe. Additionally, neither group incurred a prediction cost when the target was the most likely alternative, though L2 speakers take longer to shift their attention to the target object when predictions are not met. We argue that this reflects a slowing of lexical access rather than a specific cost of prediction. We only found prediction differences when speech rate was included in the analysis, highlighting the importance of attending to speech rate in studies using the visual world paradigm. Overall, this study supports research showing that both L1 and L2 speakers may make multiple partial predictions about upcoming information rather than predicting one specific lexical candidate while inhibiting less likely lexical candidates.

Keywords: Visual World Paradigm; prediction; prediction costs; bilingualism

A growing body of psycholinguistic research on language comprehension has found that first language (L1) speakers are not only processing words in real time, but are also actively making predictions about upcoming linguistic information (for reviews see Ferreira & Chantavarin, 2018; Huettig 2015; Huettig & Mani, 2016; Kamide 2008; Staub, 2015; Ryskin & Nieuwland, 2023). When it comes to second language speakers (L2), the picture has been a bit more opaque, but the current consensus is that L2 speakers are indeed capable of making predictions, particularly semantic predictions (for reviews see Hopp, 2022; Kaan & Grüter, 2021; Schlenter, 2022). When prediction differences arise between L1 and L2 speakers, they most likely stem from individual differences (e.g., proficiency) or methodological factors (e.g., speech rate) (see Schlenter, 2022). However, there is less research that explores what happens when a prediction is made but not met. In other words, is there a cost for L1 and/or L2 speakers when a prediction is not correct? In such a situation, a possible target has been activated, and when an alternative is heard, the original prediction must be inhibited and the new target comprehended. In this study, we investigate L1 and L2 prediction and its potential cost. Currently, it is unclear whether L1 and L2 speakers activate specific or broad lexical information when making predictions. If it is the former, a measurable cost is more likely (given that the specific prediction is not met) and if it is the latter, a measurable prediction cost is less likely (given that there may be several active potential continuations). We believe that L1 and L2 speakers may differ in the way they ultimately recover from a prediction that is not met, for example, L2 speakers may activate broader lexical information due to their uncertainty about upcoming information (thus minimizing prediction costs) relative to L1 speakers who may more narrowly activate upcoming information (Peters et al., 2018). To investigate that, we use the method of eye-tracking during listening, additionally considering the role of proficiency and speech rate.

During comprehension, predictions can be made at different linguistic levels such as syntactic structure, phonological form, meaning, etc. (Huettig et al., 2022; Pickering & Gambi, 2018) and furthermore, predictions at different levels can interact (Heilbron et al, 2022). In terms of predicting meaning, the constraining context *The tailor trimmed the...* is more likely to lead listeners to predict a plausible continuation such as *suit* relative to a neutral context sentence (e.g., *The guardian sells the....*). Research so far has yielded mixed results concerning whether there is a cost when hearing a constraining context but encountering an unpredictable but plausible word (e.g., *The tailor trimmed the tree*). In order to investigate this question, research typically compares performance in the three conditions just introduced: (1) a constraining context with predictable target word (CP), (2) a constraining context with an unpredictable target word (CU), and (3) a neutral context with the predictable target word from the CP condition (NP). We illustrate these in Table 1 with examples of the sentence types tested in the current study and the nomenclature that will be used throughout the manuscript (Frisson et al., 2017).

Table 1. Example sentences and nomenclature (target word underlined)

Condition (abbreviation)	Example sentence
Constraining context – predictable target word (CP)	The tailor trimmed the <u>suit</u> .
Constraining context – unpredictable target word (CU)	The tailor trimmed the <u>tree</u> .
Neutral context – with the same predictable target word (NP)	The guardian sells the <u>suit</u> .

L1 literature

A rapidly growing literature has investigated prediction cost in L1 speakers, with findings differing somewhat depending on the methodology used. Thus, we review these studies according to the three common methodologies: event-related potentials (ERPs)- whereby a larger late frontal positivity suggests a prediction cost, eye-tracking while reading – whereby slower

reading times suggest a prediction cost, and eye-tracking using the visual world paradigm - whereby delays in looks to the targets suggest a prediction cost. In this paper, we use the expression *prediction cost* to refer to the cost of recovery when a prediction is not met (not to be confused with the cost of *producing* a prediction).

The first set of studies investigating L1 prediction cost focuses on brain activity as assessed through ERPs, and yields mixed findings. The N400 component has been argued to reflect the ease of accessing semantic information about upcoming words (e.g., Delong et al., 2005) and has been found to inversely relate to how predictable a word is within a context (i.e., the more predictable the word, the smaller the N400 amplitude; e.g., Kutals & Hillyard, 1985; Delong et al., 2005). This suggests that the facilitation of a predicted word is related to the amount of lexical information that is predictively preactivated. Relatedly, several studies with L1 speakers have shown no differences in N400 amplitudes when processing a target word in CU sentences relative to NP sentences (e.g., Federmeier et al., 2007; Kuperberg et al., 2020). This suggests that processing a target with no semantic preactivation is similar regardless of the contextual frame; i.e., an unexpected target word is no easier/harder to process when it is in a constraining context (with a particular unmet target prediction) or a neutral context (with no particular target prediction). However, several studies have found a late frontal positivity when encountering a target word in CU sentences relative to NP sentences (e.g., Delong et al., 2012; Delong et al., 2014; Federmeier, et al., 2007; Kuperberg et al., 2020), which may reflect the cost of updating “situation model” representations (i.e., high-level meaning representations that describe a full situation; Kuperberg et al., 2020) and/or inhibiting incorrect predictions (Kuperberg et al., 2020).

Recently, Brothers et al. (2023) tested several sentence types and found smaller N400 amplitudes to CU target words that were semantically related to the expected target, suggesting that preactivated semantically related words facilitate one another. Furthermore, Brothers et al. found no evidence of late frontal positivity when integrating a CP target word relative to a target word that was the second most predicted target (based on pre-testing) into a constraining context, which supports the notion that there is no cost in processing an unexpected but semantically activated alternative. However, they did find a larger late frontal positivity to CU items where the target word was not predicted at all (based on pre-testing) relative to NP items, suggesting that the late positivity may index larger situation level updates (in the face of unexpected but plausible words) rather than lexical inhibition. Together, this suggests that multiple semantically related potential alternatives are predicted, which facilitate rather than inhibit each other, but large situation level updates may lead to measurable late costs (i.e., the better the predictions, the smaller the recovery cost).

Research using eye-tracking while reading has not found evidence of prediction costs for L1 readers. For example, Luke and Christianson (2016) had participants read text passages in which the cloze probability of each word was calculated. They found clear predictability effects across several reading measures, but interestingly when a sentence contained a word that was not the most frequently predicted by the cloze test (i.e., CU sentences), it did not evoke a measurable cost (i.e., reading measures were similar regardless of whether the word was expected or not). Additionally, they found that as a sentence became more constraining, unpredictable words were actually facilitated. However, as Frisson et al. (2017) point out, Luke and Christianson's stimuli contained relatively few highly predictable words (5%) thus limiting the power to assess prediction cost. Therefore, Frisson et al. designed a reading study that carefully controlled for

contextual constraint and predictability. Interestingly, they also found no evidence for prediction costs in CU sentences, and found a facilitation for a CU target that was semantically related to the predictable target (i.e., if the most expected continuation for the sentence *The priest wondered how to get more people to come to the* is *church*, a semantically related but less predicted continuation would be *sermon*). Taken together, this suggests that the parser may activate multiple partial predictions that could easily fit the given contextual frame without inhibiting less likely words (e.g., both *church* and *sermon* are activated in the above example), as opposed to predicting one specific lexical candidate that is the most likely continuation of the contextual frame (e.g., only *church* is activated in the above example).

Evidence that revising predictions during listening is an instantaneous process has been found using the Visual World Paradigm (VWP) with L1 speakers of Mandarin Chinese (Chow & Chen, 2020). Mandarin Chinese uses nominal classifiers. Most of these can only be used with specific nouns, but a small set are general and can be used with many nouns. Employing this latter group of classifiers, the authors created sentences in which the classifier preceding the target word was compatible with the expected target or not, thus allowing them to test whether the mismatch between a classifier and the expected noun leads to a prediction cost. For example, participants would view a leaf, bird, candy, and flower while hearing something similar to *While playing in the garden the boy gave the girl one very beautiful...* (in Mandarin) with the expected target being *flower* and the unexpected target being *leaf*. The classifiers were either specific and matching the target (*flower*), specific and matching the unexpected target (*leaf*), or general (thus allowing it to match with either *flower* or *leaf*). They found that participants would look to the expected target (*flower*) prior to hearing a classifier (thus predicting the upcoming word), but upon hearing the specific classifier matching the unexpected target (*leaf*) would, within a few

hundred ms, fixate towards the *leaf*. This type of design may be somewhat different to the previously described research because the unexpected classifier does not necessarily negate a prediction, but may signal that a prediction is likely to be incorrect and revision is necessary. Given that the prediction mismatch was immediately detected and attention was shifted to the ultimately correct object, the authors argue that this indicates instantaneous revision (and may indicate a lack of prediction cost). However, this may have been driven by the paradigm itself (given the limited number of possible referents); we will return to this point in the discussion. It should be noted that the statistics used in this study did not take into account potential non-linear patterns (often present in time-based data) and the authors were not able to pinpoint exact times at which looks to the objects diverged. In the current study we employ the VWP but use general additive mixed effects models (GAMMs) to deal with non-linearity, and divergence point analysis (DPA; Stone et al., 2020) to find an estimate of effect onset.

In sum, while L1 research on prediction cost using ERP is somewhat mixed, recent studies show that multiple interpretations are preactivated that facilitate one another, particularly when the unexpected target is semantically related to the expected target. But there may be a late cost for larger situation level updates when encountering an unexpected but plausible target word. Additionally, data from eye-tracking suggests that there are no prediction costs in language processing for L1 speakers either during listening or reading. L1 speakers quickly make many broad and partial predictions about upcoming lexical information, and if an unexpected word is encountered there does not seem to be a cost to revise and integrate the unexpected word into the sentence.

L2 literature

For L2 speakers, the general consensus for prediction is that L2 speakers are able to generate predictions similar to L1 speakers (Hopp, 2022; Kaan & Grüter, 2021; Schlenter, 2022), with any differences stemming from individual differences (e.g., proficiency, strength of stored frequency information, quality of lexical representation, processing strategies) and/or methodological factors (e.g., speech rate, time constraints). During bilingual processing, parsing a less-used second language places additional demands on cognitive processing during language comprehension (e.g., Corps et al., 2023; Ito & Pickering, 2021; Segalowitz & Hulstijn, 2009), and additionally L2 speakers have less experience with their L2, which together may lead L2 speakers to incur prediction costs differently (or to a different extent) relative to L1 speakers.

Like the L1 research, prediction cost findings for L2 speakers have also differed depending on the method used. The findings of research using ERPs have been somewhat mixed. Martin et al. (2013) found that while L2 speakers showed an effect of prediction cost when an unexpected target word was encountered, contrary to L1 speakers, they did not show evidence of this cost when they encountered the preceding article that did not agree with the expected target¹. However, more recently, Foucart and colleagues (e.g., Foucart, et al., 2014; Foucart, et al., 2015; 2016) found that L2 speakers showed a potential cost in the form of an increased negativity when they encountered an article (preceding the target) that either agreed with a constraining target word or not. Similar to the Chow & Chen study, it is currently unclear whether this type of mismatch, involving an article prediction, evokes the same cost as evoked by CU sentences, where the encountered word directly negates a prediction. However, Zirnstein et al. (2018) tested prediction costs in CU sentences with L1 and L2 speakers of English using ERP while also

¹ Note that reproducibility of this finding has been questioned (see Nieuwland et al., 2018).

testing several individual differences. They found evidence that L1 speakers of English and L2 speakers of English (L1 Mandarin) employ similar prediction mechanisms. However, these effects were modulated by individual differences, particularly inhibitory control and L2 fluency (which may explain the discrepant findings in the previous literature).

The only L2 study we are aware of using self-paced reading also suggests a prediction cost for L2 speakers. Feng and Jiang (2023) used self-paced reading to investigate prediction error costs in Chinese with L1 Chinese and L2 Chinese speakers (L1 English). They found that both groups showed a prediction cost in terms of increased reading time of CU target words relative to the same target word in a neutral context, though the L2 group was overall slower. Therefore, they argued that both L1 and L2 speakers incur a cost when encountering an unexpected word in a constraining context, and the prediction mechanisms are the same across both groups.

