Computational cognitive modeling of predictive sentence processing in a second language

Umesh Patil (umesh.patil@gmail.com)
University of Cologne
50923 Cologne, Germany

Sol Lago (sollago@em.uni-frankfurt.de)
Goethe University Frankfurt
60629 Frankfurt, Germany

Abstract

We propose an ACT-R cue-based retrieval model of the real-time gender predictions displayed by second language (L2) learners. The model extends a previous model of native (L1) speakers according to two central accounts in L2 sentence processing: (i) the Interference Hypothesis, which proposes that retrieval interference is higher in L2 than L1 speakers; (ii) the Lexical Bottleneck Hypothesis, which proposes that problems with gender agreement are due to weak gender representations. We tested the predictions of these accounts using data from two visual world experiments, which found that the gender predictions elicited by German possessive pronouns were delayed and smaller in size in L2 than L1 speakers. The experiments also found a "match effect", such that when the antecedent and possessee of the pronoun had the same gender, predictions were earlier than when the two genders differed. This match effect was smaller in L2 than L1 speakers. The model implementing the Lexical Bottleneck Hypothesis captured the effects of smaller predictions, smaller match effect and delayed predictions in one of the two conditions. By contrast, the model implementing the Interference Hypothesis captured the smaller prediction effect but it showed an earlier prediction effect and an increased match effect in L2 than L1 speakers. These results provide evidence for the Lexical Bottleneck Hypothesis, and they demonstrate a method for extending computational models of L1 to L2 processing.

1 Introduction

Although the world population is quickly becoming bilingual, there are very few computational models of bilingual sentence processing. Because most of these models were developed for technological applications—e.g., automatic translation—, this results in a scarcity of models that are cognitively realistic or even evaluable with human data (Frank, 2021; Frank et al., 2016; Hinaut et al., 2015; Hendriks and Vogelzang, 2020). However, such models are crucial to develop computational research that is informed by state-of-the-art psycholinguistic work. With this goal, we propose a computational cognitive model of bilingual processing built in an architecture, ACT-R, which is designed to model human cognition and can be evaluated with human data (Anderson, 2007; Ritter et al., 2019). The ACT-R architecture has also been used to model a number of linguistic phenomena, such as retrieval interference effects in linguistic dependency resolution (Vasishth et al., 2008), the influence of prominence on pronoun resolution (Patil et al., 2016b; Patil and Schumacher, 2022), the effect of memory load on sentence processing (van Rij et al., 2013), sentence processing in patients with aphasia (Crescentini and Stocco, 2005; Patil et al., 2016a), the interaction of sentence processing and eye movements (Engelmann et al., 2013), and incremental formal semantic processing (Brasoveanu and Dotlačil, 2020).

Accounts of bilingual processing can be divided in terms of how they explain differences between native (L1) and non-native (L2) processing. Here we focus on two different explanations of L1–L2 differences. The first, the Interference Hypothesis (IH), makes reference to the cue-based retrieval theory (Cunnings, 2017b,a). The Interference Hypothesis stipulates that memory retrieval is key for different parts of sentence processing, including the processing of non-local pronoun-antecedent dependencies like "*John noticed that* ***Richard$_i$*** *had cut* ***himself$_i$*** *with a knife*". When "*himself*" is encountered, speakers attempt to retrieve an antecedent matching the pronoun features. Retrieval success requires suppressing interfering elements that match some but not all of the relevant features (e.g., "*John*" has the appropriate gender and number features but not the syntactic ones, because it is outside the clause of the pronoun). The Interference Hypothesis proposes that L1 and L2 speakers

Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), pages 339 - 349

are similar in their likelihood of initiating retrieval operations, but that L2 speakers are more prone to interference, yielding more misretrievals (e.g., wrongly recovering "*John*" as the pronoun's antecedent).

By contrast, the Lexical Bottleneck Hypothesis (LBH) is framed within so-called capacity-based accounts, which propose that L1–L2 differences arise because speakers process an L2 in a noisier cognitive architecture, resulting in slower and more error-prone parsing (Just and Carpenter, 1992; McDonald, 2006; Hopp, 2022). The Lexical Bottleneck Hypothesis proposes that L1–L2 parsing differences are due to variability in the bilingual lexicon. Specifically, because lexical processing "precedes and feeds into syntactic processing, key characteristics of bilingual lexical processing may cause aspects of non-target parsing" (Hopp, 2018, pp. 6). With regard to grammatical features like gender—the focus of this paper—the claim is that L2 speakers fail to use this information for syntactic processing because L2 words have weaker or more unstable gender representations, making the retrieval of gender information less robust in L2 than in L1. An additional factor—not modeled here—is L1 transfer, such that L2 gender processing may be harder in syntactic contexts that differ between the L1 and the L2.

