Orthographic diversity in Large Language Models - a Case Study of Foreign Word Spelling in German

Christian Lang,
Marco Gierke,
Ngoc Duyen Tanja Tu
Leibniz-Institut für Deutsche Sprache,
Mannheim, Germany
lang@ids-mannheim.de, gierke@ids-mannheim.de, tu@ids-mannheim.de

Abstract

This study examines whether large language models (LLMs) reduce orthographic variation in AI-generated German texts. Using 48 foreign words known for variation between *f/ph* and *ee/é*, we analyze model preferences via surprisal differences and generation tasks across several LLMs. Comparing these results to corpus data, we explore whether LLM outputs reflect or diminish linguistic variability at the word level.

1 Introduction

The use of large language models (LLMs) for the production of online content—such as newspaper articles, blog posts, or social media contributions—is steadily increasing.[1] From a linguistic perspective, this raises the question of whether the increasing presence of AI-generated texts is influencing human language use.

Given the basic functionality of LLMs—where, based on an input, the next token is selected according to a probability distribution—the question emerges as to how this statistical selection process affects the inherent linguistic variability of human language. Building on this consideration, Sourati et al. (2025) demonstrate that automatically generated texts in English tend to exhibit greater stylistic homogeneity compared to human-authored texts.

2 Research Question

In this study, we examine whether a comparable trend toward homogenization can be observed at the word level. Specifically, we investigate the extent to which foreign words in German—shown in corpus studies to exhibit varying degrees of *orthographic variation* —display similar variance patterns when generated by AI.

To address this research question, we analyzed the spelling of 48 foreign words in German using various large language models (LLMs). The foreign words belong to one of two orthographic problem areas: Variation between *f* and *ph* (as in *Kalligrafie vs. Kalligraphie* [*calligraphy*]) and variation between *ee* and *é* (as in *Dekolletee vs. Dekolleté* [*decolleté*]). Not all of the variants examined correspond to the orthographic rules of German, but they can be observed in language use. The selection of these words is based on a corpus study conducted as part of orthographic research for the *Rat für deutsche Rechtschreibung* (Council for German Orthography), the results of which are available at `https://korap.ids-mannheim.de/data/variantenbeobachtung/ee_e.html`.[2] This corpus study serves as a reference against which we compare the results of our LLM analysis. In doing so, we aim to investigate orthographic variation in human vs. AI-generated texts.

3 Methodology

We conduct two experiments: (1) calculating surprisal values for orthographic variants, and (2) generating words with varying degrees of orthographic variation observed in the reference corpus, using (2a) cloze tests and (2b)

[1] cf. `https://de.statista.com/statistik/daten/studie/1498220/umfrage/nutzung-von-chatgpt-durch-journalisten-in-medienhaeusern-in-deutschland/`

[2] Not all items from the original corpus study were included in our study. We excluded particularly rare or unusual forms, as these cannot be reliably examined across all experimental paradigms used in our study.

short article generation.

For experiment 1, we constructed controlled stimulus sentences—following the recommendation by Wilcox, Futrell, and Levy (2024) to apply psycholinguistic methods in linguistic research with LLMs—, each containing one of the 48 target words in an orthographic variant. Using various LLMs, we computed surprisal values (Hale, 2001) for each variant. Larger surprisal gaps indicate a clearer model preference for one variant, suggesting a tendency toward homogeneity; smaller gaps suggest balanced probabilities and greater variation.

Since generation is also affected by hyperparameters like temperature and top_n, we validate the surprisal-based predictions with experiments 2a and 2b, which analyze actual model outputs.

For all experiments, we accessed LLMs via ChatAI (Doosthosseini, Decker, Nolte, & Kunkel, 2024) using OpenAI's chat completion API endpoint.

3.1 Experiment 1: Calculating surprisal gaps between orthographic variations

We constructed stimuli to contrast orthographic variants of our target words. The stimuli were designed to be compatible with both computational experiments and potential elicitation studies involving human participants. All stimuli followed the same structure, consisting of an introductory sentence that established the topic, followed by a second sentence containing the target word in one of its orthographic variations (see the following example).

- **orthographic variation 1:** Das Model trägt ein weit ausgeschnittenes Kleid. In ihrem **Dekolletee** glänzt eine goldene Kette.
- **orthographic variation 2:** Das Model trägt ein weit ausgeschnittenes Kleid. In ihrem **Dekolleté** glänzt eine goldene Kette.
- **translation:** *The model is wearing a wide-cut dress. A golden necklace shines in her **décolleté**.*

For both orthographic variations, we calculated surprisal on the target by summing up the logprobs, i.e. the logarithmized relative probabilities given the context, of all its subtokens and multiplying the result by -1. We then calculated the absolute difference between the surprisal values of the orthographic variations (surprisal gaps). We used the following six LLMs: *deepseek-r1*, *mistral-large-instruct*, *qwen3-32b*, *meta-llama-3.1-8b-instruct*, *llama-3.3-70b-instruct*, *llama-3.1-sauerkrautlm-70b-instruct*.

