

Human and Machine Word Associations in Slovenian: A Comparison of Associative Behaviour and Its Explanations

Klara Vrabl
University of Ljubljana
klara.vrabl@gmail.com

Mojca Brglez
University of Ljubljana
Jožef Stefan Institute
mojca.brglez@ff.uni-lj.si
mojca.brglez@ijs.si

Špela Vintar
University of Ljubljana
Jožef Stefan Institute
spela.vintar@ff.uni-lj.si
spela.vintar@ijs.si

Abstract

Word associations are one of the most established psycholinguistic methods, offering a long-standing window onto how meaning is organised in the human mental lexicon, yet it remains unclear whether large language models (LLMs) represent and structure associative word meaning in comparable ways, particularly outside English. We present WAX-SL, the first Slovenian dataset of cue–association–explanation triplets annotated with semantic relation labels, and use it to compare human associative behaviour against parallel data generated by three open-weights LLMs. All four sources are annotated with a 20-label relational taxonomy, augmented with a validity marker for output that fails the task constraints. Results show human associations are semantically more diverse and spread across all relation types, drawing more often on action, common phrases, and emotional evaluation, whereas LLM outputs concentrate on structurally transparent relations such as causal and definitional links. The differences are also qualitative: model outputs exhibit instruction violations, semantically incoherent constructions, and hallucinated content. Inter-annotator agreement is markedly lower for machine output ($\kappa = 0.32$) than for human data ($\kappa = 0.46$), reflecting the greater ambiguity of model responses. Language-specific adaptation does not straightforwardly yield more human-like behaviour. We argue that, on this task, LLMs do not echo human associative behaviour but diverge from it in systematic ways, both in the relations they express and in their markedly higher rate of invalid output.

1 Introduction

“What are the first three words you think of when you hear the word *computer*?” A person performing a free-association task might answer “*desk, work, game*”. The method places no restriction on the

Cue	Assoc. (P1)	Explanation (P2)
<i>bogat</i> (rich)	<i>debel</i> (fat)	<i>Vsaj v starih časih je bil debela lahko le bogata oseba.</i> ‘At least in the old days, only a rich person could be fat.’
	<i>zlato</i> (gold)	<i>Če imaš veliko zlata, si bogat.</i> ‘If you have a lot of gold, you are rich.’
	<i>zloben</i> (mean)	<i>Veliko bogatih ljudi so ali postanejo zlobni.</i> ‘Many rich people are or become mean.’

Table 1: Illustrative cue–association–explanation triplets for one cue, showing associations collected in phase 1 (P1), and explanations collected in phase 2 (P2), with English glosses.

response. It dates back to [Galton \(1879\)](#), who first used it to reveal spontaneous mental links, and was later taken up by [Freud \(1913\)](#) to probe emotional response; the first large-scale standardised collection came with [Kent and Rosanoff \(1910\)](#). It has since grown into a substantial research paradigm, producing large-scale datasets that give insight into the structure of the human mental lexicon across languages. Because the responses are unconstrained and spontaneous, they reveal how concepts are connected in memory — which words are close to which, and along what kinds of relations — offering a behavioural window onto semantic organisation that is hard to obtain by other means ([Clark, 1970](#); [Fitzpatrick, 2007](#)).

The emergence of large language models (LLMs) raises a question that is central to language understanding (LU): to what extent do these models replicate the associative behaviour of human speakers, and where do they diverge? The question is particularly pressing for languages other than English, where both training data and evaluation resources are limited and where assumptions underlying English-language research cannot be taken

for granted. A model may produce fluent, plausible associations while organising word meaning in a way that bears little resemblance to human cognition — a gap that fluency alone conceals.

This study sits at the intersection of these concerns. It replicates the WAX (Word Association eXplanations) framework (Liu et al., 2022) for Slovenian, a morphologically rich Slavic language with roughly two million speakers. WAX collects word associations together with short explanatory sentences and annotates each cue–association pair with a *semantic relation label*: a category, drawn from a fixed taxonomy, that names how the two words are connected (for example, *cat* → *purr* as an *Action*). Slovenian is a useful test case: well-resourced enough to make LLM generation feasible, yet sufficiently distinct from English, both linguistically and culturally, that cross-linguistic transfer is an open empirical question rather than a safe assumption. Beyond collecting human data, we generate parallel data with three LLMs and compare the two sources across lexical diversity, explanation length, and the distribution of semantic relations, with particular attention to where machine output exhibits erratic behaviour.

Crucially, the explanation layer lets us probe not only *what* associations are produced but *why* — whether the reasoning a system offers for a cue–association pair is transparent, classifiable, and human-interpretable. This connects word association to broader questions in LU about whether systems that generate linguistically plausible output also support the kind of context-sensitive, communicable reasoning that characterises human explanation (Miller, 2019).

We address three research questions:

1. To what extent do LLM-generated associations and explanations in Slovenian resemble human ones in terms of lexical diversity, explanation length, and semantic relation distribution?
2. Do the semantic relation types identified in English WAX transfer to Slovenian human data, and what cross-linguistic differences emerge?
3. What kinds of erratic or invalid output do LLMs produce on a structured association task in Slovenian, and how do these differ across models?