We evaluate the Interference Hypothesis and the Lexical Bottleneck Hypothesis by using their claims to modify an ACT-R model that was previously shown to capture L1 predictive processing (Patil and Lago, 2021). The predictions of the modified ACT-R models are evaluated against the results of two eye-tracking experiments that examined how L1 and L2 speakers use gender features to do memory retrieval and to predict upcoming referents (Stone et al., 2021b; Lago et al., under review). We show that the ACT-R version that implements the Lexical Bottleneck Hypothesis does a better job at capturing L2 gender predictions. Our results—although currently limited to gender—suggest that the Lexical Bottleneck Hypothesis provides a suitable framework to model the predictive use of morphosyntactic information in L2, and could be extended to other features such as number, case and animacy.

2 Modeling L2 processing

2.1 Starting point: The L1 model

We consider Patil and Lago's (2021) model of processing possessive pronouns as our starting point. They modeled visual-world eye-tracking data from Stone et al. (2021b) in ACT-R and the cue-based retrieval framework (CBR, henceforth) (Lewis and Vasishth, 2005; Lewis et al., 2006). Our goal is to model the L2 visual-world eye-tracking data from Lago et al. (under review) by modifying the model to reflect the processing assumptions of the IH and the LBH. The model has the following structure most of which is inherited from ACT-R and CBR.

Sentence processing takes place as an incremental word-by-word left-corner parsing. Parsing rules are part of ACT-R's procedural memory, whereas the lexical entries, syntactic phrases and the incremental parse tree (NP, DP, VP, IP, etc.) are part of ACT-R's declarative memory. Each declarative memory element, called a *chunk*, has an activation associated with it which is determined by the equation 1.[1] At each input word, parsing rules are applied on chunks that are available in short-term memory to process the word. If a required chunk is not available in short-term memory, it is retrieved from declarative memory by specifying a set of cues as feature-value pairs, a cue-based retrieval mechanism.

The speed, accuracy and success of retrieving a chunk depends on its current activation level. The activation of $chunk_i$ is influenced by its usefulness in the past (the *base level activation* B_i), relevance in the current context (the *spreading activation* received through retrieval cues which is determined by the first summation component in eq. 1), degree of match with the retrieval request (the *partial matching* determined by the second summation component in eq. 1) and stochastic noise (ϵ_i). The *strength of association* S_{ji} is calculated by eq. 2 which is influenced by fan_j, the number of chunks matching cue_j. The value for Mji is calculated by the degree of match between a retrieval cue (cue_j) and $chunk_i$. The values for W (the maximum spreading activation), P (the partial match scale) and S (the maximum associative strength) in the calculation of A_i are constants across all simula-

[1] This is a simplified version of ACT-R's activation equation and it represents how activation is calculated in CBR. The equation can be simplified further to have only one summation term but for comparability with the original ACT-R equation we have kept the two summations separate.

tions and are set as ACT-R's parameter values.

$$A_i = B_i + \sum_{cue_j} WS_{ji} + \sum_{cue_j} PM_{ji} + \epsilon_i \quad (1)$$

$$S_{ji} = S - ln(fan_j) \quad (2)$$

For modeling the visual-world eye-tracking task from Stone et al. (2021b), Patil and Lago (2021) extended the existing architecture with the following new assumptions: (i) the model predicts the target picture at each input word, (ii) prediction of the target picture is implemented as a cue-based memory retrieval, and (iii) the probability of fixating an object is determined by the activation of the chunk representing that object. To incorporate the variable influence of different retrieval cues, Patil and Lago (2021) also proposed a cue-weighting mechanism as a modification to the *strength of association* equation (as in eq. 3) such that the amount of activation spreading from cue_j to $chunk_i$ is influenced by the importance of that cue.

$$S_{ji} = weight_j S - ln(fan_j) \quad (3)$$

The next two sections describe two possible modifications of the L1 model to implement two theories of L2 processing: (i) the Interference Hypothesis, and (ii) the Lexical Bottleneck Hypothesis.

2.2 The IH model

IH proposes that L2 speakers are prone to higher interference compared to L1 speakers and that leads L2 speakers to misretrieve non-target elements more often during sentence processing. Although IH is not a computationally implemented theory, it is described in terms of the CBR framework of sentence processing, and, hence, an L1 model implemented in CBR can be straightforwardly extended to L2 processing. In ACT-R and CBR, misretrievals due to interference take place through the mechanism of partial matching (the second summation term in eq. 1). Partial matching enables non-target chunks (chunks that match some of the cues from the retrieval request but not all) to be considered in the retrieval process. Due to random fluctuation in the activation of chunks (the random noise ϵ_i eq. 1), partially matching chunks can get retrieved instead of the target chunk in some of the retrieval requests (a misretrieval). Misretrievals happen in L1 speakers as well. In fact, in psycholinguistics misretrievals due to partial matching have been suggested to explain some of the grammatical illusions such as agreement attraction and spurious NPI licensing (Wagers et al., 2009; Vasishth et al., 2008). But as per IH, misretrievals happen more often in L2 speakers.