3.2 Experiment 2: Generating orthographic variations

In order to examine how surprisal differences between orthographic variants translate into actual text generation by LLMs, we prompted different LLMs to generate the target words and counted the resulting orthographic variants. For each stimulus and each iteration, we initiated a new API call to ensure that previous outputs were not included in the model's context to avoid any accumulation of conversational history that could bias the results.

This part of our study is still in the piloting phase; at this stage, we have tested only the orthographic variations of four words using two LLMs (*llama-3.3-70b-instruct* and *llama-3.1-sauerkrautlm-70b-instruct*) with only one prompt. The selection of these four words is based on the results of experiment 1 (see section 4): Two of these words exhibit particularly pronounced surprisal gaps (*Frappee* vs. *Frappé* [*frappee*], *Grafik* vs. *Graphik* [*graphic*]), while the other two show very small surprisal gaps (*Dekolletee* vs. *Dekolleté* [*décolleté*], *Kalligrafie* vs. *Kalligraphie* [*calligraphy*]). We plan to extend this analysis to the remaining 44 words and additional models.[3] For all generation tasks, we kept the default settings for the *temperature* and *top_n* hyperparameters of the LLMs.

Since many of our target words are low-frequency and thus have a low baseline probability of being generated, we applied two techniques to elicit these words from LLMs: 1) We constructed cloze sentences based on the stimulus material described above. Instead of presenting orthographic variants, these sentences contain only the first letter of the target word.[4] We then prompted the LLMs to fill the gap. To further increase the likelihood of generating the target word, the prompt includes a list of 'banned' words that the model is instructed not to use.[5] So far, we have presented each of the four cloze sentences 100 times to the two models for completion. 2) We instructed various LLMs to generate short articles on specific topics. The topics are selected to maximize the

[3] Not all models used in experiment 1 are suitable for experiment 2, as some are designed primarily for reasoning tasks.

[4] This excludes orthographic variants that differ in their first letter.

[5] System prompt: *You are a gap-filling assistant for German texts. A word is missing in the sentence. Your task is to suggest the missing word. IMPORTANT RULES: (1) The word must begin with the same letter as the word immediately before the gap. (2) Your response must consist ONLY of the missing word — no explanations, no extra text. Example: Sentence: Es zwitschert im Baum. Der V_ baut sein Nest. Answer: Vogel.*; User prompt: *Fill in the blank: 'Das Model trägt ein weit ausgeschnittenes Kleid. In ihrem D_ glänzt eine goldene Kette.' IMPORTANT: Do not use any of the following words:* – optional list of banned words –. The prompt was in English –a possible reason for some of the orthographic variations generated– and only the example in German.

likelihood that the target words will appear in the generated text.[6] So far, we have generated a total of 50 texts with both LLMs for each of the four topics.

4 Preliminary Findings

Fig. 1 and Fig. 2 show the results of experiment 1 differentiated by orthographic problem area (Fig. 1 for *ee* vs. *é*; Fig. 2 for *f* vs. *ph*). The figures are divided into two sections based on the dominant orthographic variation in the corpus study we use as reference (see section 2). For each word, the title indicates the percentage of the dominant orthographic variation in the corpus. The two orthographic variations are coded with different colors (red for the Germanized variant, blue for the foreign-language variant). The bars show the size of the surprisal gap. Orthographically unlicensed variants are marked with an asterisk.

