

Identifying the Factors Underlying Differential Object Marking with Language Models: Analysis and Comparison of Self-Paced Reading Data and Transformer-Based Surprisal Scores

Giorgio Sciamanna^{1,2}

¹*Class of Literature and Philosophy, Scuola Normale Superiore, Piazza dei Cavalieri 7, Pisa, Italy*

²*Department of Philology, Literature and Linguistics, University of Pisa, Piazza Evangelista Torricelli 2, Pisa, Italy*

Abstract

Spanish Differential Object Marking (DOM) makes the preposition *a* obligatory with prominent (human, definite) direct objects and excluded with prototypical inanimate ones, with graded modulation by verbal affectedness. We ask whether violating DOM carries an online processing cost, and whether the surprisal of autoregressive language models tracks the same factors that govern human behaviour. We collected self-paced reading (SPR) times and acceptability judgements from 17 L1 Spanish speakers on a balanced set of 50 transitive sentences, each in six variants crossing animacy, definiteness and DOM marking, with affectedness manipulated between items. In parallel we extracted object surprisal from a range of autoregressive models of increasing complexity: a GRU and a Transformer trained from scratch, a monolingual Spanish GPT-2, and three multilingual 8B-scale LLMs (LLaMa 3.1, Qwen3, a Spanish fine-tuned LLaMa). DOM violations are reliably penalised in acceptability and slow reading on the object and downstream, and the three semantic factors interact rather than acting as a single categorical trigger. All models assign higher surprisal to violations but reproduce the human factor structure only partially; the monolingual Spanish GPT-2 is the most effective predictor of acceptability and of the spillover region, and the one whose surprisal is least driven by raw lexical frequency.

Keywords

Differential Object Marking, Spanish, surprisal, self-paced reading, large language models, psycholinguistics

1. Introduction

Spanish marks a subset of its direct objects with the preposition *a*: the object of *Juan escuchó a la vecina* ('Juan heard the neighbour') is marked, whereas that of *Juan escuchó la canción* ('Juan heard the song') is not. This Differential Object Marking (DOM) is, to a first approximation, obligatory with human objects and barred with prototypical inanimate ones, but the boundary is governed by a cluster of interacting semantic and pragmatic properties — animacy, definiteness, specificity and topicality — rather than by any single categorical feature [1, 2, 3]. A more recent strand of work has added a verbal dimension: the affectedness of the event modulates marking even when nominal properties are held constant [4, 5].

The empirical anchor for our research question is Heredero and García [6], who show that verbal affectedness gradiently modulates DOM among human indefinite objects, and whose verb classification we adopt (we return to their study in Section 2). Crucially, the existing evidence on the verbal conditioning of DOM is almost entirely offline; online processing studies of Spanish DOM have targeted animacy [7] but not affectedness, and have rarely crossed the relevant factors orthogonally. At the same time, the rapidly growing literature that uses language-model surprisal [8, 9] as a cognitive predictor of reading times [10, 11, 12] has concentrated on broad-coverage corpora and syntactic ambiguity, and has essentially never been brought to bear on a marked morphosyntactic phenomenon such as Spanish DOM.

This leaves a clear gap. No study has combined (i) an online self-paced reading experiment on Spanish DOM, designed as minimal pairs that cross animacy, definiteness and affectedness with the

Twelfth Italian Conference on Computational Linguistics (CLiC-it 2026), September 14–16, 2026, Palermo, Italy

✉ giorgio.sciamanna@sns.it (G. Sciamanna)

🌐 <https://github.com/giorgiosciamanna> (G. Sciamanna)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

presence or absence of marking, and (ii) a systematic comparison of surprisal from several language models — including models trained from scratch and a model trained only on Spanish — as predictors of the human signal. We address this gap directly. Our contribution is twofold. First, we quantify the online cost of DOM violations and the way the three semantic factors jointly modulate it. Second, we ask whether autoregressive models “understand” DOM in the operational sense that their surprisal (a) increases for violations, (b) tracks the same factors as humans, and (c) improves the prediction of human reading times and judgements beyond surface controls. The second question speaks to whether models could, in principle, stand in for human participants when probing the grammatical determinants of DOM.

Section 2 situates the study; Section 3 describes the materials, experiment, models and analysis; Section 4 and Section 5 report and interpret the results¹; Section 6 concludes.

2. Background and Related Work

2.1. DOM and its Semantic Determinants

DOM is a cross-linguistically widespread phenomenon, systematised by [1, 13], in which only a subclass of direct objects receives overt marking. A robust typological finding is that this distribution is not arbitrary but reflects scales of semantic and pragmatic prominence: the more an object resembles a prototypical subject in animacy, definiteness, specificity or topicality, the more likely it is to be marked [14, 15, 2, 16]. Aissen [2] models this as the resolution of two functional pressures, iconicity (mark prominent, atypical objects) and economy (avoid marking where it is uninformative).

For the two nominal factors we manipulate, the relevant notions are well established. Animacy is not a binary opposition but a gradient running from humans through other animates to inanimates, with several intermediate steps, and it pervasively conditions argument marking across languages [17, 14, 18]. Definiteness and specificity, for their part, are encoded in different ways from language to language — through articles, demonstratives, or bound nominal affixes (as in the Turkish specific-accusative suffix) — and govern how a referent is anchored in the discourse [19, 20, 21]; theoretically they are analysed in terms of uniqueness [22] and familiarity [23, 24], with specificity cutting across the definite/indefinite divide [25]. In Spanish, the canonical generalisation is that DOM is obligatory with human objects and excluded with prototypical inanimates, while the intermediate space is regulated by definiteness and specificity [26, 3, 27, 28].

In typological terms, the objects that attract marking are precisely those that are least prototypical as objects — animate, definite, specific or topical, and therefore closer to the profile of a prototypical subject. In the sense of linguistic markedness, they are the marked members of the object category, and many languages, Spanish among them, give them overt morphological marking [2, 15]. A diachronic perspective sharpens this. Productive DOM in Romance survives mainly at the lateral edges of the area (Spanish, Romanian) and in more conservative forms elsewhere; together with evidence from historical corpora, this suggests that in early Romance the marking was tied above all to the topicality of the object — a non-canonical, topical object within a predominantly topic-first information structure — rather than to stable semantic features such as animacy or definiteness. Only with the shift from topic-prominent to subject-prominent syntax did DOM grammaticalise around nominal prominence, leaving clitic doubling as a structural trace of its discourse origin [28, 29, 30, 31].

¹A detailed record of the analyses reported in this paper is available in the GitHub repository (see Appendix D), which includes an HTML file documenting the complete analytical workflow. The repository contains the code used for all analyses, together with a wide range of figures illustrating the results. These materials provide comparisons of human ratings across different grammatical features, as well as the outputs of the model-based raters evaluated through the methodologies described in this paper. In addition, at the end of the HTML file, reviewers can find a summary figure comparing aggregated human and model ratings across clusters of grammatical conditions. This visualization is intended solely as an illustrative aid to facilitate interpretation of the relationship between changes in language-model surprisal and corresponding changes in human judgments. Since differences between cluster-level averages have not been formally demonstrated to be statistically significant, the figure should not be interpreted as evidence of significant differences across clusters.

A more recent line shows that verbal properties also matter, in particular affectedness, the degree of persistent change an event imposes on a participant: within the transitivity framework of Hopper and Thompson [32] it is one parameter of semantic transitivity [33], and von Heusinger and Kaiser [4] and García García [5] link it to Spanish DOM. The direct precedent for our research question is Heredero and García [6]: holding the nominal frame constant (human, indefinite objects), they classify verbs as $[\pm\text{affected}]$ with an entailment-based diagnostic and show that $[\text{+affected}]$ verbs increase the choice of the *a*-marked variant. Crucially the effect is gradient — even among $[\text{+affected}]$ verbs a non-trivial share of unmarked responses persists — consistent with verbal factors acting as modulators where nominal prominence does not fully fix marking. We adopt their verb classification and extend the question from offline choice to online processing.

