

Light or Full Verb? A Minimal-Pair Dataset for Probing Phraseological Competence in Language Models

Francesca Franzon¹, Nicolas Rosàs Gómez¹, Leo Wanner^{1,2}

¹Universitat Pompeu Fabra (UPF)

²Catalan Institute for Research and Advanced Studies (ICREA)

francesca.franzon@upf.edu, nicolas.rosas@upf.edu, leo.wanner@upf.edu

Abstract

Frequent verbs such as *have* and *make* can function either as collocates in light-verb constructions or as full lexical predicates, as in *make a decision* vs. *make a cake*. Whether language models represent this distinction, and whether such representations vary across languages, remains unclear. We introduce a large-scale controlled dataset in English, Spanish, and French, comprising minimally varying sentence series in which the same context contains the same verb in light-verb and full-verb uses. Two probing experiments show that language models differentiate between these uses even in minimal contexts and exhibit separable patterns across object types. We release the dataset, generation code, and materials as a reusable resource. The framework supports extensions to broader contexts, additional verbs, and other languages.

1 Introduction

Verbs have different functions in sentences. In a *full-verb construction*, the verb is the main lexical predicate and contributes independent lexical meaning, as in *make a cake*; in a *light-verb construction* (LVC), a semantically weak or “light” verb (LV) provides verbal inflection, syntactic realization, and argument-structural support combining with a predicative noun, which contributes the main lexical content as in *make a decision*. LVCs have long been studied in syntax, lexical semantics, phraseology, typology, and computational linguistics (Jespersen, 1942; Grimshaw and Mester, 1988; Butt, 1995, 2010; Tu and Roth, 2011; Nagy et al., 2020; Fleischhauer et al., 2025), but remain underexploited as diagnostic material for evaluating language models. Existing computational work has largely treated LVCs as an identification or extraction problem: given a candidate verb–noun combination, the task is to determine whether it is an LVC. This has led to supervised and multilingual detection and classification experiments (Vincze et al., 2013; Chen

and Palmer, 2015; Vaidya et al., 2016; Cordeiro and Candito, 2019), as well as broader overviews of LVC and verbal-MWE identification (Constant et al., 2017; Tan et al., 2021). Yet identification and controlled probing address different questions: the former asks whether a verb–noun expression is an LVC, whereas the latter asks whether a model distinguishes light and full uses of the same verb under matched syntactic and contextual conditions. This matters because familiar LVCs may be recognized through distributional association, lexical memorization, or noun-category cues, without representation of the lexical-functional contrast itself. Thus, *make a decision* and *make a cake* share the same transitive frame, but only the former realizes an LVC with the predicative nominal base *decision*.

Minimal-pair benchmarks address this gap by isolating linguistic contrasts while holding other sentence properties constant. Targeted syntactic evaluation uses such pairs to probe model representations or test whether models assign higher probability to the expected alternative (Marvin and Linzen, 2018; Baroni et al., 2026). BLiMP instantiates this approach for 67 English grammatical phenomena (Warstadt et al., 2020); Jumelet et al. (2026) propose its multilingual extension. SyntaxGym provides a reproducible framework for targeted test-suite experiments based on minimal or near-minimal sentence contrasts (Gauthier et al., 2020), with applications also, e.g., to Chinese (Xiang et al., 2021) and Spanish (Táboas García and Wanner, 2025). While existing resources mostly contrast acceptable and unacceptable sentences, LVC probing requires contrasting grammatical and natural sentences that differ in the lexical-functional status of the same surface verb.

We present a multilingual dataset for controlled probing of LVCs, currently covering five highly frequent LVs in English, French, and Spanish. It extends the contrastive logic of minimal-pair benchmarks to minimal sentence series: the same sen-

tence context is paired with several objects, yielding LVC readings in half of the cases and full-verb readings in the other half. This design provides more directly comparable items than isolated minimal pairs, improves statistical power, and reduces contextual confounds. Unlike most minimal-pair resources, our dataset contrasts grammatical and natural sentences differing only in the lexical-functional status of the same surface verb. We illustrate its use through contrastive embedding analysis and surprisal-based probing in active and passive configurations, testing whether six models of different families and sizes capture both verb-to-noun and noun-to-verb expectations in LVCs.