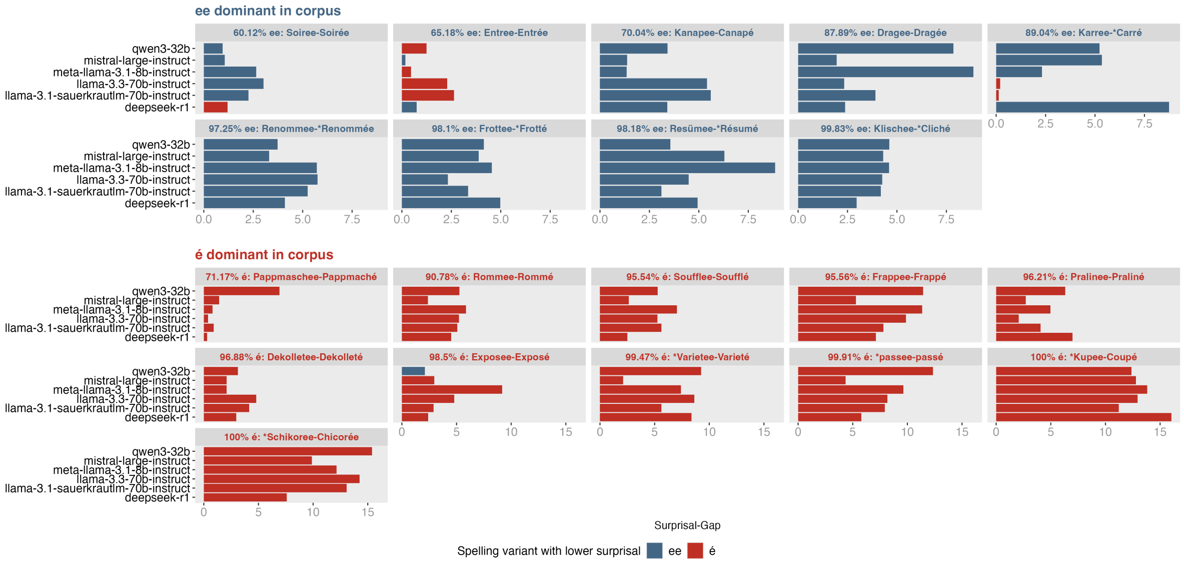

Figure 1: Surprisal gaps between orthographic variations of the orthographic problem area *ee* vs. *é*.

Fig. 1 and Fig. 2 illustrate clear differences both between the orthographic problem areas and across the LLMs. In the *ee* vs. *é* phenomenon area, the dominant spellings identified in the corpus study largely correspond to those with lower surprisal values in the LLMs, with the exception of *Entree/Entré* [*entree;entrance;fee*] (a highly ambigous word). Moreover, there are indications of a possible connection between corpus dominance and surprisal gap magnitude: higher corpus frequencies of a particular orthographic variant may correspond to greater absolute differences in surprisal values assigned by the LLMs. In other words (and bearing in mind the differences between the models), the potential for orthographic variation in AI-based generation corresponds at first glance at least partially to that of human language production in this phenomenon area. The situation differs in the *f* vs. *ph* phenomenon area. Although the corpus study shows that, with two exceptions, the *f* variant is more frequent, the LLMs often assign lower surprisal values to the *ph* spelling. Furthermore, the surprisal gaps in this area are generally more heterogeneous, both in direction and magnitude, compared to the *ee* vs. *é* area.

Preliminary generation results across both paradigms of experiment 2 indicate that orthographic variation occurs in some cases but remains generally infrequent. Variation was observed for *Dekolletee* vs. *Dekolleté* and *Frappee* vs. *Frappé*, although often with incorrect forms such as *Decolleté* or *Frappe*. No variation was found for *Kalligraphie* vs. *Kalligrafie* or *Grafik* vs. *Graphik*. These findings suggest that surprisal differences cannot be directly translated to generative behavior, and that the specific orthographic phenomenon plays a role. The models also differ: *LLaMA 3.1 Instruct* predominantly generates the low-surprisal variant, whereas the *Sauerkraut* model produces both variants as well as additional, incorrect spellings.

[6] Example prompt for eliciting the word *Dekolleté*: ‘Schreibe einen kurzen Text über elegante Damen-Mode. Gehe dabei besonders auf Oberteile und den besonderen Reiz freiliegender Hautpartien am Oberkörper ein.’ [‘*Write a short text about elegant women’s fashion. Pay particular attention to tops and the special appeal of exposed skin on the upper body.*’]