Experimental work on Spanish DOM has largely targeted acquisition and bilingualism with offline measures [34, 35, 36, 37]. The closest online study is Hopp and León Arriaga [7], who find with eye-tracking that readers are more sensitive to over-marking of inanimates than to omission with animates, but whose materials vary essentially along animacy alone.

2.2. Surprisal and Reading Times

The notion of surprisal originates in Shannon’s information theory, where the information carried by an event is its negative log probability, here $-\log p(w|c)$ for a word w in context c [38]. Hale [8] was the first to bring this measure into psycholinguistics, proposing it as an index of incremental processing difficulty, and Levy [9] developed it into expectation-based comprehension theory. The theory predicts a (roughly) linear relation between surprisal and reading time, repeatedly confirmed empirically [10, 12], with the caveat that low-level confounds such as length and frequency carry much of the variance and must be controlled [39]. The granularity at which surprisal is computed (subword vs. word) appears to matter little for aggregate fit [40, 41].

Surprisal is not the only proposed source of processing cost. Memory-based accounts such as the Dependency Locality Theory [42] attribute difficulty to the integration and maintenance of long-distance dependencies, and eye-tracking corpus studies have found that both expectation and memory contribute, sometimes independently [43]. The exact functional form of the surprisal–reading-time link is likewise debated: while the relationship is robustly positive and close to linear over the bulk of the range [10], deviations have been reported at the extremes. For our purposes the relevant point is narrower: a DOM violation creates a locally low-probability continuation, so surprisal theory makes a clear directional prediction (higher surprisal, longer reading times at and after the marked region), which we can test directly against human data.

2.3. Self-Paced Reading: Surprisal on the Human Side

On the human side, this expectation-based difficulty is measured online through reading time. Self-paced reading (SPR) records the time a reader spends on each segment as they reveal a sentence word by word, providing an implicit, incremental index of processing load [44, 45, 46, 47]: longer reading times on and just after a word index greater difficulty, and are exactly what surprisal theory predicts for a low-probability continuation. Modern practice uses a non-cumulative moving window, comprehension checks, and mixed-effects analysis with maximal random structure [48] on log-transformed times. SPR also localises where an anomaly disrupts processing and detects spillover into following regions [49]. Because reading proceeds strictly left to right, SPR is the natural human counterpart to an incremental, left-to-right language model: the reader and the model both commit to each word using only the preceding context, which is why we estimate model surprisal from autoregressive models.

2.4. Language Models as Cognitive Predictors

The probabilistic model needed to compute surprisal has evolved from *n*-grams to recurrent networks [50, 51, 52] and Transformers [53], the backbone of modern decoder-only LLMs [54, 55, 56, 57, 58]. A central and partly counter-intuitive finding of recent work is that larger, lower-perplexity models do not

necessarily predict human reading times better [59, 60, 61, 62, 63], and that surprisal may not explain the magnitude of specific syntactic effects [64]. This line of work has focused on broad-coverage corpora and ambiguity resolution, leaving open how model surprisal behaves around a marked, semantically conditioned phenomenon like DOM.

This “bigger is not better” pattern motivates our inclusion of models that span several orders of magnitude in scale and training data. At one extreme we place large multilingual decoder-only models in the 8B-parameter range; at the other, an n-gram model and small recurrent and attention-based networks trained from scratch on a modest Spanish corpus. Comparing them lets us separate two things that are easily conflated: whether a model assigns the right direction of surprisal to a DOM violation (a fairly low bar that even simple models may clear), and whether it predicts the graded, factor-specific human signal (a much higher bar). It also lets us ask whether monolingual specialisation on the target language can compensate for, or even outperform, sheer scale — a question of practical interest for low-resource and typologically specific phenomena, where large multilingual models are often the only off-the-shelf option.

3. Method

3.1. Materials

The item set comprises 50 transitive sentences, each realised in six variants (the full list is given in Appendix C). Every sentence has a fixed constituent structure — proper-name subject, past-tense verb, optional DOM marker *a*, (in)definite article + object, a locative PP (PP1) and a temporal PP (PP2) — giving a constant length of ten words, eleven when *a* is present (e.g. *Juan escuchó (a) la vecina en el balcón durante una noche*). To keep length constant we used only feminine singular objects, avoiding the masculine fusion *a + el = al*. Materials were adapted from Heredero and García [6] for the affectedness contrast and from Hopp and León Arriaga [7] for the constant-length frame.

The six variants cross three factors. Animacy: the object is human or inanimate. Definiteness: human objects are definite (*la*) or indefinite (*una*), while inanimates are always definite. Marking: the DOM preposition *a* is present or absent. We coded two categorical variables: condition (an_def, an_ind, inan_def) and correct. A sentence is correct when marking follows the prescriptive norm (present with animate objects, absent with inanimate ones) and incorrect when it omits obligatory marking or adds prohibited marking; this lets us measure the cost of a violation directly. Orthogonally, the 25 items with [+affected] verbs and the 25 with [−affected] verbs were classified a priori following the entailment diagnostic of Heredero and García [6]. This a priori split was not simply assumed: it was subsequently confirmed against the participants’ own intuitions through a dedicated affectedness test (Section 4). Because the design is built to isolate semantic contrasts, the items were not intended to be maximally natural or realistic; each item serves only as a frame within which a single feature is varied, and is compared only against its own five variants, never against other items. A further 55 fillers (one third ditransitive, one third with non-DOM uses of *a*, one third heterogeneous; a third containing errors) masked the purpose of the study.

3.2. Participants and Procedure

Seventeen L1 speakers of Spanish (aged 19–28, all Radboud University students or affiliates, no reading disorders) took part; the study was approved by the Faculty of Arts ethics committee. A post-session sociolinguistic survey classified their background (origin) as standard European Spanish (7), Latin-American (3) or Catalan–Spanish bilingual (7); no participant other than the Catalan speakers spoke a further DOM language. The experiment ran in a sound-attenuated booth in OpenSesame. In Part 1 (SPR), items and fillers were presented once each, in random order, one randomly selected variant per item per participant, in a non-cumulative moving-window format; participants advanced with a button box and answered true/false comprehension questions after roughly one sentence in three. In Part 2 (acceptability), the 50 items were rated on a 1–5 scale, with conditions re-randomised so that no item

recurred in its Part 1 condition. Finally, 20 items (10 [+affected], 10 [−affected]) were re-presented in the rephrased format of the Heredero and García [6] affectedness diagnostic and rated on the same scale, to validate the a priori verb classification against participants’ own perception.

3.3. Language Models and Surprisal

We extracted the object surprisal, $-\log p(\text{object}|\text{preceding context})$, summed over subword tokens, from seven autoregressive models of increasing complexity. Autoregressive (left-to-right) models were chosen because, in incremental reading, humans too condition only on prior material, so bidirectional models are not comparable. The four main models are decoder-only Transformers: a monolingual Spanish GPT-2 trained from scratch on the Spanish portion of OSCAR-2019 [65], and three 8B-scale multilingual LLMs — LLaMa 3.1 8B [57], Qwen3 8B [58], and a Spanish fine-tuned LLaMa. Three baselines of increasing complexity probe whether any effect is trivially due to using “some” probabilistic model: a Kneser–Ney n-gram (from which only a whole-sentence probability is available, not an object-specific score), and a GRU and Transformer trained from scratch on $\approx 2\text{M}$ tokens of OSCAR restricted to the 100k most frequent alphabetic word types, following Merx and Frank [66]; the GRU architecture follows Aurnhammer and Frank [67] and the Transformer Vaswani et al. [53]. Lexical frequency, used as a control and in a post-hoc analysis, was computed as the negative log relative frequency of the inflected verb and object forms in OSCAR.