2 Related Work

NLP has mostly focused on identification and classification of LVCs. Supervised, syntax-based, lexical-knowledge-based, multilingual, and neural approaches distinguish LVCs from non-LVC verb–noun co-occurrences (Tan et al., 2006; Tu and Roth, 2011; Vincze et al., 2013; Chen and Palmer, 2015; Vaidya et al., 2016; Cordeiro and Candito, 2019; Nagy et al., 2020), often within broader MWE processing and verbal-MWE shared tasks (Constant et al., 2017; Savary et al., 2017; Ramisch et al., 2018). This work shows that association strength is a central cue for distinguishing conventionalized verb–noun combinations from freer co-occurrences and for classifying them by syntactic and semantic patterns. Corpus-linguistic work on collocation and collostructional analysis provides complementary measures of lexical co-occurrence and word–construction attraction (Stefanowitsch and Gries, 2003; Gries, 2013). We draw on this tradition to control verb–noun conventionality, while treating LV status as a distinct lexical-functional contrast. Accordingly, lower surprisal for LVCs is interpreted not as frequency-independent categorical recognition, but as graded sensitivity to LVC-related expectancy patterns.

Recent work on MWEs in transformer models further motivates controlled LVC/FVC contrasts: MWE semantics is often captured inconsistently and may rely on surface patterns or memorized associations, while probing work on idiomaticity shows that MWE-related information can appear in contextual representations, but remains entangled with lexical and distributional properties. Surprisal-based studies of idioms and MWEs likewise suggest that conventionalized expressions may be pro-

cessed more easily because their components are more predictable in context. This motivates our active/passive design: if an LV licenses a conventional nominal base, the noun should receive lower surprisal in active configurations; conversely, if a predicative nominal base licenses a light verb, the verb should receive lower surprisal in passive configurations.

Methodologically, our work builds on targeted minimal-pair evaluation of language models (Marvin and Linzen, 2018; Warstadt et al., 2020; Gauthier et al., 2020; Xiang et al., 2021; Jumelet et al., 2026; Baroni et al., 2026), surprisal-based accounts of incremental processing (Levy, 2008; Smith and Levy, 2013; Goodkind and Bicknell, 2018), and targeted test-suite approaches using region-level surprisal or probability contrasts (Gauthier et al., 2020; Hu et al., 2020; Wilcox et al., 2021; Beyer et al., 2021; Táboas García and Wanner, 2025). Extending this work, our benchmark contrasts matched grammatical sentences, rather than acceptable–unacceptable pairs, to test whether models encode directional expectations between predicative nouns and their LVs.

3 Dataset Construction

Our seed dataset inventory includes five high-frequency LVs in each of the three languages: English: *make, take, give, have, and receive*; French: *faire, prendre, donner, avoir, and recevoir*; and Spanish: *hacer, tomar, dar, tener, and recibir*. We selected the English and French frequently occurring as LVCs in CollFrEn (Fisas et al., 2020), while the Spanish verbs were manually selected to preserve the same cross-linguistic verb classes.¹

All candidates were validated according to three criteria: the noun had to be a predicative nominal base, the verb had to be semantically light in the construction, and the combination had to instantiate a canonical LV-construction corresponding to either the Oper₁ or Oper₂ lexical functions in Explanatory Combinatorial Lexicology (Melčuk et al., 1995). Borderline cases and idiomatic expressions were excluded. Based on this LV selection, we constructed a corresponding full-verb–noun set. We selected a set of nouns denoting concrete objects

¹So far, we excluded other frequent English LVC verbs, such as *do* and *get*, and their equivalents in the other languages. These verbs are often included in LVC recognition tasks; cf., e.g., (Tu and Roth, 2011). However, many of their transitive uses are generic activity predicates and do not yield clear minimal-pair contrasts between LV and full-verb readings.

Verb	LVC use	Full-verb use
<i>make</i>	<i>make a decision</i>	<i>make a cake</i>
<i>take</i>	<i>take a walk</i>	<i>take a mug</i>
<i>have</i>	<i>have a celebration</i>	<i>have a car</i>
<i>give</i>	<i>give a look</i>	<i>give a coin</i>
<i>receive</i>	<i>receive attention</i>	<i>receive a package</i>

Table 1: Examples of LVC vs. full-verb noun contrasts.

