


Cognitive Science 45 (2021) e13019

© 2021 The Authors. *Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society (CSS).

ISSN: 1551-6709 online

DOI: 10.1111/cogs.13019

Modeling Misretrieval and Feature Substitution in Agreement Attraction: A Computational Evaluation

Dario Paape,^a  Serine Avetisyan,^a Sol Lago,^b Shravan Vasishth^a

^a*Department of Linguistics, University of Potsdam*

^b*Institute for Romance Languages and Literatures, Goethe University Frankfurt*

Received 1 February 2021; received in revised form 30 May 2021; accepted 25 June 2021

Abstract

We present computational modeling results based on a self-paced reading study investigating number attraction effects in Eastern Armenian. We implement three novel computational models of agreement attraction in a Bayesian framework and compare their predictive fit to the data using k-fold cross-validation. We find that our data are better accounted for by an encoding-based model of agreement attraction, compared to a retrieval-based model. A novel methodological contribution of our study is the use of comprehension questions with open-ended responses, so that both misinterpretation of the number feature of the subject phrase and misassignment of the thematic subject role of the verb can be investigated at the same time. We find evidence for both types of misinterpretation in our study, sometimes in the same trial. However, the specific error patterns in our data are not fully consistent with any previously proposed model.

Keywords: Agreement attraction; Eastern Armenian; Self-paced reading; Computational modeling

1. Introduction

In sentence production, agreement attraction in subject–verb dependencies refers to the observation that native speakers are more likely to produce an ungrammatical sentence like (1a) than they are to produce an equally ungrammatical sentence like (1b) (e.g., Bock & Miller, 1991; Franck, Vigliocco, & Nicol, 2002; Vigliocco, Butterworth, & Semenza, 1995).

Correspondence should be sent to Dario Paape, University of Potsdam, Department Linguistics, Haus 14, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany. Email: paape@uni-potsdam.de

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

(1)

- a. * The knife for the *cakes* are on the counter.
- b. * The knife for the *cake* are on the counter.

Likewise, in sentence comprehension, (1a) is more likely than (1b) to be accepted as grammatical (e.g., Clifton, Frazier, & Deevy, 1999; Kimball & Aissen, 1971; Wagers, Lau, & Phillips, 2009), and elicits less processing disruption during reading (e.g., Dillon, Mishler, Sloggett, & Phillips, 2013; Lago, Shalom, Sigman, Lau, & Phillips, 2015). The pattern is surprising under the assumption that adult native speakers use the rules of hierarchical syntax when computing linguistic dependencies (Everaert, Huybregts, Chomsky, Berwick, & Bolhuis, 2015), as the singular noun *knife* is the head of the subject noun phrase and should thus exclusively determine its number feature (López, 2001). Agreement attraction is also exhibited by deep learning models trained on fully grammatical input, but does not completely parallel human data (e.g., Arehalli & Linzen, 2020; Linzen & Leonard, 2018). In neural network models, agreement errors arise despite the presence of “syntax units” that encode subject–verb agreement (Lakretz et al., 2019), and have been claimed to be due to the models’ acquisition of shallow heuristics rather than syntactic rules (Bansal, Bhatt, & Agarwal, 2020), or possibly a mixture of both (see Linzen & Baroni, 2021 for a review).

How attraction errors arise in humans is also subject to debate. In order to limit the scope of the present investigation, we focus on two broad, influential classes of explanations that have been proposed in the psycholinguistics literature, and around which much of the theoretical discussion has concentrated. Additional approaches will be sketched in the general discussion. We will refer to the first class of accounts as *encoding-based*: These accounts assume that in (1a), the plural feature of the nonhead noun *cakes* is sometimes transferred to the entire subject noun phrase (NP). This feature transfer results in a representation of the stimulus sentence in which the subject of the clause is plural and thus able to license plural agreement on the verb (e.g., Pearlmutter, Garnsey, & Bock, 1999). The encoding-based view of agreement attraction was first proposed in the context of language production, which needs to posit a mechanism by which the grammatical subject can transmit its features to the to-be-produced verb in order to obtain correct agreement. The idea that features from an incorrect noun phrase can be transmitted along syntactic edges in production is supported by findings showing that the depth of syntactic embedding of the nonhead noun influences attraction rates (e.g., Franck et al., 2002). By contrast, *retrieval-based* accounts of agreement attraction assume that encountering a plural verb triggers a cue-based retrieval operation (e.g., Lewis and Vasishth, 2005; Lewis, Vasishth, & Van Dyke, 2006; McElree, 2000) that looks for a plural subject NP. Due to misretrieval of the nonsubject NP *the cakes*, the plural form is sometimes erroneously licensed in (1a) (e.g., Badecker & Kuminiak, 2007; Franck & Wagers, 2020; Tanner et al., 2017; Wagers et al., 2009).¹

One way to empirically distinguish encoding- and retrieval-based accounts is to probe readers’ interpretations of sentences like (1a). Under the encoding account, it can be assumed that the NP *the knife* is sometimes mentally represented as *the knives*, given that the entire subject phrase is taken to be plural, and that syntactic heads should have the same feature specifications as their projections. We call this pattern *number misinterpretation*. By contrast, under

the retrieval-based approach, the attractor NP *the cakes* should sometimes be misconstrued as the subject of the main verb, that is, the cakes rather than the knife should be taken to be on the counter in (1a). We call this pattern *thematic misinterpretation*. The final interpretation of the sentence could, in principle, be determined at any point in time after the critical elements—the noun phrases and the verb—have been read or heard, and could proceed either in an immediate or in a delayed fashion. Given that many psycholinguistic studies have shown interpretation to be immediate or even predictive (e.g., Altmann & Kamide, 1999; Traxler & Pickering, 1996), we will assume that interpretation occurs immediately at or shortly after the verb.

There is evidence for both thematic and number misinterpretations in the existing literature. Patson and Husband (2016) presented sentences like (1a,b) in self-paced reading and probed participants' number interpretation of the subject phrase on each trial (e.g., *Was there more than one knife?*). Results showed that plural interpretations of singular subjects were about 12% more frequent in the presence of a plural attractor than in the presence of a singular attractor, supporting the encoding account of agreement attraction. The finding was conceptually replicated by Brehm, Jackson, and Miller (2019) and Brehm, Jackson, and Miller (2021), who also found evidence of immediate misinterpretation at the critical verb. Schlueter, Parker, and Lau (2019) had subjects read sentence preambles such as *The boy by the trees are really very ...* and then select between two adjectives to finish the sentence. One adjective was semantically compatible only with the head noun of the subject NP (*boy ...chubby*) while the other was only compatible with the attractor noun (*trees ...green*). Schlueter et al. (2019) observed a small but significant increase in attractor-matching adjective choices (about 3%) in ungrammatical sentences when the attractor matched the verb in number. Assuming that selecting the attractor-matching adjective signals misinterpretation of the dependency, the finding is consistent with a retrieval-based account of agreement attraction.

Misinterpretations of agreement attraction sentences suggest that nonveridical memory traces of the sentence's semantic content are formed during processing. This fact renders language processing similar to other cognitive domains. Research on perception and memory suggests that both encoding errors and misretrievals are plausible candidate mechanisms for such errors. Replacement of the subject's number feature with that of the attractor constitutes a case of what Oberauer and Lange (2008) refer to as *feature migration* (see also Nicol, Forster, & Veres, 1997).² Such migrations are known to occur in visual processing, where they result in subjects experiencing illusory conjunctions of features (for instance, shapes and colors) that were not encountered in combination but occurred on different presentation items (e.g., Reinitz, Lammers et al., 1992; Treisman & Gelade, 1980). Illusory conjunctions occur more often within perceptual groups (Prinzmetal, 1981) and for stimuli of the same category (letters versus numbers; Esterman, Prinzmetal, & Robertson, 2004). In the realm of language processing, feature migration can be observed in production, for instance when phonemes or syllables erroneously migrate between words in an utterance, resulting in speech errors (*ad hoc* → *odd hack*; Fromkin, 1971). This phenomenon may extend to comprehension, so that the number feature of one word could be transferred to a different word. In support of this notion, Brehm and Goldrick (2017) recently provided evidence that illusory conjunctions

can be observed between sentences, arguing that they constitute a domain-general cognitive phenomenon. Furthermore, given that transmission or sharing of features between syntactic units is assumed by most grammatical formalisms anyway (e.g., Adger, 2003; Pollard & Sag, 2004), illusory conjunctions may result from an overapplication of such a mechanism (Vigliocco et al., 1995; Wagers et al., 2009).

By contrast, thematic misinterpretation due to misretrieval of the attractor NP instead of the subject NP represents a case of intrusion or *item confusion*, which is well known from serial recall tasks, with confusion of linearly adjacent items being particularly common (see Oberauer & Lange, 2008; Hurlstone, in press, for reviews). To the extent that sentence processing makes use of the same cognitive architecture and retrieval mechanisms as other tasks involving working memory (Lewis & Vasishth, 2005; Lewis et al., 2006), item confusion can be expected to occur in this domain as well (Engelmann, Jäger, & Vasishth, 2019; Mätzig, Vasishth, Engelmann, Caplan, & Burchert, 2018; Nicenboim, Vasishth, Engelmann, & Suckow, 2018; Smith, Franck, & Tabor, 2018).

One shortcoming of previous studies is that they only investigated one pattern of misinterpretation: Patson and Husband (2016) and Brehm et al. (2019) targeted only number misinterpretation while Schlueter et al. (2019) targeted only thematic misinterpretation. More recently, Brehm et al. (2021) conducted a picture-selection study in which participants could choose pictures representing number misinterpretations and thematic misinterpretations in addition to the correct interpretation. They observed number misinterpretations but almost no thematic misinterpretations. However, in Brehm et al.'s stimulus sentences the thematic interpretation was often disambiguated by plausibility (e.g., *The highway to the island(s) really was/were still under repair*). Furthermore, the presence of corresponding pictures in the display may have primed participants toward a particular type of misinterpretation. Given these caveats, additional data are needed to quantify the relative contribution of each mechanism of misinterpretation. The experiment we present next asked participants to read sentences at their own pace and answer questions about the identity and number of the subject after each trial. This allowed us to record their reading times and simultaneously test for the occurrence of both number misinterpretations and thematic misinterpretations. Anticipating our results, our data show that number misinterpretations and thematic misinterpretations can occur in nonnegligible proportions within the same data set.