The from-scratch baselines were deliberately small and trained on a controlled vocabulary so that their behaviour could be interpreted transparently. The corpus was the Spanish portion of OSCAR-2019, filtered to sentences whose tokens all fall within the 100,000 most frequent alphabetic word types; both networks were trained for next-word prediction on roughly two million tokens, following the training procedure of Merx and Frank [66]. The GRU had a 400-dimensional embedding layer, two recurrent layers of 500 units, a linear layer of 400 units with a tanh non-linearity, and a log-softmax output [67]; the Transformer used a 400-dimensional embedding, two encoder layers with eight attention heads and a 1024-unit feed-forward block, and the same log-softmax output [53]. The n-gram model used Kneser–Ney smoothing; because of its sentence-level scoring it contributes a whole-sentence probability rather than an object-specific surprisal, and is therefore expected to be the least diagnostic of the local manipulation. For the four pre-trained decoder-only models we queried the released weights directly and aggregated subword surprisal to the word level by summation, which preserves a well-founded probabilistic interpretation while aligning the model output (tokens) with the behavioural measure (words) [10, 40].

3.4. Statistical Analysis

Per-word reading times were trimmed trial by trial of implausibly fast or slow responses and the remainder log-transformed. To capture spillover we defined three target regions (object, PP1, PP2) and fitted a linear mixed model ($1\text{me}4$) per region on log RT. Fixed effects comprised surface controls (region length, presentation position `live_row`, verb and object frequency) and experimental effects (`correct` and its interactions with condition, affected, origin); only `correct` entered on its own, the semantic and sociolinguistic factors entering only in interaction with it, on the assumption that they modulate the cost of a violation rather than overall reading speed. By-subject random intercepts and slopes (for `correct`, `live_row`) were included; by-item random effects were dropped because they prevented convergence. From the maximal model we ran stepwise selection with `buildmer` [68, 48]. Acceptability was analysed analogously with cumulative link mixed models (`ordinal::clmm`); the affectedness validation used a CLMM with `affected` and `live_row`. We report the fixed effects retained after selection ($p < .05$) and, via `emmeans`, Bonferroni-corrected contrasts of correct vs. incorrect sentences, treating a contrast as significant when its 95% CI excludes zero. The from-scratch baselines were set to missing on the two out-of-vocabulary items (17, 22); other analyses use the full set. Model surprisal was evaluated with three tests: coherence (does surprisal vary with the conditions?), correlation (does adding a model’s surprisal to a surface-only baseline improve the fit to human data, as log-likelihood

gain?), and robustness (do the experimental predictors still add explanatory power once surprisal is included?). Before being applied to the real data, all three tests were first validated on simulated datasets (see the R analysis), to confirm that they recover known effects and reject their absence.

4. Results

We first report the human data (acceptability and reading times), then the model surprisal analyses. Throughout, fixed-effect significance is read from Wald p -values; for emmeans contrasts the analysis reports 95% confidence intervals but not p -values, so we judge significance by whether the interval excludes zero and state this explicitly. All 17 participants passed the comprehension check (minimum 91.5% correct), so none were excluded. The a priori affectedness classification was validated: in the rephrasing test [+affected] verbs were rated more acceptable than [−affected] ones (CLMM $\beta = 1.60$, $SE = 0.67$, $z = 2.39$, $p = .017$; EMM difference +1.6, 95% CI [0.28, 2.91]).

4.1. Acceptability Judgements

DOM violations are strongly penalised. In the CLMM (Table 1), prescriptively correct sentences are rated far higher than their violations ($\beta_{correct1} = 2.03$, $SE = 0.35$, $z = 5.83$, $p < .001$); the overall estimated marginal difference between incorrect and correct sentences is -1.55 (95% CI $[-1.86, -1.24]$). The penalty holds in every animacy/definiteness condition – animate-definite -1.11 $[-1.61, -0.62]$, animate-indefinite -1.70 $[-2.23, -1.18]$, inanimate-definite -1.84 $[-2.33, -1.35]$ – and although the violation cost is numerically smallest for animate-definite objects, the differences between conditions are not reliable (the difference-of-differences CIs include zero). Correctly marked animate-indefinite objects are nonetheless especially well accepted ($correct1:an_ind \beta = 0.98$, $p < .001$), the configuration where marking is variable but normatively expected. Affectedness interacts with correctness ($correct1:affected1 \beta = -0.72$, $p = .009$); by affectedness the violation penalty is -1.69 $[-2.11, -1.26]$ for [−affected] and -1.42 $[-1.83, -1.01]$ for [+affected] objects, a difference that is itself not significant ($[-0.29, 0.82]$). Speaker background matters most clearly: the penalty is largest for standard speakers (-2.34 $[-2.81, -1.86]$) and significantly smaller for Catalan (-0.96 $[-1.39, -0.54]$; difference vs. standard +1.37 [0.66, 2.08]) and Latin-American speakers (-1.36 $[-2.01, -0.71]$; difference +0.98 [0.07, 1.89]). Both lexical-frequency controls are positive and significant (object $p = .033$; verb $p = .009$): more frequent objects and verbs are rated slightly higher overall. Full marginal-means tables are in Appendix A.

Table 1

Offline acceptability: selected CLMM fixed effects (reference levels `correct0`, `an_def`, `origin=standard`, [+affected]). Threshold cut-points omitted. Full table in Appendix A.

Term	β	SE	z	p
<code>correct1</code>	2.031	0.348	5.834	< .001 ***
<code>obj_freq</code>	0.105	0.049	2.137	.033 *
<code>verb_freq</code>	0.184	0.070	2.630	.009 **
<code>correct1:an_ind</code>	0.977	0.264	3.700	< .001 ***
<code>correct0:catalan</code>	1.024	0.443	2.312	.021 *
<code>correct1:affected1</code>	-0.720	0.277	-2.602	.009 **

4.2. Reading Times

The reading-time models (Table 2) localise the cost of a violation in time. On the object itself, the main effect of a violation is not significant ($\beta_{correct0} = -0.19$, $p = .132$); the surface controls behave as expected (object length and `live_row` both $p < .001$). What the object region shows instead is interaction structure: violations cost more for animate-indefinite objects ($correct0:an_ind$

$\beta = 0.26, p = .045$), and affectedness modulates the penalty in opposite directions across conditions. The estimated by-condition difference-of-differences ([−affected] minus [+affected] violation cost) is significantly positive for `an_def` (+0.23, 95% CI [0.03, 0.44]) and significantly negative for `inan_def` (−0.20 [−0.40, −0.01]), the two corresponding three-way terms being significant in the model (`correct0:an_ind:affected0` $\beta = -0.21, p = .037$; `correct0:inan_def:affected0` $\beta = -0.30, p = .003$).

The clearest violation effect appears downstream, in the first PP: a DOM violation slows reading at PP1 (`correct0` $\beta = 0.063, SE = 0.016, t = 3.92, p < .001$; EMM incorrect−correct = 0.06, 95% CI [0.03, 0.10]), a clear spillover effect. By PP2 the main violation effect has dissipated ($p = .169$) and only affectedness-related interactions remain (`correct0:affected0` $p = .042$; `correct0:an_ind:affected0` $p = .018$). Collinearity diagnostics are unremarkable for the continuous predictors (all $VIF \approx 1$), with high VIFs confined to the interaction terms (expected by construction); one Catalan participant (S23) is flagged by Cook’s distance in the object and PP1 models, but removing them leaves all estimates stable (Appendix A).

Table 2

Reading-time LMMs (log RT): key fixed effects by region. Surface controls (`live_row`, `length`, `frequency`) significant where noted; full tables and marginal means in Appendix A. (* $p < .05$, ** $p < .01$, *** $p < .001$.)