We formulate three hypotheses, motivated in Section 2:

H1. The semantic relations assigned to human associations will be more diverse and more broadly distributed than those assigned to LLM associations, reflecting the wider relational knowledge that characterises human associative behaviour.

H2. The degree to which model output aligns with human data will vary across models and relate to how far each model has been trained or adapted for Slovenian.

H3. Systematic differences will be evident between the Slovenian human relation distribution and the English distribution reported by Liu et al. (2022), reflecting linguistic and cultural factors that shape associative behaviour.

Our contributions are (i) WAX-SL, the first Slovenian cue–association–explanation dataset with relation annotation;¹ (ii) a systematic comparison of human and LLM associative behaviour in a less-resourced language; and (iii) an analysis of LLM erratic behaviour that bears on what it means for these systems to represent the associative structure of word meaning.

2 Background

Word associations and the mental lexicon.

Clark (1970) argued that associations are not arbitrary but governed by systematic rules operating on syntactic and semantic features, distinguishing paradigmatic responses (same syntactic category, e.g. *dog* → *cat*) from syntagmatic ones (*dog* → *bark*). Fitzpatrick (2007) challenged the assumption that adult associative behaviour is homogeneous, showing substantial individual variation and proposing four response categories: meaning-based, position-based, form-based, and erratic. This qualitative diversity, where no single category dominates, is a substantive feature of human associative competence rather than noise. Large-scale data collection runs from Kent and Rosanoff (1910) through the USF norms (Nelson et al., 2004) to the Small World of Words project (De Deyne et al., 2019), recently extended to Slovenian as SWOW-SL (Vintar et al., 2025).

Associations in NLP. Distributional models have long been evaluated against association data. Mandera et al. (2017) found prediction-based models fit human behaviour better than count-based ones,

¹Available at <http://hdl.handle.net/11356/2358> (Vrabl et al., 2026).

while Nematzadeh et al. (2017) showed vector-space models fail to reproduce asymmetric patterns in human responses, and Günther et al. (2019) cautioned against treating such models as cognitive theories. For Slovenian specifically, Brglez et al. (2024) found low overlap between human and model associations and a disproportionate number of erratic model responses; Vintar and Javoršek (2025) found LLMs amplify emotional polarity and are less varied than humans; and Vintar and Sakšida (2023) showed the vector space of models captures little of the knowledge frames underlying human associations. Together, these suggest current models perform the task at a basic level without reproducing human richness, variability, and creativity.

Explanability. WAX (Liu et al., 2022) extended the paradigm by asking annotators to provide, alongside each association, a short explanation of why they linked the cue–association pair. Each pair is then assigned, as a separate annotation layer, a relation label from a taxonomy of 16 types in five coarse groups (Concept-Properties, Situational, Taxonomic, Linguistic, None-of-the-Above). For instance, the pair *cat* → *purr* might be explained as “a cat purrs when it is content” and labelled *Action*, while *winter* → *cold* might be explained as “winter is a cold season” and labelled *HasProperty*. This makes the reasoning behind a human association visible and classifiable. When the same pipeline is run on a language model, the explanation it produces becomes a form of self-explanation, and assessing it shifts the task from collecting associative data to evaluating machine explainability: do models account for their own associations in ways that are transparent and classifiable, or does fluent output mask a different underlying organisation?

Work on interpretability distinguishes Explainable AI (XAI), which makes model behaviour inspectable, from Explanatory AI (YAI), which, following Miller (2019), treats explanation as selective, contrastive, and social rather than a complete causal account (Sovrano and Vitali, 2024). Language models are well-suited to producing such fluent, audience-adaptive explanations. Whether that fluency corresponds to a relation the model has actually captured is, however, an empirical question, and eliciting an explanation for each pair puts it to the test at the smallest unit of word meaning: not just *what* a model connects, but whether its stated reasoning is transparent and classifiable. Both the

associations and the explanations, however, have so far been studied almost exclusively in English. Whether WAX’s relational and explanatory patterns recur in a typologically distinct, less-resourced language is the open question we address.

3 Data and Methods

Cue words. We built a list of 100 cues from the Reference List of Slovene Frequent Common Words (Holdt et al., 2020), which draws on vocabulary shared across written, spoken, and computer-mediated corpora. Following WAX, cues were sampled randomly and balanced across nouns, verbs, and adjectives; adverbs were excluded, as in the original study.

Human data (WAX-SL). The collection proceeded in two phases with nine Slovenian-speaking annotators. In phase one, each annotator provided three associations for each cue, so that every cue received responses from four annotators. In phase two, annotators wrote a 3–15 word explanation of the relationship between cue and association, required to contain both words (inflected as needed but without part-of-speech changes). This yielded roughly 1,200 cue–association–explanation triplets (see Table 2 for dataset sizes). Table 1 illustrates the two phases for a single cue.