Finally, L2 research with the VWP also suggests a prediction recovery cost for lower skilled participants (i.e., participants who identified as an L2 speaker or participants who scored low on a vocabulary test) relative to higher skilled participants (i.e., participants who identified as an L1 speaker or participants who scored high on a vocabulary test). In two interesting VWP experiments, Peters et al. (2018) investigated global and local predictions and prediction costs in lower and higher skilled speakers of English. In their first study participants heard a predictable sentence like *The pirate chases the ship* while viewing 4 images: the target (*ship*), an agent-related object (*treasure*), a verb-related object (*cat*) and an unrelated distractor (*bone*). They found that both groups made similar global predictions, that is, looks to the target *ship* before it was explicitly mentioned. They also found that both groups made local predictions, that is looks to the *cat* following *chases* despite the fact that *cat* is not likely to be the target given the agent of

the sentence (*pirate*). Additionally, they found that the lower skilled participants made more local predictions relative to the higher skilled participants. They argued that lower skilled participants may adaptatively activate broader lexical information given their uncertainty about upcoming information. If this is indeed the case, they hypothesized that lower skilled speakers may actually have reduced prediction costs relative to higher skilled speakers. To test this, they designed similar items (using the same images) in which the target word was either locally verb-related (e.g., *The pirate chases the cat*) or unrelated (e.g., *The pirate chases the bone*). As expected, they found that both groups looked to the target (the last word in the previous examples) more quickly in the locally verb-related items relative to the unrelated items. However, they found a small marginally significant effect indicating that higher skilled speakers were actually more likely to fixate on the locally verb-related object than lower skilled speakers. From these two experiments the authors argued that higher skilled speakers are able to flexibly modulate predictions based on the task (i.e., the reliability of the predictions within the experiment) relative to lower skilled speakers who make many weak predictions regardless of task. However, in this study the effect was marginal, and only local prediction costs were tested. It may be the case that during incremental processing local predictions do not incur a large cost (for example due to time-constraints).

To address the issue of confounding language with individual differences, we investigate English language proficiency across both groups (using the Oxford Placement Test-Part A; OPT) since some research has found that L2 speakers show better predictive abilities as proficiency increases (e.g., Chambers & Cooke, 2009; Dussias et al., 2013; Hopp, 2013). However, not all research has found that prediction abilities increase with proficiency (e.g., Dijkgraaf et al., 2017; Hopp, 2015; Ito et al., 2018; Kaan & Grüter, 2021; Kim & Grüter, 2020; Mitsugi, 2020;

Perdomo & Kaan, 2019). In the current study we use the VWP, given that this method provides real-time insight into the coupling of language processing and shifts of attention during listening. In addition, we investigate speech rate which varies in the current study, given that L2 speakers often identify speech rate as the source of greatest difficulty during comprehension (Graham, 2006). Additionally, in a recent study investigating the role of speech rate and prediction using the VWP, Fernandez et al. (2020) found that L1 speakers' predictive eye movements increased as speech rate increased from 3.5 to 5.5 syllables per second relative to L2 speakers who only showed predictive eye movements at 3.5 syllables per second.

Current study

Overall, evidence from both the L1 and L2 literature is mixed as to whether a prediction cost is incurred when encountering an unexpected target word in a constraining context. We believe the results from the current study will not only shed light on L1 and L2 prediction processes, but will also help elucidate between different accounts of prediction. If either group exhibits a cost when an unexpected target word is encountered (in a constraining context), this supports the account that, as context unfolds, only one specific lexical item (or the most likely item available to them) is activated while other less likely lexical items are ruled out and inhibited. Upon encountering the unexpected target word, the listener must then revise their prediction (e.g., activating the previously inhibited lexical items), leading to a measurable cost (e.g., Federmeier et al., 2007; Feng & Jiang, 2023; Foucart et al., 2014; 2015; 2016). On the other hand, if either group shows no cost when an unexpected target word is encountered (in a constraining context), this supports the account that as context unfolds, multiple lexical items are activated (and remain active) and as long as the target word is plausible there will be no cost

when the unexpected target is encountered (e.g., Chow & Chen, 2020; Frisson et al., 2016; Luke & Christianson, 2016).

In the current study we aim to test whether L1 speakers of English and sequential L2 speakers of English (with an L1 of German) incur a cost when a prediction is not met. Importantly, we control for both proficiency and speech rate by including both as predictors in our models. We split our analyses into two: first we investigate prediction in terms of eye movement behavior and timing of looks to the target in NP items. Second we turn to prediction costs, which we explore in two ways. Since cost can be understood to relate to the need to inhibit a previously activated target we analyze resolution of prediction errors in terms of eye movement behavior and timing of looks to the target in CU items. This analysis will particularly inform whether there are differences between L1 and L2 speakers in terms of reconciling an incorrect prediction. Finally, in order to assess whether making an incorrect prediction is more detrimental than not predicting at all, we analyze timing of looks to the target in CU items in comparison to NP items. This analysis will particularly inform whether there is a cost when making an incorrect prediction relative to not making a prediction at all. Eye movement behavior is analyzed using General Additive mixed models (described in detail in the analysis section) and timing using Divergent Point Analysis (described in detail in the analysis section).

In our first set of comparisons investigating prediction, we want to establish whether L1 and L2 speakers predict in the same way and within the same time frame when hearing a predictable sentence with a predictable target (CP items). We hypothesize that both L1 and L2 speakers will show similar patterns of eye movement behaviors and will predict (i.e., they will look to the target before it is explicitly spoken) within the same time frame when listening to CP items, though L2 speakers may be more impacted by proficiency and speech rate.

In our second set of comparisons testing prediction cost, we want to establish whether L1 and L2 speakers incur a prediction cost when hearing an unexpected word in a constraining context (CU items). To this end, we first look at the pattern of eye movement behaviors and timing while listening to CU items. In terms of the eye movement behavior, we hypothesize that if there is a prediction cost, looks to the expected (but incorrect) target will linger after the actual target has been heard, with shifts of attention occurring after the target word is uttered in its entirety. This comparison will also allow us investigate whether L1 and L2 speakers show similar or different patterns in the face of encountering an unexpected target, and whether they are similarly impacted by speech rate and proficiency. In terms of the timing of looks to the unexpected target while listening to CU items, we hypothesize that if either group takes longer to shift their attention this may be due to a larger prediction cost (indicating the time it takes to inhibit the incorrect target and facilitate the correct target).

In our second prediction cost analysis, we compare the timing of looks to the unexpected target in the CU items relative to the target in the NP items. In the case of the NP condition the context does not constrain the listener to a particular interpretation, while the CU condition does constrain the listener towards a particular interpretation, which is ultimately incorrect. We hypothesize that if a cost is incurred in terms of delayed looks, when a prediction is made (and ultimately not met), the cost will be apparent relative to a sentence where no prediction was made (i.e., NP items). We believe that this will provide the most compelling evidence of whether there is a prediction cost.

Methods

Participants

L1 English

Fifty participants were recruited from the University of Alberta (Canada). However, four were excluded due to exposure to a second language from birth. The remaining 46 participants reported being monolingually-raised L1 speakers of English who did not learn a second language before the age of five years. No participant reported a hearing problem and all had normal or corrected-to-normal vision. Participants were given course credit for their participation. See Table 2 for additional participant information.

L2 English

Forty-five participants were recruited from the University of Kaiserslautern-Landau. All participants reported being monolingually-raised L1 speakers of German who did not learn a second language before the age of five years and learned English through school as well as media and online sources. No participant reported a hearing problem and all had normal or corrected-to-normal vision. Participants were given course credit or 8 Euro for their participation. See Table 2 for additional participant information.

Table 2: Participant information

L1	N	Male/Female	Mean age	Mean OPT (in English)	Mean age of English acquisition
English	46	10/36	20.28 (SD=3.54)	91.17 (SD=5.74)	NA
German	45	17/28	25.75 (SD=4.86)	77.96 (SD=10.59)	9.55 (SD=1.84)

Our sample size was based on previous VWP research that has investigated L1 and L2 predictive processing, particularly Corps et al. (2022; 2023), as well as a recent semantic prediction study using webcam-based eye tracking finding that 20-30 participants (with 16 items per condition)

was sufficient to obtain 80% power (using two different power analysis approaches; Prystauka et al., 2023).

Materials




The stimuli included 107 trials: 3 practice trials, and 104 experimental items divided into two blocks (of 52 items each). There were 72 critical items (24 CP/24 NP/24 CU) and 32 fillers (which served as experimental items in a different experiment not reported here). All critical items followed the same structure: *The* [agent] [verb] *the* [critical word], all critical items are provided on OSF (<https://osf.io/3v7sd/>). The practice items and fillers followed the same structure as the critical items and the fillers consisted of 16 predictable (e.g., *The mouse nibbles the cheese*) and 16 neutral context items (e.g., *The girl touches the cauldron*).

Forty-eight trios of critical sentences were created (see Table 3). Each trio included three conditions: a constraining context with a predictable target word (CP), a constraining context with an unexpected target word (CU), and a neutral context with a predictable target word (NP). For presentation to participants, the 48 items were divided in half: list A and list B. One half of the participants saw list A's CP stimuli and list B's CU and the corresponding NP stimuli; the other half of the participants saw the opposite. Note that since the CU and NP stimuli showed no linguistic overlap they were both presented to the same participants, but if the participant saw the CU in the first block, they would see the corresponding NP in the second block (or vice versa). Therefore, each participant saw 72 critical items, 36 in each block, with each block consisting of 12 CP, 12 CU, and 12 NP items.

To further detail the stimuli, the CP and NP items were manipulated such that the critical word was the same across both sentence types, but was only predictable based on the preceding information in the CP condition. The CU items were manipulated such that the sentence

beginning was the same as the CP item, but the critical word was unexpected (but plausible) given the preceding context. For example, while it is likely that a tailor would trim a suit (CP), it is possible that a tailor could trim a tree (CU). See Table 3 for an example item, the Zipf frequency² of the agent, verb, and critical word, and the mean syllable count per word.

Table 3: Example stimuli from novel sentences and item information (standard deviation in parentheses)

Condition (abbreviation)	Example stimuli		Overall item information	
	Example sentence	Corresponding visual array	Mean Zipf frequency (agent, verb, & critical word)	Mean Syllable count
Constraining context – expected target word (CP)	The tailor trimmed the <u>suit</u> .		4.01 (0.77)	1.81 (0.77)
Constraining context – unexpected target word (CU)	The tailor trimmed the <u>tree</u> .		4.10 (0.78)	1.92 (0.84)
Neutral context – predictable target word (NP)	The guardian sells the <u>suit</u> .		3.90 (0.70)	2.74 (1.27)

Each sentence type had a corresponding visual array consisting of four images. For the example in Table 3, this included a CP and NP target image (*suit*), a CU target image (*tree*), and two distractor items (*jar*, *pot*) which are potential NP objects but not CP or CU objects (i.e., both the pot and jar are sellable but are not trimmable). Additionally, to avoid phonological overlap, none of the words for the images shared initial phonemes. All images were greyscale 300x300 pixel jpeg line drawings taken from the British English MultiPic databank (Duñabeitia et al., 2018).

² Frequency was derived from the SUBTLEX-UK database (van Heuven et al., 2014) with a Zipf value of less than 3 being considered “low frequency”.

While the linguistic content was different across corresponding CU and NP items, the images in the array were the same. Therefore, the objects in the array were randomized for the corresponding CU item (to ensure participants did not map the images to the same location across items). Note that the image location was identical across CP and NP items (though no participants saw the same CP and NP item), but image locations were rotated such that each image type occurred in each location 25% of the time. Pretesting ensured that the CP (and CU) target was selected over 97% of the time while the NP target was selected 25.05% of the time; for pre-testing information about the sentences and images, see Appendix A.

Auditory information

Sentences were recorded by a male L1 speaker of Scottish English using a Blue Yeti USB microphone at a 48,000 Hz sampling rate over Audacity® recording software (Audacity Team, 2021). A click track at 90 beats per minute was used to ensure that words of different syllable lengths were spoken within the same time frame (for ease of post-hoc acoustic manipulation). For ease of time locking the auditory stimuli, the duration of each word in the stimulus was set equal to the mean of that word across all items using Praat (Boersma & Weenink, 2021). While this may have led to slightly unnatural speech, given that the recordings fall within the typical speech rate range (typical range is variably reported anywhere from 2.5-8.0 syllables per second; e.g., Dickey et al., 2007; Hertrich et al., 2013), we do not believe this to be a major concern (though future research investigating the impact of normalization on prediction would be useful). Thus, all words across items were the same duration; see Table 4 for normalized word duration information. Given that each recording was the same length but the content varied, the speech rate of the items was not consistent; the mean speech rate across items was 3.47 (SD = 0.77) syllables per second (range 2.56 – 5.65 syllables per second).

Table 4: Normalized word durations (ms)

	1st The	Agent	Verb	2nd The	Object
Mean (ms)	93.58	612.72	602.05	130.27	464.45

The Language and Social Background Questionnaire (LSBQ)

The LSBQ (Anderson et al., 2018) was used to assess the language background of the participants.

Oxford Placement Test (OPT)

For a measure of English proficiency, the OPT (Part A) was used. This test is comprised of 50 items and assesses English morpho-syntactic knowledge. Participants read sentences and selected the most appropriate sentence continuation from three options. The score was converted into a percentage.

Apparatus

For all participants, stimulus presentation was programmed using Experiment Builder, and eye movements were recorded using an Eyelink 1000 sampling at 1000 Hz. Viewing was binocular but only the right eye was recorded, and the head was stabilized using a chin rest. Audio materials were presented via Philips Bass+ on-ear head phones. L1 participants sat approximately 50 cm from a 20' Dell monitor (model 2009W1) with a 1024 x 768 resolution and 60Hz refresh rate. L2 participants sat approximately 85 cm from a 19' Dell monitor flat screen cathode ray tube (model P1130) with a 1024 x 768 resolution and 60Hz refresh rate.

Procedure

The procedure was identical for both groups. The experiment began with participants providing their informed consent. This was followed by the eye-tracking task. Participants then completed the LSBQ and then the OPT. The study took approximately 45 minutes.