In ACT-R the frequency of misretrievals is controlled by defining the penalty to the activation of a chunk when its feature doesn't match the retrieval cue. The penalty is specified through a parameter called *maximum difference*, the highest penalty for a perfect mismatch. By default the value of *maximum difference* is -1.[2] This means that the activation penalty increases as a function of the number of cues mismatched by a chunk, making its retrieval less likely. The value of *maximum difference* can be changed to calibrate the penalty of a mismatch. Reducing this penalty leads the non-target chunks to get retrieved more often, i.e. higher misretrievals. We propose that reducing the value of *maximum difference* would be the way of extending the L1 model to L2 processing in terms of IH.

2.3 The LBH model

LBH proposes that L2 speakers fail to use grammatical features such as gender in syntactic processing because the gender representations of L2 words are weaker or more unstable, and speakers process an L2 in a noisier cognitive architecture. Although LBH is not specified in connection with a specific cognitive or sentence processing architecture, it can be realized in CBR. A possible implementation of LBH in the ACT-R and CBR frameworks could be done by: (i) having weaker representation of the gender feature in chunks representing various referents present in the input, and (ii) making the representations of the referents noisier compared to their representations in the L1 model.

In a typical CBR model the gender features have discrete values (e.g. *feminine*, *masculine* and *neuter*), and chunks denoting various referents have a certain, relatively low, activation noise associated with them (ϵ_i eq. 1). We propose the following two modifications to the L1 model for implementing LBH.

First, the gender features have values that are encoded as weaker than corresponding L1 values – *feminine-weak*, *masculine-weak* and *neuter-weak*. This leads the corresponding chunk to only *weakly* match a retrieval cue for a specific gender. For

[2] Conversely, ACT-R also provides a parameter called *maximum similarity* that specifies the least penalty for a perfect match which is set to 0 by default.

example, a chunk for a feminine referent encoding gender as *feminine-weak* will weakly match a retrieval request of type 'gender = feminine'. As a consequence, the chunk receives less spreading activation from the retrieval cue than a chunk that encodes gender clearly as *feminine*. This is equivalent to saying that the referent does not have exactly the same value of the feature as the parser expects but is similar enough to be considered in the retrieval request.

We implement this behavior partly by using ACT-R's built-in functionality of setting similarities between a retrieval cue and a feature value (the M_{ji} values in the *partial matching* component of eq. 1), and partly by modifying the *spreading activation* component in eq. 1. The *partial matching* component in the activation equation sums to a negative value since M_{ji} values vary between 0 (for a perfect match between a retrieval cue and a feature) and -1 (for a mismatch); effectively a penalty to a chunk for not fully matching a retrieval request. In ACT-R by default M_{ji}'s are either equal to the value of the parameter *maximum similarity* (0 by default) or to the value of the parameter *maximum difference* (-1 by default), but they can be set to any value between 0 and -1 to reflect the degree of similarity between a pair of values (e.g. feminine and feminine-weak or red and maroon). We propose that for the LBH model the similarity between an expected gender and the weaker value lies between the two extremes 0 and -1 but closer to 0 since a weak gender is more similar than dissimilar to the corresponding strong gender. Reciprocally, the similarity between an expected gender and any other weak gender (e.g. feminine and masculine-weak) also lies between the two extremes and, in this case, closer to -1 since it is more dissimilar than the same weaker gender but less dissimilar than a different strong gender (e.g. feminine and masculine).

We also propose that this graded similarity between a cue and a feature value also influences the *spreading activation* component. This is not part of the original ACT-R framework, so we consider a further modification to the computation of the *strength of association*, S_{ji}, as in 4–6. The *strength of association* now reflects how well the feature value matches the retrieval cue. This modification has influence on the calculations of activation only when a cue and a value don't perfectly match or mismatch, when they do, the value of activation is the same as in the original ACT-R framework. When a value perfectly matches a requested cue (i.e. M_{ji} = 0) eq. 4 reduces to eq. 3, and when a value perfectly mismatches (i.e. M_{ji} = -1) it leads to no activation spreading.

$$S_{ji} = weight_j sim_{ji} S - ln(fan_j) \quad (4)$$

$$fan_j = \sum_{chunk_k} sim_{jk} \quad (5)$$

$$sim_{jk} = (1 + M_{jk}) \quad (6)$$

To implement the LBH proposal that speakers process an L2 in a noisier cognitive architecture, we propose a second change to the L1 model in terms of its activation noise. This change is more intrinsic to ACT-R because the activation equation includes a noise term that controls the random fluctuations in the activation of chunks (ϵ_i in eq. 1). Higher noise value makes the representation of chunks noisier. We suggest that a noisier L1 model, along with weaker gender representation, should be the L2 model representing LBH.

Note that an alternative implementation of the LBH could test if both weak gender and noisier representations are necessary to capture the L2 data. Moreover, ACT-R also assumes another type of noise, the noise in procedural memory. It is conceivable that the noisier representation proposed by the LBH is realized as noisier procedural memory (e.g. Patil et al. 2016a used the noise in procedural memory to model data from patients with aphasia). However, we consider that weak gender and activation noise are the closest realization of the LBH in ACT-R, and a good starting point for modeling LBH. We leave other possible implementations for future research.

