Community-Authored Usage Glosses as Use-Conditions for Internet Memes

Victoria Sherratt **Nina Dethlefs**
Loughborough University
{v.sherratt, n.dethlefs}@lboro.ac.uk

Abstract

Convention in language normally emerges through interaction, yet meme producers who share no joint activity nonetheless converge on recognisable patterns when filling a given format. Community-authored usage glosses from meme encyclopedias document such conventions informally, describing when and how a format is deployed. We formalise these glosses as use-conditions, group them into function families and test whether they predict production regularities across grammatical, semantic, and visual layers over 1,622 documented meme formats. We find that different families bind at different layers, and the community's own functional distinctions capture genuine pragmatic constraints that emerge without interaction.

1 Introduction

Internet memes have diverse modes of participation, where users repurpose pre-existing templates, catch phrases, images, or modify copied text to communicate ideas. The conventions governing these modes of participation vary by format and are neither enforced nor formally defined; users learn what a template is for by seeing how others have used it and produce new instances consistent with the conventions they infer without interacting with other producers.

Encyclopedia sites such as KnowYourMeme[1] maintain community-authored descriptions of how meme formats are conventionally used; a template is 'used to mock' a target, 'used to express' disappointment, or 'used to describe' a person or situation. These glosses are written by participants documenting and researching formats from observing typical collective use and maintained as an ongoing editorial record of how some memes function as communicative practices.

We treat these informal, post-hoc community-glosses as a record of conventions that have emerged through participation in meme culture, and ask whether these descriptions capture real structure to serve as a basis for formalising otherwise invisible pragmatic rules that govern meme production.

We contribute a community-grounded taxonomy of documented meme use-conditions, derived by mapping usage-gloss predicates onto a small set of functional families via WordNet hypernymy.[2] Additionally, we develop a population-scale, multiply-controlled measurement demonstrating that documented function structures and binds meme production selectively at different layers, and that this structure does not survive collapse into a classical speech-act grouping (Searle, 1976).

We offer a framework for understanding meme convention based on community-documented function rather than on visible features or learned representations. By testing whether community-authored descriptions reliably predict regularities in how formats are filled, we provide evidence that these descriptions encode regularities of how meme formats are used or reshaped by users or constrain aspects of meme production.

2 Background and Related Work

Frameworks that account for convergence in language are built around interaction. Interlocutors in dialogue converge through mutual priming at multiple linguistic levels (Pickering and Garrod, 2004); convention emerges from repeated coordination between agents who adjust under feedback (Lewis, 1969; Clark, 1996), and common ground accumulates through joint activity (Clark, 1996; Hawkins et al., 2023). How coordination between pairs becomes convention shared by a population is less settled; Hawkins et al. (2023) model it as the gradual abstraction of expectations across repeated

[1] https://www.knowyourmeme.com

[2] Taxonomy available at https://github.com/vemchance/meme_conditions/.

encounters with different partners and Boyce et al. (2024) show that when participants can only broadcast without receiving responses, groups fail to converge on shared conventions.

However, in meme production users work from a shared archive of prior instances without addressing, responding to, or receiving feedback from one another, yet regularities in how they fill a given format nonetheless emerge. Participation in meme culture is itself a form of identity work; users adopt and reproduce recognisable formats to signal belonging and cultural fluency within online communities (Jenkins, 2006; Shifman, 2014; Milner, 2016) and adherence to a format's conventions is part of what makes participation legible as participation (Nissenbaum and Shifman, 2017). This accounts for why users conform once conventions exist, but not for what structure they take without interaction, or how they selectively constrain expression at scale.

The cognitive-linguistic study of memes has established that a format shapes the language placed within it. Dancygier and Vandelanotte (2017) analyse image macro memes as multimodal constructions in which images fill constructional slots, with surrounding text adjusting to the resulting multimodal frame; Dancygier and Vandelanotte (2025, 1–10) later generalise this into a grammar of memes and Zenner and Geeraerts (2018) similarly treat image macros as non-compositional constructions whose conventional meaning is partly carried by recognising the format a meme instance belongs to.

Cyberpragmatics extends relevance theory to internet-mediated communication, treating the affordances of a platform or format as constraints on the contextual information a reader brings to interpretation (Yus, 2011); Yus (2019) applies this to image macros, classifying text and image relations by how much interpretive work each asks of the reader and what it returns. Instead, the regularities we measure in this work are properties of what producers make across a population of documented formats, connected to functional categories that communities themselves use.

The broader computational landscape organises memes by computed features; perceptual hashing and visual clustering (Zannettou et al., 2018), multidimensional similarity over local and global features (Bloem and Ilievski, 2025), and learned embeddings all discover structure from the instances themselves. A substantial body of work classifies what memes communicate, from sentiment to hate speech and propaganda, leaving the conventions governing their deployment largely unaddressed (Nguyen and Ng, 2024; Wang et al., 2026).

Zhou et al. (2024) offer the closest comparative study to our work by clustering individual memes into visual templates and using the fill text to infer what each template is used for, demonstrating that template choice indexes community identity and formats diversify over time. Instead, we treat community-authored glosses written after a format has been established as use-conditions, or documentation of when and how a format is felicitously deployed rather than what it describes or states (Gutzmann, 2015).

3 Methodology

We extract community-authored usage glosses from meme encyclopedia descriptions, group them into function families, and test whether formats sharing a documented function produce more similar output than formats that do not across four layers covering function-word distributions, part-of-speech structure, semantic content, and visual appearance. Figure 6 in Appendix A.1 summarises the pipeline.

3.1 Data

KnowYourMeme (KYM) is a community meme encyclopedia documenting internet memes, with a wiki-style 'entry page' documenting each meme phenomenon and describing its origins, usage, or history, alongside a gallery of example memes for each. We use a publicly available resource of KYM which provides 16,707 entries, their metadata, editorial text, and instance galleries of meme images and OCR-extracted caption text for the following analysis (Sherratt et al., 2026).

Each entry page contains a page title used throughout as the **meme format**; a reusable image, phrase, or text pattern that many users reproduce with variation. Image galleries on each entry page contain user-contributed images as **instances** of these formats.[3] Within each entry's free-text description, editors routinely document how a meme format is conventionally deployed with phrases such as 'used to express disappointment,' 'posted in response to bad takes,' 'captioned with ironic praise.' We call these **usage glosses**, community-

[3]Example meme format: https://knowyourmeme.com/memes/surprised-pikachu

authored annotations that state a format's conventional function.

