Two ways into the hall of mirrors: Language exposure and lossy memory drive cross-linguistic grammaticality illusions in language models

Kate McCurdy and Katharina Christian and Amelie Seyfriend and Mikhail Sonkin
Saarland University

Abstract

Readers of English — but not Dutch or German — consistently show a *grammaticality illusion*: they find ungrammatical double-center-embedded sentences easier to process than corresponding grammatical sentences. If pretrained language model (LM) surprisal mimics these cross-linguistic patterns, this implies that language statistics explain the effect; if, however, the illusion requires memory constraints such as lossy context surprisal (LCS), this suggests a critical role for memory. We evaluate LMs in Dutch, German, and English. We find that both factors influence LMs' susceptibility to grammaticality illusions, and neither fully account for human-like processing patterns.

1 Introduction

Modern neural language models (LMs) produce fluent, grammatical language (Mahowald et al., 2024), but their validity as models of human linguistic cognition remains contested (Cuskley et al., 2024). One key concern is the scale of data exposure: LMs often learn from quantities of linguistic data which exceed human lifespans. This motivates research with LMs trained on human-scale data, as these models may have a greater claim to cognitive plausibility (Wilcox et al., 2025).

Another dimension of cognitive plausibility concerns language processing rather than learning. Language model surprisal robustly predicts measures of incremental language processing such as reading time (Wilcox et al., 2023). Despite this, LMs fail to fully reproduce certain processing effects which are well-established in the experimental literature, such as recovery from syntactically ambiguous "garden-path" sentences (Arehalli et al., 2022; Huang et al., 2024). Such systematic divergences from human processing further challenge LMs' cognitive plausibility. Moreover, the two issues may be connected: larger LMs trained on more data are worse at approximating reading time (Oh and Schuler, 2023), raising the possibility that non-human-scale learning contributes to non-human-like processing.

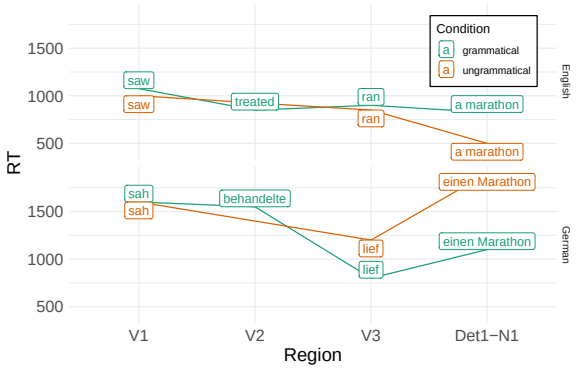


Figure 1: Example of the grammaticality illusion in reading time (RT) for (1a) vs. (1b) in English (upper), and the converse effect in German (lower), using mean RTs from Vasishth et al. (2010, Expts. 1 and 3). German speakers take longer to read the post-verbal NP ("einen Marathon") in the ungrammatical, missing-verb condition, while English speakers instead read the NP faster when the verb is missing.

One important lens on linguistic cognition comes from language *illusions*, i.e. ungrammatical or otherwise infelicitous sentences which humans should reject, but nonetheless find acceptable (Phillips et al., 2011). For example, consider the following two sentences (cf. Figure 1):

1. (a) *The painter who the doctor who the lady saw treated ran a marathon.*
   (b) **The painter who the doctor who the lady saw ran a marathon.*

(1a) contains double nested center-embedded clauses, which are rare and challenging in English (Hamilton and Deese, 1971) but indisputably grammatical. (1b), on the other hand, is ungrammatical due to missing a verb. Despite this, speakers consistently prefer sentences like (1b) to their grammatical counterparts (Gibson and Thomas, 1999;

Proceedings of the First BabyLM Workshop, pages 226–236
November 8, 2025 ©2025 Association for Computational Linguistics

Christiansen and MacDonald, 2009). This processing effect is known as the *grammaticality illusion*.

What causes the grammaticality illusion? Two competing hypotheses have been proposed. Under the *memory* account, constrained working memory causes readers to forget earlier sentence material, thereby nullifying the expectation of a third verb (Gibson and Thomas, 1999). This proposal appeals to general cognitive mechanisms; however, Vasishth et al. (2010) and Frank et al. (2016) find that the illusion appears for reading times in English, but not in German or Dutch. They posit an alternative *language statistics* hypothesis: the grammaticality illusion arises due to the relative rarity of center-embedded clauses in English. Researchers have evaluated these two accounts with computational models (Engelmann and Vasishth, 2009; Christiansen and MacDonald, 2009; Frank, 2014; Futrell et al., 2020). These simulations, however, predate today's language models (LMs), which represent linguistic distributions to unprecedented levels of precision.

In this paper, we use modern LMs to assess these two hypotheses with respect to the grammaticality illusion. LMs effectively implement the language statistics account. If the distribution of English gives rise to the illusion, then English LMs should assign higher surprisal to the post-verbal region *a marathon* in the grammatical sentence (1a) compared to the ungrammatical (1b), while German and Dutch LMs should do the converse. Moreover, this cross-linguistic divergence should hold steady, or even grow, with increased training data: if language statistics drive the effect, then higher data exposure during training should reinforce these respective language-specific outcomes.