In order to quantitatively compare the predictive performance of the encoding and retrieval-based models against our data, we implement the two competing accounts computationally. Computational modeling of cognitive processes is essential because it forces us to translate verbally expressed hypotheses into detailed implementations that are easier to understand, evaluate, and falsify (e.g., Farrell & Lewandowsky, 2010; Fum, Del Missier, & Stocco, 2007; Smaldino, 2017; Vasishth, Nicenboim, Engelmann, & Burchert, 2019). Computational models necessarily contain simplifications and are therefore unlikely to account for the entire range of available data (e.g., Box, 1979), but are nevertheless useful because they can aid theory development (Guest & Martin, 2021; Wimsatt, 1987). This is especially true when multiple models are compared to evaluate which of them best predicts the data. Here, we apply this approach to encoding- and retrieval-based models of agreement attraction.

2. Computational modeling

There are several existing implementations of attraction in subject–verb agreement: the marking and morphing model of agreement production (Eberhard, Cutting, & Bock, 2005), the retrieval-based production models of Konieczny, Schimke, and Hemforth (2004) and Badecker and Lewis (2007), the deep learning models reviewed by Linzen and Baroni (2021) and a more recent proposal by Ryu and Lewis (2021), and the diffusion model of binary grammaticality judgments proposed by Hammerly, Staub, and Dillon (2019). Additionally, the Lewis and Vasishth (2005) model has been fitted to reading time data (Dillon et al., 2013; Engelmann et al., 2019; Jäger, Mertzen, Van Dyke, & Vasishth, 2020; Nicenboim & Vasishth, 2018; Tucker, Idrissi, & Almeida, 2015), and to response accuracies (Laurinavichyute & von der Malsburg, 2020). Two major gaps exist in the previous literature on agreement attraction that are filled by the current work: First, reading time data and end-of-trial question responses from humans have, to our knowledge, never been analyzed in conjunction within the same generative model, with the exception of Nicenboim and Vasishth (2018). Second, the encoding- and retrieval-based accounts have never been quantitatively compared using human data. By addressing these gaps, we hope to be able to more closely link the online processing of agreement attraction sentences to their final interpretations, as well as investigate which of the two competing accounts best fits the data.

We focus our investigation on models of human cognition with relatively small numbers of parameters compared to current deep learning models. The deep learning approach traditionally focuses on practical utility rather than the modeling of human psychological processes (Lakretz et al., 2021). Furthermore, deep learning models are, to some extent, “black boxes,” in the sense that their behavior is not necessarily human-interpretable (Du, Liu, & Hu, 2019; Forbus, Liang, & Rabkina, 2017). However, recent studies in this domain have begun to link model behavior to plausible psychological mechanisms in humans and, possibly, human brain dynamics (Lakretz et al., 2019, 2021; Linzen & Baroni, 2021; Merks & Frank, 2021; Ryu & Lewis, 2021). Nevertheless, whereas models with parametric assumptions about response distributions can be compared relatively directly, deep learning models require additional linking assumptions, such as mapping model perplexity to human response times, which may sometimes fail (van Schijndel & Linzen, 2020). Furthermore, the performance of deep learning models depends crucially on the training set as well as on the training objective, which increases the degrees of freedom when comparing models. Finally, deep learning models require massive amounts of training data in order to derive the model predictions. Such data are usually only available for western European languages like English. Given these caveats, we leave comparisons with deep learning models to future work.

We implement three different cognitive models (two encoding models and one retrieval model) and systematically evaluate their fit to novel experimental data using two methods. First, we compare the models’ fit using k-fold cross-validation (Vehtari & Ojanen, 2012), a standard method in machine learning. This method quantitatively evaluates the posterior predictive performance of a model by repeatedly fitting subsets of the data and then checking the fit to the remaining data. Cross-validation is useful because fitting the models repeatedly to different subsets of the data guards against overfitting—that is, assuming more structure than

is necessary to accommodate a particular set of data (Hawkins, 2004)—and because each iteration can be viewed as a simulated replication (Koul, Becchio, & Cavallo, 2018). While cross-validation is not a substitute for direct replication and depends on adequate sample sizes to allow for meaningful conclusions, it is nevertheless a highly effective tool for comparing how well different models can be expected to generalize (Yarkoni & Westfall, 2017).

Second, we use visual posterior predictive checks (Gelman, Carlin, et al., 2014) to investigate the extent to which each model's predicted data match the observed data. Such checks are important because even when the estimated parameters of a model match the theoretical predictions, it is possible that new data generated from the model look different from the original data. Such a pattern indicates a lack of expected empirical coverage and failure to capture the true data-generating process (e.g., Nicenboim & Vasishth, 2018; Rubin, 1984). The posterior predictive distribution is obtained by generating a data set of the same size as the original data set for each sampled set of parameter values. The overall distribution of the generated data can then be plotted against the empirically observed data to check whether the two align; better alignment between the observed and predicted data implies better model performance.

The models were implemented only after we analyzed the reading time and response data from our experiment, finding that some patterns in the results are not compatible with existing models. The models are thus based on proposals in the literature, but we deviate from existing theories at several points. All models were implemented in Stan (Carpenter et al., 2017; Stan Development Team, 2019). An Appendix containing formal descriptions of the models, prior choices, and parameter estimation results, along with the modeling source code, is available at <https://osf.io/ykjg7>. The Appendix also contains formal descriptions and cross-validation results of two additional models, which are more complex variants of the encoding and retrieval models, respectively, but failed to outperform our best candidate models. All models reported are fully hierarchical, that is, all intercepts and slopes have both by-subject and by-item adjustments. We now describe each model, beginning with the base versions of the encoding and retrieval models, followed by an augmented encoding model, the sources of plurality (SOP) model.

2.1. *Encoding model*

Our implementation of the encoding account takes the form of a multinomial processing tree (MPT; Riefer & Batchelder, 1988). In cognitive psychology, MPTs have been used extensively to model multinomially distributed behavioral data (see Batchelder & Riefer, 1999; Erdfelder et al., 2009; Stahl, 2006 for reviews). MPTs are intended to model not the responses themselves—in our case, answers to comprehension questions about the critical dependency (*Who (singular or plural) VERBed?*)—but the latent processes involved in generating them (Ulrich, 2009). An MPT is a probability tree in which the response depends on the path taken on a given trial, with cognitive processes either occurring or not occurring at each junction, leading to different outcomes. The occurrence probability of each process can be estimated based on the observed behavioral data. Heterogeneity of parameters across individuals has been noted as an important issue in MPT-based modeling (Erdfelder et al., 2009; Matzke, Dolan, Batchelder, & Wagenmakers, 2015); our model takes parameter heterogeneity across

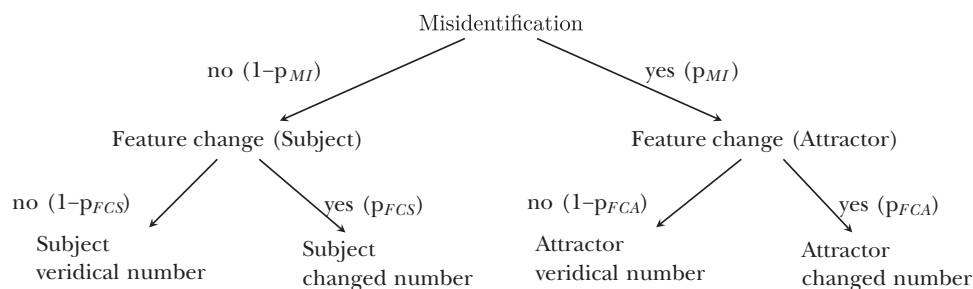


Fig. 1. The encoding-based model as a multinomial processing tree (MPT).

subjects as well as across sentences into account by assuming that the individual parameter values are sampled from multivariate normal distributions (details are given in the Appendix; see also Klauer, 2010).

Our encoding-based model of agreement attraction is shown in Fig. 1. Subjects' question responses (*Who (singular or plural) VERBed?*) are divided into five categories: subject/veridical number, subject/changed number, attractor/veridical number, attractor/changed number, and miscellaneous, which includes otherwise incorrect responses and nonresponses. Miscellaneous responses are not modeled.

The MPT consists of three distinct cognitive processes.³ The first process, which we call *Misidentification*, occurs with probability p_{MI} . If *Misidentification* occurs, the attractor is incorrectly encoded as the subject (*predication confusion*; Bock & Miller, 1991; Eberhard et al., 2005; Hupet, Fayol, & Schelstraete, 1998). In this model, *Misidentification* is assumed to be an unsystematic error and the same probability p_{MI} is thus assumed independent of the noun phrases' feature specifications (but see below for a different assumption).

If *Misidentification* occurs, so that the attractor is encoded as the subject, there is a probability p_{FCA} that its veridical number feature is changed to the opposite feature (singular to plural or plural to singular). This process is called *Feature change (Attractor)*. If *Misidentification* does not occur and the subject is encoded correctly, there is a probability p_{FCS} that its number feature is changed by the process *Feature change (Subject)*. Crucially, the probabilities p_{FCS} and p_{FCA} are assumed to be influenced by the feature match between target and attractor in the sentence: When the two noun phrases have mismatching feature specifications (one singular, one plural), the probability of feature change should increase as the feature of the other NP has some probability of migrating.

Besides the response given on a particular trial, we also model reading times in the region following the critical verb. We chose the postverbal region because this is where an attraction effect was observed in an earlier experiment (Avetisyan, Lago, & Vasishth, 2020), which had a similar setup to our current experiment. Assuming that reading times at this region index the time taken for the completion of the dependency, there should be a connection between reading times and responses to comprehension questions at the end of the trial (Nicenboim & Vasishth, 2018). The architecture of our MPT model implies that the reading times in a particular condition are a mixture of the distributions corresponding to the four different paths within the MPT (e.g., Heck & Erdfelder, 2016). In our model, we assume that the

four components (one for each response category) each have a lognormal distribution with a single shared standard deviation σ and different means μ_1, \dots, μ_4 . Each of the four means is a unique sum of costs incurred by the relevant latent processes, so that it is possible to infer the cost of each process. Our model can thus be seen as a simplified response-time extended multinomial processing trees (RT-MPT) model (Klauer & Kellen, 2018).