LLM data. We generated parallel data with three open-weights models, deployed locally and accessed via API at the Jožef Stefan Institute: **Gemma 3** (27B instruction-tuned), a Google multilingual model listing Slovenian among supported languages; **GaMS** (27B Instruct), continually pre-trained on Slovenian on a Gemma 2 base; and **Magistral Small** (24B), a reasoning-focused Mistral model trained with Reinforcement Learning from Verifiable Rewards that does *not* list Slovenian among its supported languages. For each model, we obtained three associations and explanations per cue (301 for Gemma, 300 for GaMS and Magistral),² again in two phases. Prompts were written in Slovenian, adapted from the instructions given to human annotators after early testing revealed several LLM-specific failure modes: longer prompts or prompts containing examples led mod-

²One cue (*dokončati*, ‘to finish’) received four generated associations rather than three in the Gemma output, yielding 301 rather than 300 triplets for Gemma (901 LLM triplets in total). This was an error introduced during generation; it was noticed but not corrected, and we are unable to reconstruct why it was left as is. We report the counts as actually collected.

els to reproduce the examples rather than generate original responses, and offering an “unknown word” option led models to mark every cue as unknown. The final prompts were therefore shortened and stripped of examples relative to the human instructions (full prompt text in Appendix B). All models used identical generation parameters, selected through preliminary experimentation and manual evaluation of output quality: temperature 0.5, top- p 0.3, max 120 output tokens, presence penalty 0.5, frequency penalty 1.3, seed 3.

Relation annotation. We used the extended WAX taxonomy of semantic relations from the project repository (Liu, 2022), which provides 20 relations rather than the 16 in the paper. For example, the larger taxonomy splits the *HasProperty* relation into *InherentProperty* and *PerceivableProperty* and additionally contains *LexicalSimilarity* and *SoundSimilarity*. In our experiments, we also added a special ERROR label for cases where the part of speech was not preserved in the explanatory sentence, the explanation was incomplete, or the construction was semantically incoherent, yielding 21 labels in total: the 20 relation types plus ERROR as a validity marker rather than a relation. ERROR sits within the *None-of-the-Above* coarse group, alongside the *Other* relation: *Other* marks a *valid* cue–association pair whose relationship is genuine but does not fit any of the named relation types, whereas ERROR marks output that fails the task constraints and so cannot be assigned a relation at all. The elevated *None-of-the-Above* share we report for models (Section 4) is therefore driven largely by ERROR cases. To make our data comparable to the English relation-labelled portion of WAX, we annotated a random sample of 700 of the 1,200 human triplets as well as all triplets generated by each of the models (301 for Gemma, 300 for GaMS and Magistral; Table 2). Annotation was performed on the INCEpTION platform (Klie et al., 2018); each instance was independently labelled by two annotators and curated by a third. We measured agreement with pairwise Cohen’s κ (Cohen, 1960), computed separately for human and machine data.

Annotator setup. Annotator recruitment, task allocation, and the pilot study used to refine the annotation guidelines are described in Appendix C.

Cross-linguistic mapping. To compare relation distributions across languages on comparable

	Human	LLM
Cues	100	100
Triplets	1,200	901
Relation-annotated	700	901

Table 2: WAX-SL and the parallel LLM datasets. LLM figures are pooled across the three models (301 Gemma, 300 GaMS, 300 Magistral triplets). Triplet counts are cue–association–explanation triples; relation-annotated counts are the subset labelled with the 20-label relation taxonomy (plus the ERROR validity marker).

	Human	Gemma	GaMS	Mag.
TTR	0.85	0.90	0.90	0.91
Expl. length (med.)	6.0	9.0	12.0	7.0
ERROR (%)	1	29	29	55

Table 3: Summary of results by source. Mag. = Magistral. TTR is the proportion of distinct associations among all associations; explanation length is in words.

terms, we mapped the Slovenian 20-label taxonomy onto the 16 relation types used in the published WAX-EN annotation (Liu et al., 2022) (full mapping in Appendix A). Two Slovenian labels have no English counterpart and are excluded, affecting 50 of 693 valid human-labeled Slovenian instances ($\sim 7.2\%$); the resulting comparison uses $n = 726$ English and $n = 643$ Slovenian instances over 17 shared categories.

Measures. We compare the human and model data along three dimensions. *Lexical diversity* is quantified as the type–token ratio (TTR) — the proportion of distinct associations among all associations from a given source — indicating how much of its associative vocabulary is repeated. *Explanation length* is the number of words per explanatory sentence, summarised by its median and interquartile range. *Semantic-relation diversity* is the distribution of associations over the relation taxonomy: we compare how evenly each source spreads across the 20 relation types, both at the fine-grained level and over the five coarse groups, and against the English WAX distribution. Alongside these, we report the share of invalid (ERROR) output per source separately from the relation distribution, and characterise the qualitative error types that distinguish human from machine responses.

4 Results

Vocabulary size. Lexical diversity, measured as the type–token ratio (TTR), initially appears

comparable across sources: 0.85 for human data ($n = 700$, the relation-annotated subsample), 0.90 for Gemma and GaMS, and 0.91 for Magistral ($n = 301$, $n = 300$, $n = 300$ respectively; full generated output; Table 3). This is confounded by unequal sample sizes, since TTR mechanically declines as N grows. Bootstrap subsampling at equal N (300) reverses the pattern: Human’s subsampled TTR (mean 0.927, 95% CI 0.900–0.950) exceeds all three models’ (Gemma 0.904, GaMS 0.900, Magistral 0.907) consistently from $N = 50$ to $N = 300$, though Human’s CI overlaps each model’s individually.³ This suggests the raw comparison, if anything, understated human diversity relative to the models. The frequency penalty used in LLM generation (1.3) remains a separate, uncorrected confound that may further inflate model TTR; the sample-size-corrected comparison above does not address this.