If, however, memory constraints are critical to the illusion, we may see two different patterns. Firstly, the grammaticality illusion in English may be mediated by language model capacity in the *opposite* direction (cf. Oh and Schuler, 2023). In this scenario, smaller models trained on human-scale data may show the illusion, while larger LMs consistently prefer the grammatical sequence. Secondly, the grammaticality illusion may be mediated by retention of the preceding linguistic context, such that Lossy Context Surprisal shows the effect at higher forgetting rates (LCS; Futrell et al., 2020).

Our results suggest that both language statistics and memory constraints influence how LMs process double-center-embedded sentences, with mixed implications for human sentence processing. For the language statistics hypothesis, we find robust support from Dutch, but for German and English the picture is more complicated. Larger English LMs reproduce the missing verb illusion, which decisively supports the hypothesis. Medium-sized models, however, show the opposite effect, and larger German LMs also unexpectedly show the illusion — neither of which are expected under a language statistics account. For the memory hypothesis, we find that Resource-Rational Lossy Context Surprisal (RR-LCS; Hahn et al., 2022) simulates the grammaticality illusion at higher forgetting rates in both English and German; however, we do not observe the expected language-specific interaction in effect direction. These findings highlight continuing challenges in applying LMs as cognitive models of human language processing.

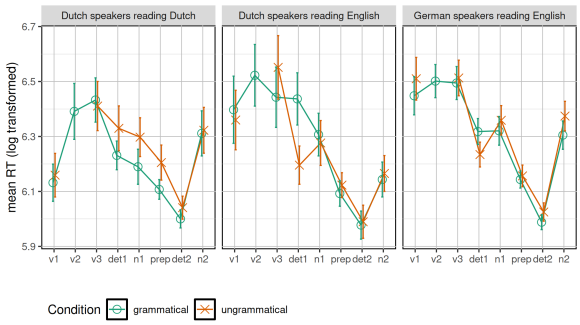


Figure 2: Grammaticality effects in reading time (RT) for Dutch and English stimuli, reproduced from Frank et al. (2016, Expts. 1–3). At the post-verbal determiner (Det1), RT is higher for ungrammatical sentences in Dutch (left), but lower in English (middle, right).

2 Related work

2.1 Grammaticality illusion effects across languages

English speakers consistently prefer ungrammatical sentences like 1b to grammatical double-center-embedded sentences like 1a. This missing verb effect has been found in both acceptability judgments (Gibson and Thomas, 1999; Frank and Ernst, 2019; Huang and Phillips, 2021) and measures of online processing such as reading time (Vasishth et al., 2010; Frank et al., 2016, 2021). Some researchers have argued that this effect reflects language-specific distributions of center-embedded sentences (e.g. Vasishth et al., 2010; Pañeda and Lago, 2024). Figure 2 presents key evidence for this **language statistics** hypothesis: the missing verb effect appears in English, but not in the related

language Dutch. Strikingly, the English reading times shown in Figure 2 come from first language speakers of Dutch and German — speakers who do not show this effect when reading their native language. The fact that the grammaticality illusion appears robustly for particular *languages*, rather than particular *speakers*, supports the language statistics hypothesis. Under this account, we would expect modern language models to show comparable language-dependent preferences: lower surprisal for grammatical Dutch and German sentences, and ungrammatical English sentences.

The main alternative hypothesis states that the grammaticality illusion is driven by domain-general **memory constraints** (Gibson and Thomas, 1999). Language-specific effects (cf. Figure 2) clearly challenge this account. In response, some recent work has explored how working memory could be mediated by particular linguistic properties. For instance, relative clauses in German and Dutch use different word order from main clauses; Bader and colleagues (Bader, 2016; Häussler and Bader, 2015) argue that this syntactic distinction may facilitate retrieval from memory in German, but not English. Huang and Phillips (2021) build on this account to characterize a similar illusion in Mandarin. Finally, Futrell and colleagues (Futrell and Levy, 2017; Futrell et al., 2020) integrate the two hypotheses with a model of memory directly mediated by language statistics, which we consider at length in the following section.

2.2 Processing illusions in language models

A number of computational modeling studies have found support for the language statistics hypothesis, although much of this previous work relies upon simulated training data. Christiansen and Chater (1999), Engelmann and Vasishth (2009), and Christiansen and MacDonald (2009) trained recurrent neural networks (RNNs) on two distinct probabilistic context-free grammar (PCFG)-generated corpora with differing relative clause distributions, reflecting corpus frequencies from German and English. These models capture human-like grammaticality preferences for double center-embedded constructions in each language respectively. Frank (2014) trained RNN language models on natural corpora from English and Dutch, and Frank et al. (2016) find that these models reproduce the language-specific grammaticality effects observed in their behavioral experiments (Figure 2). Notably, however, the English model did not fully reproduce the strength of the preference for ungrammatical sentences; it showed lower surprisal in the missing verb condition, but this difference did not reach statistical significance.