The eye-tracking task began with the standard EyeLink 9-point calibration procedure. The eye-tracking study was self-paced in that participants could take a break between trials as necessary (with a subsequent recalibration). Additionally, there was a mandatory break halfway through the study with a recalibration. The instructions for the study were provided on the screen and were verbally explained by the experimenter. Participants were told they would hear a short sentence accompanied with a visual array of 4 images, and their task was to identify the object that they believed best matched the sentence by clicking on the image using the mouse. Additionally, they would have to wait to click on the object until after the entirety of the sentence was played and a green border appeared around the array of images. They were also instructed that there was no time limit.

All trials began with a drift correct in the center of the screen. The trial proceeded when the participant fixated on the drift correct (fixation dot) and simultaneously pressed the space bar. The start of each trial began with the images being displayed for 2000 ms, after which the auditory stimulus was presented. The images remained on the screen throughout the auditory stimulus and then for an additional 2000 ms, at which point a green border appeared around the array of images and the mouse icon appeared on the screen. Participants then had to click on one of the images, which ended the trial. The drift correct would then appear indicating the start of a new trial.

Analysis

As mentioned earlier, we make two sets of comparisons in this study: to test whether our participants predicted upcoming targets, we analyzed data from the CP items and to test whether our participants incurred prediction cost, we analyzed data from the CU and NP items using two methods. For each set of comparisons, we used two statistical approaches: generalized additive

mixed models (GAMM) and divergent point analysis (DPA; Stone et al, 2020). GAMMs are a type of regression that can model non-linear time-course data (e.g., Porretta et al., 2018; Wieling, 2018; Wood, 2017). To investigate the timing of looks for both L1 and L2 speakers we used DPA which can estimate the time in which looks to one object diverge from looks to another object (e.g., at what point in the sentence *The tailor trimmed the suit* do listeners look more to the *suit* than the only other trimmable object *tree*) and allows us to compare across different groups (e.g., do L1 speakers diverge earlier than L2 speakers). Regions of interest were defined as the four objects in the array (300 x 300 pixels) plus an additional 50 pixels on all sides. All analyses were conducted using R (R core team, 2018) and all data and code is publicly available on OSF (<https://osf.io/3v7sd>).

GAMM analysis

Given that we were interested in the pattern of fixations across time, we used the empirical logit of fixation counts (Barr, 2008) as our main dependent variable in our GAMMs. The empirical logit is the log-odds ratio of looking at the target relative to not looking at the target (i.e., looking at another object in the array) in 20 ms bins, and was weighted in the models to control for eye-movement based dependences (Barr, 2008). We fit two GAMMs. In our first set of comparisons, in which we test whether our participants predicted upcoming targets, we analyzed data from the CP items from the onset of the verb to the offset of the object, plus 200ms (i.e., throughout *trimmed the suit* (+200ms) in the above example). In our second set of comparisons, in which we test whether our participants incurred prediction cost, we analyzed data from the CU items from the onset of the verb to the offset of the object, plus 200ms (i.e., *trimmed the tree* (+200ms) in the above example).

Fixed effects included language (L1/L2) as an ordered factor (with L1 serving as the baseline), the continuous predictor of English language proficiency (as measured through the OPT), and the continuous predictor of speech rate (as measured by syllables per second). All fixed effects were included as parametric components. Additionally, all fixed effects were included as non-parametric smoothed components with the addition of time (converted into ms (bin number * 20)). Three two-way interactions were included as ordered factor difference smooths: time x language, proficiency x language, and speech rate x language. And 2 three-way interactions were included using tensor product smooths: time x proficiency x language and time x speech rate x language. Additionally, we included random smooths (factor smooth interactions) of participant over time with a factor smooth, and a non-linearity penalty over the first derivative (e.g., Baayen et al., 2016; Sóskuthy, 2017). Finally, we included random smooths of item over time with a factor smooth specified by language, and a non-linearity penalty over the first derivative. To account for potential autocorrelation, we checked the residuals using the `acf_resid` function and visualizing the autocorrelation using the *itsadug* package (Van Rij et al., 2020). Autocorrelation was not present in any of our models (see <https://osf.io/3v7sd> for visualization). Before interpretation we checked the effective degrees of freedom (edf) to see whether higher basis dimensions were needed to prevent oversmoothing (Wieling, 2018), and adjusted significant edf values if necessary. Significance testing was done by checking the model output of both the parametric and smooth factors and applying a Bonferroni correction to deal with the increased likelihood of type 1 error from multiple comparisons (8 comparisons were made for each model: $.05/8$ yields a p-value of .006; see Sóskuthy, 2021).

DPA analysis

DPA is a non-parametric bootstrapping approach that deals with the inherent non-independency of fixations while decreasing the likelihood of Type 1 error and estimating confidence intervals (CI), thus allowing us to compare across groups (which is not possible using other approaches such as GAMMs). The divergence point was established with t-tests comparing looks to two objects (as outlined below) across all of the time bins until ten consecutive bins were statistically significant (i.e., 200 ms). Two thousand new data sets were then generated using non-parametric bootstrapping, by resampling the original data using participant, image type, time, and language group categories (see Stone et al., 2021). A new divergence point was estimated for each resampling and then the means and CIs were calculated.

In our first set of comparisons, we test prediction across the two groups. This is achieved by looking only at the CP items, in order to find the point at which looks to the expected target (*suit*) diverge from looks to unexpected target (*tree*). In our second set of comparisons, in which we test whether there is a prediction cost and whether this cost differs across groups, we look only at the CU items and compare fixations to the unexpected target (*tree*) and one of the distractors (*jar*). DPA cannot compute more than one divergence point, and given that CU items are constraining (e.g., *The tailor trims*), we expect looks to the *suit* to diverge from *tree* before the unexpected word is spoken (therefore we cannot calculate the timing of the second divergence point that occurs after hearing *tree*). The logic of comparing *tree* to *jar* is that, if participants are making predictions based on the onset of the sentence (*The tailor trims*), both *tree* and *jar* should be ruled out as potential targets. However, after hearing *tree*, looks should diverge from the previously disregarded *tree* and the previously disregarded *jar* (which will continue being ruled out); thus, indicating the time in which participants integrate *tree* into the sentence. Lastly, to further test prediction costs, we compare looks to the expected target (*suit*)

in the NP items to look to the unexpected target (*tree*) in the CU conditions. The logic of this comparison is that, if there is a cost to making a prediction that is not met (i.e., the CU condition), we should see this manifest relative to hearing the target word in a neutral context (i.e., the NP condition). In other words, if there is a prediction cost, looking to *tree* should take longer after hearing the constraining context (*The tailor trims the*) which led to an incorrect prediction, than looking to *suit* after hearing a neutral context where no prediction was made (*The guardian sells the*).

Results

Accuracy

Incorrectly answered items were removed from analysis. For the CP and NP items, the expected target was clicked in 98.48% of trials (thus 1.52% of trials were removed). For the CU items the CU target was correctly clicked on 88.21% of the trials (thus 11.79% of trials were removed – this consisted of 12.13% of the L1 trials and 11.39% of the L2 trials). Interestingly, in the CU items, the incorrect but predictable item was clicked on 10.70% of the trials. See Figure 1 for fixation proportions of the correctly answered trials. To explore the incorrectly answered CU items, we also visualized the 10.70% of trials in which participants clicked the incorrect predicted target (e.g., they clicked *suit* after hearing *The tailor trims the tree*; see Appendix B).

We will return to this finding in the discussion.

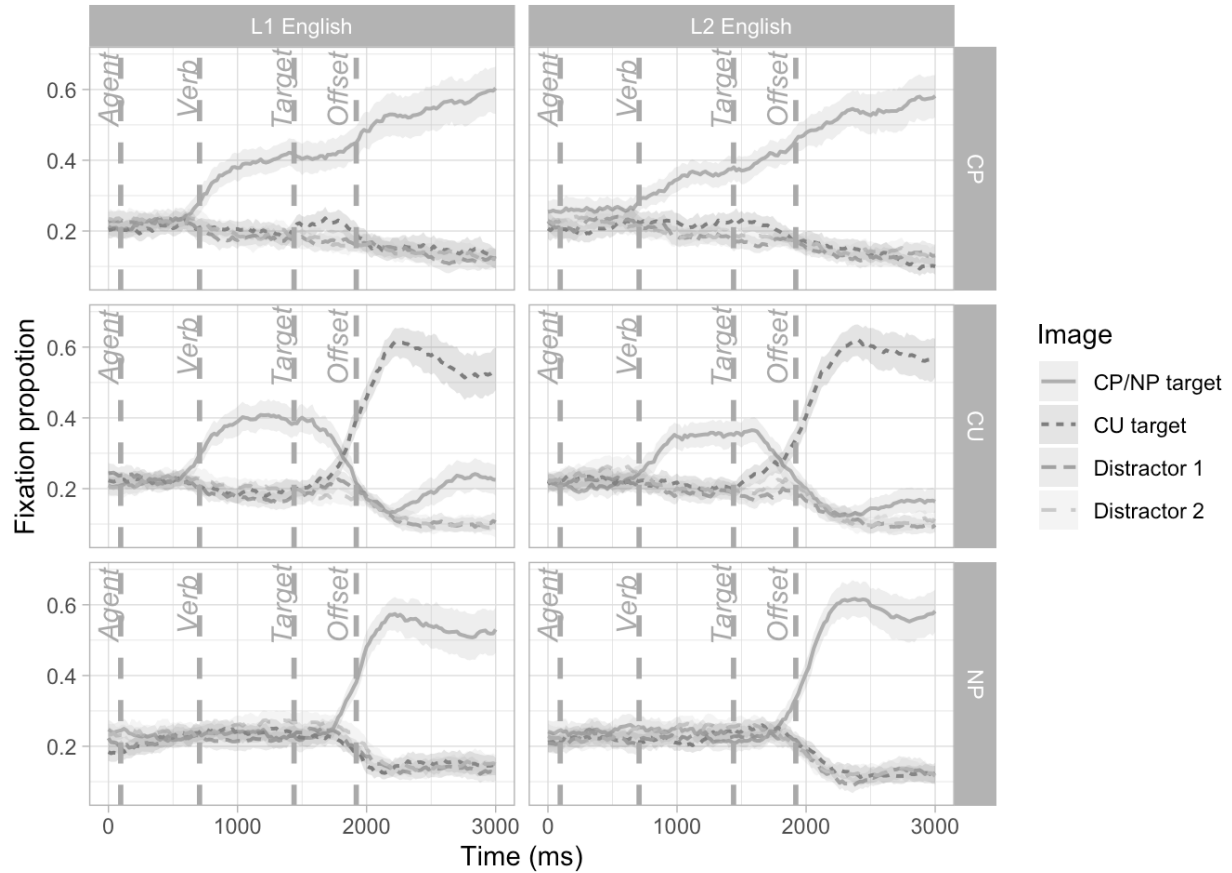


Figure 1. Fixation proportion to all objects across all sentence types for correctly answered items

Prediction comparison

GAMM

To test the overall prediction for both groups, we analyzed the empirical logit in the CP items (looks to the CP target (*suit*) vs. looks to CU target (*tree*)) from the onset of the verb to the offset of the object (+200ms), see Figure 2 for visualization.

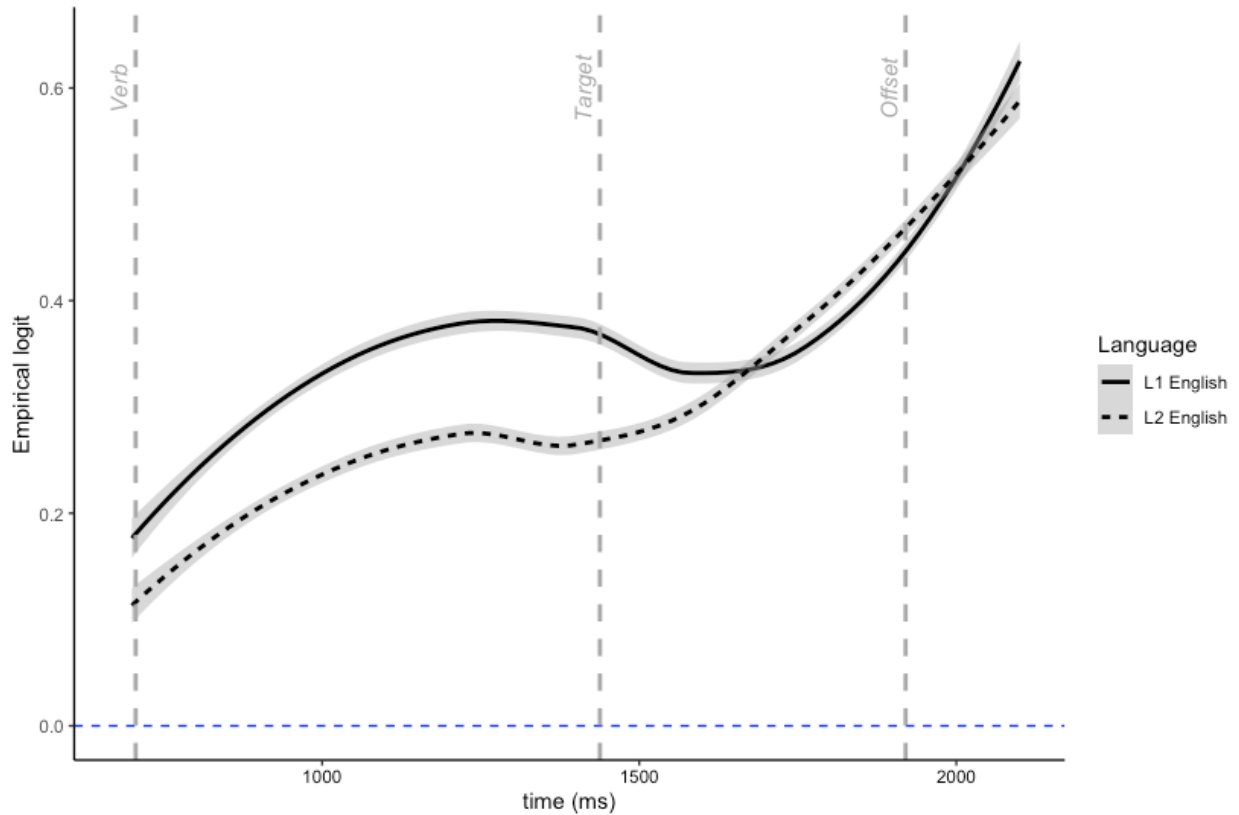


Figure 2. Empirical logit during the prediction time window for CP items (shading represents 95% confidence interval and the dotted horizontal line represents the zero line for the empirical logit)

The results from the GAMM can be seen in table 5.

