Effect of case markers during agreement production: A model comparison using Armenian forced choice data

Pranab Bagartti, Samar Husain
Indian Institute of Technology Delhi, India
{pranab.bagartti, samar} @hss.iitd.ac.in

Abstract

Agreement attraction errors, where the verb erroneously agrees with a non-subject noun, have been a useful tool for investigating processes that subserve sentence production. Research has shown that case markers play an important role in modulating such errors. These effects have been argued to arise due to an underlying cue-based retrieval system. However, subsequent research in Armenian has challenged this conclusion (Avetisyan et al., 2020), arguing against a cue-based retrieval account.

The current paper revisits the Armenian production data through computational modeling. Specifically, we implemented three distinct models and compared their predictions; we compare (a) a cue-based retrieval model, (b) a feature migration model, and (c) a case as markers for agreement prediction model. Our model comparison results show that a case as markers for agreement prediction model followed by an inference component explains the effect of case better than the cue-based retrieval model as well as the feature migration model.

Keywords: agreement attraction, case marking, cue-based retrieval, feature migration, Armenian, sentence production, computational modeling

1 Introduction

Agreement attraction errors, where the verb erroneously agrees with a non-subject/non-target noun, have been used to investigate processes that subserve sentence production (e.g., Bock and Miller, 1991; Bock et al., 2001; Bock and Eberhard, 1993; Bock and Cutting, 1992; Eberhard et al., 2005; Badecker and Kuminiak, 2007; Vigliocco and Zilli, 1999; Eberhard et al., 2005; Eberhard, 1997; Staub, 2009, 2010; Franck et al., 2002). Research has shown that case markers play an important role in modulating such errors (e.g., Nicol and Anton-Mendez, 2009; Hartsuiker et al., 2001, 2003; Badecker and Kuminiak, 2007; Gillespie, 2011; Slioussar, 2018; Gillespie and Pearlmutter, 2013). For example, when speakers of Slovak were given incomplete sentences like (1a), where the distractor noun had an overtly distinctive case marker, they produced more accurate agreement at the verb in comparison to sentences like (1b), where the case on the distractor overtly resembled the subject's. Badecker and Kuminiak (2007) proposed that such errors arise due to an underlying cue-based retrieval system. Subsequent research in Armenian (Avetisyan et al., 2020), however, has challenged such an explanation, arguing against a cue-based retrieval account. The current paper utilizes model comparison to disambiguate between these diverging claims.

1. Example preambles from Slovak:

   a. Trest za vraždu…
   punishment.M.**NOM** for murder.F.**ACC**.SG…
   'The punishment for the murder…'

   b. Trest za krádež …
   punishment.M.**NOM** for theft.F.**NOM/ACC**.SG …
   'The punishment for the theft …'

In this paper, we focus on the Armenian production data from Avetisyan et al. (2020) for two reasons. First, to the best of our knowledge, this is the only openly available production dataset on case-mediated agreement attraction with accuracy and reaction times. Second, Armenian is also a useful test case on linguistic grounds: a nominative case marker on a noun reliably co-occurs with subject-verb agreement within its clause, unlike in Slovak (Example 1), where a nominative noun can appear without triggering agreement. This tight coupling between nominative case and agreement makes it different from other languages investigated for case and agreement. Additionally, this coupling is also

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 353–364

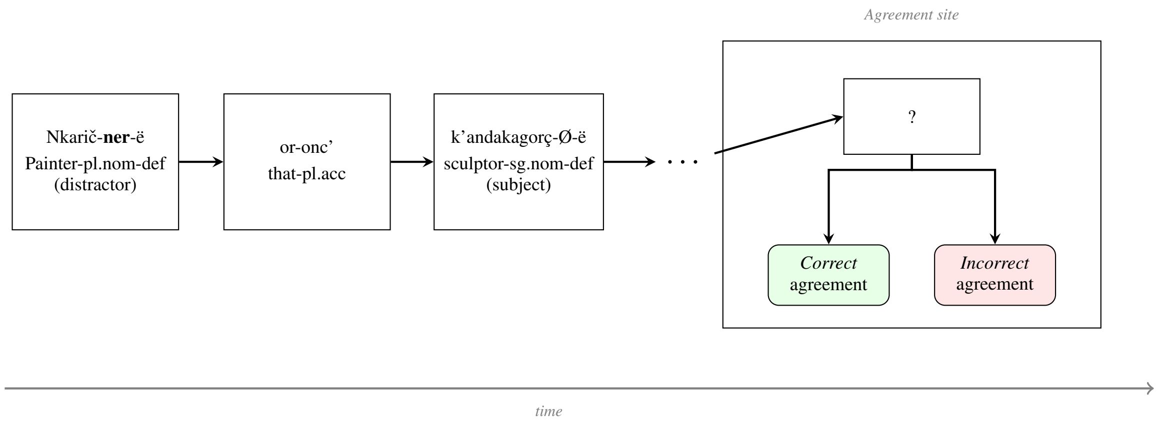

Figure 1: Schematic of the RSVP forced-choice task.

central to one of the theories that we posit moving forward (See §2.3).

In Avetisyan et al. (2020)'s task, sentences were presented one word at a time in a rapid manner (i.e., Rapid Serial Visual Presentation) followed by a forced choice between two continuations, i.e., correct or incorrect agreement (Staub, 2009, 2010). Importantly, the distractor appeared first, followed by the subject and then the agreement site where the participants choose from a correct and an incorrect response through key presses (See Figure 1).

The resulting data consisted of responses from 175 Eastern Armenian speakers. Avetisyan et al. (2020) found a limited/negligible effect of case with accuracy data but a more pronounced effect of case with their reaction time data. Does cue-based retrieval warrant such an effect?

A cue-based retrieval theory would assume that speakers use cues such as +nominative before producing an agreement. Any chunk that matches the cues gets a activation boost from the cue. In turn, the most activated chunk is retrieved. And finally, the verb shows agreement with the retrieved chunk.