Futrell et al. (2020) introduce lossy context surprisal (LCS), a model which synthesizes memory-based and language statistics accounts. The core intuition is that speakers rely on noisy memory representations to predict upcoming words, and the noisy memory is more likely to recall structures which are more frequent in their language. Therefore, even if speakers typically forget 20–30% of the words in a given sentence context, a German speaker is more likely to correctly retain multiple verb-final relative clauses than an English speaker, simply due to the greater prevalence of such constructions in German. Futrell et al. similarly evaluate their model on a PCFG-derived corpus, with the additional manipulation of a forgetting parameter. Their LCS model predicts that, at certain levels of memory loss, English comprehenders will exhibit the grammaticality illusion, whereas German comprehenders will not. The LCS model thus provides, in principle, a memory-based account for language-specific grammaticality effects (although cf. Huang and Phillips, 2021).

If the grammaticality illusion in English reflects memory constraints, however, we would not expect it to arise from modern Transformer-based language models (LMs) (Vaswani et al., 2017), which have drastically larger memory capacities than the RNNs evaluated by Frank et al. (2016). Modern LMs have shown some susceptibility to other language illusions, for instance in processing negative polarity items (Zhang et al., 2023) and number agreement (Arehalli and Linzen, 2024). In terms of memory, however, modern LMs seem to show superhuman linguistic memory in certain respects (Oh and Schuler, 2023; Oh et al., 2024). The missing verb illusion, then, presents a key test case for the language statistics hypothesis: if it truly reflects language statistics rather than memory limitations, then large modern LMs should reproduce the preference for ungrammatical double center-embeddings in English.

3 Methods

3.1 Models and measures

Surprisal We test[2] multiple pretrained language models (LMs), listed in Table 1.We select these

[2] github.com/kmccurdy/grammaticality-illusion-LMs

Language	Model	Family	Parameters	Training data	Reference
Dutch	GPT2-S Dutch	GPT	129M	33B	de Vries and Nissim (2020)
	GPT2-M Dutch	GPT	380M		
	LLaMA2 Dutch	Llama	13B	2B	Vanroy (2023)
German	GerPT2-L	GPT	876M	50B	Minixhofer (2020)
	BLOOM	GPT	6.4B	50B	Ostendorff and Rehm (2023)
	LEO-LM	Llama	7B	65B	Plüster and Schuhmann (2023)
			13B		
English	GPT-BERT-S	GPT-BERT	30M	100M	Charpentier and Samuel (2024)
	GPT-BERT	GPT-BERT	119M		
	GPT2-mini	GPT	39M	≈2.25B	Fagnou (2024)
	GPT2	GPT	137M	≈15B	Radford et al. (2019)
	GPT2-L	GPT	812M		
	GPT-Neo	GPT	2.72B	420B	Black et al. (2021)
	GPT-J	GPT	6B		
	DeepSeek[1]	DeepSeek	7B	2T	DeepSeek-AI et al. (2024)
	LLaMA2	Llama	7B	2T	Touvron et al. (2023)
			13B		
Multiple	mGPT	GPT	1.3B	≈450B EN, 100B DE, 50B NL	Shliazhko et al. (2024)
			13B		
	LLaMAX	Llama	6.74B	≈950M EN, 900M DE, 590M NL	Lu et al. (2024)
	EuroLLM	Llama	9.15B	≈2T EN, 240B DE, 100B NL	Martins et al. (2024)

Table 1: Language models used in experiments. 'Training Data' refers to language-specific training data in tokens. Note that all monolingual Dutch and German models are initialized from models pre-trained on English.

models to span a range of sizes and training regimes, but focus only on models trained on the language modeling objective, e.g. excluding instruction-tuned models.

We follow previous work (e.g. Futrell et al., 2019) in using LM surprisal to measure incremental processing difficulty. Surprisal is calculated as the negative log probability of a word[3] w_T conditioned on the sequence of preceding words:

$$-\log P(w_{T+1}|w_{1...T}) \quad (1)$$

Lossy Context Surprisal Lossy context surprisal (Futrell et al., 2020) has been proposed to model language-specific effects of constrained memory. We use a specific implementation: resource-rational lossy context surprisal (RR-LCS), proposed by Hahn et al. (2022). RR-LCS provides a fully data-driven implementation of LCS, with only one free parameter: the forgetting rate. Crucially, we can train a range of individual RR-LCS model instances at different forgetting rates to simulate different patterns of working memory engagement. As all aspects of the RR-LCS model are learned from monolingual corpora, we expect that this model is capable of learning and reproducing language-specific effects.