Table 5. GAMM mixed model output for the empirical logit during the prediction time window for CP items. Part A reports the parametric coefficients and Part B reports the smooth terms.

A. Parametric coefficients	Estimate	Std. error	t-value	p-value
Intercept	.00	.00		
Syllable count	.00	.00		
Language (L2)	0.05	0.06	0.95	0.34
OPT	0.00	0.00	5.88	< .001
B. Smooth terms	EDF	Ref.df	F-value	p-value
s(syllable count)	2.83	2.97	101.84	< .001
s(OPT)	2.09	2.11	1.08	0.34
s(time)	5.07	5.68	5.38	< .001
s(time):Language (L2)	1.00	1.00	1.48	0.22
s(OPT):Language (L2)	1.01	1.01	0.43	0.52
s(syllable count): Language (L2)	2.71	2.93	28.15	< .001

ti(time, OPT):				
Language (L2)	6.21	6.63	0.73	0.64
ti(time, syllable count):				
Language (L2)	8.16	8.85	10.71	< .001
s(time, participant)	580.97	815.00	9.42	< .001
s(time, item):Language (L2)	17.50	17.89	18.68	< .001

In terms of the parametric effects, proficiency (OPT) was significant ($t=5.88$, $p<.001$) with looks to the target (increase in empirical logit) increasing as proficiency increased. In terms of the smooth terms, there was a significant ordered factor difference smooth interaction of syllable count by language ($F=28.15$, $p<.001$), see Figure 3. The summed effect of syllable account for L2 and L1 speakers can be seen in left side of Figure 3, and the difference can be seen on the right side of Figure 3. This reveals that L1 speakers make more looks to the target at the high speech rates relative to L2 speakers.

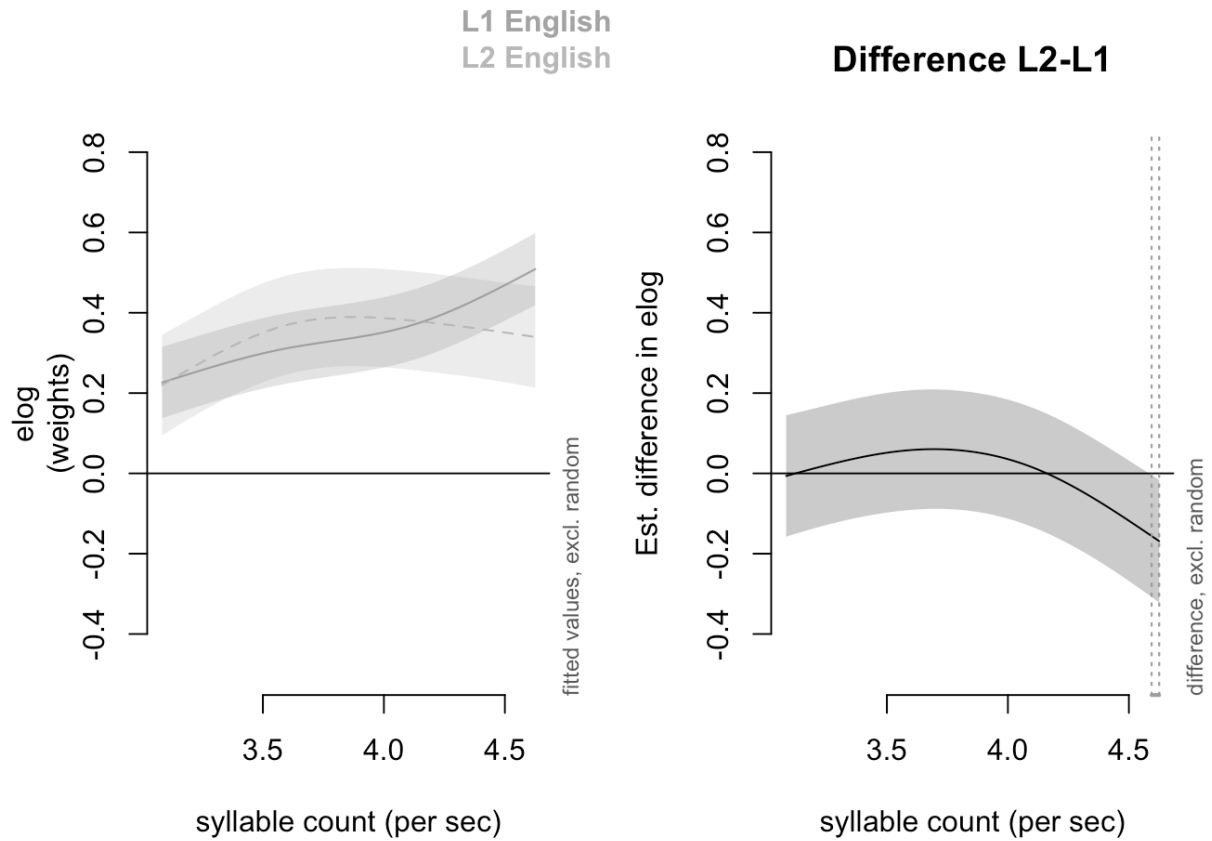


Figure 3. The solid horizontal line represents the zero effect. *Left panel.* Summed effect of syllable count for L1 English speakers (solid line) and L2 English speakers (dotted line). *Right panel.* The difference in the summed effect of syllable count for L1 and L2 speakers of English. The dashed vertical line indicated where there is a significant difference.

There was a significant tensor product smooth interaction between time, syllable count, and language ($F=10.71, p < .001$). This tensor interaction is visualized using contour plots, see

Figure 4.

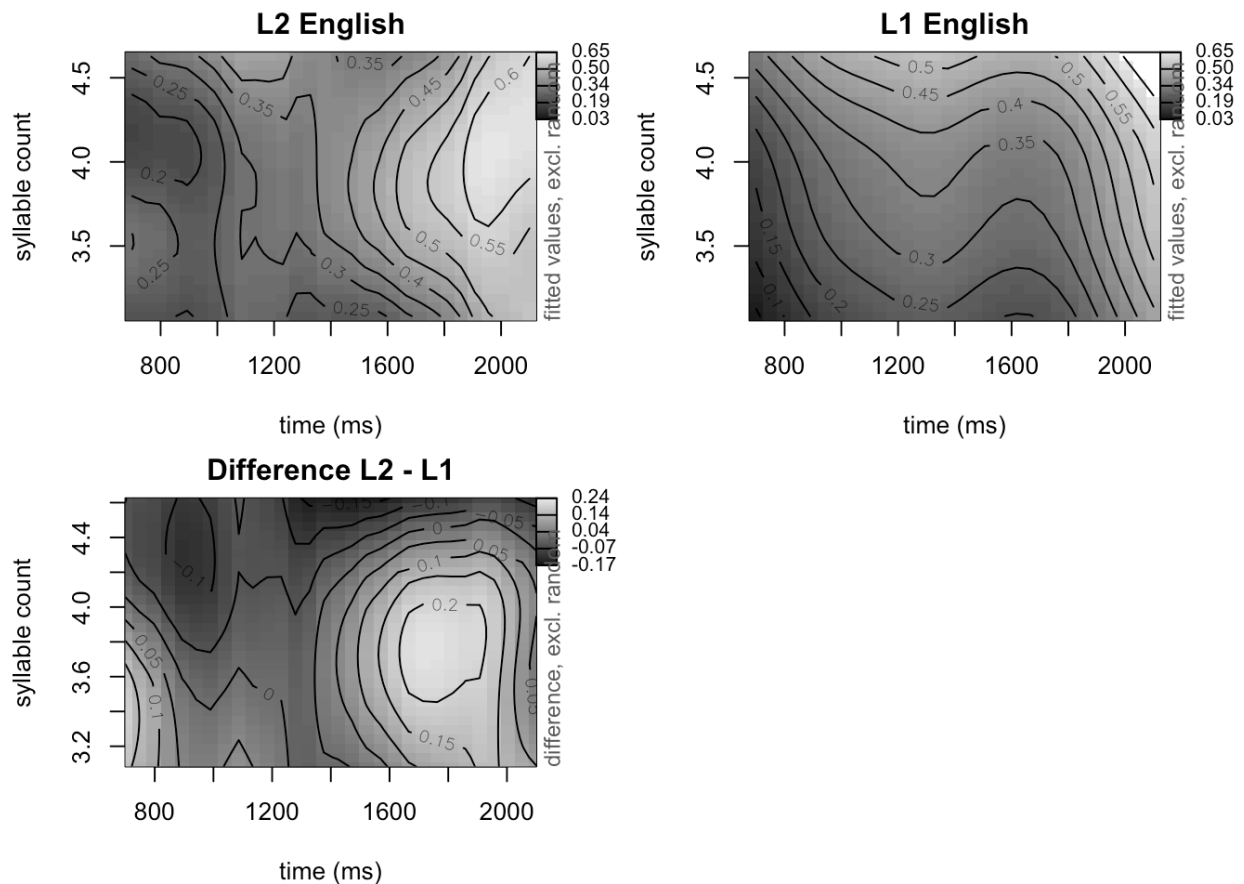


Figure 4. Contour plots for the empirical logit of the three-way interaction between time, syllable count, and language group. *Top left panel.* Contour plot for L2 speakers. *Top right panel.* Contour plot for L1 speakers. In both of the *top panels*, darker indicates less empirical logits (looks to the target) and lighter indicates greater empirical logits. *Bottom left panel.* The difference between the L2 and L1 speakers. In the *bottom panel* the negative values indicate greater empirical logits (looks to the target) by L1 speakers relative to L2 (with the difference increasing the darker it becomes), and positive values indicate greater empirical logits by L2 speakers (with the difference increasing the lighter it becomes).

Both groups show a consistent positive empirical logit across the time window and at all speech rates, indicating that at the start of the window (the verb) both groups are looking towards the CP target. At the onset of the window, L2 speakers show a similar empirical logit value across all speech rates, with the value increasing as time goes on and slightly as speech rate increases. At the onset of the window, L1 speakers show a large increase in empirical logits as the speech rate increases, and an increase as time goes on. The difference graph shows that L2 speakers make more looks to the target at lower speech rates and at the onset of the window, while L1 speakers make more looks to the target at the faster speech rates. Following the onset of the target word (~1440 ms), we see that L1 speakers only make more looks to the target than L2 speakers at the fastest speech rates.

DPA

To test for prediction, the first DPA comparison investigated the CP items only, comparing looks to the CP target relative to the CU target. The analysis revealed that the divergence occurs for L1 speakers at 720.77 ms (CI: [680,800]) and occurs for L2 speakers at 813.22 ms (CI: [700,1000]); see Figure 5. While numerically the L1 speakers have an earlier divergence point, the mean difference between groups is 92.45 ms (CI: [-60.00, 279.50]); given that the CI contains 0, we conclude that the two groups do not differ in the timing of looks to the CP target.

