Quantifying the cross-linguistic effects of syncretism on agreement attraction

Utku Turk
University of Maryland, College Park
utkuturk@umd.edu

Eva Neu
University of Massachusetts, Amherst
eneu@umass.edu

Abstract

Agreement attraction errors, in which a verb erroneously agrees with an intervening noun rather than its grammatical head, are amplified by morphological syncretism in some languages (English, German, Russian) but not others (Turkish, Armenian), a cross-linguistic pattern without a principled account. We use surprisal and attention entropy from large language models as processing proxies to investigate this variation across four languages. LLM-derived measures replicate behavioral findings in English and German (syncretism modulates attraction), align with Turkish null results (no modulation), and partially capture Russian patterns. We discuss further directions for better understanding why syncretism affects agreement attraction differently across languages.

1 Introduction

1.1 Agreement attraction and cue-based retrieval

Consider the following sentences in (1):

(1) a. * The key to the cabinet are rusty.
b. * The key to the cabinets are rusty.

Both sentences are ungrammatical because the verb incorrectly bears plural rather than singular marking. Nonetheless, psycholinguistic research has shown that speakers are more likely to produce, and consider grammatical, sentences like (1b), where the initial preamble includes a plural noun (Wagers et al., 2009). These systematic errors are known as agreement attraction: instead of agreeing with the head noun—here, *key*—, the verb erroneously agrees with an intervening attractor, *cabinets*.

One theory that explains agreement attraction effects is the cue-based retrieval account (Wagers et al., 2009; Dillon et al., 2013). Following the ACT-R model of sentence processing, this model assumes that words are stored in a content-addressable memory as a chunk that can be later retrieved using specific cues (Lewis and Vasishth, 2005). For instance, in (1b), the noun *key* might be stored as the chunk [SG, NOM, TP], encoding number, overt case marking and the dominating syntactic position. Similarly, the word *cabinets* might be stored as [PL, NOM, PP].

According to the account put forward by Wagers et al. (2009), readers form a prediction about the form in which the verb should surface based on the chunks they have processed first. If the noun that occupies the syntactic position of an agreement controller [TP] is singular, they predict that the verb surfaces with singular agreement. In cases where participants indeed see a verb with singular marking such as *is*, no disruption in reading or judgment is seen.[1]

On the other hand, if participants encounter a plural verb such as *are* that does not match with the predicted number, a repair process is initiated. In this process, participants use the relevant cues provided by the verb to search for an agreement controller. The verb *are*, which must agree with a plural noun phrase in subject position bearing nominative case, would trigger the search for a [PL, NOM, TP] chunk. In (1b), no noun phrase matches this chunk perfectly. Accordingly, most of the time participants do not find these sentences grammatical. However, occasionally but systematically, participants erroneously retrieve *cabinets*, encoded as [PL, NOM, PP], as the agreement controller due to a partial match with the cues provided by *are*. In contrast, participants are much less likely to erroneously retrieve the word *cabinet* in (1a), since the

[1] Other models of agreement attraction exist and make different predictions on processing grammatical sentences, most notably Marking and Morphing (Eberhard et al., 2005; Hammerly et al., 2019). We do not focus on these here; see Yadav et al. (2023) for an overview and computational comparison. These models are better suited to account for semantic integration (Solomon and Pearlmutter, 2004), as well as collectivity and distributivity effects (Humphreys and Bock, 2005) that are also known to modulate both agreement and agreement attraction.

features shared between the cues of *are* [PL, NOM, TP] and the chunk of *cabinet* [SG, NOM, PP] are even fewer.

1.2 The effect of syncretism

Two grammatical forms are said to be syncretic if they are realized with the same overt morphology despite bearing different syntactic and semantic features. E.g., most noun phrases in English—such as *the dog*—are syncretic between nominative and accusative case marking. An exception are some pronouns which differ in their nominative and accusative forms, as with the first person pronoun *I* (nominative) vs. *me* (accusative).

Under the cue-based retrieval model, it has been argued that in the case of syncretism, a noun phrase is encoded with all possible features that could be realized by its morphology. E.g., *the dog* would be encoded as bearing both NOM and ACC features despite the fact that it only ever bears one of them in the context of a specific sentence. In contrast, *I* would be encoded with only NOM features, and *me* with only ACC features. Given the broader assumptions about cue-based retrieval outlined above, this predicts that an attractor is more likely to interfere with the processing of subject-verb agreement if it is syncretic with the target.

By way of example, such an effect of syncretism on agreement attraction rates can be observed in a study by Hartsuiker et al. (2003) on German. The German plural determiner surfaces as *die* in both the nominative and the accusative case, but as a distinct form, *den*, in the dative. Using the paradigm in (2-3), Hartsuiker et al. showed that participants made more agreement errors while completing sentences with syncretic determiners (*die*).

(2) Die Stellungnahme zu den_{DAT} Demonstrationen
The position on the demonstrations

(3) Die Stellungnahme gegen die_{ACC} Demonstrationen
The position against the demonstrations

Under the cue-based retrieval model, this is as expected: since the determiner *die* in (3) is encoded as both ACC and NOM, *die Demonstrationen* is more likely to be misretrieved than the attractor in (2), *den Demonstrationen*, which is only encoded as DAT.

1.3 Cross-linguistic differences

While an effect of syncretism on agreement attraction can be accounted for from the perspective of cue-based retrieval, the cross-linguistic picture is more complex. Syncretism has been shown to increase agreement attraction in English (Nicol et al., 2016), Slovak (Badecker and Kuminiak, 2007) and Czech (Lacina et al., 2025), besides German as seen above. However, in Turkish (Türk and Logačev, 2024) and Armenian (Avetisyan et al., 2020), syncretism does not appear to affect error rates or reading times, respectively. Moreover, French shows a reverse pattern, with syncretic attractors producing *fewer* errors (Franck et al., 2006). Lastly, in addition to normal syncretism effects like English, Russian also shows interference from pseudoplural attractors, which are singular, but syncretic with a plural form (Slioussar, 2018).

