FAMWA: A new taxonomy for classifying word associations (which humans improve at but LLMs still struggle with)

Maria A. Rodriguez[1,2], **Marie Candito**[3],
Richard Huyghe[1]
[1] University of Fribourg,
[2] Lucerne University of Applied Sciences and Arts,
[3] LLF (Université Paris Cité / CNRS)

Abstract

Word associations have a longstanding tradition of being instrumental for investigating the organization of the mental lexicon. Despite their wide application in psychology and psycholinguistics, analyzing word associations remains challenging due to their inherent heterogeneity and variability, shaped by linguistic and extralinguistic factors. Existing word-association taxonomies often suffer limitations due to a lack of comprehensive frameworks that capture their complexity. To address these limitations, we introduce a linguistically motivated taxonomy consisting of co-existing meaning-related and form-related relations, while accounting for the directionality of word associations. We applied this taxonomy to a dataset of 1,300 word associations (FAMWA) and assessed it using various LLMs, analyzing their ability to classify word associations. The results indicate higher inter-annotator agreement with our taxonomy compared to previous studies (κ = .60 for meaning and κ = .58 for form). However, models such as GPT-4o perform only modestly in relation labeling (with accuracies of 46.2% for meaning and 78.3% for form), which calls into question their ability to fully grasp the underlying principles of human word associations.

1 Introduction

The word association task is a classic psychological experiment in which participants respond spontaneously with the first word(s) that come to mind (e.g., *cat*, *bark*, *bone*) when presented with a specific cue word (e.g., *dog*). For more than a century, word associations have been used by psychologists and psychiatrists to investigate cognitive processes, psychological behavior patterns, mental disorders, language acquisition, multilingualism, and the overall structure of the mental lexicon (see Galton 1879; Jung 1910; Kent and Rosanoff 1910; Deese 1965; Riegel and Zivian 1972; Meara 1983; a.o.). Experiments conducted across multiple languages have shown that word associations are characterized by both heterogeneity and variability. On the one hand, responses may be influenced by a wide range of relationships between cues and responses, depending on formal, semantic, and syntactic properties, as well as extralinguistic knowledge and cultural factors. On the other hand, there is considerable variation in responses across individuals, with greatly varying degrees of convergence depending on the cue word. This diversity poses challenges for linguistic analysis, and various taxonomies of word-association relations have been proposed to accurately account for word associations.

Although common elements of classification emerge from previous studies, there is no universally accepted framework for describing word-association relations, as existing taxonomies present various limitations. The categories used to analyze associations are not always explicitly defined, and some taxonomies focus only on a single aspect of word-association relations (e.g., semantic characteristics), while others merge semantic and formal relations into flat hierarchies, potentially leading to conflicting or incomplete descriptions. In addition, directionality is rarely considered, although the cue-response sequencing and the non-reciprocal nature of word associations call for a specific account of asymmetrical relations. Finally, taxonomies are seldom evaluated through inter-annotator agreement or computational models, limiting their validation and reliability.

In this study, we propose a linguistically motivated taxonomy of word-association relations based on a critical examination of previous classifications. The taxonomy includes two co-existing levels of linguistic analysis related to form and meaning, and takes into account the directionality of word associations. We apply this taxonomy

Proceedings of the 16th International Conference on Computational Semantics, pages 175–188
September 22-23, 2025,

on a dataset of 1,300 word associations in English using a well-defined annotation protocol, while assessing annotation quality through inter-annotator agreement. Furthermore, we evaluate the ability of generative language models to classify word associations according to our taxonomy, exploring how well they capture the diversity of relations in word associations, and providing a detailed analysis of model performance. Overall, the main contributions of the study include (i) a theoretical and methodological reflection on the analysis of word-association relations, (ii) the creation of a finely annotated dataset of word associations covering formal and semantic relations, and (iii) a discussion of how language models handle the heterogeneity of lexical relations within word associations.

2 Background

2.1 Linguistic description of word associations

A close examination of word associations reveals that they are not restricted to lexical relations, which alone cannot account for the full range of relationships observed between cues and responses (see, e.g., Schulte im Walde et al. (2008) for German). Across the literature, linguistic descriptions addressing specifically word associations often distinguish between syntagmatic, paradigmatic, and clang associations (see, e.g., Deese 1962; Glanzer 1962). Syntagmatic associations are observed between words that may cooccur in context (e.g., *friend-best*), whereas paradigmatic associations involve words from the same lexical class with related meanings (e.g., *certain-sure*), and clang associations are based on phonological similarities between cues and responses (e.g., *hat-cat*). Traditionally, these categories have been considered as mutually exclusive, which presents challenges when analyzing word pairs that could belong to multiple categories. More broadly, the tripartite classification between syntagmatic, paradigmatic, and clang associations has been criticized for being too coarse-grained, while still failing to account for all word associations, and relying on overly vague category definitions.

To address these challenges, more fine-grained analyses of word-association relations have been proposed. Fitzpatrick (2006, 2007) introduced a taxonomy based on 4 main categories (meaning-based, position-based, form-based, and erratic), further divided into 17 subcategories for a more detailed classification. Similarly, Santos et al. (2011) used 10 basic categories to describe response words, notably accounting for the directionality of associations (Tversky, 1977). For instance, they distinguished between "domain higher category" and "domain lower category" to differentiate cases in which the response represents a superordinate or a subordinate concept relative to the cue. However, these fine-grained taxonomies still conflate the formal and semantic aspects of word-association relations into a single classification, implying a complementary distribution that does not always apply in practice.