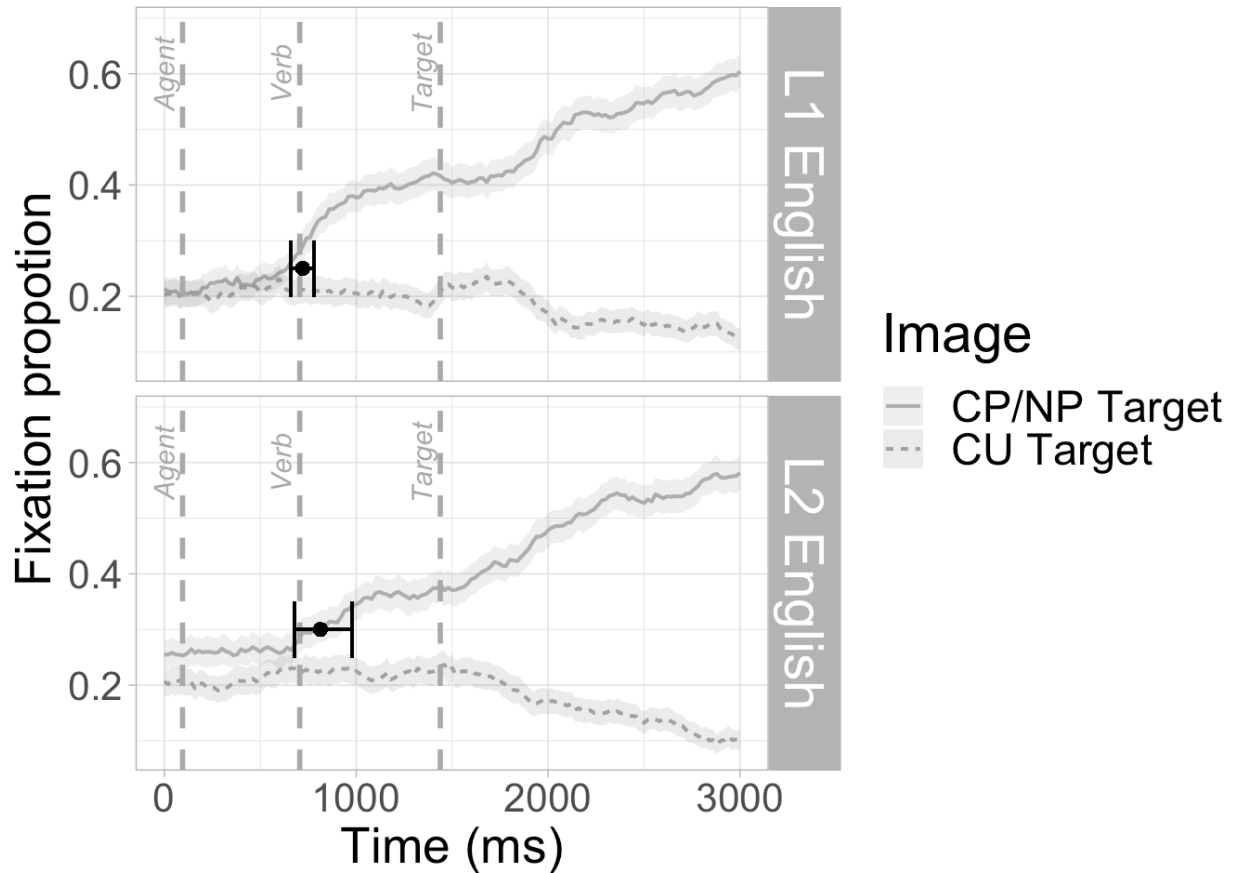


Figure 5. Divergence point and 95% confidence intervals superimposed on the fixation proportion of looks to the CP/NP target and the CU target in CP items.

Prediction cost

GAMM

To test the prediction cost for both groups, we analyzed the empirical logit in the CU items (looks to the CP target (*suit*) vs. looks to CU target (*tree*)) from the onset of the verb to the offset of the object (+200ms) thus allowing us to investigate the pattern of eye movement behavior when the unexpected target word is encountered, see Figure 6 for visualization.

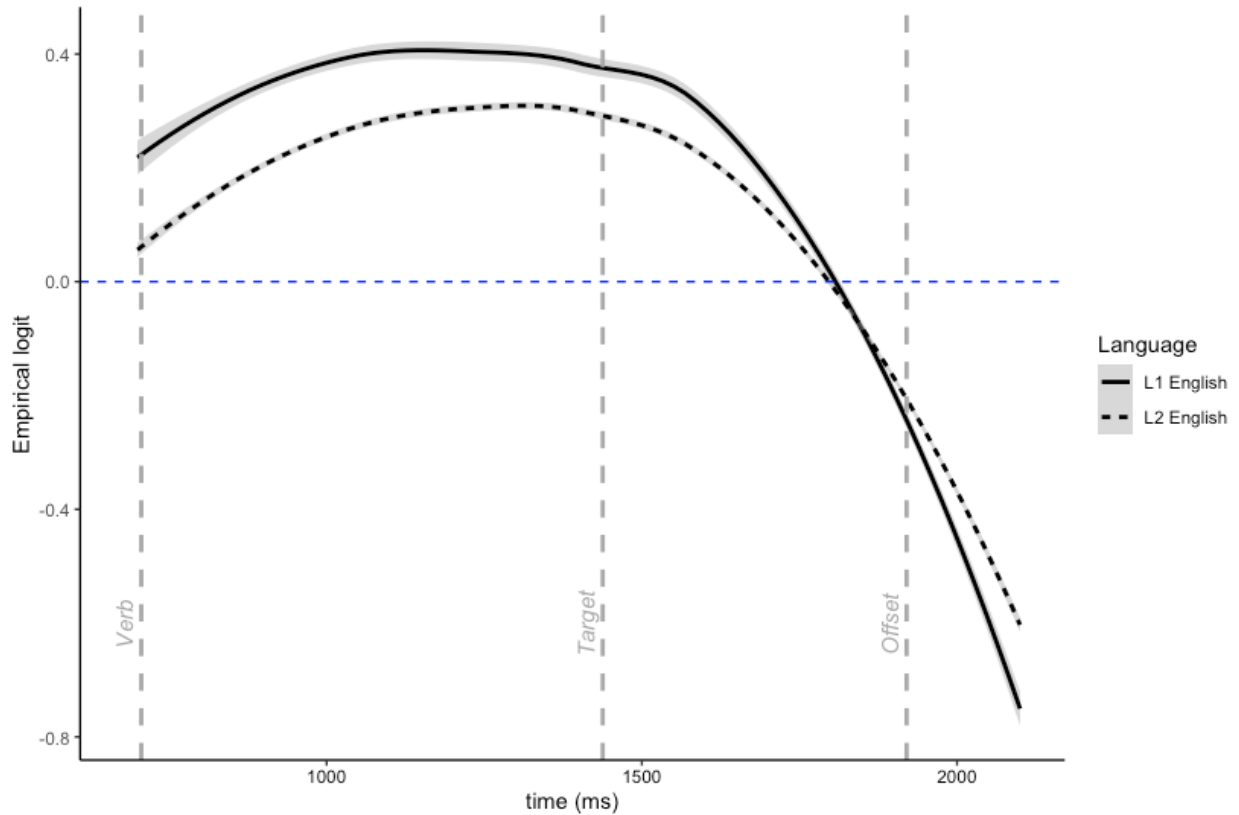


Figure 6. Empirical logit during the prediction time window for CU items (shading represents 95% confidence interval and the horizontal dotted line represents the zero mark of the empirical logits). If the empirical logit is above the zero line participants are looking more towards the CP target (*suit*) if it is below the zero line they are looking more towards the CU target (*tree*).

The results from the GAMM can be seen in table 6.

Table 6. GAMM mixed model output for the empirical logit during the prediction time window for CU items. Part A reports the parametric coefficients and Part B reports the smooth terms.

A. Parametric coefficients	Estimate	Std. error	t-value	p-value
Intercept	.00	.00		
Syllable count	.00	.00		
Language (L2)	-0.01	0.07	-0.16	0.87
OPT	0.00	0.00	3.17	<.01
B. Smooth terms	EDF	Ref.df	F-value	p-value
s(syllable count)	2.97	3.00	55.71	<.001
s(OPT)	1.00	1.01	0.01	0.93
s(time)	8.19	9.63	31.43	<.001
s(time):Language (L2)	1.00	1.00	1.03	0.31

s(OPT):Language (L2)	2.44	2.45	1.54	0.15
s(syllable count): Language (L2)	2.89	2.99	16.70	<.001
ti(time, OPT): Language (L2)	1.01	1.01	0.51	0.48
ti(time, syllable count): Language (L2)	8.03	8.78	20.36	<.001
s(time, participant)	722.57	1816.00	4.13	<.001
s(time, item):Language (L2)	70.00	99.17	5.73	<.001

In terms of the parametric effects, proficiency (OPT) was significant ($t=3.17, p<.01$) with the empirical logit increasing as proficiency increased. In terms of the smooth terms, there was a significant ordered factor difference smooth interaction of syllable count by language ($F=16.70, p<.001$), see Figure 7. The summed effect of syllable count for L2 and L1 speakers can be seen in left side of Figure 7, and the difference can be seen on the right side of Figure 7. This reveals that L1 speakers make more looks to the target at the lower speech rates relative to L2 speakers.

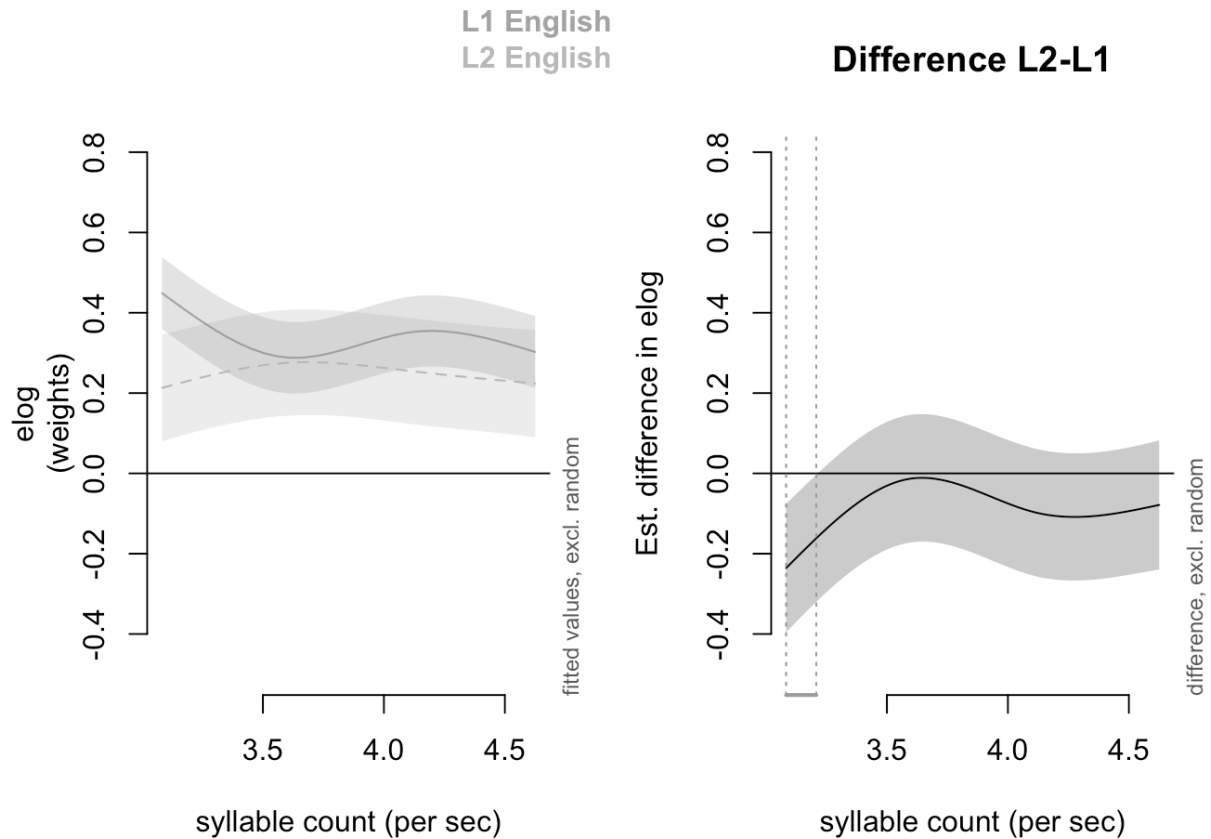


Figure 7. The solid horizontal line represents the zero effect. *Left panel.* Summed effect of syllable count for L1 English speakers (solid line) and L2 English speakers (dotted line). *Right panel.* The difference in the summed effect of syllable count for L1 and L2 speakers of English. The dashed vertical line indicated where there is a significant difference.

There was a significant tensor product smooth interaction between time, syllable count, and language ($F=20.36, p < .001$). The tensor interaction is visualized using contour plots; see Figure 8.