To the best of our knowledge, this cross-linguistic variation in the effect of syncretism has yet to receive a principled account. One tentative explanation, suggested by Dillon and Keshev (2025), is that languages might differ as to how useful morphological cues are for resolving agreement. If speakers rely heavily on morphological marking to track agreement dependencies, error rates will highly depend on whether or not target and attractor are syncretic. On the other hand, if speakers rather rely on other factors such as word order or semantics to establish a relation between the verb and its subject, syncretism should have little effect. While plausible, this explanation still lacks empirical confirmation.

Here, we use LLMs as a cross-linguistically uniform processing proxy to quantify syncretism effects on agreement attraction across four languages (English, German, Russian, Turkish), extracting surprisal at the verb and attention entropy (i.e., how diffusely the model attends to candidate nouns at the verb) over potential agreement controllers. We find that LLM-derived measures replicate behavioral data in English, German, and Turkish, and partially in Russian. The one case where our LLM results are clearly at odds with the behavioral measures—Russian pseudoplurals (Slioussar, 2018)—suggests that humans are more easily led astray by superficial morphological similarity, whereas the LLMs draw more heavily on syntactic information from the broader context of the sentence which clearly disambiguates the morphological marking (see also Bazhukov et al., 2024, for similar partial results).

2 Methods

A core finding in computational psycholinguistics is that the difficulty of processing a word, as measured by reading times, is proportional to its surprisal—the negative log probability of that word given its context (Hale, 2001; Levy, 2008). This relationship has been robustly demonstrated across reading paradigms: surprisal estimates from language models correlate linearly with self-paced reading times and eye-tracking measures such as gaze duration and total reading time (Smith and Levy, 2013; Goodkind and Bicknell, 2018). Critically, surprisal from neural language models like GPT-2 has been shown to predict human reading behavior as well as or better than traditional n-gram models (Wilcox et al., 2020; Merkx and Frank, 2021).

This general correlation between reading times and surprisal values also applies to agreement attraction data. Ryu and Lewis (2021) have argued that LLMs provide a correlate for the dynamics of memory retrieval according to the cue-based retrieval account, previously discussed in Section 1.1 (but see also Arehalli and Linzen, 2020). They showed that agreement attraction effects, as measured experimentally in reading times, can be replicated in LLMs using surprisal values, which quantify how likely a word is to appear in a context (Levy, 2008). Concretely, in an ungrammatical sentence like (1b), **The key to the cabinets are rusty*, lower surprisal at the verb corresponds to a stronger effect of agreement attraction since it indicates the possibility that the model is, like humans, led astray by the plural attractor and *illusioned* into deeming the sentence grammatical. In terms of cue-based retrieval, decreased surprisal corresponds to the facilitation due to a partial match between the agreement bearer *are* and the attractor *cabinets*. This partial match would also predict an increased retrieval of the *cabinets* as the agreement controller compared to the cases with singular attractor (1a).

In addition to surprisal, Ryu and Lewis argue that behavioral agreement attraction data also correlate with attention values in an LLM which quantify the importance of a specific word in the context (Clark et al., 2019). How much attention the model attributes, at the point of the verb, to a given noun indicates how likely it considers this noun to be the subject that is related to the verb by an agreement dependency. Accordingly, the more attention the model attributes to the intervening attractor relative to the true target of agreement, the stronger the agreement attraction effect.

Ryu and Lewis's work has focused on linking attention and surprisal in LLMs specifically to reading time data on agreement attraction. However, in experimental studies, including some that we draw on in the present work, agreement attraction is often quantified using offline behavioral measures like acceptability judgments rather than reading times. Crucially, the memory-retrieval assumptions underlying Ryu and Lewis's framework are not paradigm-specific: cue-based retrieval makes predictions about processing difficulty in general, and retrieval interference effects are routinely observed across both online and offline measures in psycholinguistics (Wagers et al., 2009; Dillon et al., 2013). Some recent work further supports the offline–LLM link directly (Lee et al., 2024; Timkey and Linzen, 2023), and we treat our measures as general-purpose proxies for retrieval competition rather than as direct models of any specific experimental paradigm. We note that some of the behavioral datasets we model come from production paradigms, while LLM-based measures are mainly used on the comprehension side (cf. Harmon and Kapatsinski, 2021). We hope to complement the current findings with self-paced reading and acceptability judgment studies to complete the cross-paradigm picture.

In the following, we describe our methodology for extracting surprisal and attention values more in detail.

Surprisal We estimated surprisal for the agreement-bearing words (auxiliary or main verbs) using GPT-2-style autoregressive language models. Because surprisal is defined as the negative log probability of a word given its preceding context, autoregressive models that generate text left-to-right provide the most theoretically appropriate estimate.

Attention We extracted attention weights from BERT-style bidirectional models, measuring the soft attention from the verb position to each noun that could either correctly (grammatical head) or erroneously (attractor) control agreement. Following methodological insights from probing studies (Voita et al., 2019; Clark et al., 2019), we identified syntactically relevant attention heads rather than aggregating across all heads and layers.