Explanation length. Explanation length differs significantly across sources (Table 3): a Kruskal-Wallis test confirms this ($H(3) = 727.3$, $p < .001$, $\epsilon^2 = 0.34$), and every pairwise comparison (Mann-Whitney U , Holm-Bonferroni corrected) is significant, indicating each source’s typical explanation length is statistically distinguishable from every other’s. Human explanations are shortest and most consistent (median 6.0 words, Interquartile Range (IQR) 3.0), Magistral somewhat longer (7.0, IQR 3.0), Gemma longer still (9.0, IQR 2.0), and GaMS longest and least consistent (12.0, IQR 5.0).

Relation distribution. The clearest human-machine difference is in valid output. Whatever a speaker produces is a valid association, and usually also a valid explanation of it, so the `ERROR` label applies to only $\sim 1\%$ of human data. For the models, by contrast, invalid output marked with the `ERROR` label is substantial — $\sim 29\%$ for Gemma and GaMS, and $\sim 55\%$ for Magistral (Figure 3) — because the models frequently break the task constraints (the failure types are detailed below). This asymmetry already marks a qualitative gap: the task is well-defined for humans but only partly tractable for the models.

³GaMS and Magistral’s sample sizes already equal 300, so their subsampled TTR equals their raw value with no sampling variability; only Human and Gemma required genuine subsampling. The apparent non-overlap involving GaMS partly reflects comparing a real distribution against a fixed point, and is not by itself strong evidence of a meaningful model difference.

The remaining, valid responses also differ significantly in how they spread across the taxonomy (Figure 1): a chi-square test across the four sources confirms the difference is unlikely to arise by chance ($\chi^2(57) = 355.8$, $p < .001$), with a moderate effect size (Cramér’s $V=0.31$) — notably larger than the cross-linguistic effect reported for H3 ($V=0.17$ – 0.23), suggesting the human-machine relational gap within Slovenian is more pronounced than the cross-linguistic gap across languages. Human responses populate all 20 relation classes, led by *Action* (14.0%), *Thematic* (9.4%), *ResultIn* (9.2%), *CommonPhrase* (8.7%), and *Location* (8.5%). At the coarse level, Situational relations are the largest group for all four sources but do not follow a simple human-vs-machine split: Gemma shows the highest Situational share of any source (71.4%), GaMS is close to human (52.8% vs. 51.1%), and Magistral is lowest (38.1%) but still substantial. Taxonomic relations, by contrast, are markedly elevated for Magistral (40.3%, versus 13.6% for humans), driven almost entirely by a single label: *Members_of_Category* accounts for 26.9% of Magistral’s valid output (36/134 instances, largely repeated templated phrasing rather than diverse relational content — consistent with the self-commentary tendencies in Table 4), versus 2.0% for humans and single digits for Gemma and GaMS. Linguistic relations (*CommonPhrase*, *Lexical*, *SoundSimilarity*) remain systematically underrepresented in machine output across all three models (0.0–6.5%) relative to humans (15.9%). The None-of-the-Above share is close to human levels for Gemma and GaMS (2.3%, 1.9% vs. 1.9%) but elevated for Magistral (9.7%), consistent with its markedly higher `ERROR` rate reported above. (The chi-square test above has 21.2% of cells with expected counts below 5 in the full 20-category table; collapsing the four sparsest relation types — *Antonym*, *SoundSimilarity*, *MadeOf*, *Time* — into a single bin reduces this to 5.9% (minimum expected count 3.3, none below 1) and yields materially the same result: $\chi^2(48) = 344.4$, $p < .001$, Cramér’s $V=0.30$.)

Cross-linguistic comparison. On the shared 17-category taxonomy (Figure 2), the two languages show close agreement in the Concept-Properties group ($\sim 19.1\%$ English, $\sim 19.0\%$ Slovenian) but diverge at the level of individual relations and other coarse groups. English’s single largest category is *Synonym* ($\sim 16.8\%$), ahead of *Action* ($\sim 14.3\%$);

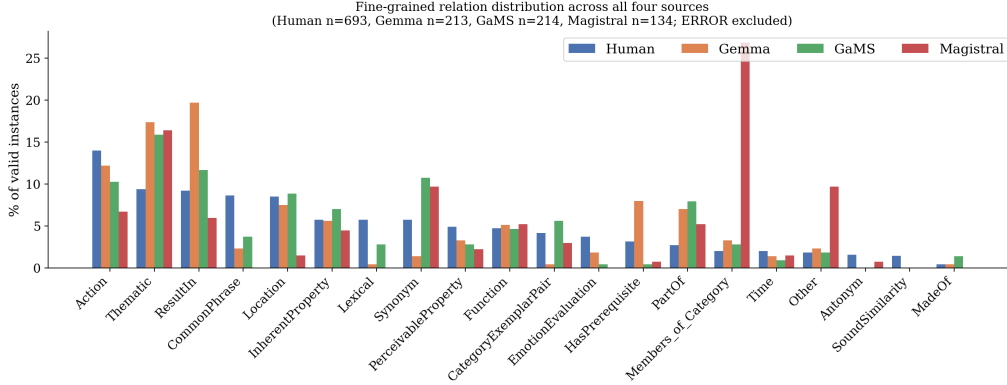


Figure 1: Fine-grained relation distribution across all four sources. Human responses populate all 20 relation classes, while model outputs concentrate on a narrow set of structurally transparent relations.