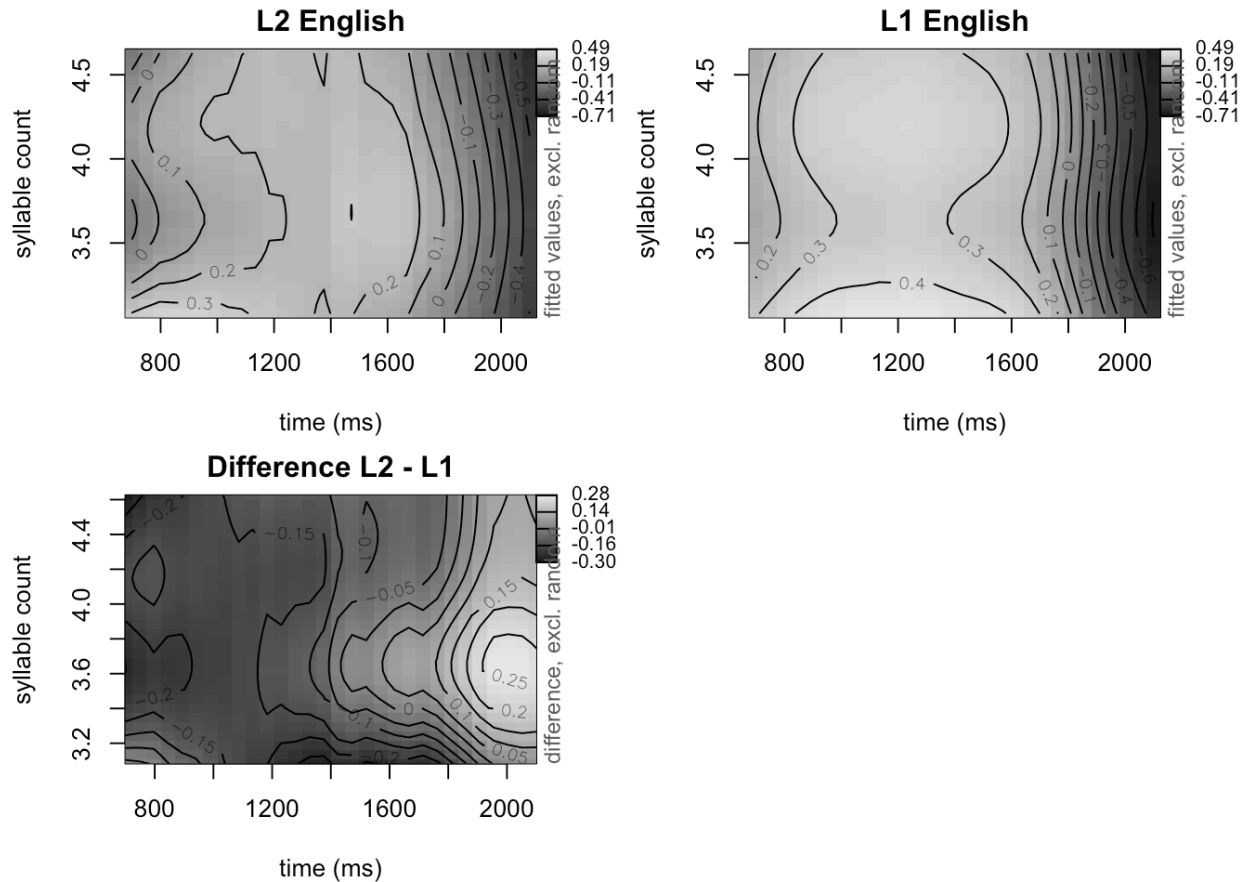


Figure 8. Contour plots for the empirical logit of the three-way interaction between time, syllable count, and language group. *Top left panel.* Contour plot for L2 speakers. *Top right panel.* Contour plot for L1 speakers. In both of the *top panels*, darker indicates less empirical logits (or more looks to the target (unexpected) image) and lighter indicates greater empirical logits (or more looks to the predicted (but incorrect) image). *Bottom left panel.* The difference between the L2 and L1 speakers. In the *bottom panel* the negative values indicate greater empirical logits (looks to the target) by L1 speakers relative to L2 (with the difference increasing the darker it becomes), and positive values indicate greater empirical logits by L2 speakers (with the difference increasing the lighter it becomes).

L2 speakers show looks to the CP target at the onset of the window, but only at the slowest speech rates (with looks to the CP target decreasing as the speech rate increases). Looks to the CP target start to decrease slightly after 1600ms and cross the 0 line at approximately 1800 ms (indicating the shift in attention from the CP target to the CU image), with looks decreasing to

the CP target (and increasing for the CU image) across all speeds for the remainder of the window. The shift in attention starts slightly earlier at the slower and faster speech rates. The L1 group shows looks to the CP target at the onset of the window across all speeds. The empirical logits start to decrease slightly at approximately 1600ms and cross the 0 line at approximately 1800ms with looks to the CP target decreasing across all speeds for the remainder of the window. The shift in attention starts slightly earlier for the middle speech rate items (~3.5-4.0 syllables per second). The difference graph shows that L1 speakers show greater looks to the CP target than L2 speakers until approximately 1600ms, when attention shifts and L1 speakers show more looks to the CU image than L2 speakers (particularly at the slower speech rates).

DPA

The second DPA comparison investigated looks to the CU target relative to a distractor³ in CU items only. The analysis revealed that the divergence occurs for L1 speakers at 1761.05 ms (CI: [1720,1840]) and occurs for L2 speakers at 1853.76 ms (CI: [1820,1920]); see Figure 9. The mean difference between groups is 92.71 ms (CI: [20, 160]); given that the CI does not contain 0, we conclude that the looks to the CU target diverged from the distractor earlier for L1 speakers relative to L2 speakers.

³ We randomly assigned the two distractor objects with the label of distractor 1 or distractor 2. Comparison were made to distractor 2.

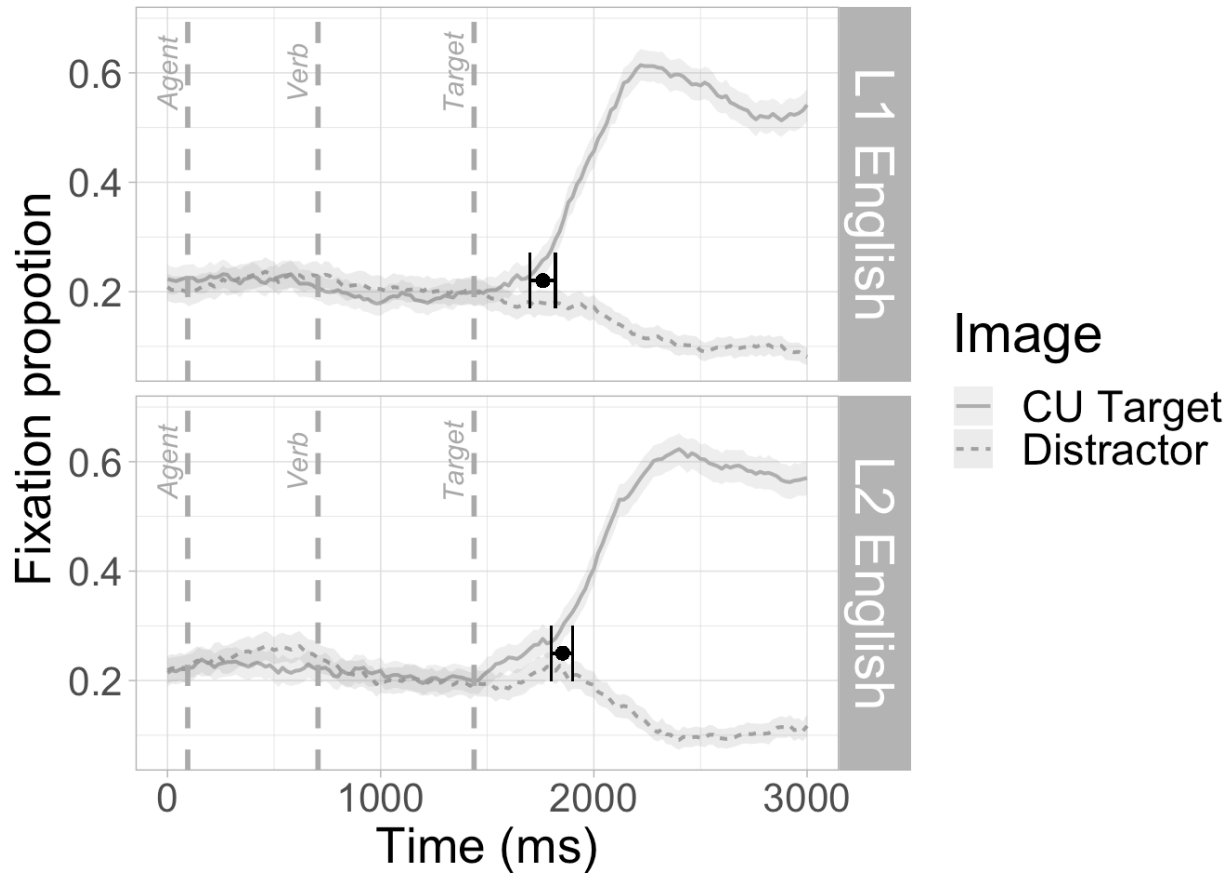


Figure 9. Divergence point and 95% confidence intervals superimposed on the fixation proportion of looks to the CU target and the distractor in CU items.

The third DPA comparison investigated looks to the CU target in the CU items relative to looks to the NP target in the NP items. Unexpectedly, this revealed a divergence between the two objects only for L1 speakers, early in the sentence; we will return to this in the discussion. L1 speakers showed a divergence at 904.85 ms (CI: [840,1060]); see Figure 10. Since the DPA cannot find more than one divergence point, we cut the window for comparison between 1240-3000 ms. Given that the target is not spoken until 1438 ms, this should encompass any differences evoked by the target word. We found no divergence during this time window suggesting no prediction cost; see Figure 11.

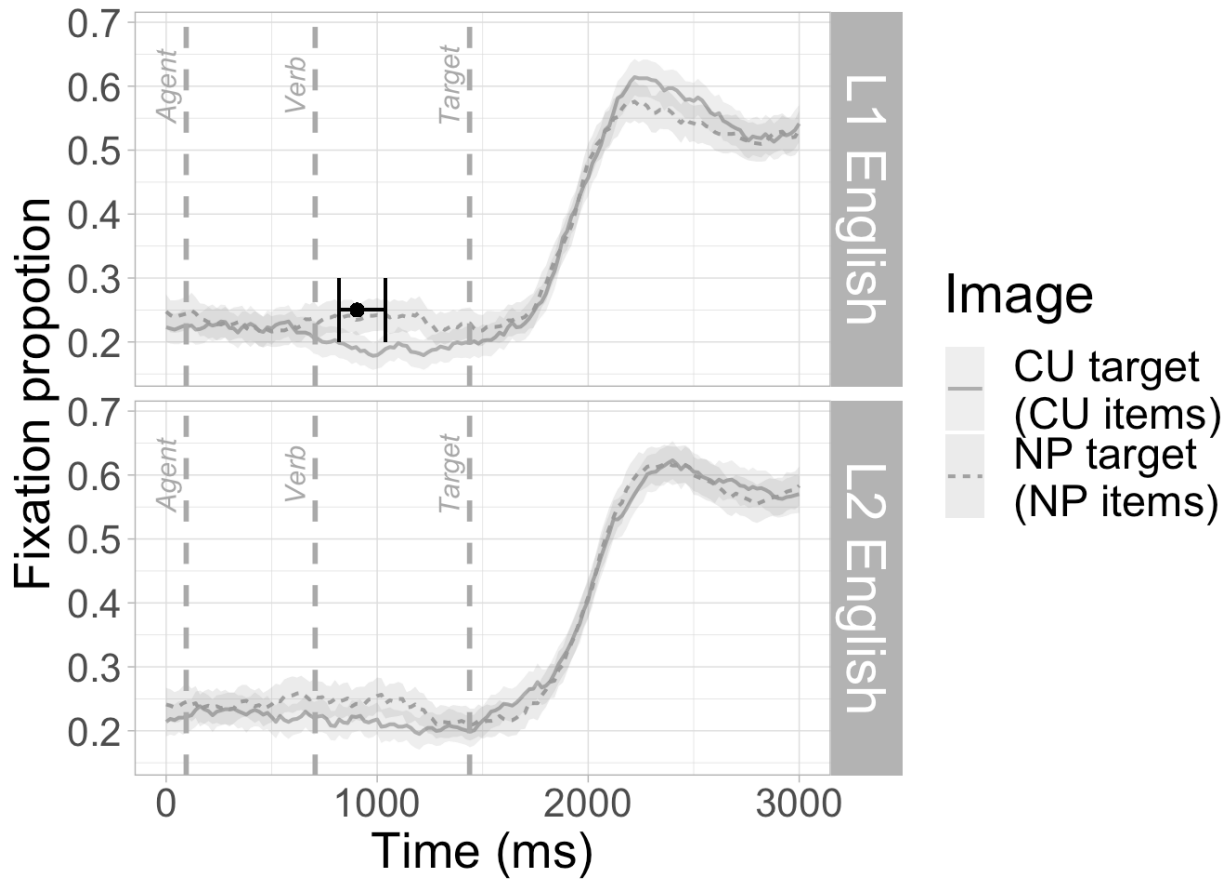


Figure 10. Divergence point and 95% confidence intervals superimposed on the fixation proportion of looks to CU target in CU items and the NP target in NP items.

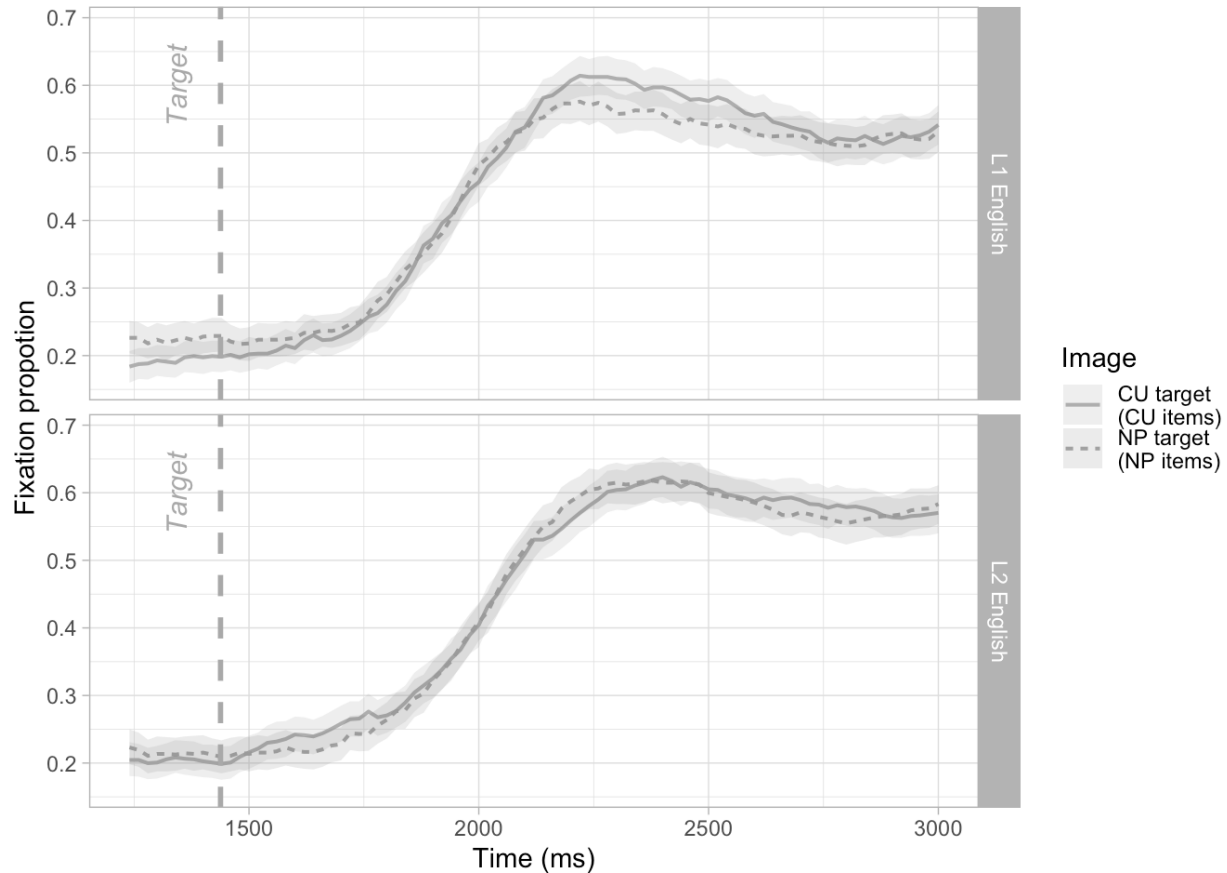


Figure 11. The fixation proportion of looks to the CU target in CU items and the NP target in NP items from the 200ms prior to the start of the target (1240ms). No divergence points were found.