3 Human data

The human data was taken from two visual world eye-tracking experiments with the same materials and design but two different groups of participants: 74 L1 German speakers (Stone et al., 2021b, Experiment 2) and 132 L2 German learners (Lago et al., under review, Experiment 2). The L2 group comprised native speakers of Spanish and English. Because they did not differ behaviorally, the comparisons below consider a unified group of L2 participants. We reanalyzed the two experiments to directly compare L1 and L2 processing.

In the experiments, L1 and L2 participants were asked to help find the belongings of two fictional characters, Martin and Sarah. They were told that

they would see images and hear instructions, and that their task was to select the object mentioned by the instruction. The instructions always contained a possessive pronoun doubly-marked for gender: the gender of the pronoun stem (*sein-*/*ihr-*) agreed in gender with the antecedent (*Martin* or *Sarah*). The gender of the pronoun suffix agreed in gender with the upcoming noun, which allowed participants to predict the identity of the target object prior to hearing it in the instruction, e.g.: 'Click on **his.**$_{\textbf{MASC}}$ blue.$_{\text{MASC}}$ button.$_{\text{MASC}}$'.

The experimental trials showed 2 colored objects: a target object (e.g. a blue button.$_{\text{MASC}}$) and a competitor of a different gender (e.g. a blue bottle.$_{\text{FEM}}$). The 96 items were distributed in two conditions (1). In the MATCH condition, the possessor and target noun had the same gender, i.e., both masculine or both feminine. In the MISMATCH condition, the possessor mismatched the gender of the target object but matched the competitor's. The results of the experiments showed that the gender of the pronoun was used predictively, such that participants showed a target-over-competitor looking preference prior to hearing the noun. In addition, there was a "match effect", with predictions starting earlier in the match than in the mismatch condition (Figure 1). We examined whether the size and onset of predictions and/or the onset of match effects differed between L1 and L2.

(1) a. **MATCH condition**
Klicke auf seinen blauen Knopf!
Click on his.MASC blue.MASC button.MASC

b. **MISMATCH condition**
Klicke auf ihren blauen Knopf!
Click on her.MASC blue.MASC button.MASC

The size of predictions was quantified as the target-over-competitor looking preference in the entire time-window before the target noun was heard (i.e., from pronoun onset to noun onset plus 200ms to account for saccade planning). The onset of prediction was quantified as the earliest point in time at which fixations to the target object significantly differed from fixations to the competitor. This timepoint, together with a 95% confidence interval, was taken as the prediction onset (Stone et al., 2021a). Our L1–L2 comparisons revealed the following differences: (i) The size of predictions was approximately 9 percentage points smaller in L2 than in L1 (henceforth **SMALLER-PREDICTION**). The target-over-competitor advantage was 58 [56, 60] % in the L2 group vs. 67 [65, 69] % in the L1 group. (ii) The onset of predictions was always later in L2 than L1 (**LATER-PREDICTION**). In the match condition, the difference in L1–L2 onsets was 211 [60, 320] ms. In the mismatch condition, the difference in L1–L2 onsets was 108 [60, 160] ms. (iii) The match effect—the difference between mismatch vs. match onsets—occurred in both groups: L1 match effect 303 [160, 400] ms and L2 match effect 200 [120, 280] ms. The match effect in onset times was numerically smaller in the L2 group (**SMALLER-MATCH**), but the between-group difference was not statistically reliable (as evidenced by the 95% CI crossing 0): 103 [-40, 220] ms.

4 Computational models

4.1 Modeling details

We generate predictions of IH model and LBH model based on the L2 modeling hypotheses and extensions proposed in sections 2.2 and 2.3. We use the L1 model reported in Patil and Lago (2021) and extend the model to capture the effects of L2 processing from Lago et al. (under review). The goal is to capture the three L2 vs. L1 effects observed in the data presented in section 3: (***SMALLER-PREDICTION***) smaller size of predictions in L2, (***LATER-PREDICTION***) later prediction onsets in L2 for both MATCH and MISMATCH conditions, and (***SMALLER-MATCH***) smaller match effect in L2.

The IH and LBH models were used to generate predicted fixation patterns from the onset of the (possessive) pronoun to the onset of noun. From these predicted fixation profiles, the three effects concerning L1-L2 differences were calculated as follows. The SMALLER-PREDICTION effect was calculated by averaging the size of predictions (i.e. mean fixation probability) in the temporal window between the onsets of the possessive pronoun and the noun across match and mismatch conditions. The effect of LATER-PREDICTION was calculated by subtracting the onset predicted by the L1 model from the onset predicted by each of the two L2 models for each condition separately. Finally, the SMALLER-MATCH effect was calculated by subtracting the prediction onsets of the match vs. mismatch conditions. All model predic-