Our approach adapts the entropy metric from Ryu and Lewis (2021), who aggregate attention

entropy across syntactically selected heads spanning all layers. Oh and Schuler (2022) instead restrict computation to the final layer, but Ryu and Lewis note that none of Oh and Schuler (2022) selected syntactic heads fall in the final layer, suggesting that intermediate layers are more informative for grammatical dependencies. We follow this insight by using probing to select the most informative layer rather than defaulting to the final one, while keeping the layer-level aggregation of Oh and Schuler (2022). Concretely, to identify which layer best tracks subject-verb dependencies, we parsed 1M-sentence corpora from the Leipzig Corpora Collection for each language using Universal Dependencies (De Marneffe et al., 2021) and selected the layer where the subject most reliably appeared among the verb's five most-attended tokens: layer 6 for English (63.1% accuracy), layer 9 for German (48.2%), layer 8 for Russian (69.3%), and layer 8 for Turkish (53.0%). Attention entropy was then computed over the mean attention distribution across all heads in that layer on the experimental stimuli.

In what follows, we focus on ungrammatical sentences and ask whether adding a plural attractor changes the measure relative to a singular attractor. We compare sentence pairs that only differ in the number marking of the attractor. Our measure of interest is $\Delta = \mu_{\text{Plural}} - \mu_{\text{Singular}}$, computed separately for syncretic and non-syncretic conditions. A more negative Δ for surprisal means the plural attractor made the ungrammatical verb less surprising — i.e., stronger agreement attraction. A more positive Δ for attention entropy means attention was more dispersed, reflecting greater retrieval competition.

3 Materials and results

For inference, we fit per-language Bayesian mixed-effects models separately for surprisal and attention entropy. Each model includes all main effects and interactions among Syncretism, Grammaticality, and Attractor Number as fixed effects. The random-effects structure accounts for by-item variability. Items were allowed to vary in their intercepts as well as their slopes for Grammaticality, Attractor Number, and their interaction, with these random effects treated as uncorrelated. Fixed effects use treatment coding with reference levels Syncretic (Syncretism), Grammatical (Grammaticality), and Singular (Attractor). Under this coding, the two-way term $\beta_{G\times A}$ indexes the Grammaticality×Attractor interaction in the Syncretic baseline, and the three-way term $\beta_{S\times G\times A}$ indexes how that interaction changes in the Non-syncretic condition (thus, the Non-syncretic interaction is $\beta_{G\times A} + \beta_{S\times G\times A}$). In the main text, we report posterior probabilities $P(\text{effect}>0)$ for (i) the two-way Grammaticality×Attractor interaction (attraction effect) and (ii) the three-way Syncretism×Grammaticality×Attractor interaction (syncretism modulation of attraction). Full model details and priors are reported in the Appendix.

3.1 Experiment 1: German

Materials For German, we modeled the data by Hartsuiker et al. (2003) summarized in Section 1.2. Recall that this experiment found that participants made more agreement errors in production when completing sentences with accusative attractors that are syncretic with the nominative (marked with the determiner *die*) compared to dative, non-syncretic attractors (marked with *den*). We created grammatical (singular) and erroneous (plural) continuations for each experimental condition and modeled these data using `bert-base-german-cased` and `dbmdz/german-gpt2`. We expected to find overall lower surprisal and higher attention to the plural attractor in the ungrammatical completions, compared to singular attractor conditions. More importantly, we further predicted that this difference would be amplified in the syncretic condition (2) compared to the non-syncretic condition (3).

Surprisal (Fig.1) Ungrammatical sentences showed higher surprisal compared to grammatical sentences as expected, and singular attractor conditions in ungrammatical sentences showed higher surprisal compared to plural attractors. The important comparison is whether the difference between plural and singular attractors is larger in syncretic than in non-syncretic items. The differences are very similar: non-syncretic $\Delta = -0.66$ (SE=0.35) and syncretic $\Delta = -0.60$ (SE=0.32), so the syncretism modulation in descriptive terms is small ($\Delta_{\text{syn}} - \Delta_{\text{non}} = 0.06$). However, there seems to be an overall syncretism effects, such that non-syncretic nouns induced more surprisal overall. Posterior probabilities still show a clear attraction interaction in the syncretic baseline ($P(\beta_{G\times A}>0)<0.01$), while the three-way term is

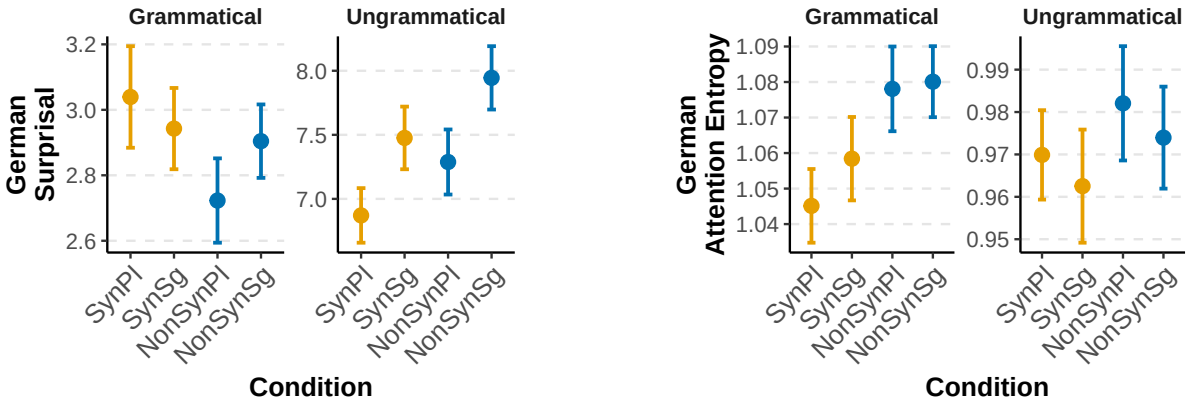

Figure 1: German model-derived surprisal and attention entropy measures (means with SE bars).

positive but uncertain ($P(\beta_{S\times G\times A}>0)$=0.822).