Slovenian shows the reverse ordering, with Action leading ($\sim 15.1\%$) and Synonym comprising only $\sim 6.2\%$ of labeled instances. Situational relations are more frequent in Slovenian overall ($\sim 55.1\%$ vs. $\sim 46.0\%$), driven by higher shares of Thematic ($\sim 10.1\%$ vs. $\sim 6.1\%$), Location ($\sim 9.2\%$ vs. $\sim 6.1\%$), and ResultIn ($\sim 10.0\%$ vs. $\sim 6.7\%$). Taxonomic relations are correspondingly less frequent in Slovenian ($\sim 14.6\%$ vs. $\sim 25.3\%$), a gap attributable almost entirely to the Synonym asymmetry rather than a broad shift across all taxonomic categories. A chi-square test on the fine-grained distribution confirms the difference is unlikely to arise by chance ($\chi^2(16)=75.2$, $p<.001$), though the associated effect size is modest (Cramér’s $V=0.23$); the same holds at the coarse-group level ($\chi^2(4)=39.6$, $p<.001$, Cramér’s $V=0.17$).

Erratic behaviour. Invalid human cases are mostly mechanical (part-of-speech changes, omitted words, unfinished sentences); the machine inventory is far wider, spanning phrase-level violations (GaMS and Magistral producing multi-word phrases despite instructions), semantic incoherence (across all three models), and hallucination (Magistral reproducing prompt material, adding examples, or commenting on its own output). These are not noise around a human-like signal but a distinct failure profile (Table 4); representative cases are shown in Table 5.

Explanation quality. Beyond length, model explanations diverge from human ones in two ways: they are longer and, for GaMS especially, more variable, padding the cue–association link with material that does not sharpen the relation; and a substantial share are invalid outright — semanti-

Type	Source
Phrase error	GaMS, Magistral
Semantically incoherent explanation	all LLMs
Part-of-speech change	all LLMs, human
Hallucination	Magistral
Self-commentary / added examples	Magistral
Mechanical	human

Table 4: Invalid-output profiles. Human invalidity is mechanical; model invalidity spans phrase-level, semantic, and hallucinatory failures.

cally incoherent sentences with no intelligible relation, or, for Magistral, hallucinated additions that reproduce prompt material, add extra examples, or comment on the model’s own generation. Human explanations, by contrast, reliably express the cue–association relationship in a way other annotators can classify, reflected in the higher agreement scores below.

Inter-annotator agreement. Agreement on human data ($\kappa = 0.46$) matches that reported for the original English WAX, but is markedly lower on machine data ($\kappa = 0.32$), reflecting the greater ambiguity and heterogeneity of model responses. On the conventional benchmarks proposed by Landis and Koch (1977), $\kappa = 0.46$ falls in the “moderate” range and $\kappa = 0.32$ in the “fair” range: neither level of agreement is high enough for the fine-grained percentages reported here (Figure 1, Figure 2) to be read as precise values rather than reasonable central estimates. In practice, this means the relation-distribution figures reflect a single curated labelling rather than a definitive ground truth; a different pair of annotators, particularly for machine output, could plausibly shift individual category percentages by a few points, especially be-

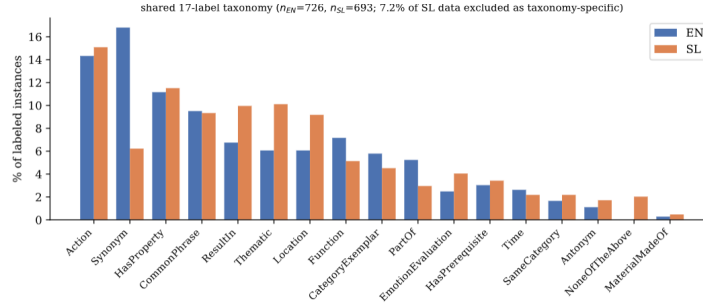


Figure 2: Relation distribution in Slovenian (WAX-SL, human) compared with English (WAX-EN) on the shared 17-category taxonomy ($n_{EN}=726$, $n_{SL}=643$; 7.2% of valid Slovenian data excluded as taxonomy-specific). Concept-Properties relations are nearly identical in proportion across languages; the clearest divergence is in Synonym labeling, which is markedly more frequent in English, while Situational relations (Thematic, Location, ResultIn) are more frequent in Slovenian.