Region	Effect	β	t	p
Object	<code>correct0</code> (main)	−0.191	−1.510	.132
	<code>correct0:an_ind</code>	0.258	2.005	.045 *
	<code>correct0:an_ind:aff0</code>	−0.212	−2.088	.037 *
	<code>correct0:inan_def:aff0</code>	−0.298	−3.024	.003 **
PP1	<code>correct0</code> (main)	0.063	3.920	< .001 ***
PP2	<code>correct0</code> (main)	−0.061	−1.376	.169
	<code>correct0:aff0</code>	0.091	2.035	.042 *
	<code>correct0:an_ind:aff0</code>	−0.153	−2.377	.018 *

4.3. Model Surprisal: Coherence

Every model assigns higher surprisal to DOM violations than to their correct counterparts: the mean incorrect−correct difference is positive and highly significant for all seven models (all $p < .001$; Table 9, Appendix B), and is largest for the monolingual Spanish GPT-2 (mean difference 4.26, $t = 21.39$). When the difference is broken down by condition (Appendix B), however, the two simplest baselines misbehave: both the n-gram and the from-scratch Transformer assign a negative difference in the `inan_def` condition (−4.09 and −0.27 respectively, both $p < .01$) — i.e. they treat the correct marked inanimate as more surprising than the incorrect unmarked one — whereas all four neural LLMs keep the difference positive across conditions. A linear model of surprisal on the conditions confirms that the `inan_def` contrast is the most consistently captured effect (significant for all models), while affectedness interactions are mostly absent and emerge significantly only for the GRU and GPT-2 (Appendix B).

4.4. Model Surprisal: Correlation and Robustness

Table 11 reports how much each model’s surprisal improves the prediction of the human measures over a surface-only baseline (log-likelihood gain). We stress that these are magnitude comparisons: the procedure yields log-likelihood gains, not significance tests, so no p -values attach to them. Three patterns stand out. (i) For offline acceptability the gains are an order of magnitude larger for the neural LLMs than for the baselines, and largest for the Spanish GPT-2 (138.7 vs. a 85.1 baseline; n-gram only 3.5). (ii) For object reading times the ordering inverts: the from-scratch RNN and Transformer give the largest gains (13.3 and 13.2), the n-gram essentially none (0.05), and the large LLMs less than the small

recurrent models. (iii) For the spillover region PP1 the from-scratch Transformer leads (9.0), while for the distal PP2 the Spanish fine-tuned LLaMa and Qwen3 lead (8.5, 8.1). The robustness test (Appendix B) shows that the experimental predictors retain explanatory power once surprisal is included, and that GPT-2 absorbs the most grammatical structure in the offline data (it leaves the smallest residual gain for the fixed effects, 67.2 vs. 85.1). A post-hoc analysis (Appendix B) regressing each model’s surprisal on lexical frequency shows that GPT-2 is the only model whose surprisal is not significantly driven by either object or verb frequency (object $p = .787$, verb $p = .074$), whereas the from-scratch Transformer is the most frequency-driven (verb frequency $t = -10.92$, $p < .001$). The full correlation table is Table 11 (Appendix B).

5. Discussion

5.1. The Online Cost of a DOM Violation

Two facts about the human data fit together. Offline, a DOM violation is heavily penalised ($\beta = 2.03$ for correctness, EMM gap -1.55 with the CI well clear of zero); online, the cost does not surface on the object that carries (or lacks) the marking – the main effect of correctness on the object is not significant ($p = .132$) – but appears one region later, as a reliable slowdown at PP1 ($p < .001$), and has faded by PP2. We read this as a delayed-integration or spillover signature: readers detect the anomaly while reading the object region, but the measurable consequence is displaced downstream, as is common in self-paced reading. The violation is therefore not merely a metalinguistic dispreference; it has an online correlate, but a temporally distributed one – evidence an offline choice task such as Heredero and García [6] cannot provide, and consistent with spillover dynamics for other argument-structure anomalies [49].

5.2. Semantic Factors Interact rather than Trigger

The most informative part of the reading-time data is not any main effect but the interaction structure. Affectedness modulates the violation cost in a condition-dependent way: on the object, the difference-of-differences shows that low affectedness increases the violation cost relative to high affectedness for animate-definite objects ($+0.23$, CI $[0.03, 0.44]$) but decreases it for inanimate-definite objects (-0.20 , CI $[-0.40, -0.01]$). The first half of this pattern is broadly compatible with the affectedness account, on which low affectedness weakens the pull toward marking and so should soften the cost of mismatching it; the second half – a reversal for inanimates – is not predicted by current theories and suggests that animacy and affectedness combine in a finer way than a single prominence scale would imply. Definiteness contributes too: correctly marked animate-indefinite objects are the most accepted configuration (correct1:an_ind $p < .001$), the region of the system where marking is variable but normatively expected. The recurring lesson is that DOM does not behave as a single categorical trigger; the three factors jointly shape the penalty, and the interactions are visible only in a controlled minimal-pair design.

The sociolinguistic result reinforces this. Catalan–Spanish bilinguals (and, more weakly, Latin-American speakers) assign a significantly smaller penalty to violations than standard speakers (difference-of-differences $+1.37$, CI $[0.66, 2.08]$). Because Catalan instantiates DOM differently and more variably, this is naturally read as contact-driven convergence relaxing sensitivity to the prescriptive Spanish norm [69].

5.3. What Model Surprisal Does and Does not Capture

At the coarsest level all seven models pass the coherence test, assigning higher surprisal to violations ($p < .001$), but with two qualifications. First, the direction is not guaranteed for the weak models. For inanimate-definite objects – where the correct form is the unmarked one and adding *a* is the violation – the n-gram and the from-scratch Transformer reverse the expected sign: they assign higher

surprisal to the correct (unmarked) object than to the incorrect *a*-marked one (mean incorrect–correct differences -4.09 and -0.27). These baselines are thus tracking surface co-occurrence statistics rather than the grammatical norm, whereas the four neural LLMs keep the difference in the expected direction (violation more surprising) across all three conditions.

Second, predictive power is strongly measure-dependent and does not track scale. For offline acceptability the multilingual LLMs and especially the Spanish GPT-2 dominate, with log-likelihood gains several times the baselines'; without time pressure, models predict global preferability well. For the online regions where semantics and spillover are decisive (object, PP1) the ordering inverts: the small from-scratch models give the largest gains and the 8B-scale LLMs do worse, mirroring the finding that larger, lower-perplexity models are not better – sometimes worse – predictors of reading times [59, 60, 61, 63]. Crucially, good prediction is not sensitivity to DOM: most models' surprisal is significantly frequency-driven, the exception being the Spanish GPT-2, which is at once frequency-independent, the best offline predictor, and the model absorbing the most grammatical structure. Specialisation on the target language, not scale, is what helps.

The upshot for our second research question is qualified. Model surprisal can stand in for the human signal where that signal is global and offline, such as overall acceptability; it is far less reliable as a proxy for the local, incremental, factor-specific effects observed in real-time reading. Concretely, a model can rank whole sentences for grammatical preferability much as humans do, and yet still fail to reproduce the way human readers weight animacy, definiteness and affectedness differently from one region of the sentence to the next as it unfolds.

Two broader implications follow. Theoretically, the animacy $\times$ affectedness reversal we observe on the object suggests that the contribution of verbal affectedness is not simply additive on top of nominal prominence: its sign appears to depend on the animacy of the object, which a single one-dimensional prominence scale cannot capture and which invites a more articulated account in which nominal and verbal cues are weighted jointly. Methodologically, the dissociation between offline and online predictability is a caution for the increasingly common practice of using LLM surprisal as a stand-in for human judgements: a model can look excellent on an aggregate, time-insensitive measure (here, acceptability, where the Spanish GPT-2's log-likelihood gain is several times the baseline's) while failing to reproduce the incremental, factor-specific signal that the same phenomenon produces in real-time reading. Researchers who adopt surprisal as a proxy should therefore match the measure to the question, and should not assume that a larger or lower-perplexity model is the safer choice – for a typologically specific phenomenon in a single language, a smaller monolingual model may be both more faithful and more interpretable.