Attention Entropy (Fig.1) Entropy shows the same pattern but much weaker: non-syncretic Δ=0.008 (SE=0.018) and syncretic Δ=0.007 (SE=0.017), with near-zero descriptive modulation (-0.001). Posterior probabilities suggest a positive attraction interaction in the syncretic baseline ($P(\beta_{G\times A}>0)$=0.969), but no reliable three-way interaction ($P(\beta_{S\times G\times A}>0)$=0.250).

Discussion Behaviorally, German is expected to show stronger attraction in syncretic (*die*) than non-syncretic (*den*) conditions. In our current model outputs, both attention and surprisal effects align with attraction predictions, but these values underpredict the syncretic-vs.-non-syncretic contrast.

3.2 Experiment 2: English

Materials For English, we modeled data from Nicol et al. (2016) and Wagers et al. (2009) using pretrained GPT2-small (Radford et al., 2019) and BERT (Devlin et al., 2019). Wagers et al. (2009) found that participants were more likely to produce agreement attraction errors with plural attractors (e.g., *cabinets*) compared to singular attractors (e.g., *cabinet*). However, as noted above, nearly all English nouns are syncretic between nominative and accusative; thus, Wagers et al.'s experiment was limited to attractors that were syncretic between nominative and accusative case marking. Nicol et al. (2016) further investigated whether the syncretism of the attractor is crucial for agreement attraction effects. They compared accusative plural attractors whose case is syncretic with the nominative (e.g., *gardens*) to genitive plural attractors which do not show this syncretism (e.g., *elves'*) and found significantly higher interference rates with syncretic attractors. In their experiment, participants produced fewer agreement attraction errors with preambles like (4c), which contain only a non-syncretic plural attractor (*elves'*), than with (4b) or (4d), which contain a syncretic plural attractor (*gardens*). Moreover, there was no significant difference between (4b) and (4d), indicating that the non-syncretic attractor did not increase interference rates when a syncretic attractor was present.

(4) a. The statue in the elf's garden ...
b. The statue in the elf's gardens ...
c. The statue in the elves' garden ...
d. The statue in the elves' gardens ...

Surprisal (Fig.2) For English, we focus on how much adding plural marking on a possible attractor (either *elf* or *garden*) changes processing in each morphology type: syncretic (*elves' garden* $\rightarrow$ *elves' gardens*) vs. non-syncretic (*elf's gardens* $\rightarrow$ *elves' gardens*). Surprisal was sensitive to grammaticality as expected. Moreover, in ungrammatical items, adding plural marking lowers surprisal much more in syncretic items ($\Delta = -2.009$, SE=0.286) than in non-syncretic items ($\Delta = -0.541$, SE=0.235). This appears as a strongly negative two-way interaction ($P(\beta_{G\times A}>0)<0.01$) plus a strongly positive three-way interaction ($P(\beta_{S\times G\times A}>0)>0.99$); i.e., the attraction pattern is clearly attenuated in non-syncretic items.

Attention Entropy (Fig.2) Entropy shows the same descriptive direction: in ungrammatical items, plural increases entropy in syncretic items (Δ=0.095, SE=0.042), but is near-zero in non-syncretic items (Δ=0.002, SE=0.049). At model level, the syncretic-baseline two-way term is strongly positive ($P(\beta_{G\times A}>0)$=0.978), while the three-way term is mostly negative

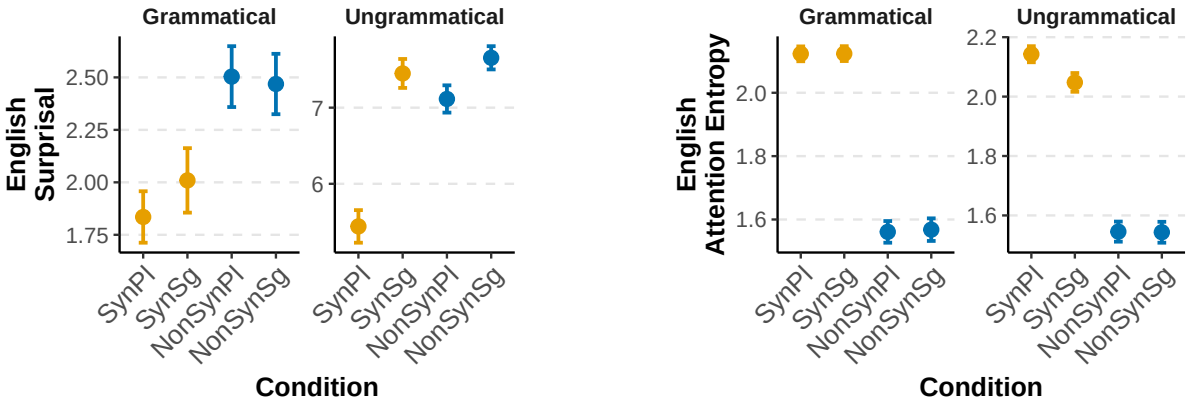

Figure 2: English model-derived surprisal and attention entropy measures (means with SE bars).

($P(\beta_{S\times G\times A}>0)$=0.090), again consistent with attenuation in non-syncretic items.

Discussion Research on English has found higher agreement attraction rates with syncretic (accusative plural) than with non-syncretic attractors (genitive plural). Our results from large language models correlate with these findings. Both surprisal and attention-related values showed a general effect of attraction results, but this effect was also modulated by the presence of syncretic nouns. The important difference in surprisal values increased with non-syncretic nouns, rendering non-syncretic plural meaningless in terms of facilitation effects. Similarly, the non-syncretic nouns also decreased attention entropy, meaning that the directed attention to the head was more concentrated, aligning with increased accuracy.