We filter available data to formats meeting three conditions, resulting in 1,622 formats with 78,014 instances for analysis: the entry must be of meme type and marked Confirmed (moderator-approved), reducing 16,707 entries to 11,582; its free-text description must contain at least one extractable usage gloss (Section 3.2), which 27.1% of Confirmed entries do; and it must have at least 20 gallery instances carrying OCR-extracted text, so that a format's linguistic profile can be reliably estimated.

KYM also records metadata about internet memes; the analysis in Section 4 also leverages tags (user-inputted labels about meme concepts) and meme types, which refer to the specific subgenre, tone, or style of meme (e.g., an image macro, copypasta, exploitable).

3.2 Extraction of Use-Conditions

We extract usage glosses automatically from entry descriptions and group the governing predicates into function families.

3.2.1 Extraction

Editors on KYM document format functions through a small number of recurring constructions. Thus, we formalise five extraction patterns from these, applied to sentence-split entry descriptions using spaCy (Honnibal et al., 2020). The primary pattern is an open 'used to/as/for/in/when' frame that extracts whatever verb the dependency parse yields after the construction, without restricting which verbs can appear; this single pattern accounts for 65% of extracted glosses. Four supplementary patterns capture constructions not headed by 'used to,' such as 'in response to,' 'captioned with,' and 'designed to express' (full specifications in Appendix A.2). Narrative sentences (origin and spread) and definitional copulas ('X is a photograph of…') are excluded as these document origin and identity rather than conventional use. This yields 4,323 usage glosses across 229 distinct predicates.

Extraction accuracy. We sampled 150 extracted glosses for precision (proportion that are genuine usage glosses) and 100 entries with no extraction for recall (proportion where a gloss was missed). Two annotators independently judged each item; precision was 74.7% (Cohen's $\kappa = 0.82$) and the recall audit found that 60.0% of unextracted entries contained at least one missed gloss ($\kappa = 0.79$) (Appendix A.4). Appendix A.4.1 compares this against an open extraction unrestricted by construction.

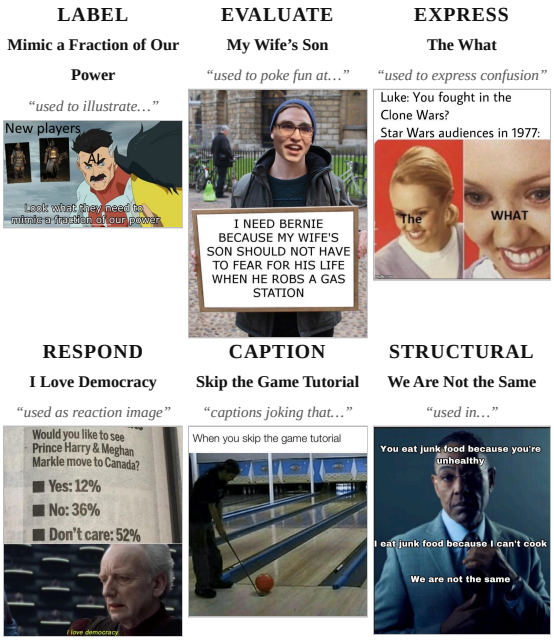

Figure 1: Example meme format and instance within each function family.

3.2.2 Function Families

The 229 extracted predicates fall into recognisable functional types, for example some document what a meme labels or categorises (*describe, refer, depict*), others what it evaluates (*mock, criticize, parody*), or what triggers its deployment (*react, respond*). To formalise this grouping without manual assignment, we use WordNet hypernymy; each family is anchored to a set of synsets that define its semantic core (e.g. EVALUATE is anchored to *mock.v.01*, *ridicule.v.01*, *disparage.v.01*, *praise.v.01*). Any predicate whose verb senses reach an anchor through hypernym closure maps to that family automatically. Six function families contain enough formats for testing and are used throughout (Table 1, Figure 1 for example formats). Full anchor sets and mapping are in Appendix A.3.

Family accuracy. The same 150-gloss sample as in Section 3.2.1 was annotated to judge whether the WordNet-assigned family matches the predicate's function in context. Overall family-label accuracy was 74.3% (Cohen's $\kappa = 0.67$), discussed further in Appendix A.4.

Family	Example predicates	Fmts	Clauses
LABEL	*describe, refer, depict*	152	309
EVALUATE	*mock, criticize, parody*	187	461
EXPRESS	*express, convey, indicate*	369	854
RESPOND	*react, respond, reply*	282	535
CAPTION	*caption, pair*	264	453
STRUCTURAL	*use in, use as, use for*	706	1,537

Table 1: Function families from community-authored usage glosses via WordNet hypernymy. *Fmts* = formats carrying at least one clause in this family (multi-label; a format may belong to more than one family). *Clauses* = total extracted usage glosses mapped to family.

3.3 Experimental Setup

Each format is represented by profiles computed over its gallery instances, and every pair of formats is compared on each of the four layers. Same-family pairs (both formats carry a given family) are compared to cross-family pairs (at most one does). The four layers are grouped into form and content:

Form. We compute two distributional profiles per format from its instance texts. Firstly, the relative frequency of function words (FW; closed-class items such as pronouns, determiners, prepositions and conjunctions); because function words carry grammatical rather than topical information, two formats can have similar profiles without sharing any subject matter. Second, the distribution of part-of-speech trigrams (POS), which captures sentence shape with vocabulary removed entirely.

Content. For semantic content (referred to as 'text'), each format is represented by the centroid (mean vector) of its instance-text embeddings using BGE-large-en-v1.5 in symmetric mode (Xiao et al., 2023). For visual appearance, each format is represented by the centroid of its instance-image embeddings using SigLIP-base (Zhai et al., 2023).

To equalise the precision of profile estimates across formats of different sizes, instances are capped at 100 per format via seeded random sampling. Site watermarks embedded in OCR text (e.g. `imgflip.com`, `9gag.com`) are stripped prior to profiling to prevent spurious similarity.

3.3.1 Test Design

We apply four tests examining the relationship between documented function and production. All use permutation (10,000 label shuffles, two-sided) to assess significance.