Activation boost from the +nominative cue increases the chance of retrieving a subject noun when no other nouns match the cues as in 1a (also See Figure 2a). When a distractor matches the +nominative cue in sentences like 1b (also See Figure 2b), the activation boost is shared with the distractor, decreasing the chance of subject retrieval.

Regardless of the claim, whether a cue-based retrieval system can in fact produce the pattern of results reported by Avetisyan et al. (2020) is ultimately a question that can only be resolved through computational simulation. Such a computational modeling based evaluation has so far not been done. While cue-based retrieval has been computationally implemented for agreement attraction during comprehension (e.g., Lewis and Vasishth, 2005; Engelmann et al., 2019; Vasishth and Engelmann,

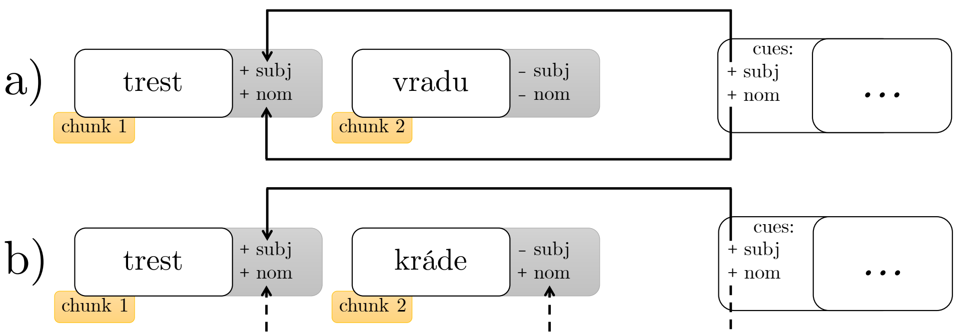

Figure 2: A schematic representation for agreement computation for 1a and 1b sentences from Slovak under a cue-based retrieval system. Here, +subj denotes a +subject feature/cue while +nom denotes a +nominative feature/cue. A +subject cue is consistent with previous implementations of cue-based retrieval for sentence processing (e.g., Engelmann et al., 2019; Yadav et al., 2022).

2021; Yadav et al., 2022), such an account remains unimplemented for agreement attraction during production.

In this paper, we present the first implementation of a cue-based retrieval model for such production data. In addition, implementing a single model in isolation has limited explanatory value. A good fit could reflect the flexibility of the model rather than the validity of its underlying mechanism. We therefore also implement other accounts proposed in Avetisyan et al. (2020), allowing us to evaluate which theoretical account best captures the observed pattern of agreement attraction in Armenian sentence production.

The rest of the paper is organised as follows. In the next section, we describe the different models under consideration. This is followed by a comparison between these models. We conclude with general discussion and limitations section.

2 Models of Subject Verb Agreement Production

The models we compare here are computational-level implementation of theoretical proposals about how the production system computes agreement; specifically, each model maps the experimental conditions in Avetisyan et al. (2020) onto a probability of correct agreement and a predicted reaction time. The free parameters of each model are estimated jointly from the accuracy and reaction time data using Hamiltonian Monte Carlo as implemented in Stan (Stan Development Team, 2026). Model comparison is done via leave-one-out cross-validation using the Pareto smoothed importance sampling estimator (Vehtari et al., 2024, 2017) in the `loo` package in R. This framework allows us to evaluate not just whether a given model can fit the data, but how well each model predicts held-out observations relative to the others.

All model priors were chosen to have similar predictions across models before looking at the data. Avetisyan et al. (2020)'s production experiment had 2×2 design with factors: case and number. Case factor had two levels, the first level had a case marked distractor while the second level had no case marking on the distractor. For the number factor, the distractor was either singular, matching with the subject's singular number, or plural, mismatching with the subject's singular number.

In a typical agreement attraction experiment, accuracy decreases when the subject and the distractor mismatches in number (e.g., Bock and Miller, 1991; Eberhard et al., 2005; Staub, 2009). When case is present on the nouns, a matching case (here, +nominative) decreases errors more than a mismatching case (e.g., Nicol and Anton-Mendez, 2009; Hartsuiker et al., 2001, 2003; Badecker and Kuminiak, 2007; Gillespie, 2011; Slioussar, 2018; Gillespie and Pearlmutter, 2013). In Avetisyan et al. (2020), there were four conditions: (a) case marked distractor with a matching number, (b) case marked distractor with a mismatching number, (c) a non case marked distractor with a matching number, and (d) a non case marked distractor with a mismatching number.

Previous results showed a decrease in accuracy in condition (d) vs conditions (a)-(c). In all the models implemented in this work, this decrease in accuracy was assumed before choosing a prior. (a)-(c) conditions were assumed to have similar accuracy.

Its important to emphasise that the priors were chosen such that the predictions of the model resembled each other. This criteria was adapted from previous model comparisons in sentence processing (e.g., Yadav et al., 2022; Husain et al., 2025). We assumed a roughly 10% decrease (this decrease reflects the case effect) in accuracy in condition (d) in comparison to conditions (a)-(c) across models while having enough room to infer from the data. Although note that this number is not chosen through a meta analysis but considering accuracies previously reported in agreement attraction studies (See Table 7 in Kandel et al., 2022).

2.1 Similarity-based interference during Retrieval: Cue-based Retrieval theory (Lewis and Vasishth, 2005)

The first account we consider is the cue-based retrieval account (Lewis and Vasishth, 2005) implemented under the ACT-R architecture (Anderson, 1996; Anderson et al., 2004). As discussed in the previous section, a cue-based retrieval account can explain the effect of case on agreement attraction. For our modeling we implementation the Lewis and Vasishth (2005) and adapted it around the two conditions shown in Figure 2a&b. Specifically, condition (a) and (b) in Table 2 resembled Figure 2a, condition (d) in Table 2 resembled Figure 2b. Condition (c) from Table 2 was taken to have similar probabilities of accuracy as condition (a) and (b) in Table 2. This is because the distractor had the same features as the the subject so the accu-

Model	Locus of interference	Type of interference
Cue-based retrieval	Retrieval	Similarity-based interference
Feature migration theory	Encoding	Similarity-based interference
Case as markers for prediction	Encoding	Prediction distortion

Table 1: Models of subject-verb agreement errors and their locus of interference.