Model parameters are constrained by theory; for instance, feature change on both noun phrases should be less likely when their features match, so the relevant parameter is constrained to the interval $[-\infty, 0]$ in the model. Furthermore, reading times should be faster when the output of the MPT matches the feature specification of the verb. For instance, if the subject is correctly encoded but the sentence is ungrammatical, reading times should be faster if the subject's number feature has been changed to match the verb's, making the sentence appear grammatical.

2.2. Retrieval model

Following Nicenboim and Vasishth (2018), we implement a modified version of the cue-based retrieval model of Lewis and Vasishth (2005) as a lognormal race of evidence accumulators (Rouder, Province, Morey, Gomez, & Heathcote, 2015; see also Van Maanen & Van Rijn, 2007). The Lewis & Vasishth (2005) model is based on the adaptive control of thought-rational (ACT-R) cognitive architecture (Anderson, Bothell, Byrne, Douglass, & Lebiere, 2004; Anderson & Lebiere, 1998). ACT-R is a unified theory of cognition (Byrne, 2012; Newell, 1990), and has been used to model phenomena ranging from between-trial effects in a Stroop task (Juvina & Taatgen, 2009) to driving a car (Salvucci, 2006; see Ritter, Tehranchi, & Oury, 2019 for a comprehensive review).

ACT-R's declarative memory is made up of chunks with continuous, stochastically fluctuating activation values. A chunk's activation determines both its probability of retrieval and its retrieval latency, which is inversely proportional to the activation. When several chunks in memory are candidates for retrieval, the chunk with the highest activation/lowest latency is retrieved, implying a race process (Nicenboim & Vasishth, 2018). In our case, the reading time for the postverbal region of the sentence is assumed to correspond to the retrieval time of the chunk with the highest activation. Furthermore, as the verbal dependency is completed at this point, the NP corresponding to the retrieved chunk should be chosen as the answer to the end-of-trial comprehension question (*Who (singular or plural) VERBed?*).

The activation value, and thus retrieval latency, of each chunk in memory is influenced by the match with the retrieval cues set by the verb. A noun phrase that matches the verb in number and/or that occupies the structural subject position should be more likely to be retrieved to complete the dependency. As activation values are subject to stochastic noise, misretrievals are expected to occur on some proportion of trials. Retrieval of the attractor and subsequent thematic misinterpretation should be more likely when the attractor matches the verb's number feature, and especially when the grammatical subject does not match the verb's number feature (that is, in ungrammatical sentences). Furthermore, reading times should be faster in ungrammatical sentences with verb-matching distractors, due to statistical facilitation from the race process (Raab, 1962).

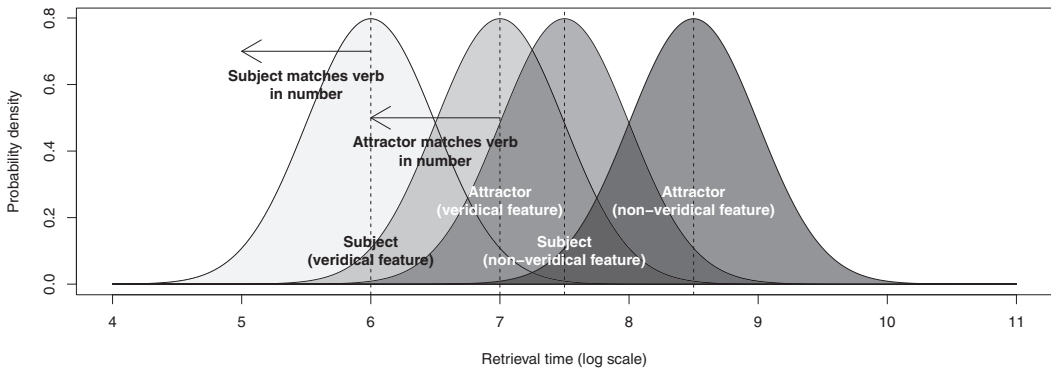


Fig. 2. Hypothetical distributions of log-scaled retrieval times for each of the candidate memory chunks. At retrieval, one value is sampled from each distribution and the candidate with the fastest retrieval time (that is, the highest activation) is accessed. For the veridical subject and attractor chunks, it is assumed that the mean latency shifts to the left when the number feature matches that of the verb, thus making retrieval more likely.

The Lewis and Vasishth (2005) model has no mechanism for encoding inference. Changes of the number feature on the retrieved NP are thus assumed to be unsystematic in the model, that is, unaffected by feature match or mismatch between the NPs in the sentence. In our modified version of the model, we assume that besides the veridical versions of the two NPs in the sentence, the corresponding nonveridical versions also compete for retrieval (see Brehm et al., 2021 for a similar proposal). The Lewis and Vasishth (2005) model does not prohibit the sentence processor from retrieving chunks from memory that did not occur in the current sentence, but their activation is assumed to be much lower. However, due to spreading activation in the mental lexicon (Van Maanen & Van Rijn, 2007), encountering a singular noun such as *knife* in a sentence could also activate the plural form *knives*, such that a corresponding chunk becomes available for retrieval.⁴

Retrieval times for the four possible chunks are assumed to be lognormally distributed with mean retrieval times μ_1, \dots, μ_4 and a shared standard deviation σ . A schematic illustration of the race process is given in Fig. 2.

2.3. Sources of plurality model

The SOP is an augmented variant of our encoding MPT model. It is based on the observation that number misinterpretations are more likely in our data when the verb's number feature mismatches the veridical number feature of the response NP. Because the encoding account posits that features can travel between elements in a sentence, it can easily allow features from verbs to migrate onto noun phrases. Recall that the encoding-based account was originally proposed in the context of sentence production, where the features of the to-be-produced verb need to be determined via the features of its subject. In comprehension, the verb is given and presumably encoded as a bundle of features in a similar way to the noun phrases in the sentence. Under the general assumption that features can migrate during encoding, this should not only be possible between the subject NP and the attractor NP, but also between the verb and any of the NPs.

Other things being equal, sentences with a singular subject, a plural attractor and a plural verb should show the highest incidence of number misinterpretations under the SOP model, as there are two sources of plural features. Sentences with one SOP (verb or attractor) should show an intermediate proportion of misinterpretations while sentences with no SOP should show few or no number misinterpretations. This is precisely the pattern in our results, which was also observed in experiments on English (Brehm et al., 2019, 2021; Patson & Husband, 2016). There is also some indication of the attractor’s number feature being more likely to be changed when it mismatches the verb’s, which further supports the notion of the verb being a source of nonveridical features.

Another change in the SOP model relates to the Misidentification stage. Eberhard et al. (2005) argue that a mechanism of predication confusion “can mimic attraction” due to the attractor occasionally being encoded as the subject (p. 555; see also Staub, 2009, 2010). The likelihood of this happening increases when the semantic match between attractor and verb is high (Pittman & Smyth, 2005; Thornton & MacDonald, 2003). Our implementation assumes that predication confusion maps onto the Misidentification stage of the encoding model and is affected by morphosyntactic match between the attractor and the verb. We hypothesize that predication confusion should occur less often in grammatical sentences and more often when the attractor matches the verb in number. In terms of thematic misinterpretation, the predictions of the augmented model are thus similar to the predictions of the retrieval account.

3. Self-paced reading study

Our experiment followed up on the studies of Avetisyan et al. (2020), who investigated agreement attraction in Eastern Armenian, a comparatively understudied language. Eastern Armenian is an Indo-European language that is written using the Armenian alphabet. Word order in Eastern Armenian finite clauses varies between SOV and SVO (Comrie, 1981; Dum-Tragut, 2009; Hodgson, 2018), and the grammatical subject controls agreement marking on the verb, which agrees with it in person and number, as in (2).

(2) Տղաները թիթեռնիկին բռնեցին:

Tġa-ner-ë t’it’erġnik-i-n bġr-ec’in.

boy-PL-DEF butterfly-ACC-DEF catch-AOR.3PL

‘The boys caught. 3PL the butterfly.’

Data from understudied languages are crucially important when investigating psycholinguistic phenomena such as agreement attraction, given that the involved mechanisms are implicitly assumed to be universal across languages (Evans & Levinson, 2009; Norcliffe, Harris, & Jaeger, 2015), and that the underlying cognitive machinery is implicitly assumed to be universal across cultures (Levinson, 2012; Norenzayan & Heine, 2005). As a case in point, Eberhard et al. (2005) fitted their marking and morphing model to English data and then tested it on Spanish data “to make a case for the crosslinguistic promise of the model and the theory behind it” (p. 551). Likewise, the cue-based retrieval model of sentence processing is explicitly presented as aiming toward crosslinguistic coverage, given that the underlying memory mechanisms are claimed to be universal (Lewis et al., 2006). Of the 53 comprehension studies

Table 1

Experimental materials used in the experiment

Grammatical, subject-attractor match, attractor-verb match			
Nkarič-Ø-ë	or-in	k'andakagorç-Ø-ë	arhamarh-ec' ...
painter- SG .NOM-DEF	that-SG.ACC	sculptor- SG .NOM-DEF	ignore-AOR.3- SG
Grammatical, subject-attractor mismatch, attractor-verb mismatch			
Nkarič-ner-ë	or-onc'	k'andakagorç-Ø-ë	arhamarh-ec' ...
painter- PL .NOM-DEF	that-PL.ACC	sculptor- SG .NOM-DEF	ignore-AOR.3- SG
Ungrammatical, subject-attractor match, attractor-verb mismatch			
Nkarič-Ø-ë	or-in	k'andakagorç-Ø-ë	arhamarh-ec'-in ...
painter- SG .NOM-DEF	that-SG.ACC	sculptor- SG .NOM-DEF	ignore-AOR.3- PL
Ungrammatical, subject-attractor mismatch, attractor-verb match			
Nkarič-ner-ë	or-onc'	k'andakagorç-Ø-ë	arhamarh-ec'-in ...
painter- PL .NOM-DEF	that-PL.ACC	sculptor- SG .NOM-DEF	ignore-AOR.3- PL
c'owc'ahandes-i	ënt'ac'k'owm	vagowc'	mekowsac'-v-el
exhibition-DAT	POST	long	ostracize-PASS-PTCP.PRF
e/en	arvestaget-ner-i	šranak-ic'.	
be-PRS.3SG/3PL	artist-PL-GEN	circle-ABL	
"The painter(s) that the sculptor ignored during the exhibition has/have long been ostracized from the art community."			