5.4. Relation to the Predictions and Limitations

Relative to the hypotheses, the predicted penalty for violations is confirmed for acceptability and, in displaced form, for reading times; the predicted attenuation for indefinite and [–affected] objects holds only in some configurations and reverses in others; and the smaller penalty for Catalan bilinguals is clearly supported. The main limitations are the modest sample (17 participants) and the need to drop by-item random effects for convergence, both of which constrain item-level generalisation; a sensitivity analysis without the one influential participant (S23) nonetheless left the estimates stable. A more substantive direction for future work concerns ecological validity: by design the items prioritise tightly controlled semantic contrasts over naturalness, so an obvious next step is to test whether the same effects hold with more natural, contextually richer materials – and with a larger, more varied participant pool that would also support by-item modelling and a direct test of whether the animacy $\times$ affectedness reversal replicates.

6. Conclusion

We combined an online self-paced reading and acceptability experiment on Spanish DOM with a comparison of object surprisal from seven language models. Three conclusions emerge. First, DOM is a

grammatically and cognitively real constraint: its violation is reliably penalised in acceptability and slows reading downstream of the marked object (spillover). Second, its semantic determinants do not act uniformly — definiteness, animacy and affectedness jointly modulate the violation cost and interact in ways that sometimes diverge from standard predictions, notably the reversal of affectedness between animate- and inanimate-definite objects — so DOM is better modelled as several interacting dimensions than as a single trigger. Third, model surprisal is informative but selective: all models flag violations as more surprising, yet they capture global offline preferability far better than the local, incremental effects that drive human online processing, and a monolingual Spanish GPT-2 outperforms larger multilingual models where DOM correctness is central. Models can thus approximate, but not yet replace, human data for probing the fine determinants of DOM.

Acknowledgments

This experiment was conducted as part of an internship supported by the Erasmus+ Traineeship Talents At Work grant issued by the Scuola Normale Superiore di Pisa. The author thanks the participants and the Centre for Language Studies at Radboud University. The entire project, including the language model experiments and all statistical analyses, was carried out as part of the author's Master's thesis in Linguistics at the University of Pisa, under the supervision of Prof. Alessandro Lenci, whom the author likewise warmly thanks.

Declaration on Generative AI

During the preparation of this work, the author used an AI assistant (Claude) for text translation, paraphrasing and rewording, formatting assistance. After using this tool, the author reviewed and edited the content and takes full responsibility for the publication's content.

References

- [1] G. Bosson, *Empirische Universalienforschung: Differentielle Objektmarkierung in den neuiranischen Sprachen*, Narr, 1985. doi:10.5281/zenodo.4697660.
- [2] J. Aissen, Differential object marking: Iconicity vs. economy, *Natural Language & Linguistic Theory* 21 (2003) 435–483. doi:10.1023/A:1024109008573.
- [3] M. Leonetti, Specificity and differential object marking in spanish, *Catalan Journal of Linguistics* 3 (2004) 75–114. doi:10.5565/rev/catjl.106.
- [4] K. von Heusinger, G. A. Kaiser, Affectedness and differential object marking in spanish, *Morphology* 21 (2011) 593–617. doi:10.1007/s11525-010-9177-y.
- [5] M. G. García, Nominal and verbal parameters in the diachrony of differential object marking in Spanish, in: I. A. Seržant, A. Witzlack-Makarevich (Eds.), *Diachrony of Differential Argument Marking*, Language Science Press, 2018, pp. 209–242.
- [6] D. R. Heredero, M. G. García, Differential object marking in Spanish: The effect of affectedness, *Caplletra. Revista Internacional de Filologia* (2023) 259–285. doi:10.7203/caplletra.74.26043.
- [7] H. Hopp, M. E. L. Arriaga, Structural and inherent case in the non-native processing of Spanish: Constraints on inflectional variability, *Second Language Research* 32 (2016) 75–108. doi:10.1177/0267658315605872.
- [8] J. Hale, A probabilistic Earley parser as a psycholinguistic model, in: *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001.
- [9] R. Levy, Expectation-based syntactic comprehension, *Cognition* 106 (2008) 1126–1177. doi:10.1016/j.cognition.2007.05.006.
- [10] N. J. Smith, R. Levy, The effect of word predictability on reading time is logarithmic, *Cognition* 128 (2013) 302–319. doi:10.1016/j.cognition.2013.02.013.

- [11] A. Goodkind, K. Bicknell, 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 (CMCL 2018)*, ACL, 2018, pp. 10–18. doi:10.18653/v1/W18-0102.
- [12] E. G. Wilcox, T. Pimentel, C. Meister, R. Cotterell, R. P. Levy, Testing the predictions of surprisal theory in 11 languages, *Transactions of the Association for Computational Linguistics* 11 (2023) 1451–1470. doi:10.1162/tac1_a_00612.
- [13] G. Bossong, DOM and linguistic typology: A personal view, in: J. Kabatek, P. Obrist, A. Wall (Eds.), *Differential Object Marking in Romance*, De Gruyter, 2021, pp. 21–36. doi:10.1515/9783110716207-002.
- [14] B. Comrie, *Language Universals and Linguistic Typology: Syntax and Morphology*, Basil Blackwell, 1981.
- [15] M. Silverstein, Hierarchy of features and ergativity, in: R. M. W. Dixon (Ed.), *Grammatical Categories in Australian Languages*, Australian National University, 1976, pp. 112–171.
- [16] K. Sinnemäki, A typological perspective on Differential Object Marking, *Linguistics* 52 (2014) 281–313. doi:10.1515/ling-2013-0063.
- [17] N. Gregorio, M. Gay, S. Goldwater, E. Ponti, The cross-linguistic role of animacy in grammar structures, in: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL, 2025, pp. 7349–7363. doi:10.18653/v1/2025.acl-long.364.
- [18] Ö. Dahl, Animacy and egophoricity: Grammar, ontology and phylogeny, *Lingua* 118 (2008) 141–150. doi:10.1016/j.lingua.2007.02.008.
- [19] M. S. Dryer, Noun phrase structure, in: *Language Typology and Syntactic Description*, Cambridge University Press, 2007, pp. 151–205.
- [20] H. Diessel, *Demonstratives*, John Benjamins, 1999.
- [21] M. Enç, The semantics of specificity, *Linguistic Inquiry* 22 (1991) 1–25.
- [22] B. Russell, On denoting, *Mind* 14 (1905) 479–493. doi:10.1093/mind/XIV.4.479.
- [23] C. Roberts, Uniqueness in definite noun phrases, *Linguistics and Philosophy* 26 (2003) 287–350. doi:10.1023/A:1024157132393.
- [24] K. von Stechow, Specificity and definiteness in sentence and discourse structure, *Journal of Semantics* 19 (2002) 245–274. doi:10.1093/jos/19.3.245.
- [25] E. Torrego, *The Dependencies of Objects*, The MIT Press, Cambridge, MA, 1998. doi:10.7551/mitpress/2337.001.0001.
- [26] M. Leonetti, Specificity in clitic doubling and in differential object marking, *Probus* 20 (2008) 33–66. doi:10.1515/PROBUS.2008.002.
- [27] A. Fábregas, Differential object marking in Spanish: State of the art, *Borealis – An International Journal of Hispanic Linguistics* 2 (2013). doi:10.7557/1.2.2.2603.
- [28] V. Escandell-Vidal, Differential object marking and topicality: The case of Balearic Catalan, *Studies in Language* 33 (2009) 832–884. doi:10.1075/sl.33.4.02esc.
- [29] G. Iemmolo, La marcatura differenziale dell’oggetto in siciliano antico, *Archivio Glottologico Italiano* (2009). doi:10.5167/UZH-81742.
- [30] G. Iemmolo, Topicality and differential object marking: Evidence from Romance and beyond, *Studies in Language* 34 (2010) 239–272. doi:10.1075/sl.34.2.01iem.
- [31] G. Iemmolo, Are dislocated direct objects clause-external? Evidence from differential object marking, in: *Annual Meeting of the Berkeley Linguistics Society*, volume 38, 2012, pp. 190–204. doi:10.3765/bls.v38i0.3329.
- [32] P. J. Hopper, S. A. Thompson, Transitivity in grammar and discourse, *Language* 56 (1980) 251–299. doi:10.2307/413757.
- [33] T. Tsunoda, Remarks on transitivity, *Journal of Linguistics* 21 (1985) 385–396. doi:10.1017/S0022226700010318.
- [34] S. Montrul, N. Sánchez-Walker, Differential object marking in child and adult Spanish heritage speakers, *Language Acquisition* 20 (2013) 109–132.
- [35] S. Montrul, A. Gürel, The acquisition of differential object marking in Spanish by Turkish speakers, in: T. Judy, S. Perpiñán (Eds.), *The Acquisition of Spanish in Understudied Language Pairings*,