Table 2: Experimental conditions in Avetisyan et al. (2020)

	Case marked	Non case marked
Matching number	(a)	(c)
Mismatching number	(b)	(d)

racy reflected by the experiment can be assumed to be the addition of the probability of correct agreement from the subject and the probability of any incorrect agreement from the distractor.

As discussed in the previous section, under a cue-based retrieval theory, chunks receive an activation boost from retrieval cues. The most activated chunk is then retrieved for agreement computation.

The activation boost S_i for chunk i is given by Eq. 1, where W_j is the weight of cue j. We assumed two cues, each with a weight of 0.5. S_{ji} denotes the strength of association between cue j and chunk i, given by Eq. 2, where MAS is the maximum associative strength set to 1 and fan_j is the number of chunks that cue j matches with.

$$S_i = \sum_j W_j S_{ji} \tag{1}$$

$$S_{ji} = MAS - \ln(fan_j) \tag{2}$$

fan_j is where the difference between the +nominative (Figure 2b) and -nominative (Figure 2a) conditions is implemented. In Figure 2a, where the distractor is -nominative, the fan_j for the +nominative cue is 1, since only the subject matches this cue. In Figure 2b, where the distractor is +nominative, the fan_j is 2, since both the subject and the distractor match this cue. This brings down the activation of the subject chunk in the +nominative distractor condition relative to the -nominative condition, increasing the likelihood of misretrieval.

Finally, the probability of retrieving the subject chunk i is given by Eq. 3, where S_i[1] is the activation of the target chunk and s is the noise parameter. The threshold parameter τ had a prior distribution of $\mathcal{N}(0.214, 0.1)$. Recall that this prior is chosen to have condition (d) show $\sim$ 10% decrease from condition (a)-(c). The predictions of the cue-based model with the given prior is shown in Figure 3.

$$P_{\text{Accuracy}} = \frac{1}{1 + e^{-\frac{(S_i - \tau)}{s}}} \tag{3}$$

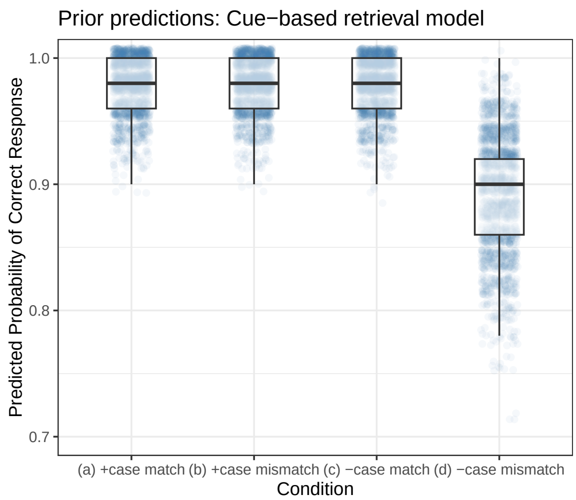

Figure 3: Possible accuracies for condition (a)-(d) in Table 2 under a cue-based retrieval theory.

The retrieval time RT_i for chunk i is given by Eq. 4. Here, F is a scaling parameter with a prior $\mathcal{N}_{lb=0.005}(0.15, 0.05)$ similar to previous implementations of cue-based retrieval model during sentence processing (e.g., Engelmann et al., 2019; Yadav et al., 2022).

$$RT_i = Fe^{-S_i} \tag{4}$$

Assuming a base reaction time β_0 needed for other processes such as integration and motor movement, the overall reaction time is given by Eq. 5.

$$RT_{total} = \beta_0 + RT_i \tag{5}$$

[1]Generally the activation of chunk is present at this part of the equation. As we set other contributing factors such as base level activation to zero, we just use S_i here.

2.2 Similarity-based interference before Retrieval: Feature migration theory (Oberauer and Lang, 2008)

Another model proposed by Avetisyan et al. (2020) to explain the case effects is encoding interference (Nairne, 1990, 2006; Oberauer and Kliegl, 2006). Such models have previously been shown to account for effects seen during sentence processing (e.g., Husain et al., 2025). Similar to the cue-based retrieval model, this model assumes similarity-based interference although at the encoding stage. In terms of case markings, when the subject and the distractor has similar case, the interference increases, thereby decreasing accuracy. In contrast to the cue-based retrieval model, this interference is active during the encoding phase. We diverge from (Avetisyan et al., 2020) by assuming a feature migration model (Oberauer and Lange, 2008) rather than a feature overwriting model. The feature migration theory gives us the possibility that accuracy can decrease in condition (d) relative to other condition. On the other hand, the feature overwriting theory posits similar accuracies (See appendix B).

We use Husain et al. (2025)'s implementation of the feature overwriting theory (Oberauer and Kliegl, 2006) as both feature migration as well as feature overwriting can be defined under a similar mathematical implementation (Oberauer and Lange, 2008). Under this implementation, it is assumed that items are stored in memory and are a source of interference to each other. This interference is modulated by the number of similar features between the items.

In the feature migration theory, interference in and of itself is defined as similarity-based feature migration, wherein, when items in memory are similar, a dissimilar feature can migrate from one item to the other. Each item is assigned an amount of Activation which goes down with increasing number of similar features with other nouns. The activation of the subject noun is given by Eq. 6. Here A_i is the activation of the subject i, n is the number of features of the subject i which is assumed[2] to be 2 in our implementation. r is the rate of feature migration due to similarity-based interference with a prior set to $\mathcal{N}_{lb=1}(1.8, 0.3)$.