Abbreviations: SG, singular; PL, plural; NOM, nominative; ACC, accusative; DAT, dative; GEN, genitive; ABL, ablative; REL, relative pronoun; POST, postposition; DEF, definite; AOR, aorist; PRS, present; PASS, passive; PTCP, participle; PRF, perfect.

reviewed by Hammerly et al. (2019), only 14 investigated languages other than English, and only two tested non-Germanic, non-Romance languages, underscoring the need for a more diverse range of data. Specifically, it has been argued that agreement attraction is less likely to occur in morphologically rich languages compared to morphologically impoverished languages such as English (Acuña-Fariña, 2012; Lorimor, Bock, Zalkind, Sheyman, & Beard, 2008), so that published effect sizes in English may give a biased picture of the true scope of the phenomenon. The morphology of Eastern Armenian is rich, with seven cases and a complex system of verb conjugation. Therefore, Eastern Armenian provides a much needed testing ground to computationally assess the size and directionality of attraction effects.

The experimental design of Avetisyan et al. (2020) used postnominal object relative clauses (see Table 1) to study agreement attraction between the clause-internal subject and the relativized object (the attractor). Attraction effects in relative clauses are known from previous studies (e.g., Bock & Cutting, 1992; Linzen & Leonard, 2018; Staub, 2009). We adapted the design of Avetisyan et al. by adding end-of-sentence comprehension probes, so that it could be inferred which of the two noun phrases in the sentence – the grammatical subject or the attractor – was interpreted as the thematic subject. The study is, to our knowledge, the first to allow subjects to type in their answers in an unconstrained fashion. This way, besides noun phrase choice, we can also investigate whether the number feature of the chosen noun phrase is changed from its veridical value. In addition, the continued study of Eastern Armenian is an opportunity to replicate the agreement attraction effect in reading times observed by Avetisyan et al. (2020).

3.1. Participants

Forty-eight native speakers of Eastern Armenian participated in the study. Five participants were excluded prior to analysis due to low question-response accuracy for the filler items (accuracy below 70%). This left data from 43 participants for the analysis (mean age 23 years; range 18–32 years). All participants provided informed consent. They received compensation in Armenian drams equivalent to 6 each.

3.2. Materials

The materials consisted of the 36 sentences used in Experiment 2 of Avetisyan et al. (2020). The factors grammaticality (grammatical vs. ungrammatical) and attractor-verb match (match vs. mismatch) were manipulated in a 2×2 design. Alternatively, the second factor can be decomposed as a manipulation of the match between subject and attractor, as shown in Table 1. The critical verb occurred inside an object relative clause whose head NP served as the attractor. The number feature of the verb either matched that of the subject (grammatical) or it did not (ungrammatical), and also either matched the number feature of the attractor (match) or not (mismatch). Stimulus sentences were semantically reversible, as judged by four Eastern Armenian native speakers.

Eighty-four grammatical fillers were used. Fillers consisted of 48 subject- and object-extracted relative clauses, as well as 36 simple sentences with (in)direct objects. The percentage of ungrammatical items was 15%. Experimental items were rotated through the experimental conditions in a Latin-square design, so that no participant saw two versions of the same item. The order of fillers and experimental items was pseudo-randomized at runtime, such that no two experimental items were adjacent.

3.3. Procedure

Sentences were presented in centered self-paced reading on a laptop computer using the Linger software (Rohde, 2003). After having read the final word of the sentence, participants were presented with an open-ended comprehension question and asked to type their response. For instance, the question for the sentence presented in Table 1 was *Ov kam ovk'er arhamarhec'in Ø?*, “Who.SG or who.PL ignored.PL Ø?” The form in which the question was given implied that the answer should specify the number of the response (singular or plural), and was intended to motivate participants to pay attention to both thematic relationships (*Who did what to whom?*) and number marking within the sentence. All questions targeted the thematic subject role (*Who (singular or plural) VERBed?*). There was no time limit for the typed responses.

3.4. Data analysis

In addition to fitting our computational models, we carried out a standard analysis using Bayesian linear mixed models (LMMs) in R (R Core Team, 2019) with the brms package (Bürkner, 2017, 2018). LMMs with full variance–covariance matrices for all random effects were fitted to the reading time data of the critical (RC verb) and postcritical (RC adverb)

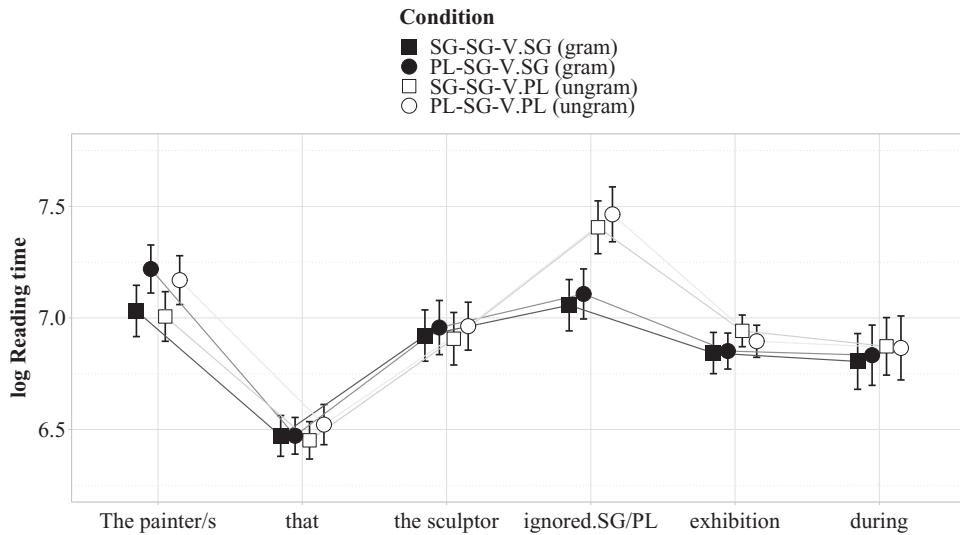


Fig. 3. Mean log reading times by region of interest. Error bars show 95% confidence intervals. The first number feature in the condition labels corresponds to the attractor noun while the second corresponds to the subject noun.

regions of the sentence, as well as to question-response times. Question-response times—that is, the time taken from the question being presented to the subject finishing the typing of their response—were recorded and analyzed because longer latencies for nonveridical compared to veridical responses may indicate post hoc pragmatic reasoning processes unrelated to online sentence processing. Bayesian generalized LMMs were fitted to the question-response data with a logit link function, also with full variance–covariance matrices for all random effects. Further information about the analyses, posterior plots for all coefficients, as well as additional information about the participant sample, are given in the Appendix.

Question responses were coded according to two factors: The noun phrase given as the response (subject = 0, attractor = 1) and the number feature of the response irrespective of noun choice (veridical feature = 0, changed feature = 1). Trials on which the response did not match any of the resulting four categories (around 16% of trials), including trials on which no response was given, were not analyzed. Thematic misinterpretation was analyzed using all remaining responses irrespective of feature change. Feature change was analyzed separately for trials on which the subject noun was chosen and trials on which the attractor noun was chosen.

3.5. Descriptive results and LMM analysis

3.5.1. Reading times

Mean log reading times for all regions between the main clause subject and the end of the relative clause are shown in Fig. 3.

Reading times at the critical verb region were shorter for grammatical than for ungrammatical sentences ($\hat{\Delta} = -500$ ms, credible interval (CrI): $[-608$ ms, -392 ms]) and showed

Table 2

Total counts and proportions of noun phrase choice and number feature changes, as well as question-response times, by condition

Noun phrase choice				
Condition	Subject	Attractor	\overline{RT} (ms)	SE
SG-SG-V.SG (gram)	249 (77%)	74 (23%)	12,540	631
PL-SG-V.SG (gram)	228 (67%)	112 (33%)	12,457	564
SG-SG-V.PL (ungram)	241 (76%)	75 (24%)	12,787	799
PL-SG-V.PL (ungram)	222 (70%)	96 (30%)	13,051	738
Number feature of chosen noun phrase				
Condition	Subject	Attractor	Veridical	Changed
	Veridical	Changed	Veridical	Changed
SG-SG-V.SG (gram)	221 (89%)	28 (11%)	69 (93%)	5 (7%)
PL-SG-V.SG (gram)	169 (74%)	59 (26%)	93 (83%)	19 (17%)
SG-SG-V.PL (ungram)	164 (68%)	77 (32%)	54 (72%)	21 (28%)
PL-SG-V.PL (ungram)	106 (48%)	116 (52%)	86 (90%)	10 (10%)

some indication of a negative interaction between grammaticality and attractor-verb match ($\hat{\Delta} = -90$ ms, CrI: $[-193$ ms, 16 ms]). Numerically, the interaction takes the form of a speedup due to subject-attractor match, or equivalently a speedup in the presence of a singular rather than a plural attractor irrespective of grammaticality (see Fig. 3). As the credible interval of the effect is very wide and includes zero, we do not speculate further on its origin.

Reading times at the postverb region were shorter for grammatical than for ungrammatical sentences ($\hat{\Delta} = -86$ ms, CrI: $[-155$ ms, -16 ms]). There is no convincing case to be made for an interaction between grammaticality and attractor-verb match at this region ($\hat{\Delta} = 36$ ms, CrI: $[-25$ ms, 96 ms]). There is thus no convincing indication of the typical asymmetrical profile of attraction effects, which are often reported in ungrammatical sentences only (e.g., Wagers et al., 2009).

3.5.2. Question responses

Table 2 shows counts and proportions of thematic misinterpretation and feature change, as well as mean question-response times, by condition.

There was no indication that the experimental manipulations affected question-response times.

For thematic misinterpretation, there was an interaction between grammaticality and attractor-verb match, which is equivalent to an effect of subject-attractor match: Thematic misinterpretation—that is, the attractor being given as the response—was less likely when the attractor matched the subject in number ($\hat{\Delta} = -0.1$, CrI: $[-0.17, -0.04]$), that is, when the attractor was singular rather than plural.

On trials where the subject was given as the response, feature change was less likely to occur when the sentence was grammatical ($\hat{\Delta} = -0.27$, CrI: $[-0.38, -0.17]$), that is, when subject and verb had the same number feature. Feature change was also less likely when subject and attractor had the same number feature ($\hat{\Delta} = -0.25$, CrI: $[-0.34, -0.16]$).

On trials where the attractor was given as the response, feature change was less likely when the attractor matched the verb in number ($\hat{\Delta} = -0.06$, CrI: $[-0.13, -0.01]$).