In contrast to standard surprisal, which conditions on an exact word sequence, LCS conditions on a noisy memory representation of the preceding context:

$$-\log P(w_{T+1}|M_T) \quad (2)$$

where M is a lossy representation generated from $w_1 \dots w_T$. At a given forgetting rate, for a given word sequence $w_{1...T}$, RR-LCS learns to stochastically retain or delete specific words from M_T. Reconstructions of the missing words are then sampled, on the basis of language statistics, from a reconstruction model, and the overall surprisal of the noisy sequence is computed using a standard pretrained language model. We refer the reader to Hahn et al. (2022) for further details. As RR-LCS is computationally expensive, we train a limited set of models for two languages, English and German,

[3]Modern LMs are typically trained on a vocabulary of subword tokens rather than words; however, this does not affect our analysis for reasons discussed in the following section.

with the subword version of RR-LCS (McCurdy and Hahn, 2024). We use BLOOM as the base LM for German, and GPT2-L as the base LM for English. We train 3 instances of the RR-LCS model at forgetting rates 20%, 30%, and so on, up to 80%.[4]

3.2 Evaluation

Stimuli We evaluate the grammaticality illusion using the Dutch, German, and English stimuli developed by Vasishth and colleagues (Vasishth et al., 2010; Frank et al., 2016). Each stimulus item appears both as a grammatical double-center-embedded construction (e.g. 1a) and with an ungrammatical missing verb (e.g. 1b). Table 2 illustrates an additional manipulation: in the English and German stimuli, subject noun phrases are either all animate, or the second noun is replaced with an inanimate object. Vasishth et al. (2010) describe this manipulation as motivated by *interference*, but do not discuss it any further. Frank et al. (2021) do not reproduce this animacy manipulation, so the Dutch stimuli only include animate subject nouns.

Critical region We focus on the determiner immediately following the third verb, for reasons illustrated by Figure 2. Across all three languages, we observe the grammaticality effect (in German and Dutch) or illusion (in English) on the post-verbal noun phrase, especially the beginning of the phrase — the determiner. This makes sense: if the reader expects a third verb, and sees a determiner instead, this mismatch in expectation should yield higher RT at this location. We also observe higher RT on the noun, but this may reflect spillover effects (e.g. Rayner, 1998). Processing difficulty appears initially on the determiner; therefore, our analysis of LM surprisal focuses on this region.

4 Results

4.1 Language model surprisal

Figure 3 shows language model (LM) surprisal results Dutch and German, including only stimuli in the animate condition. Across model sizes, Dutch LMs show a robust preference for grammatical sentences, reproducing the effect found in human reading times (Frank et al., 2016; Frank and Ernst, 2019). By contrast, larger German LMs show higher surprisal for grammatical compared to ungrammatical stimuli — in other words, they

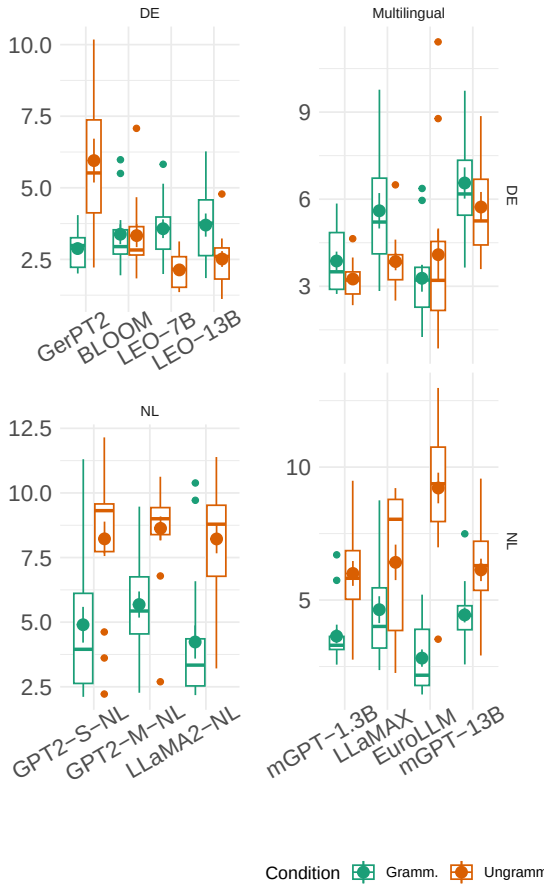


Figure 3: Language model surprisal for animate stimuli in German (upper) and Dutch (lower), for monolingual (left) and multilingual (right) models. Models are ordered left-to-right by parameter count. Dutch models consistently prefer grammatical sentences (lower surprisal), while larger German models unexpectedly show the grammaticality illusion.

show the grammaticality illusion. This is highly unexpected, as Vasishth et al. (2010) conducted multiple experiments with German speakers across different modalities, and never found a preference for ungrammatical sentences.

While Dutch LMs of all capacities prefer grammatical sentences, and German LMs flip from the grammatical to ungrammatical preference as the models grow in size, English language models (Figure 4) show an even more variable trajectory. Small models trained on human-scale data, such as GPT-BERT (the top performing model in the 2024 BabyLM competition; Charpentier and Samuel, 2024), do not consistently prefer either condition. As model size increases, we see the opposite of the illusion: GPT2 and GPT2-L assign lower surprisal to the grammatical sentence. This outcome — that

[4]We omit deletion rates of 10% and 90% for reasons of stability, as these models have high rates of invalid output.