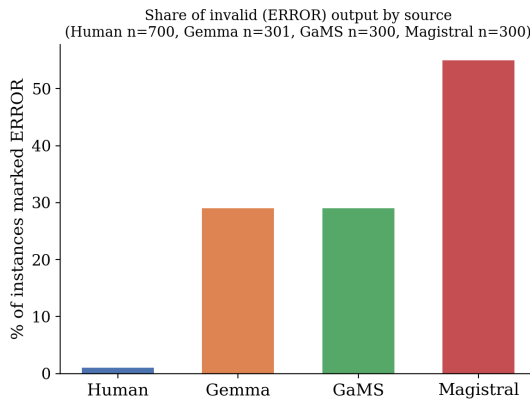


Figure 3: Share of invalid (ERROR) output by source. Gemma and GaMS sit near 29%, while Magistral reaches ~55%.

tween semantically adjacent labels (e.g., Action versus HasPrerequisite, or CategoryExemplar versus SameCategory). The large, broad-based distributional differences this paper reports between humans and models — the overall spread across relation types, the concentration of machine output on a narrower set of categories, and the elevated ERROR rates — are unlikely to be artefacts of this uncertainty, since they exceed the scale of disagreement that moderate-to-fair κ implies. Smaller differences between individual fine-grained categories, particularly among the least frequent labels, should be read with correspondingly more caution. The difficulty annotators have in assigning consistent relation labels to machine output is itself evidence about how that output is organised.

5 Discussion

Diversity of associative behaviour (H1). The results support H1. Human associations draw on

experience, social and cultural context, and practical world knowledge, populating every relation class; models that do produce valid associations concentrate on a few structurally transparent relations, a narrower spread than humans, and leave a large invalid residue — output that fails to yield a classifiable relation at all. The human-machine difference is thus twofold: a different distribution among valid responses, and a markedly higher rate of breakdown beyond them. This elevated breakdown rate echoes [Brglez et al. \(2024\)](#), who found a comparable human-machine gap in erratic responses on a Slovene word-association task with fine-tuned T5 models (11% humans vs. 47–61% across three models).

Two qualifications apply. First, associative behaviour is variable even among humans: overlap between individual respondents is only moderate, so the human distribution is itself a central tendency rather than a single norm. Second, our evidence is behavioural and does not license strong claims about internal semantic organisation; the models studied here are also relatively small, and larger or instruction-tuned commercial systems may align more closely with human responses. We therefore read the divergence as a robust behavioural difference on this task, not direct evidence about underlying representation.

The role of language-specific training (H2). H2 is only weakly testable with the present design. GaMS’s continual Slovenian pretraining does not yield more human-like behaviour than the multilingual Gemma, which sits uneasily with a simple language-specific-training account. However, the three models differ simultaneously in model lineage (Gemma 3 vs. a Gemma-2-based GaMS

Error type	Model	Output (Slovenian + gloss)	Problem
Phrase error	GaMS	<i>bos</i> → <i>Bosna in Hercegovina; Bos je uradni jezik Bosna in Hercegovina.</i> 'barefoot → Bosnia and Herzegovina; Barefoot is the official language of Bosnia and Herzegovina.'	Association is a three-word place name rather than a single word; it is linked to the cue only by shared word-initial form, and the explanation asserts a falsehood
Phrase error	Magistral	<i>umiti</i> → <i>umivna sredstva; Umi se umivajo z umivnimi sredstvi.</i> 'to wash → washing agents; Minds wash themselves with washing agents.'	Association is a two-word phrase; the explanation also misreads the cue stem <i>umi-</i> ('wash') as <i>um</i> ('mind'), yielding a nonsensical subject
Semantically incoherent	Magistral	<i>telefonski</i> → <i>zvoniti; Telefonski zvoniti je hitro in enostavno.</i> 'telephonic → to ring; Telephonic to-ring is fast and easy.'	Cue and association are juxtaposed without a well-formed syntactic link (adjective + bare infinitive), so no relation is stated
POS change	Gemma	<i>začeti</i> → <i>pobuda; Začetki pogosto potrebujejo pobudo, da se uresničijo.</i> 'to begin → initiative; Beginnings often need an initiative in order to be realised.'	The verbal cue <i>začeti</i> ('to begin') appears in the explanation as the noun <i>začetki</i> ('beginnings'), violating the requirement to preserve the cue's part of speech
Hallucination	Magistral	<i>To je primer povedi, ki opisuje vezavo med dveh izrazov.</i> 'This is an example sentence describing the link between two terms.'	The model describes the task rather than explaining the pair, reproducing prompt material instead of producing an explanation

Table 5: Representative invalid outputs by type and model, with English glosses. The human invalid set, by contrast, contains only mechanical errors.

vs. a Mistral-based Magistral), instruction-tuning regime, and training objective, not only in Slovenian adaptation, so this comparison cannot isolate language-specific training as the operative variable. Magistral’s markedly erratic behaviour (Table 3) is at least as plausibly explained by its RLVR-based reasoning-focused training, optimised for structured, verifiable tasks rather than open-ended subjective judgments, as by its lack of Slovenian support. The two explanations are confounded by design; a stronger test would compare models varying in Slovenian-specific training while holding architecture and tuning regime fixed.