3.6. Comparison of computational models using k -fold cross-validation

We now compare the predictive fit of three computational models (retrieval, encoding, and SOP) to the experimental data using k -fold cross-validation. The procedure estimates the expected log pointwise predictive density (\widehat{elpd}) for new observations given the fitted model. The data are split into k subsets, and the following procedure is carried out k times: One of the k subsets is held out while the model is fitted to the remaining data. Then, we compute \widehat{elpd} , a measure of the amount of deviation between the observed (held-out) values and the predicted values. This measure of fit is then summed across the k iterations, so that the entire data set is covered.

Abstracting away from the k folds, \widehat{elpd} is computed as follows, where n denotes the number of data points, S denotes the number of posterior draws, and $p(y_i|\theta^s)$ is the probability of obtaining a single data point y_i given the parameter draws θ^s (Gelman, Hwang, & Vehtari, 2014):

$$\sum_{i=1}^n \log \left(\frac{1}{S} \sum_{s=1}^S p(y_i|\theta^s) \right). \quad (1)$$

\widehat{elpd} can be multiplied by -2 to put the measure on the deviance scale, making it comparable to other information criteria such as AIC (Akaike, 1974) or DIC (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). Like other information criteria, \widehat{elpd} scales with the number of observations, given that every additional observation is a chance for the model to “miss.” A “perfect” model that is able to predict each data point with absolute certainty—which is unrealistic for noisy human data—would have an \widehat{elpd} of zero. However, because the comparison is quantitative rather than qualitative, we can investigate differences between models whose predictions we know a priori to be partially inaccurate for the data at hand, as is the case with our experimental data.

We set k to 10, as is standardly done (Vehtari, Gelman, & Gabry, 2017). Data points were assigned pseudo-randomly to each of the folds, taking care that subjects, items, and experimental conditions were equally represented in each fold. We report the estimated difference in \widehat{elpd} and the standard error of the difference. As a sanity check, we also compared the models using PSIS-LOO (Vehtari, Gelman, & Gabry, 2015), as implemented in the R package *loo* (Vehtari, Gabry, Yao, & Gelman, 2019). Both methods yielded similar results.

Table 3 shows the cross-validation results.

The results can be summarized as follows:

- The encoding model shows better predictive fit than the retrieval model ($\widehat{\delta elpd} = 133$, $SE = 16$; the standard error has the usual frequentist interpretation).
- The SOP model shows better predictive fit than the base encoding model ($\widehat{\delta elpd} = 39$, $SE = 10$).

Table 3

Estimates of expected log pointwise predictive density (\widehat{elpd}) for each model

Model	\widehat{elpd}	SE	n_β
Encoding	-11,102	56	6
Retrieval	-11,235	57	4
SOP	-11,063	57	15

Note. Standard errors are computed as $\sqrt{n \cdot \text{var}(\text{pointwise_elpd})}$. The larger the value of \widehat{elpd} , the better the fit. The rightmost column refers to the number of slopes in the model.

3.7. Posterior predictive checks

The cross-validation results show that the SOP model has the best predictive fit among the three models considered. Next, we consider the posterior predictive performance of each model. While cross-validation is used to estimate how well a model generalizes to new data, posterior predictive checks are used to evaluate how similar data generated from the model are to the data used for fitting (Gelman, 2004). At 4000 postwarmup samples per chain, with four chains in total, the posterior predictive distribution consists of 16,000 simulated data sets.

For log reading times at the postverb region, the posterior predictive checks are uninformative, as there is little indication of reading time differences between trials with different responses (see Appendix). Fig. 4 shows posterior predictive checks for response proportions. The width of the violin represents the number of data points falling in the respective area on the y-axis. One way of assessing a model's predictive performance is to ask whether the widest part of the violins lines up with the empirical means from the original data (white circles).

Fig. 4 reveals differences between the models concerning predicted response proportions: Matching the cross-validation results, the best predictive fit is obtained by the SOP model, which predicts that the likelihood of feature change on the singular-marked subject noun increases as the number of SOP in the sentence increases. The other models fail to account for this pattern. The base encoding model predicts fewer veridical subject answers in sentences with subject-attractor mismatch compared to subject-attractor match, and a corresponding rise in nonveridical subject answers. By contrast, the retrieval model predicts fewer veridical subject answers in ungrammatical compared to grammatical sentences, and a corresponding rise in veridical attractor answers. All models fail to systematically capture the observation that plural attractors lead to more thematic misinterpretations.

Interestingly, the retrieval model, which does not predict systematic feature migration, somewhat overestimates the proportion of nonveridical subject answers in all conditions except the PL-SG-V·PL condition, where it underestimates the proportion instead. This pattern is due to the fact that the nonveridical subject chunk has only one activation value across conditions, and that its retrieval probability depends solely on changes in the other chunks' activations.

3.8. Discussion

The results of the k-fold cross-validation and the visual posterior predictive checks show that the SOP model is better able to predict our data than the other two candidate models.

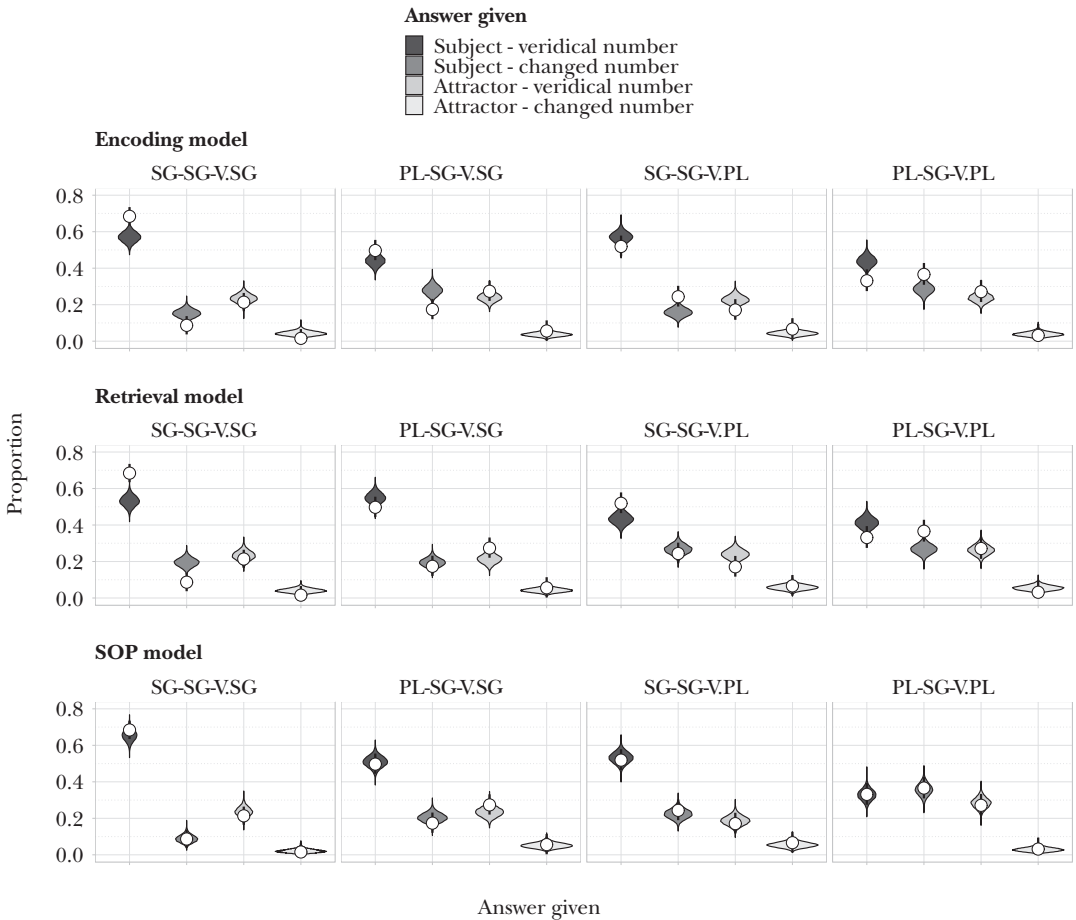


Fig. 4. Posterior predictive checks of response proportions by condition and model. Circles show the mean response proportions in the experimental data, violins show samples from the posterior predictive distribution. Error bars show 95% multinomial confidence intervals.

Moreover, the base encoding model, while showing suboptimal fit, performs better than the retrieval model.

With regard to interpretation, the encoding account predicts that responses with incorrect number features should occur more often when the two noun phrases in the sentence have different number features, irrespective of grammaticality. Indeed, our data show that the singular feature on the subject was more likely to be changed to plural when the attractor was plural (number misinterpretation), matching previous results by Patson and Husband (2016), and Brehm et al. (2019, 2021). By contrast, the retrieval account predicts subjects should misretrieve the attractor more often when it matches the verb’s number feature, especially when the subject does *not* match the verb’s number feature, that is, in ungrammatical sentences. Our data do not indicate such a pattern, explaining the model’s overall poor performance.

Both encoding and retrieval accounts predict that reading times should be faster at the critical verb in ungrammatical sentences when the attractor matches the verb in number. Our study showed a numerical tendency consistent with this prediction at the postverb region, but there was no reliable indication of an effect. For grammatical sentences, the encoding account predicts longer reading times in the presence of an attractor mismatching the subject in number. This is because the incorrect number feature is sometimes copied onto the subject, leading to the impression of ungrammaticality. By contrast, the retrieval account as implemented by Lewis and Vasishth (2005) predicts longer reading times when both subject and attractor match the verb's number feature, due to the fan effect (e.g., Anderson, 1974; Anderson & Reder, 1999; Lewis & Vasishth, 2005). There is no evidence for either of these predictions in our data.

Unexpectedly, there was a numerical speedup in the critical verb region when the subject and attractor matched in number, irrespective of grammaticality. Facilitatory effects of subject-attractor match have been observed in forced-choice production and in binary grammaticality judgments (e.g., Franck, Colonna, & Rizzi, 2015; Staub, 2009; Villata & Franck, 2020), but it is not clear why our data should pattern with binary-choice tasks rather than with other reading time results. We speculate that the result may be spurious.

Regarding interpretation, features on both the subject and the attractor were more likely to change when they mismatched the feature of the verb. This finding matches previous observations by Patson and Husband (2016) and Brehm et al. (2019, 2021). The effect can be easily accommodated by the encoding account by allowing a number feature from the verb to migrate onto a noun phrase in the sentence, analogously to feature migration between two noun phrases, as implemented in the SOP model.