[2]We take this assumption largely for simplicity as well as for consistency with the cue-based retrieval model which also assumes 2 features participating in the cue-based retrieval process mentioned in §2.1.

$$A_i = \left(1 - r\frac{C}{2n}\right)^{n-1} \quad (6)$$

The probability of a veridical agreement is a logistic function of the activation of the subject noun A_i defined as Eq. 7. Here, a_0 is the base level activation of the subject set to 1 and T is a noise parameter with a prior set to $\mathcal{N}_{lb=0.01}(0.25, 0.03)$. An increase in noise parameter T shifts the probability of accuracy to near chance level in this model.

$$P_{\text{Accuracy}} = \frac{1}{1 + e^{-\left(\frac{A_i - a_0}{T}\right)}} \quad (7)$$

Recall that the priors were chosen so that the model predicts a similar pattern of probabilities as the cue-based retrieval model across the 4 conditions (See Figure 4).

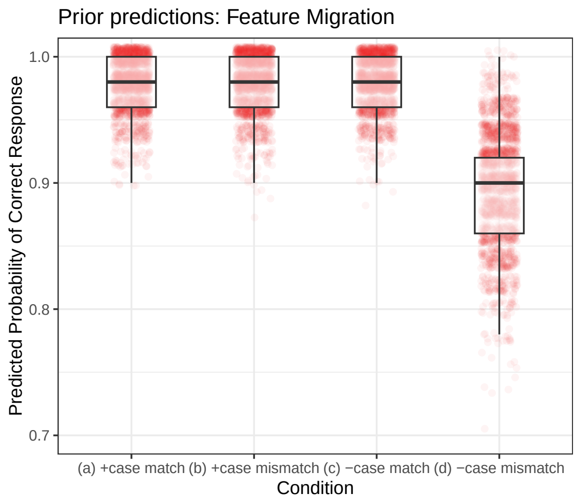

Figure 4: Possible accuracies for condition (a)-(d) in Table 2 under a Feature migration theory.

2.3 Case as Markers for Agreement Prediction

Language comprehension can potentially involve use of case markers to predict syntactic frames (e.g., Husain et al., 2014; Henry et al., 2017; Chow et al., 2018; Zafar, 2025; Yamashita, 1997). Avetisyan et al. (2020) propose that in Armenian, prediction of agreement features is contingent on case markers. The forced choice task employed in Avetisyan et al. (2020) has both a comprehension and a production element. We specifically assume that, when the subject is encountered, the system predicts the agreement features of the matrix verb based on the subject's features. Whenever an appropriate agreement site appears, these predicted features are utilised for agreement computation.

This process can clearly be present in comprehension. During production, we assume that the system can predict subject-verb agreement as soon as the subject is produced. Similarly, the system can also predict distractor-verb agreement (whenever possible) as soon as the distractor is produced. Further, we also assume that case markers denote whether such a prediction was made.

Overall, these predictions are maintained in memory and are therefore susceptible to interference from other predictions held in memory (cf. Yadav et al., 2020; Husain et al., 2025). Specifically, when a distractor's predicted agreement (when the distractor also has a nominative case) carries different number feature from the subject's, the subject's predicted agreement features can be distorted. When the distractor shares the same features as the subject, no such distortion occurs.

We implement this as follows. The degree of interference S_i from a distractor's prediction is given by Eq. 8, where p_i represents whether the distractor generates a prediction at its clause (1 if yes, 0 if no) and r is the rate of distortion, which is non-zero only when the distractor carries a different agreement feature than the subject. We set r to lie within $\mathcal{N}_{lb=0}(1.73, 0.3)$.

$$S_i = p_i \times r \tag{8}$$

The probability of retaining the veridical representation of the subject's agreement features is given by Eq. 9, where u determines the accuracy of maintaining subject prediction without any interference from distractor prediction.

$$P_{\text{Accuracy}} = \frac{1}{1 + e^{u - S_i}} \tag{9}$$

With our priors, the model roughly predicts similar probabilities as the cue-based retrieval model and the feature migration theory (See Figure 5).

2.4 A theory of agreement for non cue-based retrieval models

Recall that in a cue-based retrieval model the mechanism of retrieval is central to the theory of how agreement operates, i.e., agreement process involves retrieval of subject features. This process of retrieval then manifests as reaction time. On the other hand, there is no process posited at the agreement site for non cue-based retrieval models

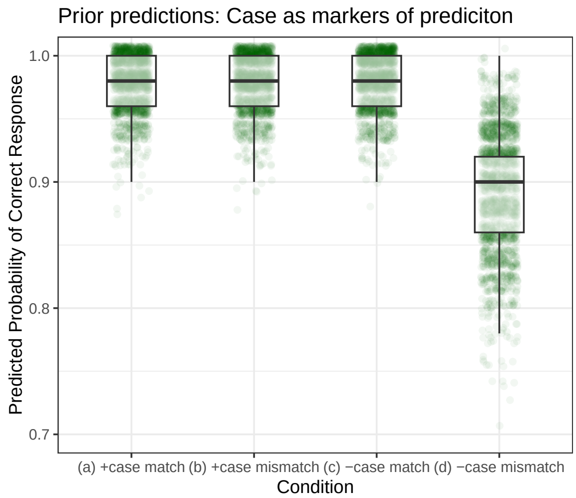

Figure 5: Possible accuracies for condition (a)-(d) in Table 2 under a case as markers for agreement prediction model.

(feature migration model in 2.2 and case as markers for agreement prediction model in 2.3). This is because these models have not been previously associated with such theories in production.

We assume that items in memory get distorted in both non-cue based retrieval models discussed in section 2.2 and 2.3. This naturalistically demands the processor to infer the correct agreement at the agreement site given the representation of either the item as mentioned in section 2.2 or prediction mentioned in 2.3 is not intact. Such an assumption is consistent with the noisy channel inference model implemented for sentence production (Ryskin et al., 2021).