Another classification approach is based on the system proposed by Wu and Barsalou (2009), originally designed to analyze concept representations, but later applied to semantic feature norming (Bolognesi et al., 2017; Vivas et al., 2022) and word associations (Liu et al., 2022; De Deyne et al., 2024). This framework distinguishes between taxonomic, situational, entity, and introspective properties, with the potential for further division into more detailed classes (see, e.g., McRae et al. 2012). However, an inherent limitation of this taxonomy is that not all relations in word associations are based on property descriptions, nor are they always determined semantically. As a result, restricting the analysis to this taxonomy may lead to an incomplete characterization of word associations, particularly by overlooking their more formal aspects.

Three key observations emerge from the discussion above. First, both formal (i.e., morphological and phonological) and semantic relations can drive word associations, and a comprehensive taxonomy should integrate both aspects to fully capture word-association relations. Second, while formal and semantic relations should be distinguished, they should not be treated as mutually exclusive, as there is no logical incompatibility in formal and semantic motivations for word associations. A multilevel analysis is necessary to reflect both the linguistic relations underlying word associations and the complexity of the cognitive processes involved. Third, a detailed analysis of word-association relations must account for both symmetrical (e.g., synonymy, phonological resemblance) and asymmetrical relations (e.g., hyponymy, morphological derivation). Given that word associations are by definition oriented from cues to responses, and reciprocity between them is rarely observed, taxonomies including directional classes are essential to provide a fine description of word associations.

2.2 Existing datasets

Word associations have been collected for various languages, on a growing scale over the years (see, e.g., Kiss et al. 1973; Nelson et al. 2004 for English). SWOW is currently the largest multilingual word-association dataset, covering 19 languages[1]. In this paper, we focus on its English part, which was collected via crowdsourcing (De Deyne et al., 2019). For each of 12,282 English cues, 100 participants were asked to answer the first 3 words coming to their mind, resulting in a dataset of over 150k unique cue-response pairs, each associated with the number of participants who answered the response at each position (hereafter R1, R2, and R3). De Deyne et al. (2019) checked that the continued response paradigm of the English SWOW and the more heterogeneous participant sample did not affect properties compared to other single-response English datasets, and observed small evidence for response chaining—cases of R2 being influenced by R1 response.

Smaller word association datasets include the labeling of cue-response pairs into categories, possibly based on participants' explanations for the associations. For example, Fitzpatrick (2006) collected single-response associations from 40 participants for 60 cues, conducted retrospective interviews, and categorized the associations according to the participants' explanations. Similarly, Liu et al. (2022) asked participants to both produce associations and explain their responses, compiling the WAX dataset, which contains 15k unique cue-response pairs and 19k cue-response-explanation triples. Among these, 1,602 triples were classified into 16 word-association categories, half manually by humans and half automatically, based on the identification of explanation patterns associated with certain labels—a method that may affect the reliability of the classification. The inter-annotator agreement for the human classification was measured but found to be only moderate (Cohen's κ = .42). A possible limitation of the explanation-based approach is that, although prompting participants to provide explanations for their associations may help clarify the cue-response relation, it can also introduce bias by making responses less spontaneous, as already observed by Woodworth (1938).

[1] https://smallworldofwords.org/en/project/home

2.3 Computational approaches to word associations

Studies on word associations using pre-trained language models have developed recently, following three lines of research. Some researchers have compared the properties of word associations with those of word embeddings. For instance, A. Rodriguez and Merlo (2020) found that the top-K neighbors of a cue encoded with BERT (Devlin et al., 2019) often contain human responses.

Computational models have also been employed to mimick the word association task. Vintar et al. (2024) prompted encoder-decoder language models to provide an unlimited list of response words given a cue in Slovene and English, from the English and Slovene SWOW datasets. Abramski et al. (2025) prompted three decoder-only large language models (LLMs) to produce three word associations, using the English SWOW cues. While both works report relatively low overlap between the human and models' associations, Abramski et al. (2025) found that human and models' responses do share semantic properties: when building a semantic network based on the associations (one network for human associations, and one network per prompted LLM), the authors report the same strong correlation level between the ease of lexical retrieval for human participants and the closeness in the semantic word association network[2], for all the four networks (one human, and three LLM-based).

A third line of research focuses on learning or using models to classify cue-response relations. For example, Liu et al. (2022) used the WAX dataset, which contains cue-response-explanation triples, and designed various tasks to assess how well language models capture the underlying relations between cues and responses. In particular, they trained relation classifiers based on BERT and BART (Lewis et al., 2020), but reported relatively low performance (weighted F1 = 48%). Similarly, De Deyne et al. (2024) prompted GPT-4 to classify a fraction of the human-labeled part of the WAX dataset (among other datasets[3]). They reported a classification F-score of 47%, indicating that the

[2] There is a negative correlation between the reaction time of participants to a lexical decision task and the distance of input-target pairs within the semantic network.

[3] Three other datasets were used: two related to concept-feature pairs, and a labeled word-association dataset cited as "Chen et al. (2024)", but whose reference is erroneous and cannot be found online. The dataset is reported to have a surprisingly high Cohen's κ (.81, twice as much as for WAX), but it cannot be found either.

Round	Sample	Meaning	Form	# pairs
1	1	.23	-	100
2	2	.39	-	100
3	3	.36	.45	50
	4	.65	.46	50
	5	.63	.58	50
4	6	.54	.55	50
	7	.54	.64	50
	8	.37	.60	50
5	9	.67	.55	50
	10	.75	.63	50
	11	.56	.62	50
3-5	3-11	.60	.58	450

Table 1: Inter-annotator agreement (Cohen's kappa) across annotation rounds for double-annotated samples.

model struggles with either the task, the taxonomy of word-association relations, or both.