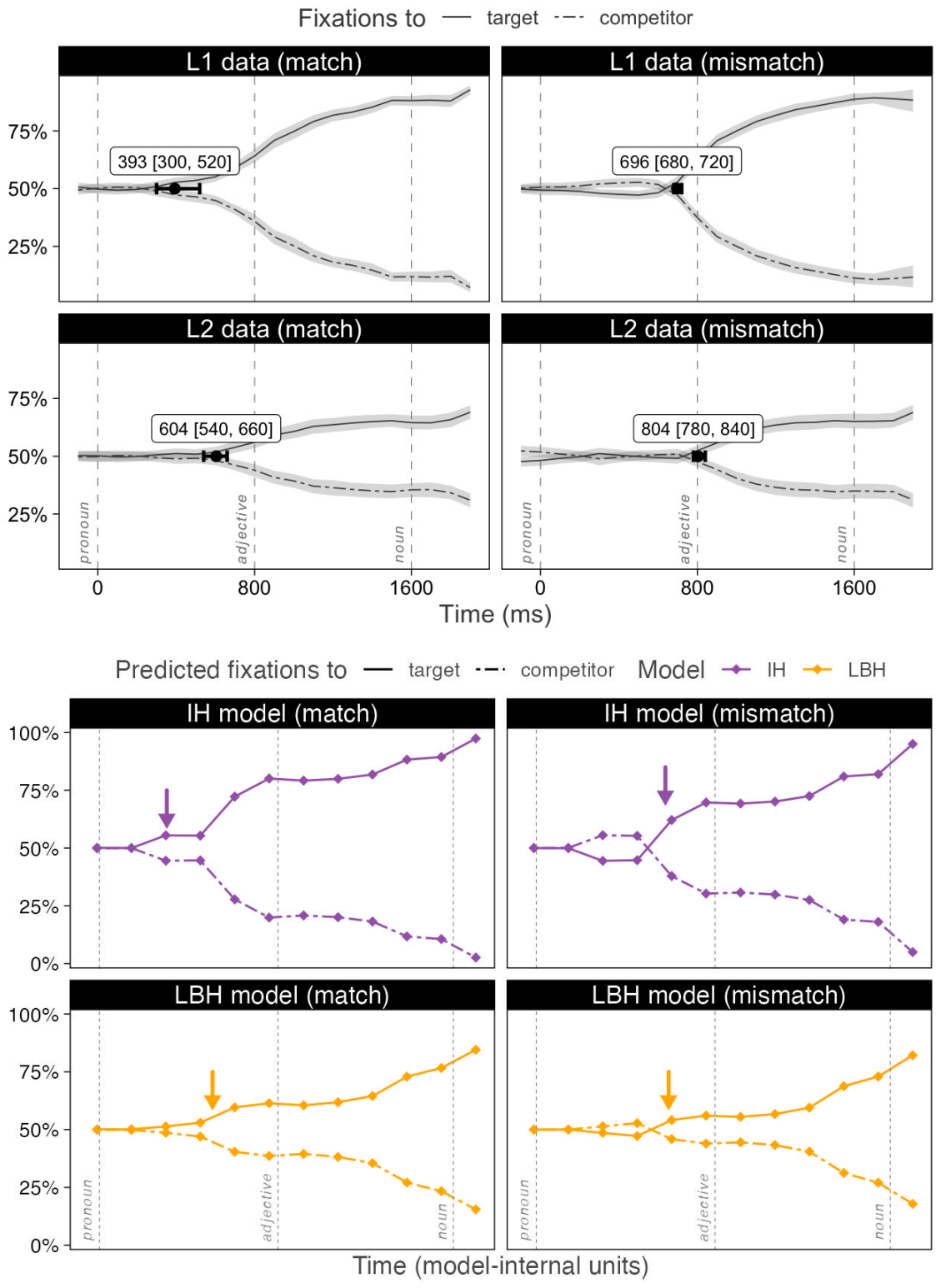

Figure 1: **Human data (top two rows)**: Fixation curves to the target and competitor object averaged across items and participants. The predictive time-window extended from the onset of the pronoun to the onset of the noun, shifted 200 ms to the right. The x-axis is time-locked to the pronoun. Estimated predictive onsets and their 95% confidence intervals (in ms) are overlaid on the fixation curves in each condition. **Model data (bottom two rows)**: Predictions of the model for fixation probabilities to the target and competitor object in the L2 groups. The x-axis reflects processing time in model-internal units. Vertical arrows show the model-predicted onsets.

tions are generated by running 100,000 simulations of each model including the L1 model. Due to ACT-R's stochastic noise component, some of the predicted the values deviate from the ones reported in Patil and Lago (2021), but the qualitative effects remain the same as reported by them. We consider our calculations of L1 predictions as reliable as theirs because we calculated the values by running a higher number of simulations (10,000 vs. 100,000). The ACT-R parameter values that were changed to implement the assumptions of the IH and LBH models are listed in Table 1. We also tested how the predictions of the two models varied as a function of variation in the values of these parameters (see section 4.3).