John Benjamins, 2015, pp. 281–308. doi:10.1075/ihll.3.11mon.

- [36] S. Bautista-Maldonado, S. Montrul, An experimental investigation of differential object marking in Mexican Spanish, *Spanish in Context* 16 (2019) 22–50.
- [37] P. Guijarro-Fuentes, T. Marinis, Acquiring phenomena at the syntax/semantics interface in L2 Spanish: The personal preposition *a*, *EUROSLA Yearbook* 7 (2007) 67–88. doi:10.1075/eurosla.7.06gui.
- [38] C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal* 27 (1948) 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.
- [39] B. Kapteijns, F. Hintz, Comparing predictors of sentence self-paced reading times: Syntactic complexity versus transitional probability metrics, *PLOS ONE* 16 (2021) e0254546. doi:10.1371/journal.pone.0254546.
- [40] S. Nair, P. Resnik, Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship?, in: *Findings of the Association for Computational Linguistics: EMNLP 2023, ACL, 2023*, pp. 11251–11260. doi:10.18653/v1/2023.findings-emnlp.752.
- [41] T. Pimentel, C. Meister, How to compute the probability of a word, 2024. doi:10.48550/arXiv.2406.14561, arXiv preprint arXiv:2406.14561.
- [42] E. Gibson, The dependency locality theory: A distance-based theory of linguistic complexity, in: *Image, Language, Brain*, 2000, pp. 95–126.
- [43] V. Demberg, F. Keller, Data from eye-tracking corpora as evidence for theories of syntactic processing complexity, *Cognition* 109 (2008) 193–210.
- [44] D. Aaronson, H. S. Scarborough, Performance theories for sentence coding: Some quantitative evidence, *Journal of Experimental Psychology: Human Perception and Performance* 2 (1976) 56–70. doi:10.1037/0096-1523.2.1.56.
- [45] M. A. Just, P. A. Carpenter, J. D. Woolley, Paradigms and processes in reading comprehension, *Journal of Experimental Psychology: General* 111 (1982) 228–238. doi:10.1037/0096-3445.111.2.228.
- [46] J. Jegerski, Self-paced reading, in: *Research Methods in Second Language Psycholinguistics*, Routledge, 2013, pp. 20–49.
- [47] C. Felser, Eye-Tracking and self-paced reading, in: G. Goodall (Ed.), *The Cambridge Handbook of Experimental Syntax*, Cambridge University Press, 2021, pp. 617–640. doi:10.1017/9781108569620.024.
- [48] D. J. Barr, R. Levy, C. Scheepers, H. J. Tily, Random effects structure for confirmatory hypothesis testing: Keep it maximal, *Journal of Memory and Language* 68 (2013) 255–278. doi:10.1016/j.jml.2012.11.001.
- [49] J. Jegerski, G. D. Keating, Using self-paced reading in research with heritage speakers, *Frontiers in Psychology* 14 (2023). doi:10.3389/fpsyg.2023.1056561.
- [50] T. Mikolov, M. Karafiat, L. Burget, J. Černocký, S. Khudanpur, Recurrent neural network based language model, in: *Proceedings of Interspeech 2010*, 2010, pp. 1045–1048. doi:10.21437/Interspeech.2010-343.
- [51] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation* 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.
- [52] K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, 2014. doi:10.48550/arXiv.1409.1259, arXiv preprint arXiv:1409.1259.
- [53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. doi:10.48550/arXiv.1706.03762, arXiv preprint arXiv:1706.03762.
- [54] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, Technical Report, OpenAI, 2019.
- [55] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, et al., Language models are few-shot learners, 2020. doi:10.48550/arXiv.2005.14165, arXiv preprint arXiv:2005.14165.
- [56] H. Touvron, T. Lavril, G. Izacard, et al., LLaMA: Open and efficient foundation language models, 2023. doi:10.48550/arXiv.2302.13971, arXiv preprint arXiv:2302.13971.

- [57] A. Grattafiori, A. Dubey, A. Jauhri, et al., The Llama 3 herd of models, 2024. doi:10.48550/arXiv.2407.21783, arXiv preprint arXiv:2407.21783.
- [58] A. Yang, A. Li, B. Yang, et al., Qwen3 technical report, 2025. doi:10.48550/arXiv.2505.09388, arXiv preprint arXiv:2505.09388.
- [59] B.-D. Oh, W. Schuler, Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?, *Transactions of the Association for Computational Linguistics* 11 (2023) 336–350. doi:10.1162/tacl_a_00548.
- [60] B.-D. Oh, W. Schuler, Transformer-based language model surprisal predicts human reading times best with about two billion training tokens, in: *Findings of the Association for Computational Linguistics: EMNLP 2023, ACL, 2023*, pp. 1915–1921. doi:10.18653/v1/2023.findings-emnlp.128.
- [61] M. van Schijndel, T. Linzen, Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty, *Cognitive Science* 45 (2021) e12988. doi:10.1111/cogs.12988.
- [62] S. Shoghi, S. Arslan, R. Bastiaanse, S. Popov, Does a walk-through video help the parser down the garden-path?, *Frontiers in Psychology* 13 (2022) 1009265. doi:10.3389/fpsyg.2022.1009265.
- [63] Y. Lin, H. Zhu, W. Schuler, Vectors from larger language models predict human reading time and fMRI data more poorly when dimensionality expansion is controlled, 2025. doi:10.48550/arXiv.2505.12196, arXiv preprint arXiv:2505.12196.
- [64] K. Huang, S. Arehalli, M. Kugemoto, C. Muxica, G. Prasad, B. Dillon, T. Linzen, Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty, *Journal of Memory and Language* 137 (2024) 104510. doi:10.1016/j.jml.2024.104510.
- [65] J. Abadji, P. O. Suárez, L. Romary, B. Sagot, Towards a cleaner document-oriented multilingual crawled corpus, 2022. ArXiv e-prints arXiv:2201.06642.
- [66] D. Merks, S. L. Frank, Human sentence processing: Recurrence or attention?, in: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL, 2021*, pp. 12–22. doi:10.18653/v1/2021.cmcl-1.2.
- [67] C. Aurnhammer, S. L. Frank, Evaluating information-theoretic measures of word prediction in naturalistic sentence reading, *Neuropsychologia* 134 (2019) 107198. doi:10.1016/j.neuropsychologia.2019.107198.
- [68] C. C. Voeten, buildmer: Stepwise elimination and term reordering for mixed-effects regression, 2025. R package.
- [69] S. Perpiñán, On convergence, ongoing language change, and crosslinguistic influence in direct object expression in Catalan–Spanish bilingualism, *Languages* 3 (2018). doi:10.3390/languages3020014.