In this paper, we propose a linguistically motivated inventory of cue-response relations meeting the requirements outlined in Section 2.1. We evaluate the relation taxonomy through inter-annotator agreement on a sample of English word associations extracted from SWOW. Additionally, we investigate word association classification using LLMs, both on our dataset and relation inventory, and on the WAX dataset and inventory (Liu et al., 2022). The relatively low performance observed for both leads us to discuss whether models have sufficient knowledge of the principles underlying human word associations.

3 A taxonomy of cue-response relations

This section presents the taxonomy we used to classify word-association relations, along with a labeled dataset of cue-response pairs. Crucially, we employed a dual-level classification, where the relation between a cue and a response is annotated for both meaning and form. We also took the directionality of the relations into account to reflect the asymmetry of the associations.

3.1 Methodology

We adopted an inductive approach to develop our taxonomic model, starting with basic linguistic categories that distinguish lexical relations, semantic features, argumental relations, and modification for the semantic part of the classification, and phonological and morphological relations for the formal part. These classes were explicitly defined and subsequently refined through multiple rounds of annotation and adjudication. Annotation guidelines were established, including a decision tree to systematize the annotation process[4].

Three expert annotators conducted double-blind annotation and adjudication over 5 rounds, on randomly selected samples from SWOW, focusing on associations between cues and R1 responses provided by at least three participants[5]. The annotation guidelines were revised after each round, and the process continued until a satisfactory inter-annotator agreement was reached. Sample sizes and inter-annotator agreement (IAA, Cohen's kappa scores) for each round of annotation are provided in Table 1.

The multi-level annotation with co-existing labels for meaning and form was introduced after Round 2, following the observation that a single label was insufficient to capture the complexity of association relations. For example, in the pair *pickup-truck*, the cue is a hyponym of the response (semantic label), and cue and response also form the compound word *pickup truck* (formal label). As can be seen in Table 1, the IAA increased significantly following the introduction of the two separate taxonomies (in Round 3), but remained stable in subsequent rounds, despite continued refinement of the annotation guidelines. Calculating IAA across all samples after taxonomy split (Rounds 3-5, Samples 3-11, totalling 450 instances), we obtained a Cohen's kappa of .60 for the Meaning taxonomy (34 labels) and .58 for Form (6 labels), representing a notable improvement over the results reported by Liu et al. (2022) for WAX ($\kappa = .42$, across 16 labels).

On top of the 450 double-annotated pairs, we sampled additional pairs from the SWOW dataset, annotated by a single expert. We obtained a dataset of 1,300 cue-response pairs, annotated for both form and meaning relations (hereafter the **F**orm **A**nd **M**eaning **W**ord **A**ssociations **(FAMWA)** dataset). The distribution of Meaning labels in FAMWA is provided in Figure 1[6]. Importantly, these 1,300 cue-response pairs were randomly sampled from the SWOW dataset, as an attempt to pre-

[4]The dataset and the guidelines are available at https://github.com/mariro8/FAMWA.

[5]We deliberately excluded words given only as second (R2) or third responses (R3) to better align with the standard single-word association task.

[6]The distribution for Form labels is quite skewed, with 1,070, 70, 98, 23, 23, and 16 items for 'none', 'compo_R+C', 'compo_C+R', 'in_mwe', 'similar', and 'morpho' labels, respectively.

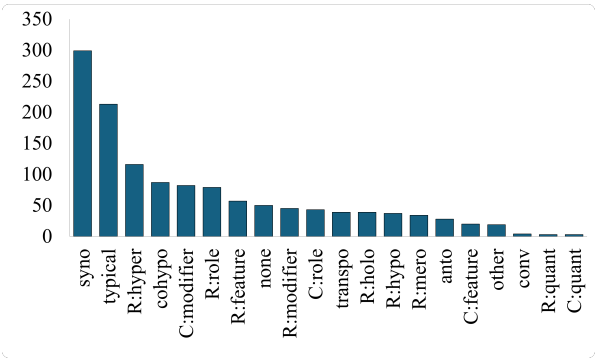

Figure 1: Distribution of Meaning labels in FAMWA. The various C/R:role_x categories are grouped into 2 single C/R:role categories, resulting in 20 labels.

serve the natural distribution of word association categories. This contrasts with the WAX dataset, as will be detailed when analyzing LLM classification in Section 4.

3.2 Resulting taxonomy

The final taxonomies for Form and Meaning consist of 34 and 6 categories, respectively (see Appendix B for the two lists of categories and their description). The Meaning taxonomy includes both lexical relations (i.e., synonymy, antonymy, hyponymy, etc.) and non-lexical relations (i.e., semantic features, semantic roles in predicate-argument relations[7], modifiers, etc.). The Form taxonomy distinguishes between phonological similarity, morphological relations (affixation or compounding), and multi-word expressions (when the cue and the response are part of a complex lexicalized expression involving other components). Both taxonomies include a "none" relation, which applies when the cue and response are unrelated in meaning or form, as well as an "other" label in the Meaning taxonomy to account for idiosyncratic relations.

Importantly, both the semantic and the formal categories are oriented, in order to capture asymmetrical relations. For instance, semantic roles were annotated when the cue is a typical argument of the response or vice-versa. A pair such as *promise-keep* was thus coded as C:role_theme, since the cue *promise* is the argument of *keep* with the role Theme, while the inverse pair would be analyzed as R:role_theme. Similarly, the sequences C+R and R+C were distinguished when cues and responses form compound words. For example, the pair *shopping-bag* was annotated as a C+R compound whereas the inverse pair would be classified as an R+C compound.