There was also an effect of subject-attractor match on thematic misinterpretation, which possibly reduces to an effect of attractor plurality: In cases of mismatch, the attractor was always plural. Plural number has been argued to be more marked, and therefore possibly more salient, than singular number (e.g., Bock & Eberhard, 1993; Tiersma, 1982). Even though the Lewis and Vasishth (2005) model does not prefer certain feature specifications over others independently of the retrieval cues, it is possible to extend the model to account for a plural markedness effect.⁵

4. General discussion

Our study was designed to tease apart encoding- and retrieval-based accounts of agreement attraction based on reading time data and responses to end-of-sentence comprehension questions. Our data show both feature substitutions on the subject noun phrase (number misinterpretations), as predicted by the encoding account, and thematic misinterpretation, as predicted by the retrieval account. Our results thus extend previous findings by Patson and Husband (2016), Brehm et al. (2019, 2021), and Schlueter et al. (2019) by showing that both types of misinterpretation can be observed within a single experiment.

Through computational modeling and quantitative model comparison, we found that the best predictive fit to our data is obtained by assuming that both attractor NPs and verbs are

Table 4

Posterior means and 95% credible intervals for effects on reading times at the postverb region in Avetisyan et al.'s (2020) Experiment 2 and in our data, along with joint estimates based on both studies

		Avetisyan et al.	Current study	Updated
Grammaticality ×	Estimate	33 ms	36 ms	34 ms
Attraction	95% CrI	[1 ms, 66 ms]	[-25 ms, 96 ms]	[4 ms, 62 ms]
Attraction	Estimate	14 ms	22 ms	17 ms
(grammatical)	95% CrI	[-30 ms, 58 ms]	(-55 ms, 100 ms]	[-23 ms, 60 ms]
Attraction	Estimate	-51 ms	-48 ms	-49 ms
(ungrammatical)	95% CrI	[-100 ms, -2 ms]	[-137 ms, 40 ms]	[-91 ms, -7 ms]

Note. Values have been back-transformed from the logarithmic scale.

possible sources of nonveridical number features on subject nouns, resulting in number misinterpretations. Meanwhile, thematic misinterpretations, which are the hallmark prediction of the retrieval account, in our experiment were mainly influenced by the attractor being plural, a finding that is not accounted for by the Lewis and Vasishth (2005) model.

4.1. On the absence of a clear attraction effect in reading times

Our reading time results did not show a clear indication of faster reading times in ungrammatical sentences in the presence of a verb-matching attractor, which is the hallmark signature of agreement attraction. The absence of a clear attraction effect in reading times may be due to a lack of statistical power (Vasishth, Mertzen, Jäger, & Gelman, 2018). Nevertheless, the estimates from our study are similar to those found in Avetisyan et al.'s (2020) recent study on Eastern Armenian, which used the same stimulus materials. Table 4 shows the means and credible intervals for the effects at the postverb region in Avetisyan et al.'s Experiment 2 versus the means and credible intervals from the current study. The final column shows the updated estimates obtained when the posterior estimates of Avetisyan et al. and their associated credible intervals are used as priors for the current experiment.

Looking at the estimates, there is a gradual measure of support for an agreement attraction effect provided by both studies, which is lower for our study than for that of Avetisyan et al., as reflected in the width of the credible intervals. Moreover, the updated estimates show that the data from our study do not result in any large shifts of the posterior means, but lead to some narrowing of the credible intervals.

The lack of a clear effect in our study may also be due to the experimental task. In order to avoid strategic reading on part of participants, most studies on agreement attraction do not use comprehension questions that target the critical subject-verb dependency (Avetisyan et al., 2020). Previous reading studies in which comprehension questions *did* target the number feature of the subject did not show consistent attraction effects: While Patson and Husband (2016) report an interaction between grammaticality and attractor number in reading times in the first of their two experiments, the interaction has the wrong shape: verb-matching attractors caused a speedup in grammatical sentences only, with no difference in ungrammatical sentences, which is inconsistent with the typical attraction pattern. Brehm et al. (2019) found no attraction effect in reading times at all. It is thus possible that explicitly

probing for the number interpretation of the subject reduces or eliminates attraction effects in reading comprehension. In order to shed light on this issue, higher-powered replication studies are necessary.

4.2. *Empirical coverage of the competing accounts, and the possibility of a hybrid model*

The retrieval-based view of agreement attraction has dominated the recent literature, largely because the encoding-based view is arguably unable to account for the grammaticality asymmetry, that is, for attraction affecting ungrammatical sentences to a larger extent than grammatical sentences (e.g., Wagers et al., 2009). However, Hammerly et al. (2019) note in a recent article that the evidence in favor of an asymmetry is somewhat equivocal, given that only 27 of 45 reviewed studies that tested for it found such an asymmetry. Similarly, Jäger, Engelmann, & Vasishth (2017) present a meta-analysis of 49 comprehension studies on subject-verb and reflexive/reciprocal dependencies, finding support for some but not all predictions of the Lewis and Vasishth (2005) model. Furthermore, similarity-based memory interference, which is the core mechanism behind agreement attraction under the retrieval-based view, has also been observed in contexts where the critical cue is not used for retrieval (see Smith, Franck, & Tabor, 2021 and references therein). The overall pattern of results thus does not uniquely favor the retrieval-based view. Given this picture, and given the evidence for number misinterpretations in English obtained by Patson and Husband (2016), Brehm et al. (2019, 2021), and in our own current study on Eastern Armenian, a credible challenge to the dominant perspective exists.

Even though the encoding account provides a better quantitative fit to the data than the retrieval account, some of our results are problematic for specific variants of the encoding-based view. The fact that feature changes were observed on the subject even though it was structurally more deeply embedded than the attractor speaks against percolation-based accounts that assume features to only migrate upwards in the syntactic tree (e.g., Franck et al., 2002; Vigliocco et al., 1995). Furthermore, the plural feature on the attractor was sometimes changed to singular, which is unexpected if singular is an unmarked default value, as assumed by some authors (Eberhard, 1997; Eberhard et al., 2005). Nevertheless, an encoding-based model without assumptions about the direction of percolation and the markedness of specific features, but with the verb as an added source of nonveridical features, is able to account for the majority of patterns in our data. The SOP model is also expected to generalize well to future data, given the cross-validation results and posterior predictive checks. In future work, we plan to extend our modeling approach to data from a more diverse set of languages to see if the predicted patterns are borne out cross-linguistically.

In addition to the observed pattern of number misinterpretation that is not accounted for by the retrieval-based account, the pattern of thematic interpretation in our data does not conform to the predictions of the retrieval account either: There was no indication that thematic misinterpretation was driven by the cue match between the attractor and the verb. We have already noted that thematic misinterpretation may be encoding-based (Bock & Miller, 1991; Eberhard et al., 2005), as also suggested by Villata, Tabor and Franck (2018). Villata et al. present experimental evidence that thematic misinterpretation is influenced by feature match

between subject and attractor even for features unrelated to retrieval, thus calling into question the role of retrieval in thematic misinterpretation. In their proposed model, the attractor sometimes ends up occupying the subject slot of the verb when feature match is high, thus predicting that the entire sentence representation is incorrectly encoded (see also Villata & Franck, 2020). The model assumes *self-organization* as the central mechanism for sentence processing, an idea that has recently started gaining traction (see, e.g., Smith et al., 2018, 2021, and references therein). A self-organized system of sentence processing attempts to form a representation by positing linkages between input words that have matching features, without central grammatical supervision. The version of self-organized sentence processing proposed by Villata et al. (2018) not only allows for the occasional formation of grammatically incorrect linkages but also for passing of features between elements of the sentence, thus covering both number misinterpretation and thematic misinterpretation, as observed in our study. Self-organization is exhibited by a wide range of physical and biological systems (see Gershenson, Trianni, Werfel, & Sayama, 2020 for a review), and is thus likely present in cognition as well (Barton, 1994), making it an attractive mechanism with potentially wide-ranging applications in sentence processing.

A further avenue to be explored is whether additional augmentations of the Lewis and Vasishth (2005) model, hybrid approaches involving both encoding and retrieval components (Yadav, Smith, & Vasishth, 2021), or future deep learning models may be successful in predicting the entire range of empirical data. For instance, the Misidentification stage of our encoding model can be interpreted as a retrieval process. This retrieval could be followed by a (re-)encoding process that can change features on the retrieved noun phrase (cf. Patson & Husband, 2016; Schlueter et al., 2019), so that the occasional faulty encoding of the subject's number feature is a *consequence* of retrieval (see also Konieczny et al., 2004). Deep learning models, by contrast, do not have a dedicated retrieval mechanism. Nevertheless, Arehalli and Linzen (2020) observed asymmetrical attraction effects in ungrammatical compared to grammatical sentences in their model, which are usually attributed to retrieval. In another line of work, Ryu and Lewis (2021) used a Transformer model to predict the upcoming verb in attraction sentences and related the model's attention mechanism to cue-based retrieval. As deep learning researchers continue to make progress in relating model behavior to human performance and discover possible shared mechanisms between their models and human cognition, it remains to be seen if and how a direct mapping to the encoding- and retrieval-based components of psycholinguistic models is possible.

4.3. *Comprehension questions, postinterpretive processing, and memory distortions*

Our study used open-ended comprehension questions to probe participants' interpretations of agreement attraction sentences, and yielded evidence that misinterpretations occur quite frequently in such a paradigm. Checking for systematic patterns of misinterpretation is crucial to disentangling the encoding- and retrieval-based accounts. However, there is a significant caveat to using comprehension questions: One cannot be certain that subjects' answers accurately reflect the thematic relations computed during the incremental processing of the sentence (see also Brehm et al., 2019; Schlueter et al., 2019; Tanner, Dempsey, &

Christianson, 2018). Bader and Meng (2018) have recently argued that answers to comprehension questions may partly reflect what Caplan and Waters (1999) call *postinterpretive* processing (see also Paolazzi, Grillo, Alexiadou, & Santi, 2019). Caplan and Waters argue that the parser generates a syntactic structure and assigns meaning to the sentence online (interpretive processing), but that afterwards additional processes come into play whose output does not always match the online interpretation.