Between-format, centroid level. For each pair of formats we compute a similarity score on each layer (negative Jensen–Shannon divergence for function words; cosine similarity for all others), producing 1,314,631 pairs with four scores each. Same-family pairs are compared to cross-family pairs. This is the primary test used in the grid (Section 4.1), the type strata (Section 4.2), and controls (Appendix A.5).

Between-format, instance level. The same comparison, but on individual instance pairs rather than format-level profiles, testing whether content binding holds meme by meme or only in the aggregate (Section 4.3.1).

Within-format consistency. For each format, the mean pairwise similarity of its own instances is computed (text and visual separately), producing one consistency score per format. Formats carrying a family are then compared to non-carriers in a one-vs-rest test. This examines whether function families differ in how much they constrain production within a meme format (Section 4.3.2).

Linguistic profiles. For each format, 16 parse-derived feature rates are computed from its instances using spaCy (e.g. subordination rate, when-clause rate, pronoun rates, sentence length). Each family $\times$ feature combination is tested one-vs-rest at the format level. This asks whether families produce distinctive grammatical profiles on specific constructions (Section 4.3.3).

3.3.2 Controls

Formats sharing a function may also share topic, inflating similarity through shared rhetorical conventions rather than shared function; we address this by restricting to tag-disjoint pairs and by a pairwise regression controlling for tag overlap, format type, and size (Appendix A.5).

Similarly, the content layers capture semantic and visual similarity by design, so function and topic are entangled. We do not attempt to separate them, however, we also control for two technical confounds on the content layers. Format-level centroids could reflect repeated template text rather than varied user production and embedding spaces may exhibit anisotropy; we verify text diversity and mean-centre embeddings to address both (Appendix A.5). The type strata (Section 4.2) separately test whether effects hold within structural types.

4 Results

Throughout, effects are reported as the difference in mean pairwise similarity between same-family format pairs and cross-family format pairs; a positive value means formats sharing a documented function are more similar on that layer than formats that do not.

We first test whether formats that share at least one function family are more similar than formats that share none, without distinguishing which family they share. Across 1,314,631 format pairs, the pooled effect on function-word similarity is -0.005 ($p = .06$) and on part-of-speech trigrams -0.006 ($p = .07$). Text and image similarity are also null. Treated as a single undifferentiated variable, shared family membership predicts nothing about convention in meme formats.

The form measures trend weakly negative rather than zero; either documented use fails to predict production, or different kinds of documented use predict different layers and pooling cancels them. The rest of this section separates the two.

4.1 Family × Layer Grid

Table 2 shows each family tested separately against the rest of the population on each of the four similarity layers. The left two columns measure grammatical form independently of topic; the right two compare format centroids on caption-text and image embeddings (content).

Family	FW	POS	Text	Visual
LABEL	**+0.042**	**+0.061**	**+0.023**	**+0.015**
EVALUATE	**+0.018**	**+0.024**	−0.000	−0.000
EXPRESS	+0.005	+0.006	**+0.010**	**+0.009**
RESPOND	+0.006	−0.008	+0.007	+0.007
CAPTION	**−0.019**	**−0.027**	−0.012†	+0.001
STRUCTURAL	−0.008	−0.007	−0.003	−0.002
Omnibus	−0.005	−0.006	+0.001	+0.002

Table 2: The family × layer grid. Each cell is the within-family minus between-family mean pairwise similarity. Bold = $p < .01$. Omnibus row pools all families. †Does not survive anisotropy control (Table 12).

LABEL and EVALUATE bind form; formats documented with these functions are grammatically more alike than those without, with LABEL the largest effect in the grid (+0.042/+0.061, both $p < .0001$), although neither binds text or image. EXPRESS and RESPOND show the opposite pattern, binding text and image but not form. LABEL is the only family that binds all four layers.

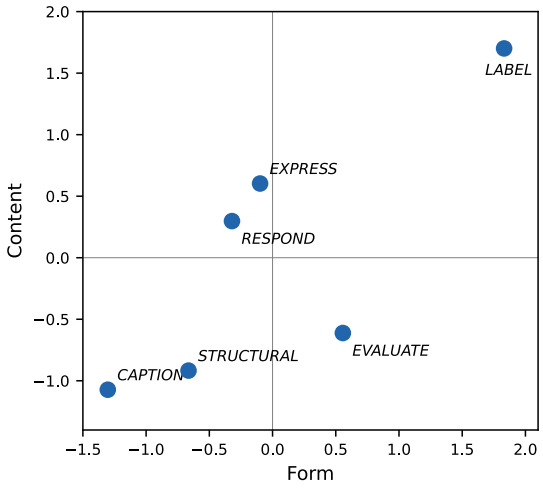

Figure 2: Family placement on form and content composite axes. Each axis is the mean of z-scored effects from two layers (form = FW neg-JSD + POS cosine; content = caption-text cosine + image cosine).

CAPTION is anti-coherent on form (-0.019/-0.027, $p < .0001$); formats documented as captioning practices are *less* formally alike than the population average. The pooled null follows from this structure, as CAPTION's negative form effects cancel the positive ones and the content-binding families contribute nothing to form. Figure 2 summarises the dissociation by plotting each family on composite form and content axes.

4.2 Type Strata

We test whether the effects previously described are carried uniformly across format types or concentrated in specific structural contexts, using the same test restricted to formats of one type. In Table 3, we report four types with at least 20 carriers in at least one family and the families with significant or interpretively relevant effects; remaining cells are null and omitted for space.

EVALUATE's form effect survives within catchphrases on both FW (+0.043, $p = .001$) and POS (+0.059, $p < .001$) but not within image macros ($p > .7$ on both). EXPRESS's text effect concentrates in image macros (+0.024, $p < .001$), more than double the pooled effect, and is also present in reactions (+0.013, $p = .038$).