Language	Animacy	Example item
English	All Animate	The dancer who the singer who the bystander admired...
	N2 Inanimate	The dancer who the shoe that the bystander admired...
German	All Animate	Der Tänzer, den der Artist, den der Zuschauer bewunderte,...
	N2 Inanimate	Der Tänzer, den der Schuh, den der Zuschauer bewunderte,...
Dutch	All Animate	De danser die gisteren de zanger die laatst de toeschouwer bewonderde...

Table 2: Example items by language and animacy. Each item is continued in both the grammatical and ungrammatical (missing verb) condition. German and English stimuli (from Vasishth et al., 2010) include an animacy manipulation, in which an inanimate object replaces the second subject noun. The inanimate condition is not included in the Dutch stimuli (from Frank et al., 2016).

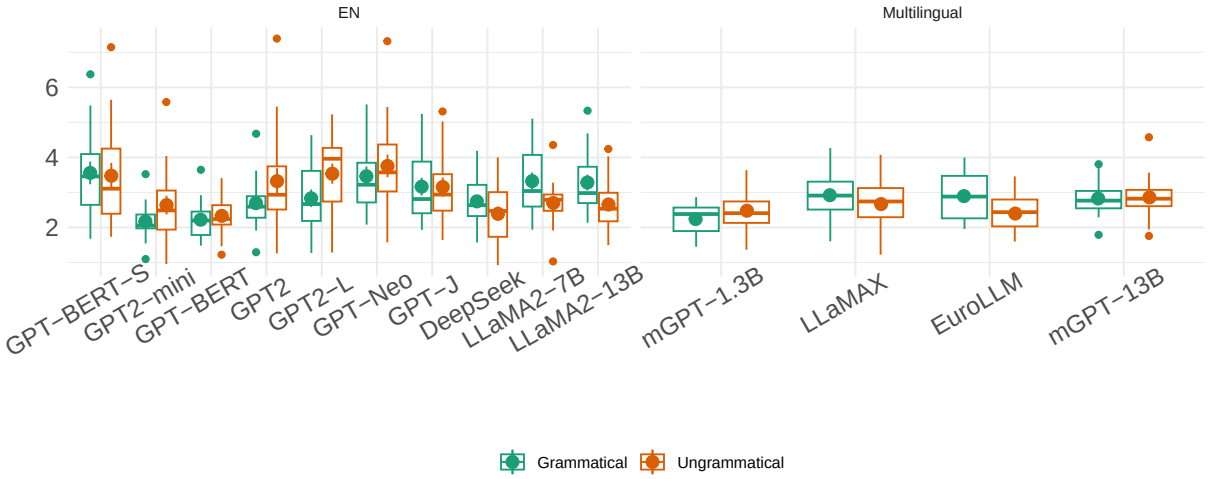


Figure 4: Grammaticality effects for animate stimuli in English LMs. Models are ordered left-to-right by size. Medium-sized models (e.g. GPT2) prefer grammatical sentences, while larger models (e.g. DeepSeek, LLaMA2) prefer ungrammatical sentences.

English LMs successfully learn the correct syntax of double center-embedded sentences — is surprising under the language statistics hypothesis: if the grammaticality illusion reflects distributional statistics of English, how do GPT2 and GPT2-L learn to predict a third verb in vanishingly rare double center embeddings? On the other hand, it's fully compatible with the memory account: as LMs grow in capacity, they can memorize linguistic events of increasing rarity (Oh et al., 2024) — to align them with human processing, we need to model memory constraints, as in (RR-)LCS.

As English LM size keeps increasing, however, the results become even more complex: larger models (e.g. GPT2-Neo and GPT-J) lose the grammatical preference, and the largest models we evaluate (Deepseek, LLaMA2) show the reverse preference — i.e. the grammaticality illusion. This outcome reverses our previous interpretation of the hypotheses. The language statistics account now looks like the decisive victor. Increased exposure to English language data leads the models to prefer ungrammatical sentences with missing verbs, and with parameter counts in the billions, their behavior is unlikely to reflect general memory limitations.

To compare grammaticality effects across models, we fit generalized linear mixed-effects models.[5] We report the t statistic as a measure of how reliably each LM distinguishes grammatical from ungrammatical sentence. In this case, as all models are evaluated on the same set of stimuli, larger values for t indicate more consistent differentiation between the two conditions; for instance, t values above 2 are often used heuristically to indicate statistical significance.

Figure 5 plots the results by model size and training data size, with separate plots for animate and inanimate stimuli. In English models, we see that training data size and model size both align with the puzzling pattern discussed above: medium-sized

[5]We use the `lme4` library (Bates et al., 2015) in R (R Core Team, 2023) with the following formula: `Surprisal ~ Condition + (1|Item)`.