Cross-linguistic variation (H3). H3 is supported, though the evidence is more specific than a uniform cross-linguistic shift. Concept-Properties relations are nearly identical in proportion across languages, suggesting this category is comparatively stable across the two languages. The clearest divergence is in Taxonomic relations, driven by Synonym labeling: English annotators labeled pairs as synonyms far more often than Slovenian annotators did, while Action led in Slovenian. Situational relations (Thematic, Location, ResultIn) are correspondingly more frequent in Slovenian. Whether this reflects genuine behavioural difference, annotator tendency, or a cue-selection arte-

fact cannot be determined here and is a target for follow-up work with matched cue sets.

Implications for language understanding. Human explanations are short, consistent, and reliably classifiable; machine explanations are longer, less consistent, and harder to agree on, with a non-trivial fraction failing entirely. This is the gap between fluency and genuine explanation: grammatical, plausible-looking output that fails to communicate the relation it purports to explain, whether by padding, incoherence, or hallucination. The word-association task makes this measurable at the smallest unit of meaning, leaving little room for an explanation to hide behind discourse-level coherence. This suggests benchmarking on outputs alone overstates how far models understand word meaning, and that explanation quality is a distinct evaluation axis — one less-resourced languages like Slovenian may expose more readily than English, pending cross-lingual evaluation of the same models.

6 Conclusion

We presented WAX-SL, the first Slovenian cue–association–explanation dataset with relation annotation, comparing human and LLM associative behaviour. Human associations are more diverse and

broadly distributed; machine output is narrower and marked by systematic errors that language-specific adaptation alone cannot resolve, supporting broader LU evaluation across less-resourced languages.

Limitations

The Slovenian dataset is smaller and less curated than the English WAX. Limited transparency of model training data constrains the interpretation of model behaviour. Inter-annotator agreement is lower for machine data than for human data, introducing uncertainty into the fine-grained relation distributions. Results for machine output were obtained under a single set of generation parameters; in particular, the frequency penalty may have inflated lexical diversity, and other settings could yield different distributions. The study covers one language pair and three models, so generalisation to other languages and models should be made with caution. The cross-linguistic comparison excludes two Slovenian relation types with no English counterpart ($\sim 7.2\%$ of valid Slovenian data); the remaining comparison uses a shared taxonomy, but cue selection differed between the original WAX-EN and WAX-SL collections, which may itself contribute to distributional differences.

Ethics Statement

Human association and explanation data were contributed by nine Slovenian-speaking annotators who took part on an informed basis; the data concern common vocabulary and contain no sensitive personal information. Relation annotation was carried out by trained annotators with curation. The released WAX-SL dataset is intended to support research on word associations and the evaluation of language models in a less-resourced language. The LLMs used were locally deployed open-weight models; no proprietary or closed systems were queried.

Acknowledgements

This work was supported by the research project GC-0002 Large Language Models for Digital Humanities (LLM4DH), funded by the Slovenian Research and Innovation Agency (ARIS). We thank the annotators who contributed to WAX-SL.

References

- Mojca Brglez, Špela Vintar, and Aleš Žagar. 2024. [How human-like are word associations in generative models? An experiment in Slovene](#). In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*. ELRA and ICCL.
- Herbert H. Clark. 1970. Word associations and linguistic theory. In John Lyons, editor, *New Horizons in Linguistics*.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Simon De Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. [The “Small World of Words” English word association norms for over 12,000 cue words](#). *Behavior Research Methods*, 51(3):987–1006.
- Tess Fitzpatrick. 2007. Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics*, 17(3):319–331.
- Sigmund Freud. 1913. On beginning the treatment (further recommendations on the technique of psychoanalysis i). In *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, volume XII. Hogarth Press.
- Francis Galton. 1879. [Psychometric experiments](#). *Brain*, 2(2):149–162.
- Fritz Günther, Luca Rinaldi, and Marco Marelli. 2019. [Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions](#). *Perspectives on Psychological Science*, 14(6):1006–1033.
- Špela Arhar Holdt, Senja Pollak, Marko Robnik Šikonja, and Simon Krek. 2020. [Referenčni seznam pogostih splošnih besed za slovenščino](#). In *Conference on Language Technologies & Digital Humanities*.
- Grace Helen Kent and Aaron Joshua Rosanoff. 1910. A study of association in insanity. *American Journal of Insanity*, 67:37–96.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.
- J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Chunhua Liu. 2022. WAX: Word association explanations – repository. <https://github.com/ChunhuaLiu596/WAX>.