4.2 Model predictions

The predictions of the two L2 models for prediction onsets and fixation probabilities are shown in Figure 1 (lower panels). The three effects observed in the data and the corresponding predictions of the two L2 models are summarized in Table 2. Both L2 models capture the SMALLER-PREDICTION effect — they show a smaller prediction size compared to the L1 model; however, numerically, the LBH model's prediction is closer to the human data. With regard to the LATER-PREDICTION effect, it is only captured by the LBH model and only in the match condition. While LBH also predicts a delayed L2 prediction onset in the mismatch condition (18 ms), visual inspection of the data revealed that the effect was driven by a few outlier simulations (around 10% of the simulations). On the other hand the IH model doesn't capture the LATER-PREDICTION effect in either the match or mismatch conditions. The SMALLER-MATCH effect is captured only by the LBH model but not by the IH model, which predicts the effect to be in the opposite direction. In both conditions the IH model in fact predicts earlier prediction onsets for L2 than L1 speakers (a negative effect).

4.3 Model predictions across parameter variation

To test if the predictions of the two models were restricted to the specific values selected for the parameters, we generated predictions of the models by varying the parameter values around the values we selected. We only varied the parameters that were modified for implementing the IH and LBH, and only within a range that was still meaningful to represent the hypotheses that were implemented. For parameter variation we randomly sampled 200 values from a uniform distribution with bounds defining a range of values around a selected parameter value. For each random value of a parameter we generated predictions by running 1000 simulations of the model.

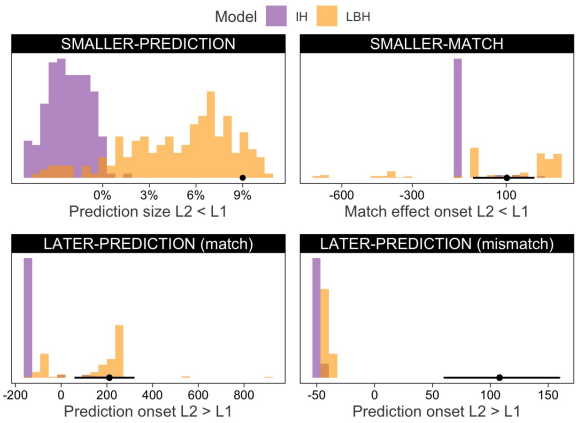

Figure 2: Distributions in terms of histograms of the effects predicted by the IH and the LBH models. The y-axis depicts the frequency of the predicted effects and it has different heights for different panels. Black dots represent the mean effects observed in the L2 data. Prediction distributions for the IH model are generated by varying ACT-R's *maximum difference* parameter, whereas those for the LBH model are generated by varying the *activation noise* parameter and the similarity values between strong and weak genders. These are the same parameters that were used to implement the respective L2 hypothesis through those models (c.f. Table 1).

For the IH model we varied ACT-R's *maximum difference* (penalty) parameter between the values of -0.7 to -0.3 ($U(-0.7, -0.3)$) because a value higher than -0.3 would be too close to the value of no penalty (i.e. 0) for a retrieval cue mismatch and a value lower than -0.7 would be too close to the default penalty (i.e. -1) in the L1's model. For the LBH model we varied ACT-R's *activation noise* parameter and the similarity values between strong and weak genders. We varied the *activation noise* in the range 0.3 to 0.7 ($U(0.3, 0.7)$), and the similarity between the strong and weak gender values of the same gender in the range -0.4 to -0.1 ($U(-0.4, -0.1)$) and between the strong and weak values of different genders in the range -0.9 to -0.6 ($U(-0.9, -0.6)$). Since the *activation noise* value for the L1 model was 0.25 we chose values higher than that, but at the same time if the *activation noise* is too high, the activation has too high impact of the random noise compared to other crucial com-

Table 1: ACT-R parameters values that were modified to model the proposals of the IH and LBH of L2 processing. The column "L1" show the original parameter values used in Patil and Lago (2021), while the columns "IH" AND "LBH" show the values modified to model the L2 data. The values that were modified are in bold face. All other ACT-R parameters had the same value as used in the L1 model.

ACT-R parameter	L1	IH	LBH
Activation noise (ANS)	0.25	0.25	**0.5**
Maximum difference (MD)	-1	**-0.5**	-1
Similarity between weak & strong gender values of the same gender	—	—	**-0.25**
Similarity between weak & strong gender values of different genders	—	—	**-0.75**

Table 2: Comparison of effects of interest in the L2 human data and in the predictions of the IH and LBH models.

Effect	Condition	Human data	IH model	LBH model
SMALLER-PREDICTION		9%	3.2%	11.1%
LATER-PREDICTION	match	211 [60, 320] ms	-58 ms	207 ms
	mismatch	108 [60, 160] ms	-1 ms	18 ms
SMALLER-MATCH		103 [-40, 220] ms	-57 ms	189 ms

ponents influencing the activation and hence the retrievals (see eq. 1). For similarity, values below -0.4 would mean that the strong and weak genders are 40% or more dissimilar, and values above -0.1 would mean they are almost similar (less than 10% dissimilar). The range for similarity between the strong and weak values of different genders was just a mirror image of the range for similarity between the strong and weak values of the same gender in the interval $[0, -1]$.[3]

The distribution of the three effects of interest for above-mentioned range of parameter values for the two L2 models are shown in Figure 2, along with the mean effects observed in the data. A visual inspection supports the generalizations drawn in section 4.2 — the LBH captures the effects SMALLER-PREDICTION, LATER-PREDICTION in the match condition (but not in the mismatch condition) and SMALLER-MATCH for most of the parameter combinations, whereas the IH qualitatively (but not quantitatively) captures the SMALLER-PREDICTION effect (since it predicts positive values for the effect) but barely captures any of the other effects.