Furthermore, we assume that probabilities posited in Eq. 7 or Eq. 9 determine the effort required to infer the veridical agreement from a potentially distorted item/prediction, i.e., when an item/prediction is less probable, the effort required is higher, leading to higher reaction times. Such relations between probability and effort has previously been posited during sentence processing (e.g., Levy, 2008, 2013). We implement an information theoretic idea of inference effort I_i in bits; see Eq. 10. Here P_{Accuracy} is probability that the correct agreement prediction has been maintained in memory.

$$I_i = -\frac{\log(P_{\text{Accuracy}})}{\log(2.0)} \tag{10}$$

Finally, the reaction time RT_i to produce agreement from prediction i is given by Eq. 11, where

β_0 represents the base reaction time for processes such as integration and motor execution, and β_1 is a scaling parameter that maps recovery effort in bits to reaction time. The β_1 was set to fall within $\mathcal{N}_{lb=0}(0, 1)$.

$$RT_i = \beta_0 + (\beta_1 \times I_i) \tag{11}$$

Along with the priors for the probabilities, it would also be ideal to have similar reaction time predictions between all three models. We, unfortunately, couldn't achieve that. Hence, our priors were mostly designed around the probabilities. A sensitivity analysis for the cue-based retrieval model with β_0 taken as 0.15, 0.2 and 0.3 found no meaningful difference in their ΔELPD values. Further more, keeping the β_0 value as a free parameter gave us unreasonable estimates (more than 0.35 or 350 ms), often absorbing the effect of other parameters. Hence, we fixed β_0 to 0.2 (0.2 seconds or 200 milliseconds) across models for model comparison.

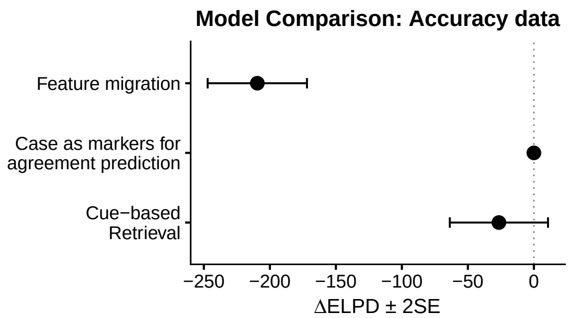

Figure 6: Expected Log Pointwise Predictive Density (ΔELPD) differences across models for the accuracy data.

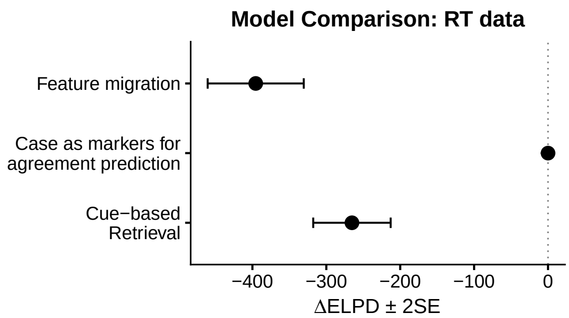

Figure 7: Expected Log Pointwise Predictive Density (ΔELPD) differences across models for the reaction time data.

3 Model Comparison

All models were implemented in Stan (Stan Development Team, 2026) and fitted using Bayesian inference via the `rstan` package in R[3]. The data comprised two dependent measures: accuracy (whether the verb was produced with veridical agreement) and reaction time. Each observation was coded for two experimental factors, case (match vs. mismatch) and number of the attractor (singular vs. plural), yielding four conditions.

Each of the three models were fitted to both accuracy and reaction time data simultaneously. All models were fitted using four parallel chains.

To compare models, we used leave-one-out cross-validation method as implemented in the `loo` package in R. LOO estimates the expected log pointwise predictive density (ELPD), which quantifies how well a model predicts held-out data. A higher ELPD indicates better predictive performance. Pareto k diagnostic values were inspected for each model to ensure the reliability of the LOO estimates (Nicenboim et al., 2025).

Model comparison was carried out separately for accuracy and reaction time. For each dependent measure, the three models were compared using the `loo_compare` function, which ranks models by their ELPD and provides the difference in ELPD along with its standard error. A model with more than 2 SE difference in ELPD is taken to have better out-of-sample predictive performance relative to the others.

For accuracy, the case as markers for agreement prediction model with inference performed best. The cue-based retrieval model was the next best with a ΔELPD of −26.5 (SE = 18.6). This difference is within two standard errors, suggesting that the two models are not reliably distinguishable in their predictive performance for accuracy. The feature migration model, however, performed substantially worse with ΔELPD values of −209.5 (SE = 18.8). This difference is well beyond two standard errors, indicating reliably poorer predictive performance.

For reaction time, the case as markers for agreement prediction model with inference was again the best model. Unlike accuracy, the difference between the case as markers for agreement prediction model and the cue-based retrieval model (ΔELPD = −265.3, SE = 26.2) was well beyond two standard errors, indicating that the case as markers

[3]Codes are available at this link.