The visual layer confirms this pattern, where EXPRESS visual binding is significant in image macros (+0.015, $p = .004$) and RESPOND similarly shows a visual effect in image macros (+0.015, $p = .008$),

Layer	Family	All	IM	Expl.	Catch.	React.
FW	EVALUATE	**+.018**	−.004	−.009	**+.043**	–
	CAPTION	**−.019**	−.003	−.018	−.032	−.015
POS	EVALUATE	**+.024**	+.005	−.024	**+.059**	–
	CAPTION	**−.027**	−.011	−.026	−.026	−.027
Text	EXPRESS	**+.010**	**+.024**	+.002	+.014	+.013
	RESPOND	+.007	+.003	+.015	+.009	+.007
	EVALUATE	−.000	−.003	−.020	**+.024**	–
Vis.	EXPRESS	**+.009**	**+.015**	+.018	+.003	+.014
	RESPOND	+.007	**+.015**	+.010	+.000	+.005
	EVALUATE	−.000	+.000	−.008	+.011	–

Table 3: Type strata across all four layers. Each cell is the one-vs-rest effect restricted to formats of that type. All = unstratified per-family effect from Table 2. Bold = $p < .01$. – = fewer than 20 carriers.

though image macros share a fixed template image by definition, so visual convergence within this type is partly expected. EVALUATE shows no text effect in the pooled grid ($p = .93$); within catchphrases a positive effect appears (+0.024, $p = .006$), though catchphrases similarly share fixed text, so text convergence within this type is partly expected.

4.3 Convention Characterisation

The previous sections test whether formats sharing a documented function are more similar than formats that do not, using format-level averages. In this section, we first examine whether the binding in Table 2 holds when we compare individual memes rather than format centroids, then examine whether function families differ in how much they constrain production within a single format or produce distinctive grammatical profiles across the instances that compose each format. Effects are reported as the difference between carrier and non-carrier format means rather than between format pairs.

4.3.1 Instance-level Binding

We re-run the content (text and visual) tests at the instance level (Table 4), comparing individual cross-format instance pairs rather than format centroids; the form layers cannot be tested at this grain because individual meme captions are typically too short to produce stable function-word or POS distributions. CAPTION and STRUCTURAL are omitted as neither has a centroid-level effect on text or visual that survives anisotropy control (Table 12) to decompose.

On text, LABEL binds at both centroid and instance level (+0.006, $p = .0002$). EXPRESS binds at the centroid level but is entirely absent at the instance level (−0.00005, $p = .97$); we return to this dissociation in Section 5.

	Text		Visual	
Family	Centroid	Instance	Centroid	Instance
LABEL	**+0.023**	**+0.006**	**+0.015**	**−0.009**
EXPRESS	**+0.010**	−0.000	**+0.009**	**+0.005**
RESPOND	+0.007	−0.001	+0.007	**+0.011**
EVALUATE	−0.000	+0.007	−0.000	**−0.007**

Table 4: Binding at two grains on text and visual layers. Centroid = between-format test on format-level centroids (from Table 2). Instance = between-format test on individual instance pairs. Bold = $p < .01$.

On visual, EXPRESS and RESPOND bind at both grains; individual memes from EXPRESS formats are visually more similar to memes from other EXPRESS formats than to the rest of the population (+0.005, $p = .001$), suggesting the conventions of this function family reach individual images but not individual captions. LABEL's visual centroid binding (+0.015, Table 2) reverses at the instance level (−0.009, $p < .001$); individual LABEL memes are too visually diverse within their own formats for the between-format convergence to survive at this grain, which we discuss further in Section 5.

4.3.2 Within-format Consistency

Table 5 shows within-format consistency, measured as the mean pairwise similarity of a format's own instances. LABEL formats have the most within-format variation on both text and visual layers, yet produce the strongest between-format signal in Table 2; individual instances within a LABEL format diverge widely, but LABEL formats as a group converge. CAPTION formats show the opposite pattern, with the highest within-format constraint on visual appearance.

Family	Text	Visual
LABEL	**−0.022**	**−0.048**
EVALUATE	+0.015	**−0.020**
EXPRESS	**−0.016**	−0.001
RESPOND	**−0.016**	**+0.016**
CAPTION	+0.014	**+0.028**
STRUCTURAL	+0.008	−0.002

Table 5: Within-format consistency by family. Effect = mean consistency of carriers minus non-carriers. Negative = more within-format variation; positive = more constrained. Bold = $p < .01$.

4.3.3 Linguistic Profiles

Table 6 shows the significant cells from a one-vs-rest permutation test across 16 parse-derived features and six families (96 cells; $p < .05$). CAPTION is the most linguistically distinctive family, with elevated subordination, when- and if-clauses, and second-person pronouns, and depressed exclamation. RESPOND shows elevated when-clauses and second-person address, whereas EVALUATE has longer sentences and depressed when-clause rates. LABEL has no significant individual features and EXPRESS has only a marginal when-clause elevation ($p = .048$).

Family	Feature	Effect	p
CAPTION	subordination	+0.027	.038
	when-clauses	+0.023	.010
	if-clauses	+0.012	.023
	2nd person	+0.005	.001
	3rd person	+0.003	.029
	exclamation	−0.015	.032
RESPOND	when-clauses	+0.039	<.001
	2nd person	+0.004	.035
EVALUATE	sentence length	+0.87	.006
	when-clauses	−0.034	.002
	2nd person	−0.004	.045
EXPRESS	when-clauses	+0.017	.048
STRUCTURAL	when-clauses	−0.023	<.001

Table 6: Significant cells ($p < .05$) from the family × feature linguistic profile grid (13 of 96 cells). Features are parse-derived rates per format (spaCy); effect = carrier mean minus non-carrier mean.

A multivariate test on the full 16-feature profile confirms grammatical distinctiveness for RESPOND ($p = .008$) beyond these individual features but is null for LABEL ($p = .20$).

5 Discussion

The pooled omnibus is null; the structure appears only when function is resolved into the community's own categories. We test whether a coarser functional grouping would serve as well by regrouping the families into Searle-style speech-act classes, which merges EVALUATE (form-binding) with EXPRESS (content-binding) into a single expressive class and produces a worse omnibus (-0.011, $p < .0001$ vs -0.005, $p = .06$; Appendix A.6).