[7]We used the semantic role tagset from Verbnet (Kipper et al., 2006), following the associated guidelines (https://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf), and annotated semantic roles with the lowest possible role in the hierarchy.

4 Ability of language models to classify associations

We now focus on examining the extent to which language-model-based systems are able to classify word associations, using human-designed linguistic classifications. Previous works have provided abundant evidence that LLMs have linguistic knowledge, and in particular lexical knowledge (see, e.g., Kello and Bruna 2024 and Hayashi 2025, who showed LLMs' ability to accurately detail lexical properties of words and to distinguish word senses in context). De Deyne et al. (2024) reported that GPT-4 performed poorly in classifying word associations with the WAX dataset and relation inventory, achieving an accuracy of only 47%. We hypothesize that a dataset with better IAA can lead to improved model performance. To test this hypothesis, we prompted various LLMs to label cue-response pairs and evaluated their performance on the FAMWA dataset. We also compared performance using the WAX dataset and taxonomy, which highlights the impact of category distribution.

4.1 Adjustments in taxonomies

To compare the classifications using taxonomies of roughly equal size, we merged some of the labels in our Meaning taxonomy—more precisely we merged all the C:role_x and R:role_x labels into C:role and R:role, respectively—and we dropped the labels with less than 10 instances[8], as well as the corresponding instances (10 instances in total). This resulted in a dataset used for classification of 1,290 instances (hereafter FAMWA-1290) with a Meaning relation inventory of 17 labels (shown as the first 17 bars in Figure 1), comparable in size to the 16 labels of the WAX taxonomy.

4.2 Models

Among the ever-growing list of available LLMs, we selected GPT-4o-mini, Llama.-3.1-70B and GPT-4o, namely three models of small, medium and large size.

[8]C/R:quant (C (resp. R) can be used as a quantifier of R (resp. C)) and conv (C and R are converse words)

4.3 Evaluation datasets

We tested the selected LLMs on FAMWA-1290 for Meaning labels and then proceeded to evaluate the best model only (GPT-4o) on other taxonomy/dataset pairs: FAMWA-1290 for Form labels, the complete labeled WAX (consisting of 1,602 instances), and the human-labeled WAX (725 instances). Notably, for more than half of the labeled instances, gold WAX labels were automatically obtained using patterns found in the explanations that participants provided during the word association task[9]. Since the automatically labeled instances are biased towards categories for which reliable patterns could be designed, they do not reflect the true distribution of categories among SWOW cue-responses. Hence, we also provide results on the human-labeled part of WAX.

4.4 Experimental protocol

We tested each model in three settings: zero-shot, few-shot (with exactly one example per label), as well as "implicit" few-shot, in which the description of a given category is accompanied by an example in parentheses. The same examples were used in few-shot and implicit few-shot settings, and throughout the experiments.

Our prompts included three distinct elements: the task description, the list of labels and their descriptions, and the input/output format. A fourth section is added in the few-shot setting, providing one example of input and expected output per label, in the desired format. We performed preliminary tests with a few formulations, for the task description and the input/output format (see the variants of each section in Appendix C). In these preliminary experiments, we tested 3 variants for the input/output format section (see Table 8 in Appendix C), which resulted in marginal performance differences. We then retained 4 formulations for the task description section, and a single formulation for the other sections of the prompt, resulting in 4 prompt variants which we tested in a systematic way for all the models, datasets and settings.

We used a zero temperature for all the experiments, hence forcing the models to always generate the most probable token at each position. When parsing the models' answers, we removed any symbols and converted the text to lowercase to match the answer with the label names. We counted the answers containing no known labels as incorrect.

In the few-shot and implicit few-shot settings, we removed the instances used as examples from the testing instances.

Labels: FAMWA Meaning			
Instances: FAMWA-1290			
Models	Zero	Impl.	Few
GPT-4o-mini	32.1 (4)	27.2 (2)	36.6 (1)
Llama-3.1-70B	41.0 (3)	41.2 (3)	42.3 (3)
GPT-4o	45.1 (3)	**46.2 (2)**	46.0 (1)
Labels: WAX			
Instances: full WAX (1602)			
GPT-4o	52.1 (3)	53.8 (4)	**57.5 (1)**
Instances: WAX human-labeled (725)			
GPT-4o	41.7 (1)	45.1 (3)	**46.1** (3)
Labels: FAMWA Form			
Instances: FAMWA-1290			
GPT-4o	**78.3 (1)**	76.6 (3)	75.0 (3)

Table 2: Accuracies for cue-response pair classification across datasets, models and settings (**zero**-shot, **implicit** few-shot, and **few**-shot). The best accuracy for the 4 prompt variants is reported, with their preferred task variant in the prompt indicated in parentheses (see Table 6 in Appendix C).

4.5 Results

The results of the experiment are provided in Table 2. Analysis of the results on FAMWA-1290 for Meaning labels reveals a consistent and expected trend: performance improves systematically with increasing **model size** across all prompt settings (zero, implicit few-shot, and few-shot). However, even in the best-performing setup—the implicit few-shot configuration—the accuracy reaches only 46.2%, showing that the overall performance of the best model is still limited.