3.3 Experiment 3: Russian

Materials For Russian, we modeled data from Slioussar (2018) using `deepvk/bert-base-uncased` and `ai-forever/rugpt3small_based_on_gpt2`. In some respect, Russian aligns with German and English in that speakers are more likely to wrongly consider a sentence with a singular head, a plural attractor and a plural verb grammatical if the case marking on the attractor is ambiguous. Slioussar (2018) showed this by comparing accusative attractors such as *polja* ('field.ACC.PL'), which are syncretic with their nominative form (5), to genitive attractors such as *večerinok* ('party.GEN.PL'), which are not ambiguous with nominative marking (6). They found that the syncretic attractors resulted in higher error rates.

(5) Trassa čerez polja/pole byli ...
highway.NOM.SG across field.ACC.PL/SG were ...
'The highway across the fields/field were ...'

(6) Komnata dlja večerinok/večerinki byli ...
room.NOM.SG for party.GEN.PL/SG were
'The room for parties/party were ...'

However, Russian also exhibits another syncretism that affects agreement attraction rates, which differs from the patterns described so far. In (6), the genitive singular form of 'party', *večerinki*, is syncretic with the nominative plural form, whereas the genitive plural form *večerinok* does not exhibit a syncretism with the nominative. This pattern is widespread in Russian nominal paradigms, so any resulting effects cannot be attributed to rarity. Slioussar found that with genitive attractors as in (6), participants produced *more* agreement attraction errors with singular than with plural attractors, contrary to the usual pattern. The effect was robust in both production and comprehension, and visible in reading times. What this suggests is that attractors can be misretrieved upon encountering a plural verb not only if they truly bear plural marking, but also if they are singular but syncretic with a plural form.

Following Slioussar's results, we expect to find a slightly different picture compared to English and German. We expect to see a two-way interaction that signals attraction effects in syncretic cases, and a three-way interaction showing a clear difference between syncretic and non-syncretic cases. However, we also expected to see a reversed attraction effect only in non-syncretic (genitive) cases, corresponding to higher interference rates from genitive singular than genitive plural attractors. To this end, we fitted an additional model only for the non-syncretic cases to see a non-interaction or reversed

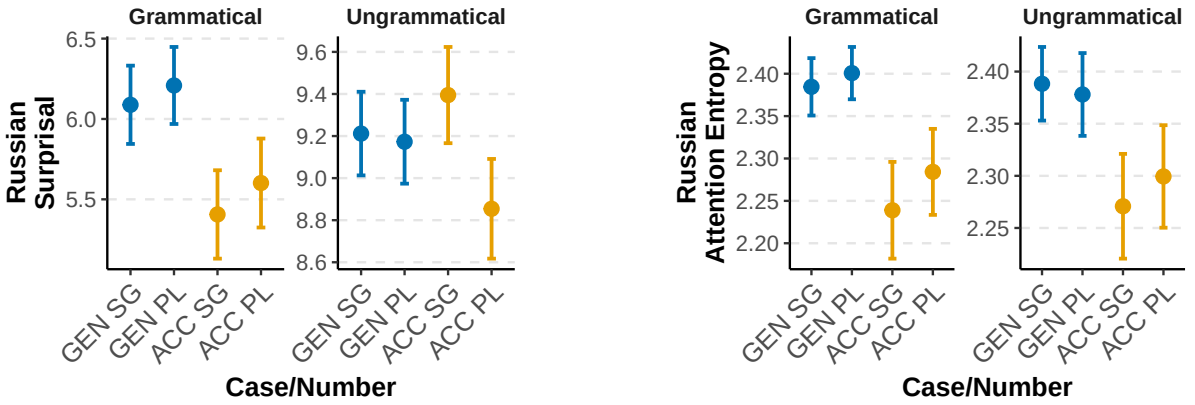

Figure 3: Russian model-derived surprisal and attention entropy measures (means with SE bars).

interaction.

Surprisal (Fig.3) In Russian syncretic (ACC) ungrammatical items, adding a plural attractor lowers surprisal relative to singular ($\Delta_{\text{Pl}-\text{Sg}} = -0.541$, SE=0.329). For this syncretic baseline case, the main model gives a strongly negative two-way interaction ($\beta_{G\times A} = -0.730$, $P(\beta_{G\times A}>0)<0.01$). In non-syncretic (GEN) ungrammatical items, the descriptive contrast is much smaller and close to reversal: $\Delta_{\text{Pl}-\text{Sg}} = -0.039$ (SE=0.282). The main-model three-way interaction is strongly positive ($\beta_{S\times G\times A}=0.568$, $P(\beta_{S\times G\times A}>0)>0.99$), and the non-syncretic-only model gives a negative two-way interaction ($\beta_{G\times A} = -0.159$, $P(\beta_{G\times A}>0)<0.01$), consistent with the reversal pattern, though the magnitude is very small compared to Slioussar's (2018) results.

Attention entropy (Fig.3) In Russian syncretic (ACC) ungrammatical items, adding a plural attractor slightly increases entropy relative to singular ($\Delta_{\text{Pl}-\text{Sg}}=0.029$, SE=0.070). For this syncretic baseline case, the main-model two-way term is weak and not clearly directional ($\beta_{G\times A} = -0.016$, $P(\beta_{G\times A}>0)=0.183$). In non-syncretic (GEN) ungrammatical items, the descriptive effect is reversed and near zero ($\Delta_{\text{Pl}-\text{Sg}} = -0.010$, SE=0.053; equivalently $\Delta_{\text{Sg}-\text{Pl}} = +0.010$). The main-model three-way interaction is also weak ($\beta_{S\times G\times A} = -0.010$, $P(\beta_{S\times G\times A}>0)=0.352$), and the non-syncretic-only two-way term is similarly weak ($\beta_{G\times A} = -0.025$, $P(\beta_{G\times A}>0)=0.138$).