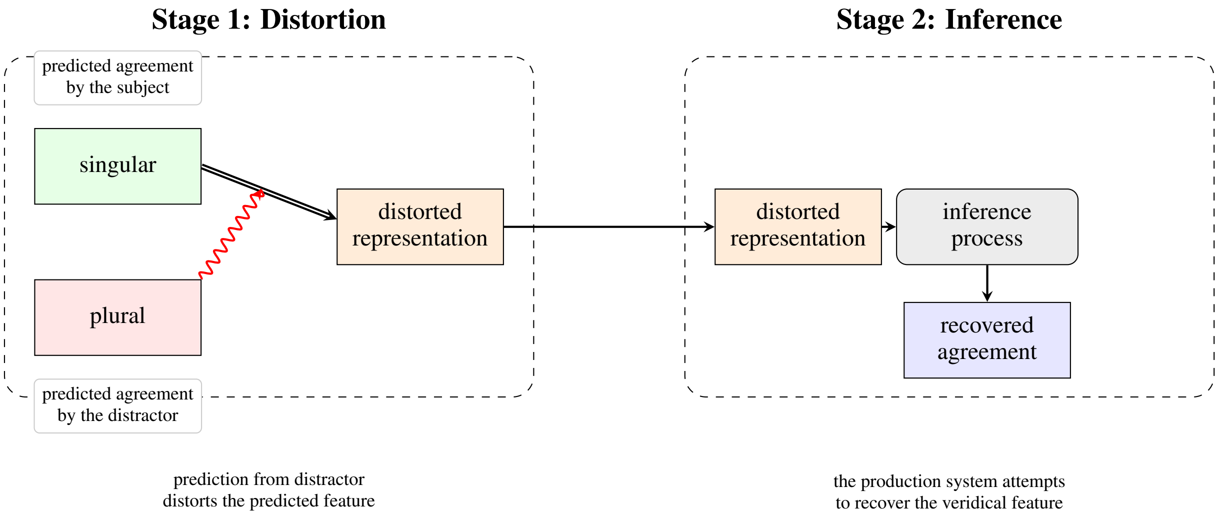

Figure 8: A schematic of the two-stage model. In Stage 1, the predicted agreement feature is susceptible to distortion from interference by the predicted agreement for the distractor clause. In Stage 2, the production system infers the veridical agreement feature from the potentially distorted representation. The effort required for this inference determines the reaction time

for agreement prediction model reliably outperforms the cue-based retrieval model for reaction time. The feature migration model with inference still performed the worst (ΔELPD = −395.5, SE = 32.5).

4 General Discussion

The current paper achieves two goals. First, it implements three distinct computational models that can explain the role of case during agreement production, focusing on the phenomenon of agreement attraction.

Second, it compared these different models and found that a 'case as markers for agreement prediction' model followed by an inference component explains the effect of case-marker during agreement in Armenian better than a cue-based retrieval model. Specifically, our model comparison results illustrate a language production system where agreement computation is largely predictive in nature. These predictions are in turn susceptible to distortion in memory; given such distortions, the processor infers the original agreement features (See Figure 8). Additionally, our results highlight that using both accuracy as well as reaction time data provides greater insights to understanding agreement processes (also see, Staub, 2010).

4.1 Prediction and incrementality

The scope of planning during sentence production is thought to be minimal in nature (Kempen and Hoenkamp, 1987; Levelt et al., 1999; Levelt, 1999). While some proposals argue that the scope of planning is radically incremental (e.g., Konopka, 2012; Paul, 1970; Bock and Irwin, 1980; Ferreira, 1994), others have argued that the scope of planning is larger (e.g., an entire clause, a verbal unit, etc.)(e.g., Momma et al., 2020; Momma and Yoshida, 2023; Bock and Levelt, 1994; Bock and Cutting, 1992). There is very little work that has explored agreement computation in the context of planning scope. The current work contributes towards this discussion as well.

Our results provide some evidence for a production system where the agreement representation is in fact planned (via prediction). Specifically, our results point to a scenario where the decision to predict agreement is contingent on the use of nominal case-marker. In other words, the presence of a nominative case marker on a noun triggers a prediction that agreement will need to be computed within that clause. When a distinct prediction is generated by the distractor, it interferes with the agreement prediction generated by the subject, giving rise to attraction errors. The absence of a nominative case marker, on the other hand, does not generate such a prediction, thereby making the inference process easier. As such, while the current work suggests that agreement representations are planned prior to the articulation of the verb, the precise nature of such representations needs to be investigated further.

4.2 Cross-linguistic validity

An important point needs to be made regarding the cross-linguistic validity of our case as markers for agreement prediction model. The case as markers for agreement prediction model entails that whenever there is a nominative case, a prediction for agreement is made. This holds for a language like Armenian but need not be true for other languages where a nominative noun can appear without any agreement within its clause. For example, in the Slovak (shown in example 1), having a nominative case would not automatically mean that there will be agreement within its clause. Hence, the case as markers for agreement prediction model only stays relevant for languages where nominative case predicts agreement. In other words, how language properties of a language dictate the prediction mechanism need to be investigated in future studies.

Regardless, the case as markers for agreement prediction model can also apply to other features which predict agreement. For example, in Hindi, a distractor-induced agreement attraction has been attributed to a 'feature' of the distractor when it has agreement within its clause (See controller coding hypothesis in Bhatia and Dillon, 2022). Interestingly this presence of a feature that cause interference and potential for agreement prediction is confounded in their explanation. In other words, a predictive system could also explain this data. Within a prediction based account, whenever a prediction for agreement is made at the distractor, inference of the subject agreement becomes difficult. This can effectively replace the 'feature' which was previously attributed in Bhatia and Dillon (2022) to cause interference. Overall, these cross-linguistic considerations suggest that while the specific cue that drives prediction may vary across languages, the underlying mechanism of prediction followed by inference may be a more general property of the agreement production system in languages with similar case and agreement properties.

4.3 Agreement attraction during production

The agreement attraction phenomenon during production has generally been evaluated either under an encoding based interference account (e.g., Bock and Miller, 1991; Bock et al., 2001; Eberhard et al., 2005; Franck et al., 2002) or a retrieval based account (e.g., Badecker and Kuminiak, 2007; Bhatia and Dillon, 2022; Kandel et al., 2022). Our results suggests that apart from a simple interference/encoding-based constraints, the role of prediction and inference for such computations are equally important (cf. Ryskin et al., 2021), if not more. On the other hand, there is a strong evidence for the interference/encoding-based constraints on agreement computation during sentence comprehension (for an overview see Yadav et al., 2022). This suggests that agreement computation processes during comprehension might be distinct from how such computations happen during sentence production.

5 Conclusion

The current paper investigated the role of case during agreement attraction in sentence production. Through computational modeling, we compared different accounts that could explain the effect of case on agreement errors. Our results show that a case as markers for agreement prediction followed by an inference component provides a better explanation than a cue-based retrieval model, which was previously argued to account for such effects.