LABEL and EVALUATE's form binding holds after controlling for shared topic, format type and size (Appendix A.5). The two LABEL formats

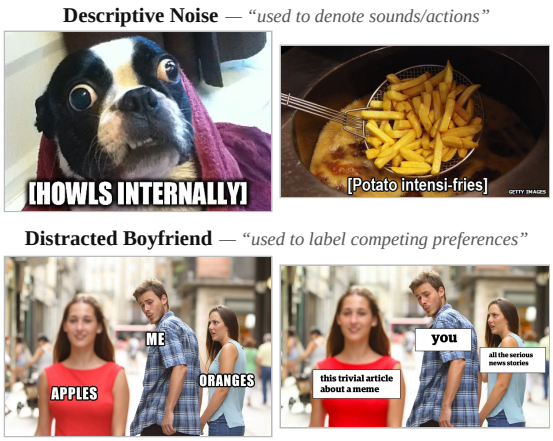

Figure 3: Two LABEL formats producing short categorising text despite different topics and visual diversity.

in Figure 3 address different topics using different conventions; *Descriptive Noise* applies bracketed notation to visually diverse submitted images whilst *Distracted Boyfriend* applies noun-phrase labels to a fixed template. In both, users map a scenario onto the format's visual structure by naming its parts; the text categorises rather than narrates, and this shared act of categorisation through naming produces short labelling fragments across 152 LABEL formats.

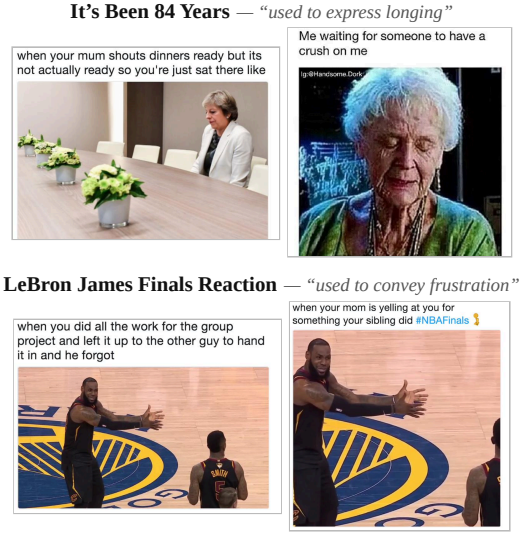

Figure 4: Two EXPRESS formats. Individual captions narrate different situations; format centroids converge within the same emotional domain.

On content, EXPRESS conventionalises differently. Formats documented as expressing emotions bind text and image at the centroid level but text binding is entirely absent at the instance level

(-0.00005, $p = .97$). In Figure 4, captions within each format are about different everyday situations; any two cross-format captions are no more similar than a random pair, but users of each format tend to write about the same kinds of situations, so the format-level averages converge. The convention constrains what users write *about* without constraining how they write or what any individual caption says.

On the visual layer, this convention reaches individual images ($+0.005$, $p = .001$); EXPRESS formats look more alike meme by meme even where their captions do not. EXPRESS formats in Figure 4 use different template images, but other formats in this family reuse a single image with open white space for user-inputted text; in both cases, the convention of writing about similar situations is maintained and anchored to what the chosen format's image helps express.

Philosoraptor — *"paired with captions depicting…"*

Figure 5: CAPTION format *Philosoraptor* with a fixed if-clause conditional frame and fixed template image.

CAPTION shows a third pattern; it is anti-coherent on aggregate form (-0.019/-0.027, both $p < .0001$) yet the most linguistically distinctive family in the profiles (Table 6). In Figure 5, *Philosoraptor* uses an if-clause conditional frame; other CAPTION formats use when-clause setups or substitution frames, and all exhibit the grammatical features that define the family whilst diverging enough in surface syntax to register as anti-coherent in the aggregate. These are what Nissenbaum and Shifman (2018) call 'expressive repertoires' that simultaneously enable and constrain expression; the usage gloss documents *how* to participate rather than what to say. CAPTION also has the highest within-format visual consistency ($+0.028$, $p < .01$); the template image is fixed but the caption slot is open, a population-scale instance of the varied formal realisations of a single constructional meaning that Dancygier and Vandelanotte (2025, 151–170) describe for individual formats.

The type strata similarly reinforce this; EVALUATE's form effect survives within catchphrases ($+0.043$, $p = .001$) but is absent in image macros, whilst EXPRESS's text binding concentrates in image macros ($+0.024$, $p < .001$). Function families cut across format types, and what the convention can constrain depends on what the format structurally affords; catchphrases carry fixed text that gives form constraints something to bind, whilst image macros provide open caption slots where content constraints operate, a production-side counterpart to the interface constraints cyberpragmatics identifies on the interpretation side (Yus, 2011).

6 Conclusion

The text on example memes in the figures throughout Section 5 is free text rather than fills from a pre-defined template, yet the regularity is in the choices users make when the format leaves them free to choose. Community-authored usage glosses encode pragmatic constraints on meme production selectively rather than uniformly, in interaction with what the format structurally affords; users select formats whose visual and structural properties support the communicative act the documentation describes, and the community's own functional distinctions preserve this structure.

Grouping formats by documented function is inherently different from grouping by computed similarity; function organises formats that are *used* in the same way. CAPTION formats are formally dissimilar to one another yet share a documented mode of participation, so a grouping built from formal similarity could not assemble this set; EXPRESS formats produce no individual caption pair above chance yet converge at the distributional level on the emotional domain their documentation selects for. Neither pattern would surface from clustering by visual or textual features, yet both are measurable across a population of documented formats, extending qualitative observations that individual formats shape the language placed within them to a population-level account (Dancygier and Vandelanotte, 2025; Zenner and Geeraerts, 2018).

For computational approaches, which largely organise memes by computed features and classify what memes communicate, community-documented function offers a complementary organising principle to visual or semantic similarity, describing how a population deploys a format rather than what its instances contain. Memes can

then be studied through such conventions, for example by sourcing data for content moderation by documented use (e.g. formats 'used to mock'), layering function over existing content analysis, examining how formats constrain expression, or analysing why users select formats by function.

7 Limitations

The analysis is limited to English-language memes from a single source, and function families are derived from KYM's editorial conventions which may not generalise to undocumented formats or non-English meme cultures; what Nissenbaum and Shifman (2018) describe as repertoires varying by culture and language implies that communities with different documentation practices could produce different families.

The extraction covers 27% of confirmed entries, and the 20-instance floor further restricts the population to formats with sufficient gallery content for stable profiling. With 1,314,631 format pairs, permutation significance is easy to reach and individual effects are small; the contribution is the selective pattern across families and layers rather than the magnitude of any single cell.