[3]As another approach one could also vary the values of these three parameters for a broader range of the intervals. Although these values might not represent either of the theories, they are informative to find the broadest range of values for which the current implementation does not break. Moreover, it is also possible to test other parameters in ACT-R that do not represent either of the L2 hypotheses, the "hyperparameters", to see if they influence predictions. Due to time constraints, we restricted our simulations to narrower intervals around the chosen values.

5 Discussion

We proposed computational cognitive models of two main theories of L2 processing — the Interference Hypothesis and the Lexical Bottleneck Hypothesis. Both are verbally stated theories of processing differences between L2 and L1 speakers, and ours is, to our knowledge, the first computational cognitive realization of those theories. The theories were implemented by extending an existing L1 processing model (Patil and Lago, 2021). We used visual-world eye-tracking data from a predictive sentence processing task to test the models. The results showed that the LBH performed better than IH in capturing the three key effects observed in the data. With the exception of one effect, the IH predicted effects that were opposite to the ones observed in human speakers. Overall the LBH appears to be a more likely explanation of L2 sentence processing as far as the predictive use of gender in processing is concerned. Therefore, we propose that the well-attested difficulty shown by L2 speakers in using gender predictively (as compared to L1 speakers) is more likely attributable to problems in how L2 speakers represent gender information in a non-native language (Gollan et al., 2008; Kroll and Gollan, 2014; Hopp, 2018) and/or to difficulties in using this information as quickly as L1 speakers (Grüter et al., 2017; Kaan, 2014).

An important qualification is that the two implementations evaluated here did not model the potential effect of L1 transfer. Recall that the L2 group consisted of both Spanish and English learners of

German. Because the majority of nouns used in the human experiments had the same gender across Spanish and German, and because there was no evidence of between-group differences, we think that the current dataset is not suitable for modeling L1 transfer effects. Research using other datasets will be relevant to address the role of L1 transfer, which is hypothesized to play a role in the Lexical Bottleneck Hypothesis (Hopp, 2018, 2022). The role of L1 transfer in the IH is less clear, but it may affect the current implementation if, for example, both L1- and L2-based gender features are available for retrieval in the memory chunks corresponding to the objects on-screen.

The effects reported in the L1 and L2 data were possibly born out of retrieval interference during predictive processing (Patil and Lago, 2021). Hence we expected the IH account to capture the effects better as the IH is rooted in the cue-based retrieval framework of sentence processing, and cue-based retrieval theory has rendered explanation to various psycholinguistic phenomenon through retrieval interference. A possible reason for the IH predicting opposite patterns to the ones observed in the data could because retrieval interference as per the cue-based retrieval theory can lead to two opposite processing phenomena — inhibitory vs. facilitory processing — depending on the context (Dillon et al., 2013; Patil et al., 2016b; Parker et al., 2017). The precise nature of the interference effect in a given context can only be predicted through an actual implementation of the model. Our results emphasize the importance of computationally formalizing the predictions of the cue-based retrieval theory in particular (Vasishth et al., 2019), and of verbal theories in cognition in general (Guest and Martin, 2021).

Although the LBH model captured crucial patterns in the differences between L2 and L1 processing, one serious limitation of the model (and also of the IH model) was in terms of capturing the effect of delayed prediction onsets (LATER-PREDICTION) in the mismatch condition for L2 speakers. Since both the L2 models failed at capturing this effect, we think it is also unlikely that a combination of the two models would be able to capture this effect. This also implies that the gender prediction in L2 speakers possibly also involves a process that cannot be explained by either of the hypotheses. We think a computational implementation of another L2 processing hypothesis, in combination with the LBH model, might help capture this effect.

Acknowledgements

The research for this project has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): Project-ID 281511265 – SFB 1252 "Prominence in Language" in the project C07 "Forward and backward functions of discourse anaphora" at the University of Cologne, Department of German Language and Literature I, Linguistics, and Project-ID 317308350 – "AGREE: Agreement in native and second language processing". We furthermore thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing time on the DFG-funded (Funding number: INST 216/512/1FUGG) High Performance Computing (HPC) system CHEOPS as well as support.

Supplementary files

The model code and supplementary files are available at: `https://osf.io/p28k6/`

References

John R. Anderson. 2007. *How can the human mind occur in the physical universe?* Oxford series on cognitive models and architectures. Oxford University Press, New York, NY, US.

Adrian Brasoveanu and Jakub Dotlačil. 2020. *Computational Cognitive Modeling and Linguistic Theory*. Language, Cognition, and Mind. Springer International Publishing, Cham.

Cristiano Crescentini and Andrea Stocco. 2005. Agrammatism as a failure in the lexical activation process. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Lawrence Erlbaum Associates.