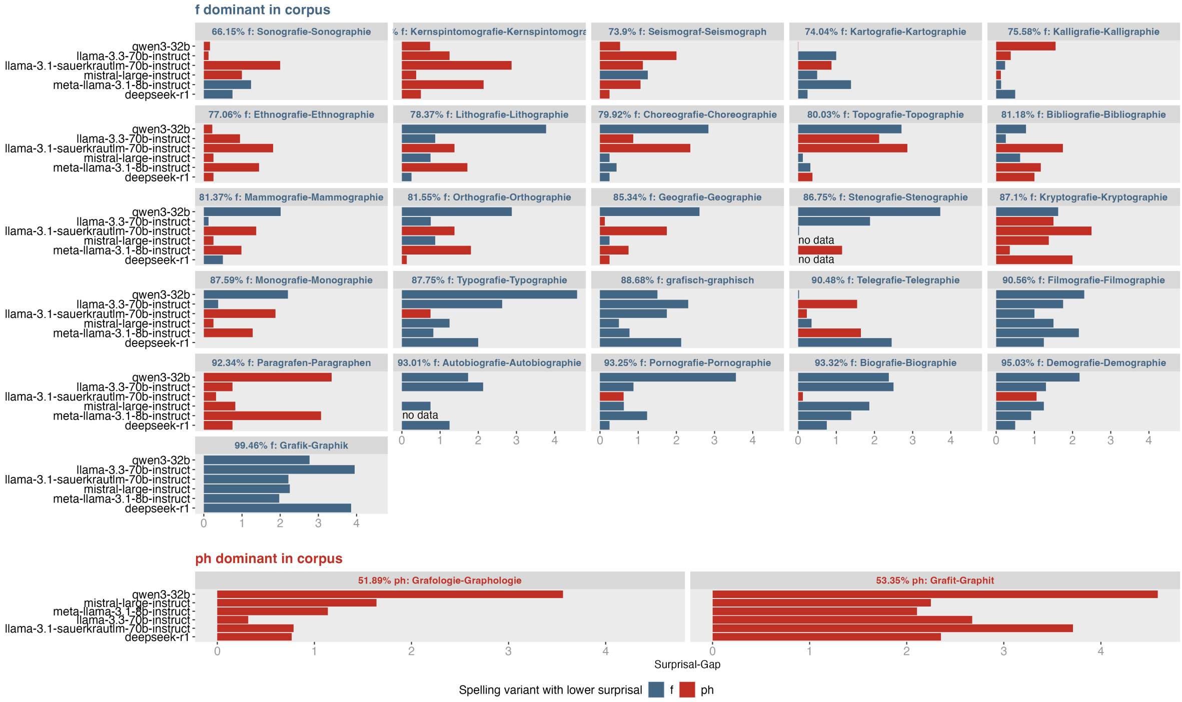

Figure 2: Surprisal gaps between orthographic variations of the orthographic problem area *f* vs. *ph*.

5 Impact

To understand how LLM-generated texts may influence orthographic conventions, their effects can be examined across three levels: the micro level of individual writing, the meso level of usage consolidation in model texts, and the macro level of codification. At the *micro level*, orthographic variation often occurs in transitional zones of the writing system, such as the integration of foreign words. Even when such variants are semantically equivalent, they can carry pragmatic or indexical meaning (cf. Sebba 2007, p. 32). Brommer and Frick (2023) illustrate this with orthographic variants for the Ukrainian capital. Individual choices may, at the *meso level*, be aggregated in writing practices, particularly in mass-reach press texts. As outlined in Ammon (1995)'s model of the *social force field of a standard variety*, such *model texts* shape shared expectations of correctness. If co-produced by LLMs, these texts may lend normative power to AI-generated variants. At the *macro level*, codification bodies such as the Council for German Orthography base their decisions on corpus data, which also draw from press texts (cf. Diewald, Gierke, Kupietz, and Lüngen 2024). Across all three levels, LLMs may influence the usage-based dynamics of orthographic variation and standardization. While many anticipate a homogenizing effect, our findings point in the opposite direction: some LLMs frequently produce rare variants. If this pattern persists at scale, it could increase orthographic variation—not despite AI, but because of it.

References

Ammon, U. (1995). *Die deutsche Sprache in Deutschland, Österreich und der Schweiz: Das Problem der nationalen Varietäten*. Berlin: De Gruyter. doi: 10.1515/9783110872170

Brommer, S., & Frick, K. (2023). Kiew, Kyiv oder Kyjiw? Positionierung durch Begriffsverwendung in der schweizerischen, bundesdeutschen und österreichischen Berichterstattung zum Ukrainekrieg. *Aptum*, *19*(2+3), 262–271. doi: 10.46771/9783967693843_18

Diewald, N., Gierke, M., Kupietz, M., & Lüngen, H. (2024). Das Orthografische Kernkorpus (OKK) in DeReKo: Zusammensetzung, Analyse- und Zugriffsmöglichkeiten über KorAP. In S. Krome, M. Habermann, H. Lobin, & A. Wöllstein (Eds.), *Orthographie in Wissenschaft und Gesellschaft* (pp. 329–344). Berlin, Boston: De Gruyter. doi: 10.1515/9783111389219-017

Doosthosseini, A., Decker, J., Nolte, H., & Kunkel, J. M. (2024). *Chat AI: A Seamless Slurm-Native Solution for HPC-Based Services.* Retrieved from `https://arxiv.org/abs/2407.00110`

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics.* Retrieved from `https://`

aclanthology.org/N01-1021/

Sebba, M. (2007). *Spelling and Society: The Culture and Politics of Orthography around the World.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511486739

Sourati, Z., Karimi-Malekabadi, F., Ozcan, M., McDaniel, C., Ziabari, A., Trager, J., … Dehghani, M. (2025). *The Shrinking Landscape of Linguistic Diversity in the Age of Large Language Models.* Retrieved from https://arxiv.org/abs/2502.11266

Wilcox, E. G., Futrell, R., & Levy, R. (2024, 10). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, *55*(4), 805-848. Retrieved from https://doi.org/10.1162/ling_a_00491 doi: 10.1162/ling_a_00491