Overall, this work puts forward a two-stage account of agreement production, comprising prediction and inference, that advances our understanding of how case information modulates agreement errors. Future work comparing the current models as well as additional models across languages with different case-agreement configurations will be critical in evaluating the generality of this account.

6 Limitations

6.1 Case vs. Subjecthood

One important caveat regarding Avetisyan et al. (2020)'s production data is that every time the distractor case was nominative, it was also the subject within its own clause. On the other hand, when it was not nominative, it was an object within its own clause. Critically, objects never show agreement in Armenian. Such a confound is important to address because subjecthood can impact agreement attraction in some languages, e.g., Romanian, (Bleotu and Dillon, 2024), but not in others, e.g., Hindi, (Bhatia and Dillon, 2022). Regardless of whether this effect was driven by case or subjecthood, the case as markers for agreement prediction model can be adapted for both scenarios. Subjecthood might mark a prediction of agreement in place of case in Armenian. As mentioned in §4.3, the current prediction based theory can also extend to

Hindi (Bhatia and Dillon, 2022). In Hindi, instead of case or subjecthood, whether the distractor had agreement within its own clause predicted agreement attraction. It would be insightful to compare these models on Hindi as well as Romanian data in future work.

6.2 Role of Paradigm

A forced choice paradigm as implemented in Avetisyan et al. (2020) has both a comprehension as well as a production component. Specifically, the rapid serial visualization of words constitutes the comprehension aspect of the task while the choice for agreement constitutes the production aspect. An important test for these results would be to replicate these results through other paradigms (e.g., preamble repetition paradigm, scene description paradigm).

6.3 A note on condition (c)

While agreement errors during comprehension have previously been implemented computationally, models of agreement errors during production have not. The present work addresses this gap and, as such, stands as a starting point for future computational modeling of agreement errors in production. This framing has necessarily entailed certain simplifications, most notably in the implementation of condition (c). As discussed in §2, we assumed the same accuracy for condition (c) as for conditions (a) and (b), and further assumed that the underlying process was similar across the three conditions. This choice was motivated by the desire to maintain a consistent set of assumptions applicable across all models, particularly in the absence of clear empirical or theoretical guidance during implementation. In future work, we plan to develop models that more faithfully/fruitfully capture the processes specific to condition (c).

References

John R. Anderson. 1996. ACT: A simple theory of complex cognition. *American Psychologist*, 51(4):355–365. Place: US.

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. 2004. An Integrated Theory of the Mind. *Psychological Review*, 111(4):1036–1060. Place: US.

Serine Avetisyan, Sol Lago, and Shravan Vasishth. 2020. Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, 112:104087.

William Badecker and Frantisek Kuminiak. 2007. Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of Memory and Language*, 56:65–85.

Sakshi Bhatia and Brian Dillon. 2022. Processing agreement in Hindi: When agreement feeds attraction. *Journal of Memory and Language*, 125:104322.

Adina Camelia Bleotu and Brian Dillon. 2024. Romanian (subject-like) DPs attract more than bare nouns: Evidence from speeded continuations. *Journal of Memory and Language*, 134:104445.

J. Kathryn Bock and David E. Irwin. 1980. Linearization in sentence production: Evidence from comprehension. *Journal of Verbal Learning and Verbal Behavior*, 19(4):467–479.

Kathryn Bock and J. Cooper Cutting. 1992. Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, 31(1):99–127. Place: Netherlands.

Kathryn Bock and Kathleen M. Eberhard. 1993. Meaning, sound and syntax in english number agreement. *Language and Cognitive Processes*, 8(1):57–99.

Kathryn Bock, Kathleen M. Eberhard, J. Cooper Cutting, Antje S. Meyer, and Herbert Schriefers. 2001. Some Attractions of Verb Agreement. *Cognitive Psychology*, 43(2):83–128.

Kathryn Bock and Willem JM Levelt. 1994. Language production: Grammatical encoding. In Morton Ann Gernsbacher, editor, *Handbook of psycholinguistics*, pages 945–984. Academic Press.

Kathryn Bock and Carol A Miller. 1991. Broken agreement. *Cognitive Psychology*, 23(1):45–93.

Wing-Yee Chow, Andrew Nevins, and Manuel Carreiras. 2018. Effects of subject-case marking on agreement processing: ERP evidence from Basque. *Cortex*, 99:319–329.

Kathleen Eberhard, John Cutting, and Kathryn Bock. 2005. Making Syntax of Sense: Number Agreement in Sentence Production. *Psychological review*, 112:531–59.

Kathleen M. Eberhard. 1997. The Marked Effect of Number on SubjectVerb Agreement. *Journal of Memory and Language*, 36(2):147–164.

Felix Engelmann, Lena A. Jger, and Shravan Vasishth. 2019. The Effect of Prominence and Cue Association on Retrieval Processes: A Computational Account. *Cognitive Science*, 43(12):e12800.

Victor S. Ferreira. 1994. The choice of structural transitions in oral delivery of pennsylvania dutch. *Journal of Memory and Language*, 33(6):715–736.

Julie Franck, Gabriella Vigliocco, and Janet Nicol. 2002. Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, 17(4):371–404. _eprint: https://doi.org/10.1080/01690960143000254.

Maureen Gillespie. 2011. Effects of Semantic Integration and Advance Planning on Grammatical Encoding in Sentence Production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Maureen Gillespie and Neal J. Pearlmutter. 2013. Against structural constraints in subjectverb agreement production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2):515–528.

Robert J. Hartsuiker, Inés Antón-Méndez, and Marije Van Zee. 2001. Object Attraction in Subject-Verb Agreement Construction. *Journal of Memory and Language*, 45(4):546–572.

Robert J. Hartsuiker, Herbert J. Schriefers, Kathryn Bock, and Gerdien M. Kikstra. 2003. Morphophonological influences on the construction of subject-verb agreement. *Memory & Cognition*, 31(8):1316–1326.