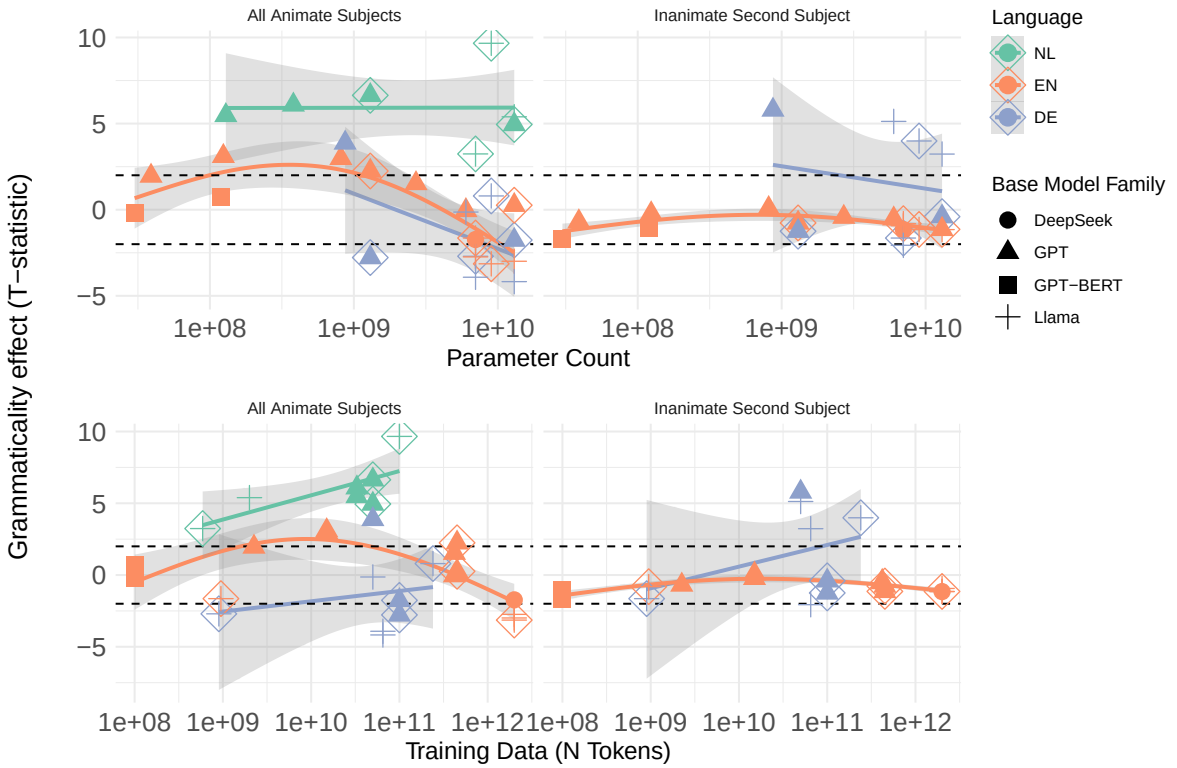


Figure 5: Summary of grammaticality effect (measured by t-statistic) by language and model size in pre-trained LMs, plotted by model size (upper) and training data size (lower). Diamond outline indicates multilingual model. Negative t-value indicates preference for the ungrammatical sentence, i.e. the grammaticality illusion. Dotted lines indicate approximate significance threshold ($|t| > 2$). Trend line fit by generalized additive model (GAM) with 3 basis dimensions. Increased training data exposure drives cross-linguistic divergence: German and Dutch models increasingly prefer grammatical sentences, while English models increasingly prefer ungrammatical sentences.

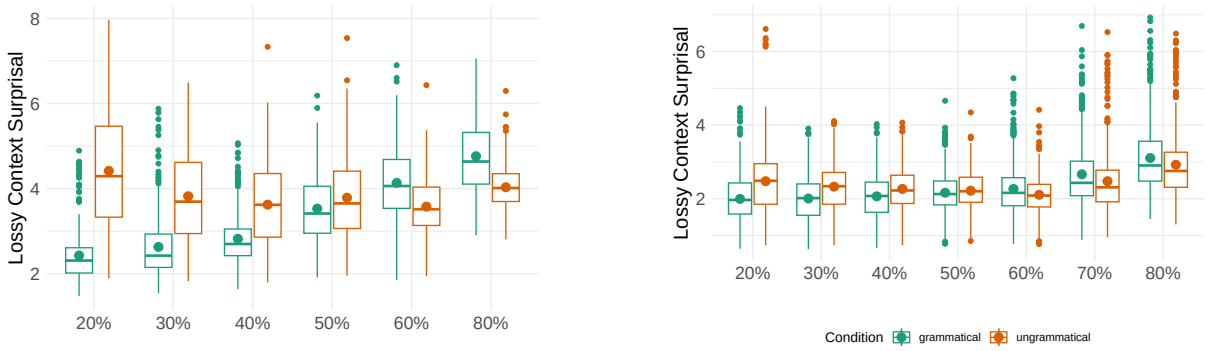
models prefer grammatical sentences, while larger models prefer ungrammatical sentences. In German and Dutch, however, we see that model size may be a misleading measure. A clearer relationship emerges with training data size: the more Dutch or German data an LM trains on, the stronger its preference for grammatical sentences. This outcome appears compatible with the language statistics hypothesis once again — it seems that all models develop stronger human-like language-specific preferences with increased exposure to data from the relevant language. For German, however, we still have the core mystery of how any LM learns the ungrammatical preference in the first place, given that this preference has never been found in experiments with German speakers.