- Chunhua Liu, Trevor Cohn, Simon De Deyne, and Lea Frermann. 2022. [WAX: A new dataset for word association explanations](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. [Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation](#). *Journal of Memory and Language*, 92:57–78.
- Tim Miller. 2019. [Explanation in artificial intelligence: Insights from the social sciences](#). *Artificial Intelligence*, 267:1–38.
- Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. [The University of South Florida free association, rhyme, and word fragment norms](#). *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.
- Aida Nematzadeh, Stephan C. Meylan, and Thomas L. Griffiths. 2017. [Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words](#). In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.
- Francesco Sovrano and Fabio Vitali. 2024. [Explanatory artificial intelligence \(YAI\): Human-centered explanations of explainable AI and complex data](#). *Data Mining and Knowledge Discovery*, 38(5):3141–3168.
- Špela Vintar, Mojca Brglez, Ela Novak, and Simon De Deyne. 2025. [The SWOW-SL project: Building the word association data set for Slovene. Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave](#), 13(1):104–119.
- Špela Vintar and Jan Jona Javoršek. 2025. [The truth is no diaper: Human and AI-generated associations to emotional words](#). In *Proceedings of the Sixteenth International Conference on Computational Creativity (ICCC’25)*, Campinas, Brazil.
- Špela Vintar and Amanda Saksida. 2023. [The anatomy of specialized knowledge: Comparing experts and non-experts through associations, frames and language models](#). *Lexicographica*, 39(1):165–190.
- Klara Vrabl. 2026. [Medjezikovna primerjava človeških asociacij z asociacijami velikega jezikovnega modela](#). Master’s thesis, University of Ljubljana. Accessed 7 September 2026.
- Klara Vrabl, Špela Vintar, and Mojca Brglez. 2026. [Slovenian word associations with explanations wax-sl](#). CLARIN.SI repository. Accessed 9 September 2026.

A Cross-linguistic relation taxonomy mapping

Slovenian	English (WAX-EN)	Coarse group	Relationship
EmotionEvaluation	EmotionEvaluation	Concept-Properties	identical
PartOf	PartOf		identical
InherentProperty	HasProperty		split of one EN category
PerceivableProperty	MaterialMadeOf		renamed
MadeOf			
Action	Action	Situational	identical
Function	Function		identical
HasPrerequisite	HasPrerequisite		identical
Location	Location		identical
ResultIn	ResultIn		identical
Thematic	Thematic		identical
Time	Time		identical
Antonym	Antonym	Taxonomic	identical
CategoryExemplarPair	CategoryExemplar		renamed
Members_of_Category	SameCategory		renamed
Synonym	Synonym		identical
CommonPhrase	CommonPhrase	Linguistic	identical
Lexical	—	Linguistic	no EN counterpart (excluded)
SoundSimilarity	—	Linguistic	no EN counterpart (excluded)
Other	None-of-the-Above	None-of-the-Above	residual category, both taxonomies

Table 6: Mapping between the 20-label Slovenian relation taxonomy and the 16-label taxonomy used in the published WAX-EN annotation (Liu et al., 2022), grouped by coarse category following the extended taxonomy’s own documented coarse groupings (Liu, 2022). Lexical and SoundSimilarity do not appear in the 16-label taxonomy of Liu et al. (2022); their coarse-group assignment here is not our own inference but follows the crowdsourcing task documentation in the project repository (Liu, 2022). ERROR is a Slovenian-only validity marker, not a relation type, and is not included here (see §3, Relation annotation).

B LLM prompts

The following prompts were used to generate associations (Phase 1) and explanations (Phase 2) from all three LLMs, with {word} and {answer} substituted with the cue and association respectively. Both prompts were shortened and stripped of examples relative to the human instructions after early testing showed that longer prompts, or prompts containing examples, led models to reproduce the examples rather than generate original responses.

Phase 1 (association generation).

Generiraj tri asociacije za besedo {word}. Zapiši samo 3 besede, loči jih s podpičjem. Asociacija ne sme biti fraza. Ne ponavljaj izhodiščne besede.

English translation: “Generate three associations for the word {word}. Write only 3 words and separate them with semicolons. The association must not be a phrase. Do not repeat the stimulus word.”

Phase 2 (explanation generation).

Napiši kratko poved (3–15 besed), ki opisuje povezavo med {word} in {answer}. Poved mora vključevati tako {word} kot {answer}. Da bo poved slovnično pravilna, lahko besedi postaviš v druge oblike (npr. v množino, ju sklanjaš/spregaš), ampak ne spreminjaj njune besedne vrste. Ne generiraj dodatnega besedila, samo poved.

English translation: “Write a short sentence (3–15 words) that describes the relationship between {word} and {answer}. The sentence must include both {word} and {answer}. To make the sentence grammatically correct, you may change their inflected forms (e.g., pluralise them or adjust case/conjugation), but do not change their part of speech. Do not generate any additional text, only the sentence.”

C Annotator recruitment and task allocation

The nine annotators who elicited the human associations and explanations (paid students of linguistics or digital linguistics) each provided responses for 45–48 cues; full elicitation instructions are provided in Vrabl (2026). For relation annotation, five annotators were drawn from this same pool by availability. The human data was split into 35 files of 20 triplets each (one reserved for INCEpTION training; agreement was calculated over the remaining 34), each labelled independently by two annotators with a third serving as curator, confirming agreements and selecting the more appropriate label on disagreement. For the LLM-generated data, three annotators participated: two from the human relation-annotation stage and one new volunteer briefed on the schema beforehand. A pilot study with first-year Digital Linguistics students, run as a class exercise, refined the annotation guidelines before the main annotation began; these remained accessible on INCEpTION throughout.