KYM's glosses are post-hoc editorial descriptions rather than exhaustive specifications, and null effects may reflect gaps in documentation rather than absence of convention; however, the production signatures associated with each family could be used to predict family membership for formats lacking glosses, testing whether the pattern strengthens as coverage expands.

Not all effects are uniform; some families show clean binding on specific layers whilst others show effects that concentrate in particular format types, reverse at finer grains, or do not survive anisotropy controls. OCR-extracted text is treated as a single block without positional decomposition; separating slot-specific contributions (e.g. top text vs bottom text in image macros) could clarify whether form binding reflects the full caption or specific slots, and could help disambiguate families whose effects are uneven across format types.

Instance-level timestamps are similarly absent from the resource; the results establish that documented function predicts production regularities cross-sectionally, and temporal decomposition would allow testing whether conventions tighten as formats accumulate instances or vary across a format's lifespan.

References

Peter Bloem and Filip Ilievski. 2025. Clustering internet memes through template matching and multi-dimensional similarity. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):259–276.

Veronica Boyce, Robert D. Hawkins, Noah D. Goodman, and Michael C. Frank. 2024. Interaction structure constrains the emergence of conventions in group communication. *Proceedings of the National Academy of Sciences*, 121(28):e2403888121.

Herbert H. Clark. 1996. *Using Language*. "Using" Linguistic Books. Cambridge University Press.

Barbara Dancygier and Lieven Vandelanotte. 2017. Internet memes as multimodal constructions. *Cognitive Linguistics*, 28(3):565–598.

Barbara Dancygier and Lieven Vandelanotte. 2025. *The Language of Memes: Patterns of Meaning Across Image and Text*. Cambridge University Press.

Daniel Gutzmann. 2015. *Use-Conditional Meaning: Studies in Multidimensional Semantics*. Oxford University Press.

Robert D. Hawkins, Michael Franke, Michael C. Frank, Adele E. Goldberg, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. 2023. From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review*, 130(4):977–1016.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Henry Jenkins. 2006. *Convergence Culture: Where Old and New Media Collide*. NYU Press.

David Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press.

Ryan M. Milner. 2016. *The World Made Meme: Public Conversations and Participatory Media*. The MIT Press.

Khoi P. N. Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267, Miami, Florida, USA. Association for Computational Linguistics.

Asaf Nissenbaum and Limor Shifman. 2017. Internet memes as contested cultural capital: The case of 4chan's /b/ board. *New Media & Society*, 19(4):483–501.

Asaf Nissenbaum and Limor Shifman. 2018. Meme templates as expressive repertoires in a globalizing world: A cross-linguistic study. *Journal of Computer-Mediated Communication*, 23(5):294–310.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

John R. Searle. 1976. A classification of illocutionary acts. *Language in Society*, 5(1):1–23.

Victoria Sherratt, Suzanne Elayan, and Nina Dethlefs. 2026. SemioMeme: A symbolic–subsymbolic knowledge graph dataset for multimodal meme analysis. *Proceedings of the International AAAI Conference on Web and Social Media*, 20(1):2921–2935.

Limor Shifman. 2014. *Memes in Digital Culture*. The MIT Press.

Bingbing Wang, Jingjie Lin, Zhixin Bai, Zihan Wang, Shengzhe Sun, Zhengda Jin, Zhiyuan Wen, Geng Tu, Jing Li, Erik Cambria, and Ruifeng Xu. 2026. Internet meme on social media: A comprehensive review and new perspectives. *Information Fusion*, 130:104102.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general Chinese embedding. *Preprint*, arXiv:2309.07597.

Francisco Yus. 2011. *Cyberpragmatics: Internet-Mediated Communication in Context*, volume 213 of *Pragmatics & Beyond New Series*. John Benjamins, Amsterdam.

Francisco Yus. 2019. *Multimodality in Memes: A Cyberpragmatic Approach*, pages 105–131. Springer International Publishing, Cham.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, page 188–202, New York, NY, USA. Association for Computing Machinery.

Eline Zenner and Dirk Geeraerts. 2018. *One does not simply process memes: Image macros as multimodal constructions*, pages 167–194. De Gruyter, Berlin, Boston.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952.

Naitian Zhou, David Jurgens, and David Bamman. 2024. Social meme-ing: Measuring linguistic variation in memes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3005–3024, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

A.1 Pipeline

Figure 6 summarises the pipeline as two independent strands. The text of each entry supplies the extracted glosses and their family assignments, whilst the gallery instances supply the four production profiles; family labels and pairwise similarities meet only in the final test, which compares same-family to cross-family pairs.

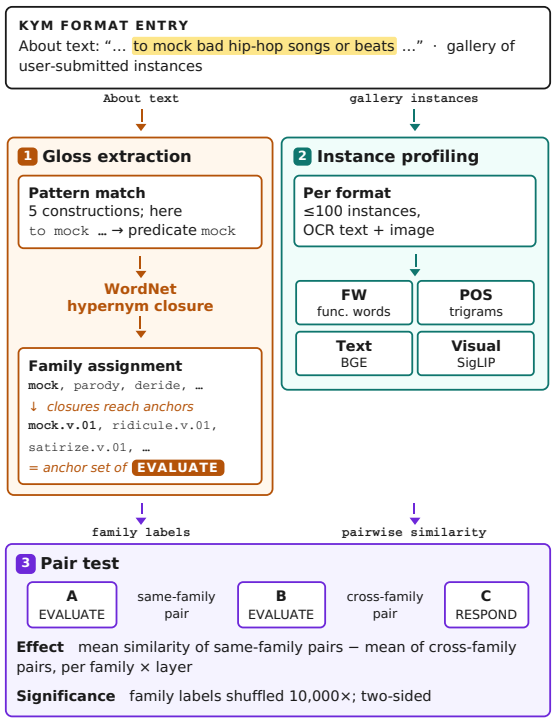

Figure 6: Processing pipeline from KYM entries to the pair test, traced through one extracted gloss.

A.2 Extraction Pattern Specification

Five patterns extract usage glosses from sentence-split entry descriptions parsed with spaCy `en_core_web_sm` (Honnibal et al., 2020). The patterns were identified by surveying how KYM editors document format function.