Nick Henry, Holger Hopp, and Carrie N. Jackson. 2017. Cue additivity and adaptivity in predictive processing. *Language, Cognition and Neuroscience*, 32(10):1229–1249.

Samar Husain, Apurva, Ishita Arun, and Himanshu yadav. 2025. The effect of similarity-based interference on bottom-up and top-down processing in verb-final languages: Evidence from Hindi. *Journal of Memory and Language*, 143:104627.

Samar Husain, Shravan Vasishth, and Narayanan Srinivasan. 2014. Strong Expectations Cancel Locality Effects: Evidence from Hindi. *PLOS ONE*, 9(7):e100986.

Margaret Kandel, Cassidy Rae Wyatt, and Colin Phillips. 2022. Agreement attraction error and timing profiles in continuous speech. *Glossa Psycholinguistics*, 1(1).

Gerard Kempen and Edward Hoenkamp. 1987. An incremental procedural grammar for sentence formulation. *Cognitive science*, 11(2):201–258.

Agnieszka E Konopka. 2012. Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. *Journal of Memory and Language*, 66(1):143–162.

W. J. Levelt, A. Roelofs, and A. S. Meyer. 1999. A theory of lexical access in speech production. *The Behavioral and Brain Sciences*, 22(1):1–38; discussion 38–75.

Willem JM Levelt. 1999. Producing spoken language: A blueprint of the speaker. *The neurocognition of language*, 83:122.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In *Sentence processing*, pages 78–114. Psychology Press.

Richard L. Lewis and Shravan Vasishth. 2005. An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval. *Cognitive Science*, 29(3):375–419.

Michael McCloskey. 1991. Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, 2(6):387–395.

Masaaki Momma, L Robert Slevc, and Colin Phillips. 2020. The timing of verb selection in english active and passive sentences. *Journal of Memory and Language*, 111:104066.

Masaaki Momma and Masaya Yoshida. 2023. The architecture of sentence production. *Annual Review of Linguistics*, 9:219–238.

James S. Nairne. 1990. A feature model of immediate memory. *Memory & Cognition*, 18(3):251–269.

James S. Nairne. 2006. Modeling distinctiveness: Implications for general memory theory. In *Distinctiveness and memory*, pages 27–46. Oxford University Press, New York, NY, US.

Bruno Nicenboim, Daniel J Schad, and Shravan Vasishth. 2025. *Introduction to Bayesian data analysis for cognitive science*. CRC Press.

Janet L. Nicol and Ines Anton-Mendez. 2009. The effect of case marking on subject-verb agreement errors in English.

Klaus Oberauer and Reinhold Kliegl. 2006. A formal model of capacity limits in working memory. *Journal of Memory and Language*, 55(4):601–626.

Klaus Oberauer and Elke B. Lange. 2008. Interference in verbal working memory: Distinguishing similarity-based confusion, feature overwriting, and feature migration. *Journal of Memory and Language*, 58(3):730–745.

Hermann Paul. 1970. *Prinzipien der Sprachgeschichte (Principles of Language History)*. Max Niemeyer Verlag. Original work published 1880.

Rachel Ryskin, Leon Bergen, and Edward Gibson. 2021. Agreement errors are predicted by rational inference in sentence processing. Technical report, PsyArXiv. Type: article.

Natalia Slioussar. 2018. Forms and features: The role of syncretism in number agreement attraction. *Journal of Memory and Language*, 101:51–63.

Stan Development Team. 2026. Stan Users Guide.

Adrian Staub. 2009. On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language*, 60(2):308–327.

Adrian Staub. 2010. Response time distributional evidence for distinct varieties of number attraction. *Cognition*, 114(3):447–454.

Shravan Vasishth and Felix Engelmann. 2021. *Sentence Comprehension as a Cognitive Process: A Computational Approach*. Cambridge University Press. Google-Books-ID: 6KZKzgEACAAJ.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. 2024. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58.

Gabriella Vigliocco and Tiziana Zilli. 1999. Syntactic Accuracy in Sentence Production: The Case of Gender Disagreement in Italian Language-Impaired and Unimpaired Speakers. *Journal of Psycholinguistic Research*, 28(6):623–648.

Himanshu Yadav, Garrett Smith, Sebastian Reich, and Shravan Vasishth. 2022. Number feature distortion modulates cue-based retrieval in reading. Technical report, PsyArXiv. Type: article.

Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. 2020. Word order typology interacts with linguistic complexity: A cross-linguistic corpus study. *Cognitive science*, 44(4):e12822.

Hiroko Yamashita. 1997. The Effects of Word-Order and Case Marking Information on the Processing of Japanese. *Journal of Psycholinguistic Research*, 26(2):163–188.

Mudafia Zafar. 2025. *Planning scope in production how grammar shapes incrementality*. Ph.D. thesis, Indian Institute of Technology Delhi.

A Codes for the paper

Codes are available at this link.

B Feature migration over feature overwriting

A feature overwriting account assumes that features shared between items in memory get degraded. Such an assumption does not align well with the base assumptions for our conditions. Under a feature overwriting model, the feature being degraded would be case (the shared feature), and not number. This in turn entails no predicted difference in number agreement between condition (c) and condition (d).

A feature migration model, on the other hand, fits better with our assumptions for the conditions. Although both condition (c) and condition (d) involve similarity-based interference, only condition (d) has a dissimilar number feature on the distractor that can migrate to the subject noun. Hence, feature migration model can explain our data but not a feature overwriting model.

C Model flexibility and degrees of freedom

A model is not a theory (McCloskey, 1991). A good fit to the data can come from the flexibility a model's free parameters afford rather than the validity of the underlying mechanism it posits. The three models we compared slightly differ in the number of free parameters they contain. The cue-based retrieval model has two (τ, F), the feature migration model has three (r, T, β_1), and the case as markers for agreement prediction has three (r, u, β_1). Notably, the case as markers for agreement prediction model is not the most flexible of the three in terms of number of parameters. This makes it unlikely that its better performance is solely an artifact of having more degrees of freedom.