A. Supplementary Human Data Tables

This appendix reports the full statistical output summarised in Section 4. Cut-points of the cumulative-link models are omitted from the fixed-effect tables. *emmeans* contrasts are reported with 95% confidence intervals (Bonferroni-corrected).

Diagnostics. For the continuous predictors collinearity is negligible ($VIF \approx 1$ for *live_row*, *length* and *frequency* in all models); high VIFs are confined to the interaction terms, as expected by construction.

Table 3

Offline acceptability CLMM: full fixed effects.

Term	β	SE	z	p
correct1	2.031	0.348	5.834	< .001 ***
obj_freq	0.105	0.049	2.137	.033 *
verb_freq	0.184	0.070	2.630	.009 **
correct0:an_ind	0.386	0.248	1.557	.119
correct1:an_ind	0.977	0.264	3.700	< .001 ***
correct0:inan_def	-0.434	0.257	-1.686	.092
correct1:inan_def	0.295	0.248	1.187	.235
correct0:catalan	1.024	0.443	2.312	.021 *
correct1:catalan	-0.349	0.442	-0.788	.430
correct0:latin	0.648	0.560	1.157	.247
correct1:latin	-0.331	0.566	-0.584	.559
correct0:affected1	-0.454	0.274	-1.655	.098
correct1:affected1	-0.720	0.277	-2.602	.009 **

Table 4Acceptability: emmeans contrasts (incorrect – correct), 95% CI. Difference-of-differences vs. the reference level given in *italics*.

Grouping	Level	Estimate	95% CI
Overall	—	-1.55	[-1.86, -1.24]
Condition	an_def	-1.11	[-1.61, -0.62]
	an_ind	-1.70	[-2.23, -1.18]
	inan_def	-1.84	[-2.33, -1.35]
Δ	<i>an_ind – an_def</i>	-0.59	[-1.39, 0.21]
	<i>inan_def – an_def</i>	-0.73	[-1.50, 0.04]
Origin	standard	-2.34	[-2.81, -1.86]
	catalan	-0.96	[-1.39, -0.54]
	latin	-1.36	[-2.01, -0.71]
Δ	<i>catalan – standard</i>	1.37	[0.66, 2.08]
	<i>latin – standard</i>	0.98	[0.07, 1.89]
Affected	-affected	-1.69	[-2.11, -1.26]
	+affected	-1.42	[-1.83, -1.01]
Δ	<i>+aff – -aff</i>	0.27	[-0.29, 0.82]

Table 5

Object reading-time LMM: full fixed effects (log RT).

Term	β	SE	t	p
(Intercept)	6.330	0.203	31.187	< .001 ***
live_row	−0.004	0.001	−4.193	< .001 ***
object_len	0.034	0.007	4.673	< .001 ***
correct0	−0.191	0.126	−1.510	.132
c1:std	−0.053	0.224	−0.236	.816
c0:std	−0.047	0.225	−0.211	.835
c1:cat	−0.277	0.224	−1.237	.231
c0:cat	−0.070	0.225	−0.310	.760
c1:an_ind	−0.020	0.123	−0.159	.874
c0:an_ind	0.258	0.129	2.005	.045 *
c1:inan_def	−0.179	0.120	−1.487	.137
c0:inan_def	0.212	0.125	1.690	.091
c1:aff0	−0.094	0.068	−1.382	.167
c0:aff0	0.139	0.071	1.962	.050
c1:std:an_ind	−0.021	0.135	−0.152	.879
c0:std:an_ind	−0.228	0.143	−1.601	.110
c1:cat:an_ind	0.041	0.135	0.303	.762
c0:cat:an_ind	−0.181	0.142	−1.271	.204
c1:std:inan_def	0.003	0.132	0.025	.980
c0:std:inan_def	0.058	0.137	0.421	.674
c1:cat:inan_def	0.245	0.132	1.862	.063
c0:cat:inan_def	−0.189	0.137	−1.373	.170
c1:an_ind:aff0	0.025	0.098	0.250	.803
c0:an_ind:aff0	−0.212	0.102	−2.088	.037 *
c1:inan_def:aff0	0.137	0.095	1.444	.149
c0:inan_def:aff0	−0.298	0.099	−3.024	.003 **

Table 6

PP1 and PP2 reading-time LMMs: full fixed effects (log RT).

Reg.	Term	β	SE	t	p
PP1	(Intercept)	7.240	0.078	93.354	< .001 ***
	live_row	−0.002	0.001	−3.457	.003 **
	correct0	0.063	0.016	3.920	< .001 ***
	obj_freq	0.008	0.004	1.985	.048 *
PP2	(Intercept)	7.471	0.055	134.864	< .001 ***
	live_row	−0.002	0.001	−4.338	< .001 ***
	correct0	−0.061	0.045	−1.376	.169
	c1:an_ind	−0.034	0.042	−0.813	.417
	c0:an_ind	0.060	0.044	1.350	.177
	c1:inan_def	0.058	0.044	1.302	.193
	c0:inan_def	0.062	0.044	1.408	.160
	c1:aff0	0.031	0.044	0.710	.478
	c0:aff0	0.091	0.045	2.035	.042 *
	c1:an_ind:aff0	−0.090	0.063	−1.418	.157
	c0:an_ind:aff0	−0.153	0.064	−2.377	.018 *
	c1:inan_def:aff0	−0.097	0.061	−1.597	.111
	c0:inan_def:aff0	−0.037	0.062	−0.603	.547

Table 7

Object reading-time emmeans (incorrect – correct) by condition and affectedness, 95% CI, and the affectedness difference-of-differences.

Condition	Affected	Estimate	95% CI
an_def	+	–0.12	[–0.27, 0.03]
an_ind	+	0.01	[–0.13, 0.16]
inan_def	+	0.14	[0.00, 0.29]
an_def	–	0.11	[–0.03, 0.26]
an_ind	–	0.01	[–0.15, 0.17]
inan_def	–	–0.06	[–0.20, 0.08]
Δ (an_def)		0.23	[0.03, 0.44]
Δ (an_ind)		0.00	[–0.21, 0.21]
Δ (inan_def)		–0.20	[–0.40, –0.01]

Table 8

Descriptive reading-time means (ms) per sentence type, and data exclusions per region. c: 1 = correct, 0 = violation. “*type” = prescriptively incorrect.

Type	c	Object	PP1	PP2
an. def. marked	1	605.7	1438.8	1710.0
an. indef. marked	1	573.7	1408.8	1530.4
inan. (def.) unmarked	1	549.2	1467.5	1647.4
*an. def. unmarked	0	604.2	1533.1	1638.9
*an. indef. unmarked	0	573.5	1496.4	1628.1
*inan. (def.) marked	0	580.5	1554.8	1697.5

B. Supplementary Language Model Tables

Model labels: n-gram (KenLM), GRU and Transformer (from scratch), GPT-2 (monolingual Spanish), LLaMa 3.1 8B, Qwen3 8B, LLaMa (Spanish fine-tuned). The from-scratch baselines were evaluated without two out-of-vocabulary items.

Table 9

Coherence test: mean object-surprisal difference (incorrect – correct) per model, one-sample t -test. All differences positive and significant.

Model	Mean diff.	t	p
n-gram (KenLM)	2.94	5.56	< .001 ***
GRU (from scratch)	0.38	9.83	< .001 ***
Transformer (scratch)	0.63	8.18	< .001 ***
GPT-2 (Spanish)	4.26	21.39	< .001 ***
LLaMa 3.1 8B	2.21	15.28	< .001 ***
Qwen3 8B	3.48	12.71	< .001 ***
LLaMa (Spanish ft.)	2.74	11.95	< .001 ***

Table 10

Coherence test by condition: mean surprisal difference (incorrect – correct) per model, one-sample t -test. Negative values (in bold) indicate that the model treats the correct variant as more surprising.