Discussion Our LLM based values show clear attraction effect within syncretic items. In non-syncretic cases, we find that it is heavily attenuated, similar to the English findings. While this overall effect is expected, Russian behavioral data also show a reversed effect within the GEN conditions in which the singular attractor leads to higher interference rates. However, this reversal is not fully supported, with much weaker non-syncretic (GEN) contrasts and weak non-syncretic-only reversed-attraction effects. Entropy remains less decisive overall.

3.4 Experiment 4: Turkish

Materials For Turkish, we modeled data from Türk and Logačev (2024), who compared agreement attraction rates between sentences such as (7) and (8), building on work by Lago et al. (2019). We used `dbmdz/bert-base-turkish-128k-cased` and `redrussianarmy/gpt2-turkish-cased`. The head noun (here, *eğitmen-i* 'instructor' or *hoca-sı* 'teacher') bears a possessor morpheme and is preceded by a genitive-marked possessee (here, *teknisyen-(ler)-in* 'of the technician(s)').

(7) Teknisyen-(ler)-in eğitmen-i … koştu-(lar).
tech-(PL)-GEN instructor-POSS ran-(PL)
'The instructor of the technician(s) … $\text{ran}_{\text{SG/PL}}$.'

(8) Teknisyen-(ler)-in hoca-sı … koştu-(lar).
tech-(PL)-GEN teacher-POSS ran-(PL)
'The teacher of the technician(s) … $\text{ran}_{\text{SG/PL}}$.'

Since genitive case also surfaces on the subjects of embedded clauses in Turkish, speakers might misinterpret the possessee as an embedded subject and, if it is plural, misretrieve it when encountering a plural verb. Crucially, after a consonant as in (7), the possessor morpheme on the head is realized as *-I*, which is syncretic with accusative case. Vowel-final nouns as in (8), on the other hand, take the possessor morpheme *-sI*, which is not syncretic but can only surface on nominative subjects.[2] Türk

[2]Note that in the Turkish materials, the attractor precedes

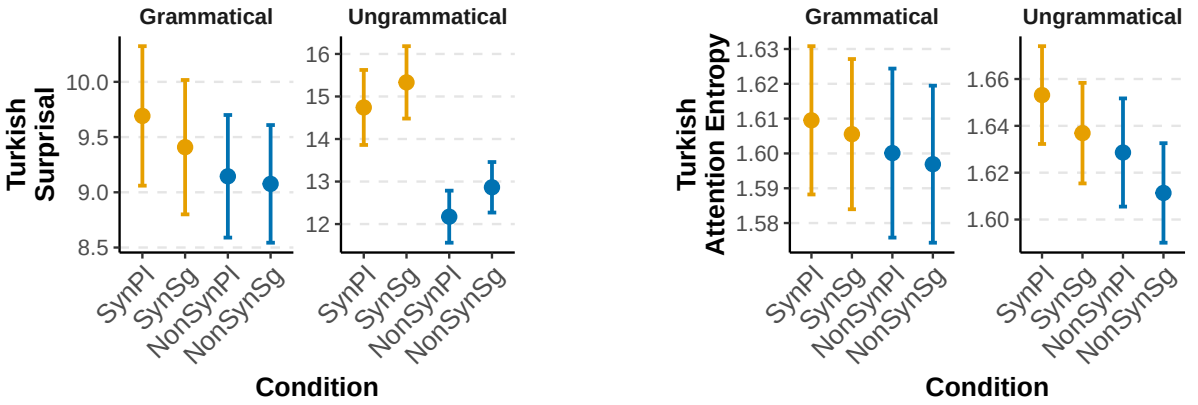

Figure 4: Turkish model-derived surprisal and attention entropy measures (means with SE bars).

and Logačev (2024) expected that the unambiguous nominative case marking on the head as in (8) (*hoca-sı*) would diminish interference rates from the genitive attractor *teknisyen-ler-in*, but this prediction was not borne out. Unlike in the languages seen so far, syncretism does not appear to affect agreement attraction rates in Turkish.

Following their results, we expect that plural attractors overall decrease surprisal and increase attention entropy in ungrammatical sentences. However, this effect should not be modulated by syncretism.

Surprisal (Fig.4) For Turkish, the key test is whether plural-attractor effects differ by syncretism. They do not: adding a plural attractor lowers surprisal in both conditions, with the effects being of very similar size (non-syncretic $\Delta = -0.691$, SE=0.853; syncretic $\Delta = -0.590$, SE=1.230). Bayesian estimates show weak directional evidence for the syncretic-baseline two-way term ($P(\beta_{G\times A}>0)$=0.194), and more importantly, no evidence for a three-way interaction ($P(\beta_{S\times G\times A}>0)$=0.332).

Attention entropy (Fig.4) Entropy shows the same qualitative conclusion. The plural-attractor increase is almost identical in the two conditions (non-syncretic Δ=0.017, SE=0.031; syncretic Δ=0.016, SE=0.030). However, posterior probabilities show no directional evidence for both interaction terms ($P(\beta_{G\times A}>0)$=0.782, $P(\beta_{S\times G\times A}>0)$=0.550).

Discussion While our surprisal results align with the behavioral Turkish data—namely, less surprisal in agreement attraction cases—the attention values are inconclusive. As for the syncretism, we show that it does not modulate attraction related facilitation within surprisal, and plural-attractor effects are comparable in syncretic and non-syncretic conditions. Given that there is a general sensitivity in LLM measures to attraction effects, the lack of syncretism modulation is consistent with the behavioral data.

