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

Kate McCurdy, Katharina Christian, Amelie Seyfried, and Mikhail Sonkin

Introduction & Research Question

The Grammaticality Illusion:
Speakers of **English** — but not **German** or **Dutch** — find ungrammatical double-center-embedded sentences *easier to process* than grammatical ones:
Grammatical: The painter who the doctor who the lady saw **treated** ran a marathon.
Ungrammatical: *The painter who the doctor who the lady saw ran a marathon.

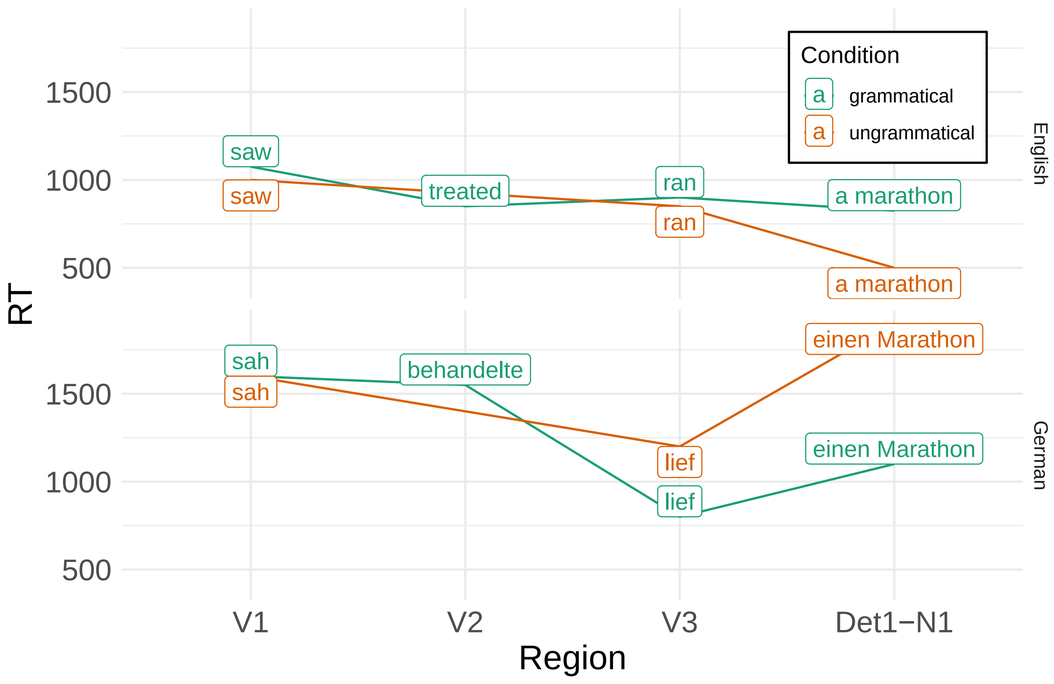

Fig. 1: Grammaticality illusion in reading time (RT) in English (upper), and converse in German (lower), using mean RTs from Vasishth et al. [1].

Competing hypotheses:
1. **Language statistics**: English double center-embeddings are rare.
2. **Memory constraints**: Earlier subject is forgotten.

Research question: Can modern language models (LMs) help us understand which account is correct?

Methods

Language Models Evaluated:
- **Dutch:** GPT2-S/M, LLaMA2 (33B-2B tokens training)
- **German:** GerPT2-L, BLOOM, LEO-LM (50B-65B tokens)
- **English:** GPT-BERT to LLaMA2 (100M-2T tokens)
- **Multilingual:** mGPT, LLaMAX, EuroLLM

Measures:
1. **Surprisal:** Negative log probability of target word given context $-\log P(w_{T+1}|w_{1\ldots T})$
2. **Lossy Context Surprisal (RR-LCS) [2]:** Stochastically deletes/reconstructs context words at various forgetting rates (20%-80%) $-\log P(w_{T+1}|M_T)$ where M_T is lossy representation

Stimuli: Double-center-embedded sentences in Dutch, German, and English [1,3] with grammatical vs. ungrammatical (missing verb) conditions
Critical region: Post-verbal determiner (where RT effects are strongest)

Conclusions

Partial support for both **language statistics** (training data effect in Dutch and English; BUT: why would German models ever show the illusion?) and **memory constraints** (higher forgetting rate causes grammaticality illusion in both German and English; BUT: illusion arises at the same forgetting rate, when we expect language-specific interaction).

Take-home message: The grammaticality illusion in LMs can be driven by language statistics *or* memory limitations.