Model	Cond.	Mean diff.	t	p
n-gram	an_def	6.52	14.19	< .001 ***
	an_ind	6.72	11.66	< .001 ***
	inan_def	−4.09	−8.45	< .001 ***
GRU	an_def	0.20	2.86	.007 **
	an_ind	0.57	8.73	< .001 ***
	inan_def	0.39	6.87	< .001 ***
Transf.	an_def	0.83	9.38	< .001 ***
	an_ind	1.38	17.04	< .001 ***
	inan_def	−0.27	−3.27	.002 **
GPT-2	an_def	5.00	13.89	< .001 ***
	an_ind	4.33	14.18	< .001 ***
	inan_def	3.46	10.37	< .001 ***
LLaMa	an_def	2.38	11.88	< .001 ***
	an_ind	2.65	9.31	< .001 ***
	inan_def	1.62	6.77	< .001 ***
Qwen3	an_def	4.57	9.54	< .001 ***
	an_ind	4.24	9.18	< .001 ***
	inan_def	1.67	4.67	< .001 ***
LLaMa ft.	an_def	3.62	9.19	< .001 ***
	an_ind	3.26	7.61	< .001 ***
	inan_def	1.38	4.94	< .001 ***

Missing items per model (%). n-gram 0.0; GRU 4.0; Transformer 4.0; GPT-2 7.4; LLaMa 7.4; Qwen3 7.4; LLaMa ft. 7.4. Coherence linear model: regressing each model’s surprisal on condition (and its interactions) confirms the patterns described in Section 4.

Table 11

Correlation test: log-likelihood gain from adding each model's object surprisal to the surface-only baseline, by human measure. Descriptive magnitude comparisons, not significance tests (no *p*-values). base = surface-only reference.

Model	Offline	Object	PP1	PP2
base	85.07	17.33	7.59	5.87
n-gram	3.51	0.05	0.13	0.26
GRU	44.28	13.28	6.06	1.08
Transformer	46.02	13.17	9.05	1.01
GPT-2 (Spanish)	138.71	6.45	4.72	7.16
LLaMa 3.1 8B	133.89	8.22	1.34	7.07
Qwen3 8B	133.93	6.13	0.97	8.10
LLaMa (Spanish ft.)	128.95	9.28	2.16	8.49

Table 12

Robustness test: log-likelihood gain from the experimental fixed effects once each model's surprisal is included (descriptive magnitude, not a significance test). base = surface-only reference.

Model	Offline	Object	PP1	PP2
base	85.07	17.33	7.59	5.87
n-gram	82.03	17.55	7.49	5.65
GRU	84.42	18.26	6.63	6.16
Transformer	82.24	18.31	4.62	5.81
GPT-2	67.22	17.71	3.84	5.51
LLaMa	71.48	17.33	6.26	5.59
Qwen3	72.27	17.41	6.66	5.14
LLaMa ft.	76.44	17.84	5.59	4.81

Table 13

Post-hoc: each model's object surprisal regressed on object and verb (negative-log) frequency. GPT-2 is the only model with no significant frequency effect. The n-gram (sentence-level only) is excluded.

Model	Predictor	β	t	p
GRU	object_freq	0.01	0.77	.442
	verb_freq	-0.09	-6.89	< .001 ***
Transf.	object_freq	-0.05	-2.76	.006 **
	verb_freq	-0.24	-10.92	< .001 ***
GPT-2	object_freq	0.02	0.27	.787
	verb_freq	0.17	1.79	.074
LLaMa	object_freq	-0.01	-0.17	.868
	verb_freq	-0.19	-3.12	.002 **
Qwen3	object_freq	0.05	0.71	.475
	verb_freq	-0.18	-2.13	.034 *
LLaMa ft.	object_freq	0.25	3.72	< .001 ***
	verb_freq	0.17	2.14	.033 *

C. Experimental Item Set

The 50 experimental items are listed below in their base form (subject, verb, human-object/inanimate-object, locative PP, temporal PP); each was realised in six variants crossing animacy, definiteness and DOM marking.

[+affected] items.

- *Juan golpeó actriz/piedra en la calle en diez segundos*
- *Carlos asaltó conductora/camioneta en el barrio en pocos instantes*
- *Simón pateó jugadora/pelota en la cancha en un instante*
- *Pablo secuestró turista/camioneta en el aeropuerto durante una hora*
- *Jorge arrolló ciclista/valla en la calle en un segundo*
- *Manuel abrazó señora/planta en el parque en pocos instantes*
- *Francisco besó novia/cruz en la iglesia durante unos segundos*
- *Tomás empujó vieja/silla a la puerta en pocos segundos*
- *Helena colgó criminal/fotografía en el patio en cinco minutos*
- *Carlos quemó bruja/página en la plaza en pocos minutos*
- *Mateo eliminó rehén/carpeta en el banco en pocos segundos*
- *Valeria neutralizó criminal/bomba en la frontera en cinco minutos*
- *Paula abatió contrincante/resistencia en el ring en pocos minutos*
- *Martina atropelló anciana/carreta en el aparcamiento en pocos minutos*
- *Carlos derribó futbolista/barrera en el campo durante el partido*
- *Lucía destrozó oponente/mesa en el bar en diez minutos*
- *Valentina venció adversaria/resistencia en la guerra en dos años*
- *Luis zarandéó mujer/muñeca en la casa durante algunos minutos*
- *María transportó mochilera/guitarra en la furgoneta durante un día*
- *Lucas arrastró víctima/carreta en la calle por un tiempo*
- *Claudia maltrató compañera/camiseta en la casa durante varios años*
- *Hugo apaleó rebelde/muñeca en la calle durante varios minutos*
- *Carla asedió contrincante/ciudad en la región durante bastante tiempo*
- *Daniel remolcó enferma/motocicleta en la calle durante algunos minutos*
- *Coral guió turista/decisión en el palacio en una hora*

[-affected] items.

- *Pedro observó secretaria/planta en la oficina durante una hora*
- *Juan escuchó vecina/alarma en el balcón durante una noche*
- *Felipe anunció cantante/actuación en el escenario en cinco minutos*
- *Anna comprendió maestra/tarea en la escuela en dos minutos*
- *Alfonso miró granjera/estrella en la colina durante dos horas*
- *Carlos esperó profesora/furgoneta en la esquina durante una hora*
- *Miguel recibió invitada/caja en la casa en un día*
- *Martín buscó médica/escalera en el hospital durante dos horas*
- *Marco imaginó amiga/fiesta en la playa durante dos minutos*
- *Carmen odió vendedora/música en el cine durante la película*
- *Carlos vigiló niña/operación en la habitación durante unas horas*
- *Irene oyó niña/discusión en la sala durante una noche*
- *Isabel vió estudiante/tormenta desde su ventana durante una mañana*
- *Unai amó jovencita/vida en su pueblo durante varios años*

- *Valeria admiró cantante/escenografía en el teatro durante el concierto*
- *Adrián percibió mujer/sombra detrás de él durante algunos minutos*
- *Emma investigó empresaria/epidemia en la ciudad durante varias semanas*
- *Ismael encontró conocida/dificultad en la carretera en pocas horas*
- *Sofía conoció chica/verdad en el colegio en dos días*
- *Martín detectó intrusa/enfermedad en el hospital en dos minutos*
- *Elena localizó amiga/bodega en la ciudad en dos minutos*
- *Alejandro identificó agresora/amenaza en el aeropuerto en dos horas*
- *Marta descartó aspirante/sospecha en la selección en un día*
- *Samuel descubrió infiltrada/tabla en el museo en pocas horas*
- *Alba mencionó actriz/ciudad tras varios minutos*

D. Online Resources

The scripts and data used in this paper are available via [GitHub](#).