Concerning **prompt settings**, providing examples, either in implicit few-shot or few-shot, systematically elicited better results than zero-shot. This suggests that, given the technical nature of the labels, the models struggle to "understand" them and their descriptions, and benefit from the inclusion of examples. In general, the implicit few-shot setting provided slightly lower performance than the few-shot approach, suggesting that the models take

[9]For instance, searching the pattern "opposite" within the explanations allowed Liu et al. (2022) to automatically classify 76 instances into the Antonym category.

advantage of examples presented in the expected input/output format. Still, this pattern did not apply to the FAMWA-1290 dataset with GPT-4o, since the best settings for Meaning and Form were the implicit few-shot and zero-shot, respectively.

We also compared the automatic classification of cue-response pairs between WAX and FAMWA-1290 Meaning inventories[10]. However, the comparison is limited by differences in evaluation instances and category distribution in the datasets. The classification task proved to be easier with the full set of WAX instances than with the FAMWA-1290 instances, since GPT-4o in few-shot setting achieved an accuracy of 57.5% on full WAX (vs. 46.2% on FAMWA-1290 Meaning labels in implicit few-shot). Yet, the performance dropped to 46.1% on the human-labeled part of WAX, which more accurately represents the distribution of categories found in SWOW cue-response pairs, similar to the performance on the FAMWA-1290 Meaning instances[11].

The accuracy was higher for the predictions of the FAMWA-1290 Form labels (78.3%), but this is largely attributable to the highly imbalanced distribution of classes, as will be discussed in the analysis of the performance across categories.

Performance across categories The overall accuracies presented in the previous section conceal substantial variation in both model performance and category prevalence. In this section, we investigate these differences on FAMWA-1290, as a dataset that reflects the category distribution observed in SWOW. Focusing on GPT-4o in the implicit few-shot setting, as the best-performing model and configuration, we examine the F1-score and number of instances for each of the FAMWA-1290 Meaning labels in Figure 2. The results show substantial variation across categories, ranging from F1 = 77.4 for the Antonym label to null performance for several others. It is worth noting that this performance ranking does not correspond to the frequency ranking in FAMWA (see Figure 1). For instance, Antonym was predicted more accurately than Synonym despite being ten times less frequent in FAMWA, while R:role, which appears

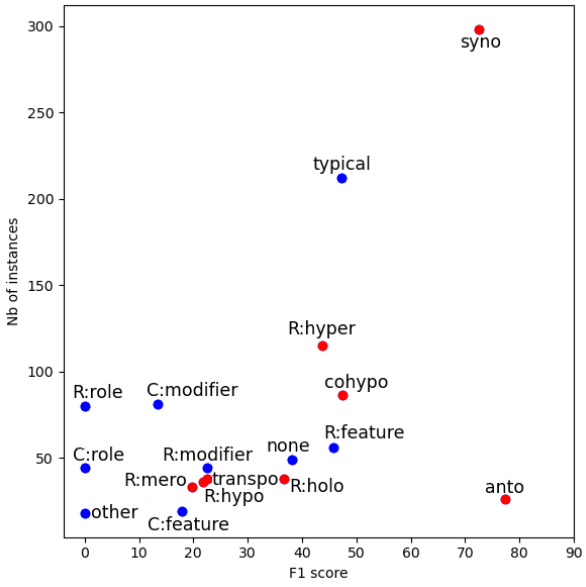

Figure 2: GPT-4o performance in the implicit setting on the FAMWA-1290 Meaning labels (excluding 17 instances used in prompt examples and 3 unpredicted relations). Lexical relations are shown in red and non-lexical relations in blue.

three times more often than Antonym (80 vs. 26 instances), was not predicted at all by GPT-4o.

The best classified relations were lexical ones (Antonym, Synonym, as well as R:hyper and Cohyponym to a lesser extent). Hyponymy was easier to detect when the response was a hypernym of the cue (F1 = 43.7 for R:hyper) than vice-versa (F1 = 21.8 for R:hypo), which is surprising given the reciprocal nature of the relation. This underscores the usefulness of using oriented categories when analyzing model performance. The Typical and R:feature categories are the only non-lexical relations that were predicted with moderate success (F1 $>$ 40%), whereas the performance on all other relations remains poor (F1 $<$ 40%). Moreover, in symmetric relations such as R:feature and C:feature, the R:x categories were consistently predicted more accurately than the C:x categories—for example, the prediction was better for R:feature and R:modifier than for C:feature and C:modifier. This suggests increased difficulty when the response is more central than the cue[12].

Turning to the breakdown by Form labels (Table 3), the model mostly predicted the absence of

[10]Note the WAX inventory does include a category "common phrase" which pertains to a formal classification, but most labels in WAX are semantic, hence the WAX inventory is more comparable to FAMWA Meaning than to FAMWA Form.

[11]De Deyne et al. (2024) obtained 47.1% on a fraction of the human-labeled WAX dataset when prompting an underspecified version of GPT-4.

[12]The breakdown per WAX label is provided in Table 9 in Appendix D, together with rough mappings with the FAMWA labels when applicable. It too shows varying performance across categories, and the same two best predicted categories (Antonym and Synonym).

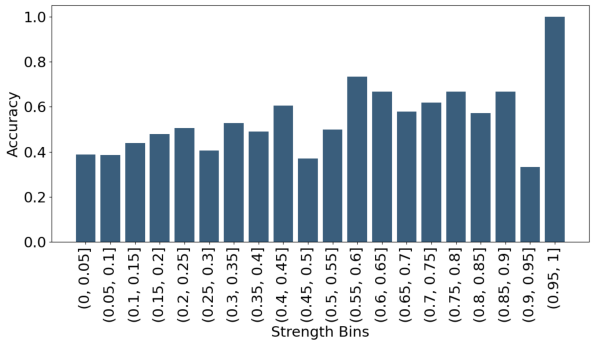

Figure 3: Accuracy of GPT-4o predicted Meaning labels, on FAMWA-1290, broken down across bins of associative strength of the pairs.

a formal relation—unsurprisingly given its prevalence in the dataset—but performed poorly across all other categories. We note though a relative ability to detect morphological relations, counterbalanced by a tendency to overpredict this category (P = 23.1, R = 1.0, F1 = 37.5). Moreover, identifying compounds proved more challenging for R+C compounds (F1 = 25.5) than for C+R compounds (F1 = 45.8), highlighting the model's difficulty in learning a pattern where the linguistic order (R then C) differs from that presented in the prompt (C then R).