Discussion

Research has found that both L1 and L2 speakers are able to make predictions about upcoming information before it is encountered (e.g., L1- Ferreira & Chantavarin, 2018; Huettig, 2015; Huettig & Mani, 2016; Kamide 2008; Staub, 2015 / L2 - Hopp, 2022; Kaan & Grüter, 2021; Schlenter, 2022). When differences arise between L1 and L2 speakers in terms of prediction, they have been argued to be quantitative in nature and to stem from individual differences (e.g., proficiency) or methodological factors (e.g., speech rate) (Kaan, 2014; Schlenter, 2022). The literature on whether L1 or L2 speakers incur a cost when hearing a constraining context but encountering a plausible but unpredicted word is less clear. Some L1 research has shown a cost

that may indicate situation level updating when an unexpected but plausible word is encountered (e.g., Brothers et al., 2023; Kuperberg, 2020) while other research has not shown a cost (e.g., Chow & Chen, 2020; Frisson et al., 2016; Luke & Christianson, 2016). Likewise, some L2 literature has shown a cost (ERP-Foucart et al., 2014; 2015; 2016 / self-paced reading- Feng & Jiang, 2023) while other research has not shown a cost, particularly when individual differences are taken into account (ERP- Zirnstein, et al., 2018 / VWP- Peters et al., 2018). In the current study we aimed to shed light on this discrepancy by comparing prediction and prediction costs using the VWP for both L1 and L2 speakers of English, while investigating individual differences (English proficiency) and a potential methodological factor (speech rate). We believe this could provide insight into whether only one lexical item is activated as context unfolds (leading to a cost when a prediction is not met), or multiple lexical candidates are activated (leading to no cost when a prediction is not met). Additionally, we used two statistical approaches that are uniquely suited for this type of data: GAMMs and DPA.

Our first set of analyses focused on whether L1 and L2 speakers make predictions in a similar way and within a similar time frame while exploring proficiency and speech rate. For these analyses, we tested the pattern of looks to the CP target (*suit*) during a CP item (*The tailor trims the suit*). As expected, we found that speakers with higher proficiency made more looks to the target (regardless of whether they were an L1 or L2 speaker), and that differences between L1 and L2 speakers occurred in relation to speech rate, with L1 speakers making more predictions at the faster speech rates relative to L2 speakers, but L2 speakers making more predictions at the slower speech rates relative to L1 speakers. Yet overall, we found that both L1 and L2 groups showed similar looks to the target across time with no difference in the timings of looks. Both groups made similar predictions soon after hearing the verb, and showed similar

trajectories across time. The three-way interaction between speech rate, time, and language group again revealed that L1 speakers made more looks to the target particularly at the faster speech rates.

From these analyses we argue that L2 speakers make predictions in a similar way and within a similar time-frame relative to L1 speakers. Interestingly, we found that proficiency did not interact with language, with time, or with language and time, suggesting that prediction skills do not change with proficiency (in line with other studies such as Dijkgraaf et al., 2017; Hopp, 2015; Ito et al., 2018; Kaan & Grüter, 2021; Kim & Grüter, 2020; Mitsugi, 2020; Perdomo & Kaan, 2019). However, we note that the items used in the present study are particularly easy and therefore do not rule out the possibility that proficiency may come into play with more difficult items (e.g., syntactic prediction). Differences between L1 and L2 speakers seemed only to arise within the context of speech rate. However, we argue some caution in interpreting these results because this study did not directly manipulate speech rate and also had few very fast or slow speech rates, which may have skewed the results. We encourage future research with material directly designed to test the impact of speech rate.

Our second set of analyses focused on whether L1 and L2 speakers incur a prediction cost when hearing an unexpected word in a constraining context. For these analyses we first tested the pattern of looks to the CU target (*tree*) during a CU item (*The tailor trims the tree*) to test whether there are differences between L1 and L2 speakers in terms of reconciling an incorrect prediction. We found that both groups make predictions based on context at the onset of the item (i.e., looks to the CP item *suit*), but are able to quickly shift their attention when their predictions are not met. Additionally, the DPA showed that L1 speakers look to the CU image

significantly earlier than L2 speakers. We also found an impact of speech rate but again urge some caution when interpreting this finding given that (1) speech rate was not directly manipulated, (2) there are few instances of very fast and very slow speech rate, and (3) this analysis is collapsed across the whole time window (including the shift in attention, where we expect a decrease in looks to target as time goes on). However, what is clear from this and the previous analysis is that speech rate is impacting both L1 and L2 speakers predictive behaviors. Carefully controlled studies are needed to better understand these effects.

While the two groups exhibit some differences in the timing of looks to the CU target in the CU items, we believe that the underlying prediction mechanisms are the same. We argue that later looks to the CU target by L2 speakers may reflect a general slowing of lexical access (e.g., Shook et al., 2015) rather than a cost. This is supported by the findings from our second prediction cost comparison of looks to the NP target (*suit*) in the NP condition (*The guardian sells the suit*) relative to the looks to the target in the CU condition. In the NP condition the context does not constrain the listener to a particular interpretation, while the CU condition does constrain the listener towards a particular interpretation, which is ultimately incorrect. If a cost was incurred when a prediction was made (and ultimately not met), the cost would be apparent relative to a sentence where no prediction was made (i.e., NP items). When comparing looks to the respective targets in the CU and NP items, we see that they are visually identical (particularly from the onset of the target). The DPA revealed no divergence between looks to the target across the two items from 200ms before the target until the end of the analyzed window (3000 ms) for either L1 or L2 speakers. Therefore, we argue that neither L1 nor L2 speakers incur a prediction cost, and that integrating an unexpected word after predicting another ultimately

incorrect word is no different than integrating a word into a sentence with no constraining context (in line with Chow & Chen, 2020; Frisson et al., 2016; Luke & Christianson, 2016).

One question that may arise is: why is there no difference in timing between L1 and L2 speakers in looks to the CP target in the CP items, but there is a delay for the L2 speakers in looks to the CU target in the CU items? We argue that this is not evidence of a prediction cost for L2 speakers; rather, it is evidence of slower lexical access (e.g., Shook et al., 2015). In situations where L2 speakers build up a prediction (due to constraining context) they show no delay when that predicted word is encountered, given that the target word is the most activated. However, when there is no such prediction made (due to neutral context), L2 speakers may have several candidates (potentially already activated). Thus, when the target word is heard, it takes them longer to shift their attention to the correct candidate because of a general slowing of lexical activation (this is also evidenced by the fact that there is no difference in the timing of looks between the NP and CU items). This was further supported by a post-hoc DPA analysis of the NP items only comparing looks to the target relative to a non-target object (all non-target objects were randomly assigned a label) in which L2 speakers' looks to the target diverged later relative to L1 speakers⁴. L2 speakers seem to take slightly longer when they have to choose between several activated lexical candidates, whether that is without a prediction being made (neutral context) or when a prediction is made but is ultimately incorrect. We also note that the difference in the timing of looks in the CU items for L2 relative to L1 speakers is less than

⁴ The L1 group showed a divergence 90.45 ms (CI: [40, 140]) earlier than the L2 group; given that the CI does not contain 0, we conclude that the looks to the NP target diverged from the non-target earlier for L1 speakers relative to L2 speakers. See OSF for analysis and visualization (https://osf.io/3v7sd/?view_only=b625d20b77564d4eb7ce1862f9cd7734).

100ms, so they are still able to do this very quickly. If this line of reasoning is correct, this would suggest that prediction leads to quicker lexical access and more L1-like processing. This, of course, needs to be tested further (e.g., if this extends to more difficult constructions) but it would indicate prediction as a potential area for training in language learning (e.g., Hopp, 2016; Schremm, et al., 2017).

We believe that it is important to highlight how our findings generalize to other paradigms and discuss how the paradigm may impact participant behavior. The VWP involves a limited set of referents, and in the case of the current study, they were presented for 2000ms prior to the onset of the sentence, giving participants plenty of time to identify and activate relevant information and spatially encode image location on the screen. This may therefore support easy revision and instantaneous shifts in attention when a prediction is not met given that the two potential referents following the verb (e.g., *tree* and *suit*) are already active and spatially encoded (similar to Chow & Chen, 2023). While this may lead to issues with generalizability, we believe this may explain our lack of prediction cost. Brothers et al. (2023) found a larger late frontal positivity to CU items when the target word was unexpected and highly unlikely relative to NP items, reflecting a situation level update; but when the unexpected word was highly likely (i.e., the *second* most predicted target), they found no difference in late frontal positivity. We argue our results are in line with their second finding. That is, while *tree* is not the expected target of *The tailor trimmed the* it is the second best option based on which lexical items that are currently active by virtue of being images on the screen in front of the participant (i.e., *tree*, *jar*, and *pot*). Our findings could therefore also be understood to reflect lack of prediction cost for processing a highly likely alternative (likely based on the paradigm as opposed to abstract semantics). This is also similar to the findings of Luke and Christianson (2016) and Frisson et

al., (2017) in that when a CU target is semantically related (or in this case, the most likely alternative) there is no measurable cost. Together, we believe that this supports the argument that listeners make several partial predictions. Whether a cost can be evoked with VWP items designed to evoke a situation level update remains to be seen, but it would lend further support to the findings of Brothers et al. (2023).

An unexpected finding in the comparison of CU to NP items, is that for L1 speakers there is a divergence in the NP items following the verb (at ~905 ms; the verb occurs between 700-1300ms) relative to the CU items. It is unclear if this is driven by a decrease in looks to the CU target or whether it is indicative of a local prediction by L1 speakers (e.g., Peters et al., 2018), in that they look to an object that can be *sold* (i.e., in *The guardian sells the ...*), but since there is not enough information to commit to an interpretation (as all the items can be sold), they continue to shift their attention elsewhere. This is not seen in the CU item (*The tailor trimmed the tree*) since the context (*The tailor trimmed the*) has constrained the interpretation and the listener has committed to the predicted target after hearing the verb (even if it is ultimately incorrect). This pattern is not seen in L2 speakers, or at least to a lesser extent (see Figure 1), which may suggest less sensitivity to local predictions for L2 speakers, or perhaps these local predictions are impacted by proficiency (Peters et al., found that less skilled speakers made more local prediction) and speech rate (which we do not test in the DPA analysis). We believe this would be an interesting avenue for future research.

Another point for discussion relates to the eye-movement pattern for the CU items in which the participants answered incorrectly (i.e., they heard *The tailor trims the tree* but they clicked on *suit*). Note that we interpret this anecdotally, given that only 10% of the items were

answered incorrectly (with the CP target). Interestingly, we do not see either group dismissing the CU target completely and only focusing on the CP target. Rather both groups show a decrease in looks to the CP target after hearing the target followed by an increase in the CU target (for both groups there is a point at which the participants are focusing on the CU target more than the CP target) but ultimately the participants shift their attention back to the incorrect CP target. This pattern may indicate some conflict (for example from mis-hearing, not paying attention, not believing what they heard) and the participant ultimately picks the most likely target based on what they did process. This may indicate “good enough” processing (e.g., Ferreira & Patson, 2007) in which participants will make shallower analyses even if it leads to the incorrect interpretation. Or it may support the ‘noisy-channel’ hypothesis (e.g., Gibson, et al., 2013; Levy, 2008), in which a listener will use their own knowledge to infer the most sensible meaning in the face of uncertainty (in the case of mis-hearing, not paying attention, etc.). Again, we encourage future research investigating this further.