4 General discussion

This study aimed to establish a baseline for future work on syncretism and morphological encoding in memory retrieval by identifying correlates between human behavioral data and LLM-based measures, following recent methodological advances (Wilcox et al., 2020; Ryu and Lewis, 2021). Across three of the four languages examined, model-derived measures matched the effects attested in the behavioral literature. Surprisal provided a clearer and more consistent signal, while attention entropy was less decisive: the expected patterns were present but weaker and less certain. This divergence is not unexpected, given that attention-based measures are more indirect and less established than surprisal as proxies for retrieval competition, but it points to room for improvement in how subject-tracking attention is extracted and interpreted.

The one clear exception is Russian. Surprisal successfully captured the standard attraction effect in the accusative (syncretic) condition, but failed to

the head noun, whereas in the English, German, and Russian materials, the attractor intervenes between the head and the verb. At least for Russian and German, and possibly Turkish, alternative orders can be constructed (e.g., via topicalization or focus movement) without changing the structure drastically. Sturt and Kwon (2024) found that distractor position modulated the size of attraction effects in English; however, whether similar positional effects hold across languages and whether LLMs mirror human behavior under these rarer configurations remains an open question.

reproduce the reversed pattern in the genitive condition, where behavioral data show that a genitive singular attractor—syncretic with the nominative plural—generates more interference than a genuine genitive plural. One possible explanation is that LLMs are less sensitive than human comprehenders to spurious morphological overlaps between two forms that share no underlying feature ([GEN, SG] and [NOM, PL]) and/or rely more heavily on syntactic information to disambiguate these forms. Resolving this question requires targeted probing of how models represent syncretism with and without feature overlap, which we leave for future work.

More broadly, we hope that in the long term, this study contributes to understanding of *why* syncretism modulates agreement attraction differently across languages. Our results suggest that LLMs are sensitive to the same syncretisms that drive attraction in human experiments, even if through different underlying mechanisms. A promising direction for future work is to ask whether languages differ systematically in how much weight they place on morphological cues when tracking dependencies. Syncretism effects may be strongest in languages where morphological distinctions are few but highly informative (English, German) or where case encodes a dense bundle of features (fusional languages like Russian and Czech). In agglutinative languages like Turkish and Armenian, where morphological information is more transparent and compositional, speakers and models alike may rely less on the presence or absence of a specific syncretic overlap, leading to the null effects observed here.

References

Suhas Arehalli and Tal Linzen. 2020. Neural language models capture some, but not all, agreement attraction effects. In Proceedings of the Annual Meeting of the Cognitive Science Society.

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

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

Maxim Bazhukov, Ekaterina Voloshina, Sergey Pletenev, Arseny Anisimov, Oleg Serikov, and Svetlana Toldova. 2024. Of models and men: Probing neural networks for agreement attraction with psycholinguistic data. In Proceedings of the 28th Conference on Computational Natural Language Learning, pages 280–290.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does BERT look at? An analysis of BERT's attention. arXiv Prepr. arXiv:1906,04341.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. Computational Linguistics, 47:255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.4805.

Brian Dillon and Maayan Keshev. 2025. Syntactic dependency formation in sentence processing: A comparative perspective. In Sjef Barbiers, Norbert Corver, and Maria Polinsky, editors, Cambridge Handbook of Comparative Syntax. Cambridge UP.

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

Kathleen M Eberhard, J Cooper Cutting, and Kathryn Bock. 2005. Making syntax of sense: number agreement in sentence production. Psychological Review, 112:531.

Julie Franck, Glenda Lassi, Ulrich H Frauenfelder, and Luigi Rizzi. 2006. Agreement and movement: A syntactic analysis of attraction. Cognition, 101:173–216.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics, pages 10–18.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. Second meeting of the North American Chapter of the Association for Computational Linguistics.

Christopher Hammerly, Adrian Staub, and Brian Dillon. 2019. The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. Cognitive Psychology, 110:70–104.

Zara Harmon and Vsevolod Kapatsinski. 2021. A theory of repetition and retrieval in language production. Psychological Review, 128:1112.

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

Karin R Humphreys and Kathryn Bock. 2005. Notional number agreement in English. Psychonomic Bulletin & Review, 12:689–695.

Radim Lacina, Anna Laurinavichyute, and Jan Chromỳ. 2025. Only case-syncretic nouns attract: Czech and Slovak gender agreement. Journal of Memory and Language, 143:104623.

Sol Lago, Martina Gračanin–Yuksek, Duygu Fatma Şafak, Orhan Demir, Bilal Kırkıcı, and Claudia Felser. 2019. Straight from the horse's mouth: Agreement attraction effects with Turkish possessors. Linguistic Approaches to Bilingualism, 9:398–426.

Eun-Kyoung Rosa Lee, Sathvik Nair, and Naomi Feldman. 2024. A psycholinguistic evaluation of language models' sensitivity to argument roles. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3262–3274.

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

Richard L. Lewis and Shravan Vasishth. 2005. An activation–based model of sentence processing as skilled memory retrieval. Cognitive Science, 29:375–419.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics.

Janet Nicol, Andrew Barss, and Jason E Barker. 2016. Minimal interference from possessor phrases in the production of subject-verb agreement. Frontiers in Psychology, 7:548.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9324–9334.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and Others. 2019. Language models are unsupervised multitask learners. OpenAI Blog, 1(8):9.

Soo Hyun Ryu and Richard L Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. arXiv Prepr. arXiv:2104,12874.

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

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. Cognition, 128:302–319.