There was conceivably a lot of room for postinterpretive processing in our task, given that the median response time for comprehension questions in our experiment was about 10.5 seconds (bootstrapped CrI: [10.2 s, 11 s]). Participants incorrectly chose the attractor in about 21% of cases overall (CrI: [14%, 30%]), with an estimated increase of about 10% (CrI: [4%, 17%]) when the attractor was plural. By contrast, Schlueter et al. (2019) used a more implicit task (adjective-noun matching) with a time-out of three seconds in order to minimize postinterpretive processing. They observed overall high accuracy (about 90%) on the adjective-matching task, as well as a smaller increase in thematic misinterpretations (about 3%). Brehm et al. (2021) also observed very few instances of thematic misinterpretation; however, in their sentences the thematic misinterpretation was often implausible. Speculatively, the larger effect in our study may be partly due to postinterpretive strategies, or due to the increased plausibility of the incorrect interpretation, given that our sentences were pre-tested to be semantically reversible.

In our modeling, we have assumed that the final interpretation of the sentence, including the identity and number feature of the verbal subject, is determined at the postverb region, based on the fact that Avetisyan et al. (2020) observed the agreement attraction effect in this region. This is, of course, a simplifying assumption. There may be inter-individual as well as trial-by-trial variability with regard to when the interpretation is determined: immediately at the verb, in any of the following regions, at the end of the sentence, or during the typing of the question response. However, immediacy of interpretation is supported by the fact that in their picture-selection study, Brehm et al. (2021) observed differences between conditions in the proportion of looks to incorrect pictures shortly after the verb was processed. Furthermore, our data showed no indication that question-response times differed between conditions, which is compatible with the notion that participants had already derived an interpretation at this point in time. Nevertheless, future work should aim to investigate more thoroughly how the dynamics of interpretation unfold over time, and to what extent postinterpretive processing may play a role.

More generally, nonveridical representations of linguistic input, constructed either online or postinterpretively, should be viewed against the background of research on memory distortions. End-of-trial comprehension probes constitute a recall task, with feature substitutions and thematic misinterpretations constituting recall errors. It is known that illusory recall and recognition of target stimuli can be induced through the presentation of stimuli that share features with the target (e.g., Reinitz, Lammers et al., 1992), and that illusory recall can lead to subsequent confident recognition (Roediger & McDermott, 1995). Such errors are likely due to the application of “reconstructive” rather than purely reproductive memory processes, resulting in the creation of novel memory traces from partially recalled experiences (Bartlett, 1932). Reconstruction is known to be affected by general knowledge in addition to actual

experience (e.g., Hemmer & Steyvers, 2009; Kolodner, 1983), which may extend to experience with grammatical sentences in which the number features of a verb and its subject match.

In the realm of sentence processing, reconstructive processes have been proposed that “repair” unexpected input – such as an ungrammatical sentence – to match expectations (e.g., Fillenbaum, 1974; Levy, 2008), or allow for post hoc reconstruction of sentence meaning after encountering a comprehension question (Wonnacott, Joseph, Adelman, & Nation, 2016). Brehm et al. (2019, 2021) have suggested that such processes may play a role in agreement attraction. Despite the wealth of research on “false memories” (Koriat, Goldsmith, & Pansky, 2000), it is still unclear which component processes are involved in their generation, and how general these processes are (Gallo, 2010). Future work on agreement attraction may shed light on the extent to which subjects generate nonveridical representations with incorrect number features or thematic relations, for instance by using a paraphrasing task.

5. Conclusion

We have presented the first comprehension study of agreement attraction that combined reading time data with open-ended end-of-trial comprehension questions about the critical dependency, finding that both misrepresentations of the subject’s number feature and mis-retrievals of the attractor occur during processing. We have also demonstrated the value of computational implementation and quantitative model comparison in uncovering the processes underlying agreement attraction in comprehension, finding that an encoding-based model yields a better fit to our data. The fit is further improved when the verb is considered as a source of nonveridical features encoded on its argument noun phrases. We have suggested avenues for future work on the phenomenon, and put particular emphasis on the need to establish links with the broader literature on human memory.

Acknowledgments

This work was supported partly by the University of Potsdam. The second author was funded through an Erasmus Mundus Joint Doctorate (EMJD) Fellowship for “International Doctorate for Experimental Approaches to Language and Brain” (IDEALAB) under grant no. 2012-0025-EMII-EMJ. For helpful comments and criticism, we are grateful to the Daniel Schad, Garrett Smith, Anna Laurinavichyute, and the audience at AMLaP 2019 in Moscow. Special thanks go to Bruno Nicenboim for many insightful discussions about the modeling approach and advice on the implementation.

Open Access funding enabled and organized by Projekt DEAL.

Notes

1. It is not clear whether cue-based retrieval is the basic mechanism for computing subject–verb agreement or whether it is only triggered as a last resort in ungrammatical sentences (Wagers et al., 2009). However, because memory retrieval in ungrammatical sentences is assumed in either case, we do not attempt to resolve this debate here.

2. A further plausible mechanism for encoding interference that we do not consider here is *feature overwriting*, in which features that are shared by multiple memory traces are lost with some probability (Neath, 2000; Nairne, 1990; Oberauer & Kliegl, 2006). Feature overwriting in the context of agreement attraction has been implemented by Vasishth, Jäger, and Nicenboim (2018).
3. MPT models often explicitly include parameters related to guessing (Erdfelder et al., 2009), but we refrain from including guessing in our model. Given that there are two possible guessing processes ('subject guesses' and 'number guesses'), which may be associated with different biases, such a model would have too many parameters to be identifiable.
4. For our current purpose, we consider only nonveridical chunks derived by replacing the number feature of the veridical chunk. In order to more exhaustively evaluate the proposal, one would also need to consider other possible retrieval features, such as case and gender, as well as semantic features.
5. See Phillips (2013) for a similar point.

References

- Acuña-Fariña, J. C. (2012). Agreement, attraction and architectural opportunism. *Journal of Linguistics*, 48(2), 257–295.
- Adger, D. (2003). *Core syntax: A minimalist approach*. Oxford, England: Oxford University Press.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 6(4), 451–474.
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060.
- Anderson, J. R., & Lebiere, C. J. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.
- Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, 128(2), 186–197.
- Arehalli, S., & Linzen, T. (2020). Neural language models capture some, but not all, agreement attraction effects. Available at: psyarxiv.com/97qcg <https://doi.org/10.31234/osf.io/97qcg>
- Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, 112, 104087.
- Badecker, W., & Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of Memory and Language*, 56(1), 65–85.
- Badecker, W., & Lewis, R. (2007). A new theory and computational model of working memory in sentence production: Agreement errors as failures of cue-based retrieval. *Paper presented at the 20th Annual CUNY Sentence Processing Conference*, San Diego, La Jolla, CA.
- Bader, M., & Meng, M. (2018). The misinterpretation of noncanonical sentences revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(8), 1286–1311.
- Bansal, H., Bhatt, G., & Agarwal, S. (2020). Can RNNs trained on harder subject–verb agreement instances still perform well on easier ones? *Proceedings of the Society for Computation in Linguistics*: Vol. 4, Article 38.
- Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

- Barton, S. (1994). Chaos, self-organization, and psychology. *American Psychologist*, 49(1), 5–14.
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6(1), 57–86.
- Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, 31(1), 99–127.
- Bock, K., & Eberhard, K. M. (1993). Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes*, 8(1), 57–99.
- Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, 23(1), 45–93.
- Box, G. E. (1979). Robustness in the strategy of scientific model building. In R. Launer & W. G. N. (Eds.), *Robustness in statistics* (pp. 201–236). New York: Academic Press.
- Brehm, L., & Goldrick, M. (2017). Distinguishing discrete and gradient category structure in language: Insights from verb-particle constructions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(10), 1537–1556.
- Brehm, L., Jackson, C. N., & Miller, K. L. (2019). Speaker-specific processing of anomalous utterances. *Quarterly Journal of Experimental Psychology*, 72(4), 764–778.
- Brehm, L., Jackson, C. N., & Miller, K. L. (2021). Probabilistic online processing of sentence anomalies. *Language, Cognition and Neuroscience*, 1–25. <https://doi.org/10.1080/23273798.2021.1900579>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28.
- Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R package brms. *R Journal*, 10(1), 395–411.
- Byrne, M. D. (2012). Unified theories of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(4), 431–438.
- Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, 22(1), 77–94.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1). <https://doi.org/10.18637/jss.v076.i01>
- Clifton, C., Frazier, L., & Deevy, P. (1999). Feature manipulation in sentence comprehension. *Italian Journal of Linguistics*, 11(1), 11–40.
- Comrie, B. (1981). *The languages of the Soviet Union*. Cambridge: Cambridge University Press.
- Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2), 85–103.
- Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68–77.
- Dum-Tragut, J. (2009). *Armenian: Modern Eastern Armenian*. Amsterdam: John Benjamins.
- Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and Language*, 36(2), 147–164.
- Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, 112(3), 531–559.
- Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, 43(12), e12800.
- Erdfelder, E., Auer, T.-S., Hilbig, B. E., Abfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology*, 217(3), 108–124.
- Esterman, M., Prinzmetal, W., & Robertson, L. (2004). Categorization influences illusory conjunctions. *Psychonomic Bulletin & Review*, 11(4), 681–686.
- Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429–448.
- Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, 19(12), 729–743.

- Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, 19(5), 329–335.
- Fillenbaum, S. (1974). Pragmatic normalization: Further results for some conjunctive and disjunctive sentences. *Journal of Experimental Psychology*, 102(4), 574–578.
- Forbus, K. D., Liang, C., & Rabkina, I. (2017). Representation and computation in cognitive models. *Topics in Cognitive Science*, 9(3), 694–718.
- Franck, J., Colonna, S., & Rizzi, L. (2015). Task-dependency and structure-dependency in number interference effects in sentence comprehension. *Frontiers in Psychology*, 6, 349.
- Franck, J., Vigliocco, G., & Nicol, J. (2002). Subject–verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, 17(4), 371–404.
- Franck, J., & Wagers, M. (2020). Hierarchical structure and memory mechanisms in agreement attraction. *PLoS One*, 15(5), 1–33.
- Fromkin, A. (1971). The non-anomalous nature of anomalous utterances. *Language*, 47, 27–52.
- Fum, D., Del Missier, F., & Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research*, 8(3), 135–142.
- Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, 38(7), 833–848.
- Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13(4), 755–779.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6), 997–1016.
- Gershenson, C., Trianni, V., Werfel, J., & Sayama, H. (2020). Self-Organization and Artificial Life. *Artificial Life*, 26(3), 391–408.
- Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*. <https://doi.org/10.1177/1745691620970585>
- Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology*, 110, 70–104.
- Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1), 1–12.
- Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, 23(5), 1440–1465.
- Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, 1(1), 189–202.
- Hodgson, K. (2018). Word order, information structure and relativization strategies in Eastern Armenian. *International workshop OV-IS 2018: OV basic word order correlates and information structure*, INALCO, Paris.
- Hupet, M., Fayol, M., & Schelstraete, M.-A. (1998). Effects of semantic variables on the subject–verb agreement processes in writing. *British Journal of Psychology*, 89(1), 59–75.
- Hurlstone, M. (in press). Serial recall. In J. Kahana & A. D. Wagner (Eds.), *The Oxford handbook on human memory*. Oxford, England: Oxford University Press.
- Jäger, L. A., Engelmann, F., & Vasisht, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, 94, 316–339.
- Jäger, L. A., Merten, D., Van Dyke, J. A., & Vasisht, S. (2020). Interference patterns in subject–verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, 111, 104063.
- Jovina, I., & Taatgen, N. (2009). A repetition-suppression account of between-trial effects in a modified Stroop paradigm. *Acta Psychologica*, 131(1), 72–84.
- Kimball, J., & Aissen, J. (1971). I think, you think, he think. *Linguistic Inquiry*, 2(2), 241–246.
- Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, 75(1), 70–98.

- Klauer, K. C., & Kellen, D. (2018). RT-MPTs: Process models for response-time distributions based on multinomial processing trees with applications to recognition memory. *Journal of Mathematical Psychology*, 82, 111–130.
- Kolodner, J. L. (1983). Reconstructive memory: A computer model. *Cognitive Science*, 7(4), 281–328.
- Konieczny, L., Schimke, S., & Hemforth, B. (2004). An activation-based model of agreement errors in production and comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Chicago, IL (Vol. 26).
- Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology*, 51(1), 481–537.
- Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. *Frontiers in Psychology*, 9, 1117.
- Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement attraction in Spanish comprehension. *Journal of Memory and Language*, 82, 133–149.
- Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., & Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*. <https://doi.org/10.1016/j.cognition.2021.104699>
- Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). *The emergence of number and syntax units in LSTM language models*. arXiv:1903.07435v2.
- Laurinavichyute, A., & von der Malsburg, T. (2020, February). *Semantic attraction in sentence comprehension*. Available at: psyarxiv.com/hk9nc
- Levinson, S. C. (2012). The original sin of cognitive science. *Topics in Cognitive Science*, 4(3), 396–403.
- Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, HI (pp. 234–243).
- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3), 375–419.
- Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10), 447–454.
- Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1), 195–212.
- Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of CogSci* (pp. 692–69).
- López, L. (2001). Head of a projection. *Linguistic Inquiry*, 32(3), 521–532.
- Lorimor, H., Bock, K., Zalkind, E., Sheyman, A., & Beard, R. (2008). Agreement and attraction in Russian. *Language and Cognitive Processes*, 23(6), 769–799.
- Mätzig, P., Vasishth, S., Engelmann, F., Caplan, D., & Burchert, F. (2018). A computational investigation of sources of variability in sentence comprehension difficulty in aphasia. *Topics in Cognitive Science*, 10(1), 161–174.
- Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, 80(1), 205–235.
- McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, 29(2), 111–123.
- Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.
- Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, 18(3), 251–269.
- Neath, I. (2000). Modeling the effects of irrelevant speech on memory. *Psychonomic Bulletin & Review*, 7(3), 403–423.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, 99, 1–34.

- Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science*, 42, 1075–1100.
- Nicol, J. L., Forster, K. I., & Veres, C. (1997). Subject–verb agreement processes in comprehension. *Journal of Memory and Language*, 36(4), 569–587.
- Norcliffe, E., Harris, A. C., & Jaeger, T. F. (2015). Cross-linguistic psycholinguistics and its critical role in theory development: Early beginnings and recent advances. *Language, Cognition and Neuroscience*, 30(9), 1009–1032.
- Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin*, 131(5), 763–784.
- Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, 55(4), 601–626.
- Oberauer, K., & Lange, E. B. (2008). Interference in verbal working memory: Distinguishing similarity-based confusion, feature overwriting, and feature migration. *Journal of Memory and Language*, 58(3), 730–745.
- Paolazzi, C. L., Grillo, N., Alexiadou, A., & Santi, A. (2019). Passives are not hard to interpret but hard to remember: Evidence from online and offline studies. *Language, Cognition and Neuroscience*, 34(8), 991–1015.
- Patson, N. D., & Husband, E. M. (2016). Misinterpretations in agreement and agreement attraction. *Quarterly Journal of Experimental Psychology*, 69(5), 950–971.
- Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, 41(3), 427–456.
- Phillips, C. (2013). Some arguments and nonarguments for reductionist accounts of syntactic phenomena. *Language and Cognitive Processes*, 28(1–2), 156–187.
- Pittman, C., & Smyth, R. (2005). The effect of the predicate on agreement error rates. In *Proceedings of the 2005 Canadian Linguistics Association Annual Conference*, Ontario.
- Pollard, C., & Sag, I. (2004). *Head driven phrase structure grammar*. Chicago, IL: University of Chicago Press.
- Prinzmetal, W. (1981). Principles of feature integration in visual perception. *Perception & Psychophysics*, 30(4), 330–340.
- R Core Team (2019). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>
- Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, 24(5), 574–590.
- Reinartz, M. T., Lammers, W. J., & Cochran, B. P. (1992). Memory-conjunction errors: Miscombination of stored stimulus features can produce illusions of memory. *Memory & Cognition*, 20(1), 1–11.
- Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, 95(3), 318–339.
- Ritter, F. E., Tehranchi, F., & Oury, J. D. (2019). ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(3), e1488.
- Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(4), 803–814.
- Rohde, D. (2003). *Linger* (version 2.94). Available at: <http://tedlab.mit.edu/~dr/Linger/>
- Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 80(2), 491–513.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151–1172.
- Ryu, S. H., & Lewis, R. L. (2021). *Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention*. arXiv:2104.12874v1.
- Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors*, 48(2), 362–380.
- van Schijndel, M., & Linzen, T. (2020). *Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty*. Available at: psyarxiv.com/sgbqy <https://doi.org/10.31234/osf.io/sgbqy>

- Schlueter, Z., Parker, D., & Lau, E. F. (2019). Error-driven Retrieval in Agreement Attraction rarely leads to Misinterpretation. *Frontiers in Psychology*, 10. <https://doi.org/10.3389/fpsyg.2019.01002>
- Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology* (pp. 311–331). New York: Routledge.
- Smith, G., Franck, J., & Tabor, W. (2018). A self-organizing approach to subject–verb number agreement. *Cognitive Science*, 42, 1043–1074.
- Smith, G., Franck, J., & Tabor, W. (2021). Encoding interference effects support self-organized sentence processing. *Cognitive Psychology*, 124, 101356.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.
- Stahl, C. (2006). Multinomiale Verarbeitungs-Baummodelle in der Sozialpsychologie. *Zeitschrift für Sozialpsychologie*, 37(3), 161–171.
- Stan Development Team (2019). *RStan: The R interface to Stan* (R package version 2.19.2). Available at: <http://mc-stan.org/>
- Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language*, 60(2), 308–327.
- Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition*, 114(3), 447–454.
- Tanner, D., Dempsey, J., & Christianson, K. (2018). Does attraction lead to systematic misinterpretation of NP number? Probably not. In *Poster presented at the 31st CUNY conference on human sentence processing*, Davis, CA.
- Tanner, D., Grey, S., & van Hell, J. G. (2017). Dissociating retrieval interference and reanalysis in the P600 during sentence comprehension. *Psychophysiology*, 54(2), 248–259.
- Thornton, R., & MacDonald, M. C. (2003). Plausibility and grammatical agreement. *Journal of Memory and Language*, 48(4), 740–759.
- Tiersma, P. M. (1982). Local and general markedness. *Language*, 58(4), 832–849.
- Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, 35(3), 454–475.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.
- Tucker, M. A., Idrissi, A., & Almeida, D. (2015). Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology*, 6, 347.
- Ulrich, R. (2009). Uncovering unobservable cognitive mechanisms: The contribution of mathematical models. In F. Rösler, C. Ranganath, B. Röder, & R. H. Kluwe (Eds.), *Neuroimaging of human memory: Linking cognitive processes to neural systems* (pp. 25–41). Oxford, England: Oxford University Press.
- Van Maanen, L., & Van Rijn, H. (2007). An accumulator model of semantic interference. *Cognitive Systems Research*, 8(3), 174–181.
- Vasishth, S., Jäger, L. A., & Nicenboim, B. (2018). *Feature overwriting as a finite mixture process: Evidence from comprehension data*. arXiv:1703.04081v2.
- Vasishth, S., Merten, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103, 151–175.
- Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, 23(11), 968–982.
- Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models* (R package version 2.1.0).
- Vehtari, A., Gelman, A., & Gabry, J. (2015). *Pareto smoothed importance sampling*. arXiv:1507.02646v7.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432.
- Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6, 142–228.

- Vigliocco, G., Butterworth, B., & Semenza, C. (1995). Constructing subject–verb agreement in speech: The role of semantic and morphological factors. *Journal of Memory and Language*, 34(2), 186–215.
- Villata, S., & Franck, J. (2020). Similarity-based interference in agreement comprehension and production: Evidence from object agreement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(1), 170–188.
- Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in Psychology*, 9, 2.
- Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2), 206–237.
- Wimsatt, W. C. (1987). False models as means to truer theories. In M. Nitecki & A. Hoffmann (Eds.), *Neutral models in biology* (pp. 23–55). Oxford, England: Oxford University Press.
- Wonnacott, E., Joseph, H. S., Adelman, J. S., & Nation, K. (2016). Is children’s reading “good enough”? Links between online processing and comprehension as children read syntactically ambiguous sentences. *Quarterly Journal of Experimental Psychology*, 69(5), 855–879.
- Yadav, H., Smith, G., & Vasisht, S. (2021). *Feature encoding modulates cue-based retrieval: Modeling interference effects in both grammatical and ungrammatical sentences*. Available at: psyarxiv.com/76aex <https://doi.org/10.31234/osf.io/76aex>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122.