Results: Language Model Surprisal

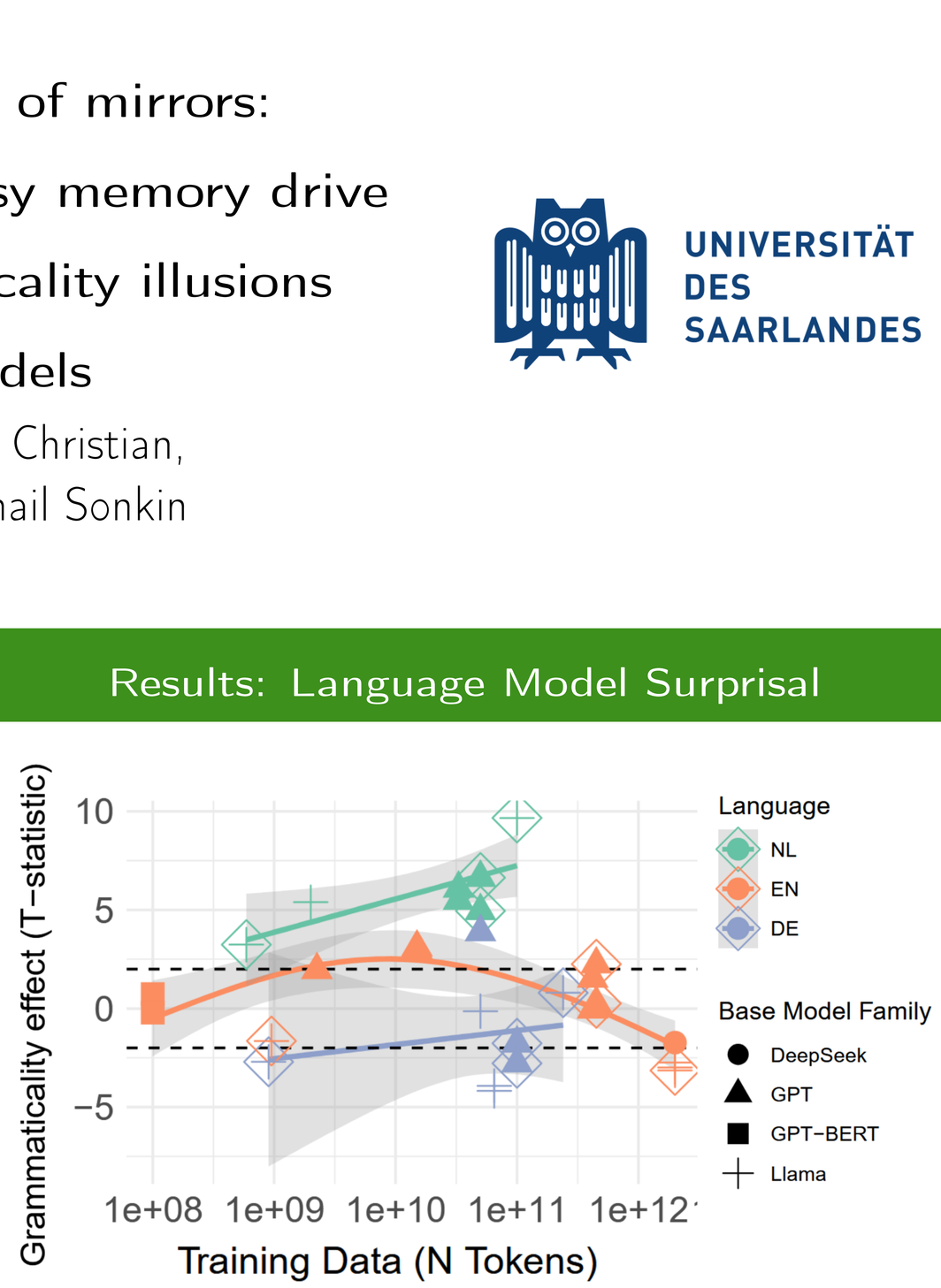

Fig. 2: LM grammaticality effect (t-statistic; positive = prefer grammatical; negative = prefer ungrammatical, i.e. illusion) by language and training data size.

Result summary by language. L, M, S refer to model size (in parameters). "Gramm. illusion?" = "Do LMs from this language prefer ungrammatical sentences?" "Human-like?" = "Does this preference match behavioral RT patterns from this language's speakers?"

	NL	DE	EN
Gramm. illusion?	No	Most models L, not M models	
Human-like?	Yes	No	L, yes; M, no

Key finding: Training data size predicts effect strength
- More data → stronger language-specific preferences
- Dutch/German: increasingly prefer grammatical
- English: increasingly prefer ungrammatical

Results: Lossy Context Surprisal

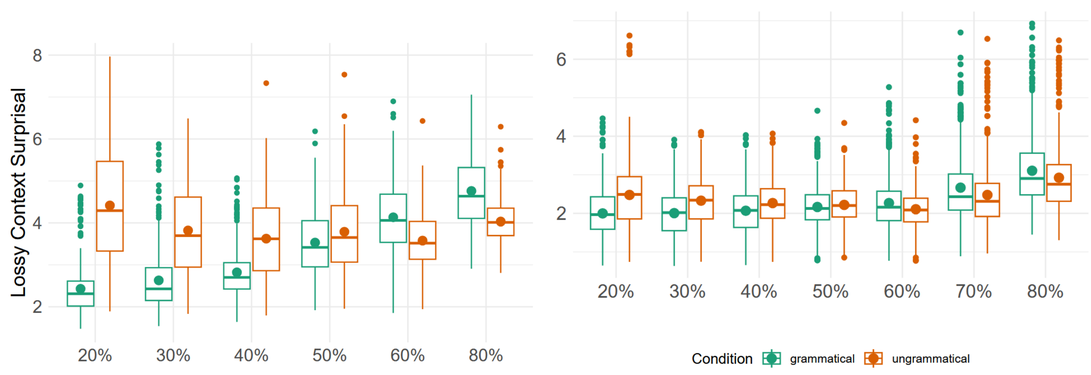

Fig. 3: LCS for grammatical (green) vs. ungrammatical (orange) in German (left; base LM BLOOM) and English (right; base LM GPT2-L).

Key findings:
- Both models switch to prefer ungrammatical sentences at 60% forgetting rate
- Captures language-specific magnitude but not directional differences
- Interpretation: Memory limitations can produce illusion, but do not explain cross-linguistic variation

References

[1] Vasishth, Suckow, Lewis, Kern (2010) LangCogProc [2] Hahn, Futrell, Levy, Gibson (2022) PNAS [3] Frank, Trompenaars, Vasishth (2016) CogSci