Ian Cunnings. 2017a. Interference in native and non-native sentence processing. *Bilingualism: Language and Cognition*, 20(4):712–721.

Ian Cunnings. 2017b. Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, 20(4):659–678.

Brian Dillon, Alan Mishler, Shayne Sloggett, and Colin Phillips. 2013. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103.

Felix Engelmann, Shravan Vasishth, Ralf Engbert, and Reinhold Kliegl. 2013. A framework for modeling

the interaction of syntactic processing and eye movement control. *Topics in Cognitive Science*, 5(3):452–474.

Stefan L. Frank. 2021. Toward computational models of multilingual sentence processing. *Language Learning*, 71(S1):193–218.

Stefan L. Frank, Thijs Trompenaars, and Shravan Vasishth. 2016. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, 40(3):554–578.

Tamar H. Gollan, Rosa I. Montoya, Cynthia Cera, and Tiffany C. Sandoval. 2008. More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, 58(3):787–814.

Theres Grüter, Hannah Rohde, and Amy J. Schafer. 2017. Coreference and discourse coherence in l2: The roles of grammatical aspect and referential form. *Linguistic Approaches to Bilingualism*, 7(2).

Olivia Guest and Andrea E. Martin. 2021. How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, 16(4):789–802. PMID: 33482070.

Petra Hendriks and Margreet Vogelzang. 2020. Pronoun processing and interpretation by l2 learners of italian: Perspectives from cognitive modelling. *Discours : A journal of linguistics, psycholinguistics and computational linguistics*, 26.

Xavier Hinaut, Johannes Twiefel, Maxime Petit, Peter Dominey, and Stefan Wermter. 2015. A recurrent neural network for multiple language acquisition: Starting with english and french. In *Proceedings of the 2015 International Conference on Neural Information Processing Systems (NIPS 2015), Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, volume 1583, Montreal, CA.

Holger Hopp. 2018. The bilingual mental lexicon in l2 sentence processing. *Second Language*, 17:5–27.

Holger Hopp. 2022. Second language sentence processing. *Annual Review of Linguistics*, 8(1):235–256.

Marcel A. Just and Patricia A. Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1):122–149.

Edith Kaan. 2014. Predictive sentence processing in l2 and l1: What is different? *Linguistic Approaches to Bilingualism*, 4:257–282.

Judith F. Kroll and Tamar H. Gollan. 2014. *Speech planning in two languages: What bilinguals tell us about language production.*, Oxford library of psychology., pages 165–181. Oxford University Press, New York, NY, US.

Sol Lago, Kate Stone, Elise Oltrogge, and João Veríssimo. under review. Possessive processing in bilingual comprehension. Submitted to Language Learning.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29:1–45.

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10):447–454.

Janet L. McDonald. 2006. Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, 55(3):381–401.

Dan Parker, Michael Shvartsman, and Julie A. Van Dyke. 2017. The cue-based retrieval theory of sentence comprehension: New findings and new challenges.

Umesh Patil, Sandra Hanne, Frank Burchert, Ria De Bleser, and Shravan Vasishth. 2016a. A computational evaluation of sentence processing deficits in aphasia. *Cognitive Science*, 40(1):5–50.

Umesh Patil and Sol Lago. 2021. Prediction advantage as retrieval interference: an ACT-R model of processing possessive pronouns. In *Proceedings of the 19th International Conference on Cognitive Modeling*, pages 213–219. University Park, PA: Applied Cognitive Science Lab, Penn State.

Umesh Patil and Petra B. Schumacher. 2022. Modeling prominence constraints for German pronouns as weighted retrieval cues. In *Proceedings of the 20th International Conference on Cognitive Modeling*.

Umesh Patil, Shravan Vasishth, and Richard L. Lewis. 2016b. Retrieval interference in syntactic processing: The case of reflexive binding in english. *Frontiers in Psychology*, 7:329.

Frank E. Ritter, Farnaz Tehranchi, and Jacob D. Oury. 2019. Act-r: A cognitive architecture for modeling cognition. *WIREs Cognitive Science*, 10(3):e1488.

Kate Stone, Sol Lago, and Daniel J. Schad. 2021a. Divergence point analyses of visual world data: applications to bilingual research. *Bilingualism: Language and Cognition*, 24(5):833–841.

Kate Stone, João Veríssimo, Daniel J. Schad, Elise Oltrogge, Shravan Vasishth, and Sol Lago. 2021b. The interaction of grammatically distinct agreement dependencies in predictive processing. *Language, Cognition and Neuroscience*, 36(9):1159–1179.

Jacolien van Rij, Hedderik van Rijn, and Petra Hendriks. 2013. How wm load influences linguistic processing in adults: A computational model of pronoun interpretation in discourse. *Topics in Cognitive Science*, 5(3):564–580.

Shravan Vasishth, Sven Brüssow, Richard L. Lewis, and Heiner Drenhaus. 2008. Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, 32(4):685–712.

Shravan Vasishth, Bruno Nicenboim, Felix Engelmann, and Frank Burchert. 2019. Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, 23(11):968–982.

Matthew W. Wagers, Ellen F. Lau, and Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237.