Eric S Solomon and Neal J Pearlmutter. 2004. Semantic integration and syntactic planning in language production. Cognitive Psychology, 49:1–46.

Patrick Sturt and Nayoung Kwon. 2024. Agreement attraction in comprehension: do active dependencies and distractor position play a role? Language, Cognition and Neuroscience, 39:279–301.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 8705–8720.

Utku Türk and Pavel Logačev. 2024. Agreement attraction in Turkish: The case of genitive attractors. Language, Cognition and Neuroscience, 39:448–454.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5797–5808.

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

Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In Proceedings of the Annual Meeting of the Cognitive Science Society.

Himanshu Yadav, Garrett Smith, Sebastian Reich, and Shravan Vasishth. 2023. Number feature distortion modulates cue-based retrieval in reading. Journal of Memory and Language, 129:104400.

A Model Specification and Full Results

All Bayesian models were fit with the `brms` package in R (Gaussian family). Priors: $\text{Normal}(0, 2)$ on all fixed effects, $\text{Student-}t(3, 0, 2.5)$ on the intercept, $\text{Exponential}(1)$ on random-effect SDs and residual σ. Sampling: 4 chains $\times$ 12,000 iterations (2000 warmup), `adapt_delta`$=0.95$, `max_treedepth`$=12$, seed 1234. The main model formula was:

```
value ~ Syn * Gram * Attr + (1 + Gram *
            Attr || Item)
```

Treatment coding: Syncretic, Grammatical, and Singular as reference levels, so the intercept represents the expected value for grammatical, syncretic, singular-attractor items. All Russian fits were restricted to singular-head items. Code and data for our paper can be found at `www.github.com/utkuturk/scil_TurkNeu2026`.

Table 1: Full posterior summaries for the **Surprisal** models. Terms are shown as rows; languages as column groups. $\hat{\beta}$ is the posterior mean and P is $P(\text{effect}>0)$. S = Syncretism (Non-syncretic vs. Syncretic), G = Grammaticality (Ungrammatical vs. Grammatical), A = Attractor (Plural vs. Singular). **Bold**: $P>0.89$ or $P<0.11$.

	English		German		Russian		Turkish	
Term	$\hat{\beta}$	P	$\hat{\beta}$	P	$\hat{\beta}$	P	$\hat{\beta}$	P
Intercept	2.091	$>$ **0.999**	2.944	$>$ **0.999**	5.411	$>$ **0.999**	9.793	$>$ **0.999**
S	0.351	**0.972**	−0.044	0.339	0.611	**0.953**	−0.795	**0.015**
G	5.358	$>$ **0.999**	4.460	$>$ **0.999**	3.917	$>$ **0.999**	5.481	$>$ **0.999**
A	−0.220	**0.092**	0.085	0.718	0.191	$>$ **0.999**	0.099	0.596
$S\times G$	−0.160	0.244	0.512	$>$ **0.999**	−0.803	**0.009**	−1.739	$<$ **0.001**
$S\times A$	0.270	0.885	−0.272	**0.053**	−0.071	0.207	−0.022	0.482
$G\times A$	−1.731	$<$ **0.001**	−0.690	$<$ **0.001**	−0.730	$<$ **0.001**	−0.488	0.194
$S\times G\times A$	1.131	$>$ **0.999**	0.218	0.822	0.568	$>$ **0.999**	−0.307	0.332

Table 2: Full posterior summaries for the **Attention Entropy** models. Terms are shown as rows; languages as column groups. $\hat{\beta}$ is the posterior mean and P is proportion of posterior samples that are bigger than 0 $P(\text{effect}>0)$. S = Syncretism (Non-syncretic vs. Syncretic), G = Grammaticality (Ungrammatical vs. Grammatical), A = Attractor (Plural vs. Singular). **Bold**: $P>0.89$ or $P<0.11$

	English		German		Russian		Turkish	
Term	$\hat{\beta}$	P	$\hat{\beta}$	P	$\hat{\beta}$	P	$\hat{\beta}$	P
Intercept	2.106	$>$ **0.999**	1.059	$>$ **0.999**	2.237	$>$ **0.999**	1.597	$>$ **0.999**
S	−0.537	$<$ **0.001**	0.022	**0.994**	0.147	**0.987**	−0.006	0.281
G	−0.076	**0.012**	−0.096	$<$ **0.001**	0.032	**0.956**	0.031	$>$ **0.999**
A	0.000	0.513	−0.013	**0.074**	0.045	**0.988**	0.004	0.643
$S\times G$	0.049	0.867	−0.011	0.163	−0.027	0.166	−0.018	0.120
$S\times A$	−0.008	0.411	0.011	0.831	−0.029	0.150	−0.001	0.471
$G\times A$	0.093	**0.978**	0.021	**0.969**	−0.016	0.183	0.012	0.782
$S\times G\times A$	−0.083	**0.090**	−0.010	0.250	−0.010	0.352	0.002	0.550

Table 3: Russian non-syncretic-only models (GEN condition only). No Syncretism factor; intercept is the mean for grammatical, singular-attractor items. $\hat{\beta}$ is the posterior mean and P is proportion of posterior samples that are bigger than 0 $P(\text{effect}>0)$. G = Grammaticality (Ungrammatical vs. Grammatical), A = Attractor (Plural vs. Singular). **Bold**: $P>0.89$ or $P<0.11$

	Surprisal		Entropy	
Term	$\hat{\beta}$	P	$\hat{\beta}$	P
Intercept	6.054	$>$ **0.999**	2.380	$>$ **0.999**
G	3.119	$>$ **0.999**	0.004	0.613
A	0.117	**0.976**	0.016	0.821
$G\times A$	−0.159	$<$ **0.001**	−0.025	0.138