Label	P	R	F1	Nb
none	90.3	88.0	89.1	1062
compo_C+R	63.6	35.7	45.8	98
morpho	23.1	1.0	37.5	15
similar	33.3	30.4	31.8	23
compo_R+C	34.1	20.3	25.5	69
in_mwe	5.6	17.4	8.4	23

Table 3: Performance of GPT-4o in zero-shot setting, for each of the FAMWA-1290 Form labels.

Performance across associative strengths We additionally examined whether cue-response pairs that are frequently provided by humans are easier for the models to classify. We used the concept of **associative strength** (De Deyne et al., 2019), defined for a cue-response pair (c, r) as the number of participants who gave r (as R1) in response to c, normalized for the total number of participants who provided at least one response for the cue c.

Figure 3 shows the performance of GPT-4o for the FAMWA-1290 Meaning labels, broken down by associative strength. There is no clear correlation between associative strength and classification accuracy, which varies across bins. However, bins with an associative strength above .55 generally exhibit higher accuracy compared to those with lower strength. Interestingly, this shift aligns with lexical relations surpassing non-lexical relations within the gold data[13]. Yet, it should be noted that the number of instances per bin, from where this shift happens up to the strongest bin, is lower than 20 (with the last 5 bins having between 3 and 10 instances). Consequently, any conclusions based on these bins should be interpreted cautiously due to the limited sample size.

5 Conclusion

Our efforts to develop a linguistically motivated taxonomy of word-association relations proved effective, as we achieved higher inter-annotator agreement than comparable studies using different analytical frameworks. Integrating a dual-level analysis of formal and semantic relations, while also accounting for directionality in associations, is not only more satisfactory from the perspective of linguistic description, but also ensures greater stability and consistency in annotation quality, at least with expert annotators. Nevertheless, the observed agreement remains moderate, highlighting the inherent challenge of producing a metalinguistic analysis of word associations. This difficulty is frequently noted by researchers who attempt to classify word associations, and it contrasts with the naturalness of the word association task itself, which is effortless as it relies only on the existence of the mental lexicon. Arguably, the basic task of generating word associations and the metalinguistic task of analyzing them involve fundamentally distinct cognitive processes and engage contrasting aspects of the language faculty.

While expert human annotators achieve only moderate agreement, even advanced models like GPT-4o exhibit mediocre performance in analyzing semantic relations between cues and responses, despite their well-documented linguistic capacities. This is congruent with the previous conclusion that the word association task and its analysis leverage different types of capabilities. Moreover, LLMs' ability to classify word associations is not improved by refinements in descriptive frameworks and varies considerably across relation classes. The

[13]The bin with strength between .90 and .95 is the only exception, both in accuracy and number of non-lexical relations.

limited performance of LLMs in labeling word associations should be analyzed in light of their ability to produce them. The inherent heterogeneity and variability of word associations pose challenges not only for their metalinguistic analysis, but also for their generation by language models. Future research should explore this generative capacity in greater depth, for example through detailed analysis of the LLM-generated association norms reported in LWOW (Abramski et al., 2025).

Acknowledgments

We are grateful to three anonymous reviewers for their careful reading and constructive comments that helped us refine the arguments presented in this paper.

References

Maria A. Rodriguez and Paola Merlo. 2020. Word associations and the distance properties of context-aware word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 376–385, Online. Association for Computational Linguistics.

Katherine Abramski, Riccardo Improta, Giulio Rossetti, and Massimo Stella. 2025. The "LLM World of Words" English free association norms generated by large language models. *Scientific Data*, 12(1):803.

Marianna Bolognesi, Roosmaryn Pilgram, and Romy Van Den Heerik. 2017. Reliability in content analysis: The case of semantic feature norms classification. *Behavior Research Methods*, 49:1984–2001.

Simon De Deyne, Chunhua Liu, and Lea Frermann. 2024. Can GPT-4 recover latent semantic relational information from word associations? A detailed analysis of agreement with human-annotated semantic ontologies. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 68–78, Torino, Italia. ELRA and ICCL.

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*, 51:987–1006.

James Deese. 1962. Form class and the determinants of association. *Journal of verbal learning and verbal behavior*, 1(2):79–84.

James Deese. 1965. *The structures of associations in language and thought*. The John Hopkins Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tess Fitzpatrick. 2006. Habits and rabbits: Word associations and the l2 lexicon. *EUROSLA yearbook*, 6(1):121–145.

Tess Fitzpatrick. 2007. Word association patterns: unpacking the assumptions. *International Journal of Applied Linguistics*, 17:319–331.

Francis Galton. 1879. Psychometric experiments. *Brain*, 2(2):149–162.

Murray Glanzer. 1962. Grammatical category: A rote learning and word association analysis. *Journal of verbal learning and verbal behavior*, 1(1):31–41.

Yoshihiko Hayashi. 2025. Evaluating LLMs' capability to identify lexical semantic equivalence: Probing with the word-in-context task. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6985–6998, Abu Dhabi, UAE. Association for Computational Linguistics.