Pattern 1: open *used to* frame (primary). Matches *used* + *to*/*as*/*for*/*in*/*when*, with optional preceding modifiers (*is*, *commonly*, *often*, etc.). When the preposition is *to*, the governed verb is extracted via dependency parse, walking forward past the particle and any adverbs to the first content word. No verb restriction is used. For other prepositions, the construction itself is recorded as the predicate (*use_as*, *use_in*, etc.). This pattern accounts for 2,792 of 4,323 kept clauses (65%).

#	Gloss (from About Text)	Predicate
1	'is often used to **illustrate** how one party is angry'	*illustrate*
1	'commonly **used as** a reaction image'	*use_as*
1	'**used in** imageboards and forum conversations for trolling purposes'	*use_in*
2	'in **response** to discussion threads or posts that are perceived as derailing'	*response*
2	'**reacts** to a variety of unrelated statements'	*react*
3	'**captioned** with the phrase Tony use to eat cheeseburgers'	*caption*
3	'**paired** with captions that followed the White People When phrasal template'	*pair*
4	'designed to **express** disappointment or frustration'	*express*
4	'to **show** comradery and a willingness to keep going in life'	*show*
5	'to **mock** bad hip-hop songs or beats'	*mock*
5	'to **parody** videos of an Indonesian commercial'	*parody*

Table 7: Example extracted glosses. Bold marks the matched trigger or predicate.

Pattern 2: reactive constructions. Matches *in/as a* + *response/reaction/reply* + *to*; *react(ing/s) to*; *used when*; *posted* + *in response/when/after*.

Pattern 3: caption constructions. Matches *captioned* + *with/to/as*; *paired with*; *with captions*. Fixed predicates (*caption*, *pair*).

Pattern 4: expressive extension. Matches *to* + one of seven verbs (*express*, *convey*, *indicate*, *signal*, *denote*, *show*, *communicate*), extending reach to constructions not headed by *used to*.

Pattern 5: evaluative extension. Matches *to* + one of eleven verbs (*mock*, *criticize/criticise*, *ridicule*, *deride*, *troll*, *parody*, *satirize/satirise*, *make fun of*, *poke fun at*, *celebrate*, *praise*).

Exclusions and clause boundaries. Sentences containing narrative vocabulary (*began*, *originated*, *gained*, *went viral*, *was uploaded*) are excluded. Within Pattern 1, *used to be/have* is excluded to filter the past-habitual reading or functional use of memes. Extracted clauses run from the match to the sentence end or to a clause boundary (semicolons; colons; commas before *which*, *who*, *although*, etc.).

A.3 Function Families

Each extracted predicate is mapped to a function family via WordNet hypernymy (WordNet 3.0). A predicate's verb senses are collected and their full hypernym closures computed; if any synset in the closure matches a family's anchor set, the predicate maps to that family. Table 8 lists the anchor synsets.

Family	Anchor synsets
RESPOND	*react.v.01, answer.v.01*
EVALUATE	*mock.v.01, mock.v.02, ridicule.v.01, disparage.v.01, satirize.v.01, praise.v.01, knock.v.06*
EXPRESS	*express.v.01, express.v.02, convey.v.01, carry.v.04, indicate.v.03, sign.v.05, bespeak.v.01*
LABEL	*label.v.01, describe.v.01, portray.v.02, denote.v.01, denote.v.02, name.v.01, name.v.02, mention.v.01, picture.v.02, typify.v.02*
DIRECT	*request.v.01, request.v.02, solicit.v.01, ask.v.02, urge.v.01, recommend.v.01, encourage.v.03, invite.v.04, demand.v.01*

Table 8: WordNet anchor synsets per family.

Constructional rules. Two families are defined by syntactic construction rather than WordNet path. Predicates recording the frame itself (*use_as*, *use_for*, *use_in*, *create*) map to STRUCTURAL; predicates from the caption extraction pattern (*caption*, *pair*) map to CAPTION. Four reactive predicates (*response*, *reaction*, *use_when*, *post*) map to RESPOND by the same mechanism.

Supplement. A ratified list of 14 verbs overrides WordNet where its paths fail or mislead: EVALUATE gains *insult*, *shame*, *celebrate*, *commemorate*, *mimic*, *joke*; EXPRESS gains *complain*, *demonstrate*, *display*, *express*, *show*; LABEL gains *characterize*, *call*, *claim*. Two verbs (*hold*, *censor*) are pruned to OTHER.

Multi-hit resolution. When a predicate's closure reaches anchors in more than one family, the reading that constrains the most about the context of use is preferred: RESPOND specifies deployment context, EVALUATE specifies stance toward a target, DIRECT specifies an act toward an addressee, LABEL specifies referential function, and EXPRESS specifies only that an emotional state is conveyed. Thus, priority follows this order. A clause-level override also applies, so any clause whose complement head noun is *reaction* maps to RESPOND regardless of the predicate's primary assignment.

DIRECT. A seventh family, DIRECT (*request*, *ask*, *urge*, *tell*, *direct*, *lure*; 16 clauses, 10 formats), is included for descriptive completeness but falls below the contrast threshold and is excluded from all statistical tests.

Unmapped predicates. Of 4,323 extracted clauses, 4,165 (96.3%) map to a family. The remaining 158 clauses span 113 low-frequency predicates (most frequent: *make*, 15 clauses) whose WordNet paths do not reach any anchor.

A.4 Audit Procedure and Inter-Annotator Agreement

Procedure. Two annotators independently judged three samples drawn from the extraction output. The clause-level sample (n=150) was drawn uniformly from all extracted clauses; for each clause, annotators recorded whether the clause is a genuine usage gloss and whether the WordNet-assigned family is correct given the predicate's function in context. The entry-level sample (n=100) was drawn uniformly from confirmed entries with no extracted clause; for each entry, annotators recorded whether the About text contains a usage gloss the extractor missed, and quoted the missed text where present. Annotators worked from the clause text and the entry title; the recall audit additionally exposed the full About text. Annotators were not shown each other's judgements.

Agreement. Cohen's κ on the genuine-gloss judgement was 0.82 (simple agreement 93.3%); on the family-label judgement, $\kappa = 0.67$ (87.3%); on the missed-gloss judgement, $\kappa = 0.79$ (90.0%). Family-label agreement is lower because several clauses sit on family boundaries that have to be resolved by context (a clause matching the LABEL pattern may describe an evaluative function; a STRUCTURAL match may describe a genuine communicative one).