Lastly, we think it is important to briefly discuss potential trial order effects in the current study. Research has found that when predictive cues are no longer reliable, language users may stop using them (e.g., Brother et al., 2019). Given that participants in the current study encounter more CU items as the experiment goes on, they may change their behaviors or stop predicting all together. Therefore, we ran a post-hoc analysis investigating whether L1 and L2 speakers showed differing eye-movement patterns over the course of the study (see Appendix C for additional information, analysis, and visualization). We found that L1 speakers showed a clear decrease in prediction as the trial number increased (i.e., as the encountered more items) while the L2 speakers did not show a clear pattern of decrease (or increase) as the trial number increased. This suggests that L1 speakers may adaptively adjust their prediction behavior based

on utility (Kuperberg & Jaeger, 2016) while L2 speakers may place less weight on prediction success (leading to more variable effects of trial order). We hesitate to interpret these findings in more depth given that the current items were not designed to test utility, but we encourage future research with and explicitly designed materials and *apriori* hypotheses to further investigate this compelling finding.

Conclusions

In this study we investigated prediction and prediction costs in L1 and L2 speakers of English, while also exploring the effect of English language proficiency and speech rate. The literature is mixed as to whether predictions come at a cost, particularly when they are not met, and it is not clear if L1 and L2 speakers show similar patterns. We found that both L1 and L2 speakers make predictions in the same way and at the same time, supporting the growing consensus that L1 and L2 speakers' prediction mechanisms are the same (e.g., Kaan, 2014). In terms of prediction costs, we found similar eye-movement trajectories but L2 speakers took longer to shift their attention than L1 speakers. We argue that this may be due to a general slowing in lexical access for L2 speakers rather than evidence for prediction costs. This is further supported by the fact that there were no timing differences between processing the target word in a neutral context relative to the unexpected target word in a constraining context. While we found that increased proficiency led to increased looks to the target, we did not find evidence that proficiency impacted predictive behaviors across time or across L1 and L2 groups. However, both groups were impacted by speech rate, with L1 speakers making more predictions at fastest speech rates relative to L2 speakers, and L2 speakers making more predictions at the slowest speech rates relative to L1 speakers. On the one hand, we hesitate to make too much of

these findings since our study did not directly manipulate speech rate. We thus encourage future research with items specifically designed to test speech rate. On the other hand, it is clear that speech rate plays an important role and should be at minimum reported in the literature (Fernandez et al., 2020).

Overall, we argue that both L1 and L2 speakers make predictions in the same way, and that there are no costs associated with making a prediction that is not met. The VWP may have at least in part contributed to the lack of cost, given that participants activate and spatially encode the four visual objects displayed in the array, making the integration of the unexpected target less costly given the is the most likely alternative in the array (similar to what has been found in previous research: VWP – Chow & Chen, 2023; reading– Frisson et al., 2017; and ERP – Brothers et al., 2023). Our findings supports the growing body of literature that during processing we may make several partial predictions about upcoming information without inhibiting less likely words (e.g., Chow & Chen, 2020; Brothers et al., 2023; Frisson et al., 2016; Luke & Christianson, 2016) rather than predicting one specific lexical candidate while inhibiting less likely candidates.

References

- Anderson, J. A. E., Mak, L., Keyvani Chahi, A., & Bialystok, E. (2018). The Language and Social Background Questionnaire: Assessing degree of bilingualism in a diverse population. *Behavioral Research Methods*, *50*, 250–263. 10.3758/s13428-017-0867-9
- Audacity® software is copyright © 1999-2021 Audacity Team. Version 3. The name Audacity® is a registered trademark.
- Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. N. (2016). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. *arXiv preprint arXiv:1601.02043*.
- Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. 10.1016/j.jml.2007.09.002
- Boersma, P. & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.44.
- Brothers, T., Dave, S., Hoversten, L. J., Traxler, M. J., & Swaab, T. Y. (2019). Flexible predictions during listening comprehension: Speaker reliability affects anticipatory processes. *Neuropsychologia*, *135*, 107225. 10.1016/j.neuropsychologia.2019.107225
- Brothers, T., Morgan, E., Yacovone, A., & Kuperberg, G. (2023). Multiple predictions during language comprehension: Friends, foes, or indifferent companions?. *Cognition*, *241*, 105602. 10.1016/j.cognition.2023.105602
- 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(4), 1029–1040.

10.1037/a0015901

Chow, W.-Y., & Chen, D. (2020). Predicting (in)correctly: Listeners rapidly use unexpected information to revise their predictions. *Language, Cognition and Neuroscience*, 35(9), 1149–1161. 10.1080/23273798.2020.1733627

Corps, R. E., Brooke, C., & Pickering, M. J. (2022). Prediction involves two stages: Evidence from visual-world eye-tracking. *Journal of Memory and Language*, 122, 104298. <https://doi.org/10.1016/j.jml.2021.104298>

Corps R.E., Liao, M., & Pickering, M.J. (2023). Evidence for two stages of prediction in non-native speakers: A visual-world eye-tracking study. *Bilingualism: Language and Cognition*, 26(1), 231–243. doi:10.1017/S1366728922000499

DeLong, K. A., Groppe, D. M., Urbach, T. P., & Kutas, M. (2012). Thinking ahead or not? Natural aging and anticipation during reading. *Brain and Language*, 121(3), 226–239. 10.1016/j.bandl.2012.02.006

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, 61, 150–162. 10.1016/j.neuropsychologia.2014.06.016

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8), 1117–1121. 1038/nm1504

Dickey, M. W., Choy, J. J., & Tompson, C. K. (2007). Real-time comprehension of wh-movement in aphasia: Evidence from eyetracking while listening. *Brain and Language*, 100, 1– 22. <https://doi.org/10.1016/j.bandl.2006.06.004>

- 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. 10.1017/S1366728916000547
- Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816. 10.1080/17470218.2017.1310261
- Dussias, P. E., Valdés, J. R. K., Guzzardo, R. E. T., & Gerfen, C. (2013). When gender and looking go hand in hand: Grammatical gender processing in L2 Spanish. *Studies in Second Language Acquisition*, *35*, 353–387. 10.1017/S0272263112000915.
- Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84. 10.1016/j.brainres.2006.06.101
- Feng, L., & Jiang, N. (2023). Prediction error cost exists in reading processing of Chinese native speakers and advanced Chinese L2 learners. *Frontiers in Psychology*, *14*(1134229). 10.3389/fpsyg.2023.1134229
- Fernandez, L. B., Engelhardt, P. E., Patarroyo, A. G., & Allen, S. E. (2020). Effects of speech rate on anticipatory eye movements in the visual world paradigm: Evidence from aging, native, and non-native language processing. *Quarterly Journal of Experimental Psychology*, *73*(12), 2348–2361. 10.1177/1747021820948019

- 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.
10.1177/0963721418794491
- Ferreira, F., & Patson, N. D. (2007). The ‘good enough’ approach to language comprehension. *Language and Linguistics Compass*, 1(1-2), 71-83. 10.1111/j.1749-818X.2007.00007.x
- 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(5), 1461–1469. 10.1037/a0036756
- Foucart, A., Ruiz-Tada, E., & Costa, A. (2015). How do you know I was about to say “book”? Anticipation processes affect speech processing and lexical recognition. *Language, Cognition and Neuroscience*, 30(6), 768-780. 10.1080/23273798.2015.1016047
- Foucart, A., Ruiz-Tada, E., & Costa, A. (2016). Anticipation processes in L2 speech comprehension: Evidence from ERPs and lexical recognition task. *Bilingualism: Language and Cognition*, 19(1), 213–219. 10.1017/S1366728915000486
- Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, 95, 200–214. 10.1016/j.jml.2017.04.007
- Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20), 8051-8056. 10.1073/pnas.1216438110
- Graham, S. (2006). Listening comprehension: The learners’ perspective. *System*, 34, 165–182.

10.1016/j.system.2005.11.001

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(32), e2201968119.
<https://doi.org/10.1073/pnas.2201968119>

Hertrich, I., Dietrich, S., & Ackermann, H. (2013). How can audiovisual pathways enhance the temporal resolution of time-compressed speech in blind subjects? *Frontiers in Psychology*, *4*, 10.3389/fpsyg.2013.00530

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, *29*, 33–56. 10.1177/0267658312461803

Hopp, H. (2015). Semantics and morphosyntax in predictive L2 sentence processing. *International Review of Applied Linguistics in Language Teaching*, *53*(3), 277-306.
<https://doi.org/10.1515/iral-2015-0014>

Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research*, *32*(2), 277–307. 10.1177/026765831562496

Hopp, H. (2022). Second language sentence processing. *Annual Review of Linguistics*, *8*(1), 235-256.
10.1146/annurev-linguistics-030821-054113

Huetig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118-135. 10.1016/j.brainres.2015.02.014

- Huetting, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, 31(1), 19-31. 10.1080/23273798.2015.1072223
- Huetting, F., Audring, J., & Jackendoff, R. (2022). A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition*, 224, 105050. 10.1016/j.cognition.2022.105050
- Ito, A., & Pickering, M. J. (2021). Automaticity and prediction in non-native language comprehension. In E. Kaan and T. Grüter (Eds.), *Prediction in Second Language Processing and Learning* (pp. 25–46). John Benjamins.
- Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, 98, 1-11. 10.1016/j.jml.2017.09.002.
- Kaan, E. & Grüter, T. (2021). Prediction in second language processing and learning: Advances and directions. In: Kaan, E. and Theres Grüter (Eds) *Prediction in second language processing and learning*, John Benjamin.
- Kamide, Y. (2008). Anticipatory processes in sentence processing. *Language and Linguistics Compass*, 2/4, 647-670. 10.1111/j.1749-818X.2008.00072.x
- Kim, H., & Grüter, T. (2020). Predictive processing of implicit causality in a second language: A visual-world eye-tracking study. *Studies in Second Language Acquisition*, 1-22. 10.1017/S0272263120000443
- Kuperberg, G. R., Brothers, T., & Wlotko, E. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, 32(1), 12–35. 10.1162/jocn_a_01465

- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience*, 31(1), 32–59.
10.1080/23273798.2015.1102299
- Levy R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceeding of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 234–243). Stroudsburg, PA: Association for Computational Linguistics.
- Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, 88, 22–60. 10.1016/j.cogpsych.2016.06.002
- Martin, C. D., Thierry, G., Kuipers, J.-R., 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. 10.1016/j.jml.2013.08.001
- Mitsugi, S. (2020). 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*, 58(3), 323-349. 10.1515/iral-2017-0118
- Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... & Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, 7, e33468. 10.7554/eLife.33468
- Perdomo, M., & Kaan, E. (2019). Prosodic cues in second-language speech processing: A visual world eye-tracking study. *Second Language Research*, 37, 349-375. 10.1177/0267658319879196
- Peters, R., Grüter, T., & Borovsky, A. (2018). Vocabulary size and native speaker self-identification influence flexibility in linguistic prediction among adult bilinguals. *Applied Psycholinguistics*, 39(6), 1439–1469.1017/S0142716418000383

- Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological bulletin*, *144*(10), 1002–1044. 10.1037/bul0000158
- Porretta, V., Kyröläinen, A.J., van Rij, J., Järvikivi, J. (2018). Visual World Paradigm Data: From Preprocessing to Nonlinear Time-Course Analysis. In: Czarnowski, I., Howlett, R., Jain, L. (eds) Intelligent Decision Technologies 2017. IDT 2017. *Smart Innovation, Systems and Technologies*,73. Springer, Cham. 10.1007/978-3-319-59424-8_25
- Prystauka, Y., Altmann, G.T.M. & Rothman, J. (2023). Online eye tracking and real-time sentence processing: On opportunities and efficacy for capturing psycholinguistic effects of different magnitudes and diversity. *Behavioral Research Methods*. 10.3758/s13428-023-02176-4
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.Rproject.org/>.
- Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences*, *27*(11), 1032–1052. 10.1016/j.tics.2023.08.003
- Schlenter, J. (2023). Prediction in bilingual sentence processing: How prediction differs in a later learned language from a first language. *Bilingualism: Language and Cognition*, *26*(2), 253-267. 10.1017/S1366728922000736
- Schremm, A., Hed, A., Horne, M., & Roll, M. (2017). Training predictive L2 processing with a digital game: Prototype promotes acquisition of anticipatory use of tone-suffix associations. *Computers & Education*, *114*, 206-221.
- Segalowitz, N., & Hulstijn, J. H. (2009). Automaticity in bilingualism and second language learning. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of Bilingualism: Psycholinguistic Approaches* (pp. 371– 388). Oxford University Press.

- Shook, A., Goldrick, M., Engstler, C., & Marian, V. (2015). Bilinguals show weaker lexical access during spoken sentence comprehension. *Journal of Psycholinguistic Research*, *44*(6), 789–802. 10.1007/s10936-014-9322-6
- Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv*. 10.48550/arXiv.1703.05339
- Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, *84*, 101017. 10.1016/j.wocn.2020.101017
- Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language & Linguistics Compass*, *9*, 311-327. 10.1111/lnc3.12151
- Stone, K., Lago, S., & Schad, D. (2021). Divergence point analyses of visual world data: Applications to bilingual research. *Bilingualism: Language and Cognition*, *24*(5), 833-841. 10.1017/S1366728920000607
- van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190. <https://doi.org/10.1080/17470218.2013.850521>
- Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. 10.1016/j.ijpsycho.2011.09.015
- Van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2020). *Itsadug: interpreting time series and autocorrelated data using GAMMs*. R package version, 2.4.1.
- Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of*

Phonetics, 70, 86–116. 10.1016/j.wocn.2018.03.002

Wood, S. (2017). *Generalized Additive Models: an introduction with R*. 2nd edition. Boca Raton: CRC press.

Zirnstien, M., van Hell, J. G., & Kroll, J. F. (2018). Cognitive control ability mediates prediction costs in monolinguals and bilinguals. *Cognition*, 176, 87–106. 10.1016/j.cognition.2018.03.001