Carl G Jung. 1910. The association method. *The American journal of psychology*, 21(2):219–269.

Cristopher Kello and Polyphony J. Bruna. 2024. Emergent mental lexicon functions in ChatGPT. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, pages 5452–5459.

Grace Helen Kent and Aaron Joshua Rosanoff. 1910. A study of association in insanity. *American Journal of Psychiatry*, 67(1):37–96.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

George R Kiss, Christine Armstrong, Robert Milroy, and James Piper. 1973. An associative thesaurus of English and its computer analysis. *The computer and literary studies*, 153.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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)*, pages 106–120.

Ken McRae, Saman Khalkhali, and Mary Hare. 2012. Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy.

Paul Meara. 1983. Word associations in a foreign language. *Nottingham Linguistics Circular*, 11(2):29–38.

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.

Klaus F Riegel and Irina WM Zivian. 1972. A study of inter-and intralingual associations in English and German 1. *Language Learning*, 22(1):51–63.

Ava Santos, Sergio E Chaigneau, W Kyle Simmons, and Lawrence W Barsalou. 2011. Property generation reflects word association and situated simulation. *Language and Cognition*, 3(1):83–119.

Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327.

Špela Vintar, Mojca Brglez, 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*, pages 42–48.

Leticia Vivas, M Yerro, Sofía Romanelli, A García Coni, Ana Comesaña, F Lizarralde, I Passoni, and J Vivas. 2022. New Spanish semantic feature production norms for older adults. *Behavior Research Methods*, pages 1–17.

Sabine Schulte im Walde, Alissa Melinger, Michael Roth, and Andrea Weber. 2008. An empirical characterisation of response types in german association norms. *Research on Language and Computation*, 6(2):205–238.

R. S. Woodworth. 1938. Experimental psychology: Association. *American Psychological Association*, pages 340–367.

Ling-ling Wu and Lawrence W Barsalou. 2009. Perceptual simulation in conceptual combination: Evidence from property generation. *Acta psychologica*, 132(2):173–189.

Limitations

While this study provides insights into the classification of word associations, some limitations should be acknowledged.

First, we used exclusively English, which restricts the generalizability of the findings to other languages. Word associations are known to be influenced by linguistic and cultural aspects, thus, it can be interesting to explore whether the proposed taxonomy and observed patterns hold across different languages.

Second, the dataset used was relatively small, consisting of only 1,300 instances. While this allowed for detailed annotation and analysis, expanding the dataset would improve the robustness of the taxonomy and enable more reliable evaluation of model performance.

Finally, this study was conducted on a small scale due to the limited availability of expert annotators and the highly time-consuming nature of the annotation process. The reliance on a small group of experts, while ensuring high-quality annotations, may introduce biases or limit the diversity of perspectives in the classification process.

A Examples of form-meaning annotation for word associations

cue	response	meaning	form
banana	yellow	R:feature	none
capita	per	none	compo_R+C
stone	wall	C:modifier	compo_C+R
rareness	rarity	syno	morpho
meet	greet	typical	in_mwe
weight	height	cohypo	similar
lottery	win	C:role_theme	none

Figure 4: Examples of Meaning and Form annotations for word associations.

B Form taxonomy and Meaning taxonomy

We detail the FAMWA inventory for Form (Table 4) and Meaning (Table 5) labels with their descriptions. Short labels were used in the human annotation while their extended form was used for prompts.

C Prompts for labeling word associations

We provide below the various forms of prompts we tested on Llama-3.1-8B/70B. A single prompt is made of a description of task (4 variants shown in Table 6), the list of labels with their description (Table 7), and the description of the expected output format (Table 8).

D GPT-4o performance on WAX human-labeled dataset broken down per WAX label

We detail the performance of GPT-4o on the human-labeled part of the WAX dataset (725 instances). We provide the WAX label and the corresponding FAMWA label, if a mapping is possible, with the F1-score performance and the total number of instances per class (see Table 9).

Short labels	Prompt labels	Definitions	Examples
compo_C+R	compound cue response	the sequence C R forms a compound	hermit-crab
compo_R+C	compound response cue	the sequence R C forms a compound	rights-human
in_mwe	multiword expression	R and C belong to the same multiword expression, containing other elements	pedal-metal
morpho	morphological relation	R and C belong to the same derivational or inflectional paradigm	gave-gift
similar	similar	C and R are similar in graphical or phonological form but not morphologically related	hat-cat
none	no relation	No formal relationship	roof-house

Table 4: FAMWA Form inventory: short labels, corresponding labels used in prompts, short definitions, and corresponding examples.

Short labels	Prompt labels	Definitions	Examples
syno	synonym	The cue and the response are synonyms	belly-stomach
anto	antonym	The cue and response are antonyms	large-small
transpo	transposition	The cue and the response are synonyms but they have a different part-of-speech	smelly-stink
cohypo	cohyponym	The cue and the response have a close common hypernym but they are not synonyms	weight-height
typical	typical	There is an obvious (and implicit) predicate that links the cue and the response or when the eventuality denoted by the cue typically cooccurs with that denoted by the response (or vice versa)	honey-bee
R:hyper	hypernym	The response is a hypernym of the cue	labrador-dog
R:hypo	hyponym	The response is a hyponym of the cue	pets-cat
R:holo	holonym	The response is a holonym of the cue	roof-house
R:mero	meronym	The response is a meronym of the cue	universe-stars
C:feature	response's feature	The cue is a semantic feature of the response	green-grass
R:feature	cue's feature	The response is a semantic feature of the cue	sauna-hot
C:modifier	response's modifier	The cue is used as a modifier of the response	metallic-paint
R:modifier	cue's modifier	The response is used as a modifier of the cue	debt-school
C:role_x	response's argument	C is an argument of R with semantic role x	pillow-sleep
R:role_x	cue's argument	R is an argument of C with semantic role x	hike-woods
C:quant	response's quantifier	C can be used as a quantifier of R	bunch-grapes
R:quant	cue's quantifier	R can be used as a quantifier of C	item-one
conv	converse	C and R are converse words	prey-predator
other	other relation	The cue and the response are in a semantic relation of different type than those listed above	trip-vacation
none	no relation	No semantic relation between cue and response	shall-we