Finally, we conduct a comparative statistical analysis using the reading time (RT) data released by Frank et al. (2016) (Figure 2). For each LM, we fit a linear mixed effects model[6] to asses how well its surprisal value predicts reading times on critical regions in Dutch and English. Model fit was compared using Akaike's An Information Criterion (AIC). In Dutch, all LMs showed roughly the same goodness of fit, within a limited AIC range. For Dutch speakers reading English, bigger models were better, with both LLaMA models and EuroLLM showing a relative reduction of roughly 5 units.

4.2 Resource-Rational Lossy Context Surprisal

Our aim with RR-LCS is to test whether simulating lossy memory can also induce the grammaticality illusion for an LM that does not have it at the outset. Moreover, we expect RR-LCS to capture language-specific effects. Futrell et al. (2020) demonstrate that LCS can handle such effects in principle, by showing that the model predicts different directional outcomes for English and German — specifically, the grammaticality illusion arises at a relatively low forgetting rate for English, while

[6]Formula: `RT ~ Surprisal + (1|Subject) + (1|Item)`



(a) RR-LCS in German, base model BLOOM.

(b) RR-LCS in English, base model GPT2-L.

Figure 6: Grammaticality effects for resource-rational lossy context surprisal (RR-LCS; Hahn et al., 2022) at a range of forgetting rates for German and English. While the magnitude of the grammaticality effect differs across languages, both models switch from preferring grammatical to ungrammatical sentences at 60% forgetting.

the same forgetting rate for German predicts the opposite effect. Although the previous experiment yielded some unexpected outcomes in both German and English, our selected base LMs — Bloom in German, and GPT2-L in English — prefer grammatical sentences from the start. Thus, we expect that increasing the forgetting rate under RR-LCS will cause this preference to change, and we expect that this change will occur at an earlier forgetting rate for English than for German.

We find that increasing the forgetting rate in RR-LCS successfully yields the grammaticality illusion for both English and German (Figure 6). Unexpectedly, the illusion arises at the same forgetting rate for both languages, even though they start from very different points. The German model strongly prefers grammatical sentences for all forgetting rates up to 50%, then at 60% switches to favoring the missing verb condition. The English GPT2-L model also prefers grammatical sentences at lower forgetting rates, although the preference is weaker — then, as for the German model, at a 60% forgetting rate it switches to preferring the ungrammatical construction.

This outcome is somewhat challenging to interpret. On the one hand, it is consistent with the perspective that general memory limitations can drive the grammaticality illusion. Previous computational work has demonstrated this broad conclusion (Futrell and Levy, 2017; Futrell et al., 2020), but relied upon simulated datasets based on corpus statistics. To the best of our knowledge, our work is the first to show that data-driven approximations to noisy contexts can also produce the grammaticality illusion with surprisal calculations from modern neural language models. Neural models trained on the language modeling objective mirror human language processing in many respects, suggesting possible broader cognitive implications.

On the other hand, this outcome does not fully reproduce a key modeling aim of LCS as presented by Futrell et al. (2020), which is to account for how language-specific effects in different directions can arise under the same memory constraints. Our RR-LCS model captures language-specific differences in effect *magnitude*, as the preference for grammatical sentences is weaker in English at lower forgetting rates; however, it does not reflect the directional interaction. This could, however, reflect a technical failure on our part, or other limitations that may be resolved in future work.

5 Conclusions

In this work, we investigated whether the grammaticality illusion seen in behavioral studies of English speakers reflects language-specific statistics or memory limitations. Using Dutch, German, and English LMs, we found evidence for both. Dutch results match the language statistics account, and large English models reproduce the illusion — but mid-sized English and large German models show the reverse, which the statistics account cannot explain. Resource-Rational Lossy Context Surprisal produces the illusion at high forgetting rates in English and German, but misses a key language-specific difference in effect direction. While neither factor fully captures the relevant human patterns, we find that both influence how language models process complex sentences.

Limitations

One key aspect missing from our analysis is a lossy context model trained on Dutch. We did not anticipate that Dutch models would show such a robust grammaticality preference relative to German models, such that Dutch provides a better test case for the interaction of language statistics with RR-LCS. The time and computational cost of training RR-LCS models prevented us from conducting this analysis.

Another limitation of the paper is its focus on only Germanic languages, when similar grammaticality illusions have been found for a typologically diverse range of languages, such as Spanish, French, Mandarin, and Korean. Our analysis focuses on languages for which a range pre-trained language models are available, but this criterion likely reflects and reinforces broader inequalities in which languages are researched.

Acknowledgments

The first author is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Suhas Arehalli and Tal Linzen. 2024. Neural Networks as Cognitive Models of the Processing of Syntactic Constraints. *Open Mind*, 8:558–614.