Per-family accuracy. Family-label accuracy varied across the assigned families (Table 9). EVALUATE, RESPOND, and CAPTION assignments were judged correct in over 90% of cases. STRUCTURAL and OTHER were lower (56% and 46% respectively); both are absorbing categories in the taxonomy, and clauses in them were often re-labelled by annotators into a function family. LABEL (n=6) is too small to estimate stably.

Interpretive limits of the recall audit. The recall audit returned a 60% missed-gloss rate, with annotator disagreement on 10% of entries. Some missed glosses are clearly marked usage statements outside the pattern set ('Banana For Scale: to indicate the true-to-life size of another object'); others sit on the boundary between use-condition and production description ('featured in My Reaction When style posts'), and the annotators disagreed on whether to count them.

Family	n	Precision	Family accuracy
EVALUATE	13	100.0%	100.0%
RESPOND	21	83.3%	95.2%
CAPTION	21	78.6%	92.9%
EXPRESS	29	93.1%	75.9%
LABEL	6	100.0%	66.7%
STRUCTURAL	47	51.1%	56.4%
OTHER	11	54.5%	45.5%

Table 9: Per-family precision and family-label accuracy, combined across annotators. DIRECT (n=2) omitted.

A.4.1 Extraction Coverage

Patterns 4 and 5 use closed verb lists, so verbs outside those lists are invisible unless caught by the open *used to* frame (Pattern 1). To assess the impact of this restriction, we compared the five-pattern extraction against an open alternative that matches any verb predicated of the meme as grammatical subject, applying the same WordNet taxonomy (Appendix A.3) to both. Table 10 reports format-level agreement.

Family	Both	Pattern	Open	Retention
LABEL	141	11	71	92.8%
EXPRESS	275	94	54	74.5%
EVALUATE	87	100	14	46.5%
STRUCTURAL	672	34	267	95.2%
RESPOND	50	232	16	17.7%
CAPTION	5	259	2	1.9%

Table 10: Format-level agreement between the five-pattern and open extractions under the same WordNet taxonomy. *Pattern* = formats found only by the five-pattern approach; *Open* = formats found only by the open approach; *Retention* = proportion of five-pattern formats also present under open extraction. Counts over the 5,472-format analysis population.

For LABEL and EXPRESS, whose membership is most directly shaped by the verb-list restriction, retention is high (93% and 75% respectively). RESPOND and CAPTION show low retention (18% and 2%) because the five-pattern extraction captures noun-headed and participial constructions (*in response to*, *captioned with*) that the open alternative misses; the 232 and 259 Pattern-only formats are additional coverage. EVALUATE retains 47% because Pattern 5 matches evaluative verbs without requiring the *used to* frame.

A.5 Controls

Table 11 reports two controls on the form-binding families. Restricting to the 93.8% of format pairs sharing no editorial tags removes topic overlap by construction; MRQAP controls simultaneously for tag overlap, shared format type, and format size. All form effects survive both controls.

Family	Raw	Zero-tag	MRQAP β
LABEL	**+0.042**	**+0.040**	**+0.066**
EVALUATE	**+0.018**	**+0.018**	**+0.031**
CAPTION	**−0.019**	**−0.021**	**−0.038**

Table 11: Form controls (FW neg-JSD). Raw = uncontrolled effect from Table 2. Zero-tag = restricted to pairs sharing no editorial tags. MRQAP β = standardised within-family coefficient controlling for tag overlap, format type, and size. Bold = $p < .01$.

Table 12 reports anisotropy controls on the text and visual columns. Mean-centring preserves text binding for LABEL, EXPRESS, and RESPOND, but reverses the sign of CAPTION's text effect; CAPTION's apparent text anti-coherence in Table 2 does not survive. EVALUATE shows a sign reversal on visual, moving from null to a significant positive effect under mean-centring (+0.025, $p < .0001$).

	Text		Visual	
Family	Raw	Centred	Raw	Centred
LABEL	**+0.023**	**+0.009**	**+0.015**	**+0.014**
EXPRESS	**+0.010**	**+0.003**	**+0.009**	**+0.004**
RESPOND	+0.007	**+0.003**	+0.007	**+0.014**
CAPTION	**−0.012**	+0.008†	+0.001	**+0.012**
EVALUATE	−0.000	+0.002	−0.000	**+0.025**†
STRUCTURAL	−0.003	−0.001	−0.002	+0.000

Table 12: Embedding controls (anisotropy). Raw = uncontrolled effect from Table 2. Centred = after removing the global centroid. Bold = $p < .01$. †Effect changes direction after centring.

A.6 Searle Crosswalk

Table 13 maps community families onto Searle-style speech-act classes (Searle, 1976). EVALUATE and EXPRESS both merge into *expressive* (550 formats); CAPTION and STRUCTURAL are not speech acts and are pooled into a residual *affordance* class. RESPOND maps to *responsive* (reactive deployment), which is not a Searle primitive.

Table 14 reports per-class form effects. The assertive class (= LABEL) retains the strongest positive effect, as expected. The expressive class

Community family	Searle class	Fmts
LABEL	assertive	152
EVALUATE	expressive	187
EXPRESS	expressive	369
RESPOND	responsive	282
CAPTION	affordance	264
STRUCTURAL	affordance	706
DIRECT	directive	10

Table 13: Mapping from community function families to Searle-style classes.

Searle class	Fmts	FW effect
assertive	152	**+0.042**
expressive	550	**+0.011**
responsive	282	+0.006
directive	10	−0.006
affordance	933	**−0.017**
Omnibus	–	**−0.011**

Table 14: Per-class form effects under the Searle mapping (FW neg-JSD). Bold = $p < .01$.

(+0.011, $p = .003$) is significant but weaker than EVALUATE alone (+0.018), because it merges EVALUATE (form-binding) with EXPRESS (not form-binding). The affordance class is significantly anti-coherent (−0.017, $p < .0001$), driven by CAPTION. The pooled Searle omnibus is −0.011 ($p < .0001$), worse than the community omnibus (−0.005, $p = .06$); the coarser grouping amplifies the negative affordance signal without preserving the family-level distinctions that produce positive effects.