Table 5: FAMWA Meaning inventory: short labels, corresponding labels used in prompts, short definitions and corresponding examples. C/R:quant and conv labels were discarded in the evaluation.

Tasks	Description
Variant 1	Objective: Given a word association, consisting of a pair of cue and response, label the semantic relation between these pairs with a label based on the specified criteria.
Variant 2	Objective: Given a word association task where a cue word elicits a response word, classify the semantic relation between the cue word and the response word using one of the labels described in the specified criteria.
Variant 3	Objective: You're a linguist interested in semantic relations between words. Given a pair of words, a cue word and a response word, classify the semantic relation between the cue and the response using one of the labels described in the specified criteria.
Variant 4	Objective: You're a linguist interested in semantic relations between words. Given a pair of words, composed by a cue and a response, classify the pair into its corresponding semantic relation using the labels described in the specified criteria.

Table 6: Variants for the first section of the prompts: description of the task to perform.

Criteria:
Synonym: The cue and the response are synonyms (CUE:recycle, RESPONSE:reuse)
Antonym: The cue and the response are antonyms (CUE:outside RESPONSE:inside)
Hypernym: The response is a hypernym of the cue (CUE:piano, RESPONSE:instrument)
Hyponym: The response is a hyponym of the cue (CUE:mammal, RESPONSE:human)
Meronym: The response is a meronym of the cue (CUE:face, RESPONSE:nose)
Holonym: The response is a holonym of the cue (CUE:plant, RESPONSE:garden)
Transposition: The cue and the response are synonyms but they have a different part-of-speech (CUE:anger, RESPONSE:mad)
Cohyponym: The cue and the response have a common hypernym but they are not synonyms (CUE:discourse, RESPONSE:conversation)
Response's argument: The cue is a syntactic argument of the response (CUE:rabbit, RESPONSE:hop)
Cue's argument: The response is a syntactic argument of the cue (CUE:filled, RESPONSE:cup)
Response's feature: The cue is a semantic feature of the response (CUE:explosive, RESPONSE:dynamite)
Cue's feature: The response is a semantic feature of the cue (CUE:sunset, RESPONSE:orange)
Response's modifier: The cue is used as a modifier of the response (CUE:custard, RESPONSE:pudding)
Cue's modifier: The response is used as a modifier of the cue (CUE:friend, RESPONSE:best)
Typical: There is an obvious (and implicit) predicate that links the cue and the response or when the eventuality denoted by the cue typically cooccurs with that denoted by the response (or vice versa) (CUE:incense, RESPONSE:church)
No relation: No semantic relation between the cue and the response (CUE:rally, RESPONSE:pep)
Other relation: The cue and the response are in a semantic relation of different type than those listed in our labels (CUE:saucy, RESPONSE:sauce)

Table 7: Section 2 of the prompts, listing the labels, each accompanied by a description, and an example. The examples are provided only in the "implicit few-shot" setting.

Type format	Description
Format 1	Input and Output format: The input follows the format: 'Input: CUE:cue_word, RESPONSE:response_word' where cue_word is the cue word and response_word is the response word. The output follows the format: 'Output: CUE:cue_word, RESPONSE:response_word, LABEL:label' where cue_word is the cue word and response_word is the response word and label is one of the labels in the specified criteria. Generate only the content without explanations following strictly the output format.
Format 2	Input and Output format: The input follows the format: 'CUE:cue_word, RESPONSE:response_word' where cue_word is the cue word and response_word is the response word. The output follows the format: 'CUE:cue_word, RESPONSE:response_word, LABEL:label' where cue_word is the cue word and response_word is the response word and label is one of the labels in the specified criteria. Generate only the content without explanations following strictly the output format.
Format 3	Input and Output format: The input is a pair of words that follows the format: 'CUE:cue_word - RESPONSE:response_word' where cue_word is the cue word and response_word is the response word. As output, return the corresponding label. Generate only the content without explanations following strictly the output format.

Table 8: Variants for the Section 3 of the prompts: desired input/output formats. The Format 3 variant was retained after tests on the Llama models.

Label (Corresp.)	**F1**	**Nb**
antonym (= Antonym)	55	8
synonym (= Synonym)	78	122
material made of ($\subset$ R/C:modifier)	40	2
has property ($\subset$ R/C:feature)	69	81
location ($\subset$ R/C:role)	34	43
category exemplar ($\subset$ R:hyper+R:hypo)	35	42
function	44	52
part of ($\subset$ (R:holo+R:mero))	37	38
common phrase	52	69
action ($\subset$ R:role)	27	104
emotion evaluation	26	18
time	46	19
result in	25	49
has prerequisite	23	22
thematic ($\subset$ R:role)	23	44
same category (= cohyponym)	18	12

Table 9: GPT-4o F1-scores and number of instances in the human-labeled WAX dataset (for each WAX label). Rough correspondence to FAMWA labels is indicated in parentheses, when applicable.