Markus Bader. 2016. Complex center embedding in German – The effect of sentence position. In Sam Featherston and Yannick Versley, editors, *Quantitative Approaches to Grammar and Grammatical Change: Perspectives from Germanic*, pages 9–32. De Gruyter Mouton.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. BERT or GPT: why not both? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Morten H Christiansen and Nick Chater. 1999. Toward a Connectionist Model of Recursion in Human Linguistic Performance. *Cognitive Science*, 23(2):157–205.

Morten H. Christiansen and Maryellen C. MacDonald. 2009. A Usage-Based Approach to Recursion in Sentence Processing. *Language Learning*, 59(s1):126–161.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The Limitations of Large Language Models for Understanding Human Language and Cognition. *Open Mind*, 8:1058–1083.

Wietse de Vries and Malvina Nissim. 2020. As good as new. how to successfully recycle english gpt-2 to make models for other languages. *Preprint*, arXiv:2012.05628.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, and 69 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

Felix Engelmann and Shravan Vasishth. 2009. Processing grammatical and ungrammatical center embeddings in English and German: A computational model. In *Proceedings of the Ninth International Conference on Cognitive Modeling*, Manchester, UK.

Erwan Fagnou. 2024. Gpt-2 mini. `https://huggingface.co/erwanf/gpt2-mini`.

Stefan Frank. 2014. Modelling reading times in bilingual sentence comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36. Issue: 36.

Stefan L. Frank and Patty Ernst. 2019. Judgements about double-embedded relative clauses differ between languages. *Psychological Research*, 83(7):1581–1593.

Stefan L. Frank, Patty Ernst, Robin L. Thompson, and Rein Cozijn. 2021. The missing-VP effect in readers of English as a second language. *Memory & Cognition*, 49(6):1204–1219.

Stefan L. Frank, Thijs Trompenaars, and Shravan Vasishth. 2016. Cross-Linguistic Differences in Processing Double-Embedded Relative Clauses: Working-Memory Constraints or Language Statistics? *Cognitive Science*, 40(3):554–578. Publisher: John Wiley & Sons, Ltd.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, 44(3):e12814. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12814.

Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain. Association for Computational Linguistics.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward Gibson and James Thomas. 1999. Memory Limitations and Structural Forgetting: The Perception of Complex Ungrammatical Sentences as Grammatical. *Language and Cognitive Processes*, 14(3):225–248. Publisher: Routledge _eprint: https://doi.org/10.1080/016909699386293.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119. Publisher: Proceedings of the National Academy of Sciences.

Helen W. Hamilton and James Deese. 1971. Comprehensibility and subject-verb relations in complex sentences. *Journal of Verbal Learning and Verbal Behavior*, 10(2):163–170.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Nick Huang and Colin Phillips. 2021. When missing NPs make double center-embedding sentences acceptable. *Glossa: a journal of general linguistics*, 6(1). Publisher: Open Library of the Humanities.

Jana Häussler and Markus Bader. 2015. An interference account of the missing-VP effect. *Frontiers in Psychology*, 6. Publisher: Frontiers.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540. Publisher: Elsevier.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Kate McCurdy and Michael Hahn. 2024. Lossy Context Surprisal Predicts Task-Dependent Patterns in Relative Clause Processing. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 36–45, Miami, FL, USA. Association for Computational Linguistics.

Benjamin Minixhofer. 2020. GerPT2: German large and small versions of GPT2.

Byung-Doh Oh and William Schuler. 2023. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency Explains the Inverse Correlation of Large Language Models' Size, Training Data Amount, and Surprisal's Fit to Reading Times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Malte Ostendorff and Georg Rehm. 2023. Efficient Language Model Training through Cross-Lingual and Progressive Transfer Learning. *arXiv preprint*. ArXiv:2301.09626 [cs].

Claudia Pañeda and Sol Lago. 2024. The Missing VP Illusion in Spanish: Assessing the Role of Language Statistics and Working Memory. *Open Mind*, 8:42–66.

Colin Phillips, Matthew W. Wagers, and Ellen F. Lau. 2011. 5 Grammatical Illusions and Selective Fallibility in Real-Time Language Comprehension. In *Syntax and Semantics*, volume 37, pages 147–180. Emerald Group Publishing, Bingley.

Björn Plüster and Christoph Schuhmann. 2023. LeoLM: Igniting German-Language LLM Research | LAION.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422. Place: US Publisher: American Psychological Association.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-Shot Learners Go Multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79. Place: Cambridge, MA Publisher: MIT Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. ArXiv:2307.09288 [cs].

Bram Vanroy. 2023. Language resources for Dutch large language modelling. *arXiv preprint arXiv:2312.12852*.

Shravan Vasishth, Katja Suckow, Richard L. Lewis, and Sabine Kern. 2010. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4):533–567.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Ethan Gotlieb Wilcox, Michael Y. Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

Yuhan Zhang, Edward Gibson, and Forrest Davis. 2023. Can Language Models Be Tricked by Language Illusions? Easier with Syntax, Harder with Semantics. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 1–14, Singapore. Association for Computational Linguistics.