Condition	Example
LVC-active	<i>During the day, they made a decision.</i>
Full-active	<i>During the day, they made a cake.</i>
LVC-passive	<i>During the day, a decision was made.</i>
Full-passive	<i>During the day, a cake was made.</i>

Table 2: Examples of matched active and passive sentence-level contrasts for English.

that could occur after each verb. The constructions in the full-verb condition had to be grammatical and natural, instantiate the same broad syntactic frame of the LVC condition, and involve a noun with which the verb expresses an independent lexical meaning. Table 1 exemplifies the resulting contrasts for English; for French and Spanish examples, see Appendix A.1.

For each verb, a set of natural-language sentence contexts was generated and paired with the verb + object constructions. Sentence contexts vary across several grammatical properties, including voice, tense, length, and number of specifications, allowing different degrees of stimulus variation and experimental control; cf. Table 2 for some examples.²

4 Experiments and their Results

In what follows, we present the results of the evaluation of the proposed dataset, obtained with minimal contexts in experiments carried out with six different models: google/gemma-3-270m (Gemma Team, Google DeepMind, 2025), google/gemma-2-2b (Gemma Team, Google DeepMind, 2024), Qwen/Qwen3.5-0.8B (Qwen Team, 2026), Qwen/Qwen2.5-3B (Qwen Team, 2024), facebook/opt-1.3b (Zhang et al., 2022), and EleutherAI/pythia-410m (Biderman et al., 2023). The models were chosen to span parameter scale, model family, training generation, and language orientation. This allows us to ask whether the contrast is reflected robustly in the contextual representations, and whether cross-linguistic

²The dataset, along with the code and materials used for its generation are released with documentation for reuse at https://github.com/XplainLing/LVC_sentences_database. Details on the dataset creation and composition are provided in Appendix A.

separability varies with model type and scale.

We evaluate the proposed dataset along two strands. First, we use it for surprisal-based probing of language models, testing whether models assign different expectations to LVC and full-verb configurations when the surface verb and syntactic frame are controlled. Second, we conduct a contrastive embedding analysis that examines whether object-level contextual representations separate LVC objects from full-verb objects under matched verbal and sentential contexts. Together, these experiments are intended to assess whether the light/full-verb distinction is reflected in model expectations and model-internal representations.

4.1 Surprisal-Based Probing

For a token w_t in context $w_{<t}$, *surprisal* is defined as the negative log-probability assigned to that token by a language model:

$$S(w_t) = -\log P(w_t \mid w_{<t}).$$

We report surprisal in nits. For multi-token critical regions, we use the maximum surprisal score across the tokens. The surprisal experiment follows the SyntaxGym evaluation format (Gauthier et al., 2020): each test item consists of matched conditions and manually defined critical regions. We compute token-level surprisals for the critical regions using autoregressive language models and compare differences across verbs and classes. In active voice configurations, the critical region is the noun following the verb. This tests whether the preceding verb and context make an LVC-compatible nominal base more or less expected than a full-verb object. In passive voice configurations, the critical region is the verb following the nominal base. This tests whether a predicative nominal base increases the expectation for its conventional LV. For each condition, we compare surprisal values across matched LVC and full-verb sentences.

We compared surprisal between LVCs and full verbs using linear mixed-effects models with random intercepts for sentence context and verb+object. Across languages, models and voice, LVCs showed significantly lower surprisals [Holm-adjusted $p \leq .033$]; full results are reported in Appendix B.1. Figure 1 shows the contrastive surprisal scores for Qwen2.5-3B, whose patterns are broadly representative across models, in active and passive voice configurations for all three languages; cf. Appendix B.1 for the other five models. In active voice, the critical regions in the LVC condition

receive lower surprisal than their full-verb counterparts, except for *receive* and its French and Spanish equivalents, *recevoir/recibir*. This suggests that the model assigns higher contextual probability to nominal bases occurring with conventional LVs such as *give/donner/dar*, *have/avoir/tener*, *make/faire/hacer*, and *take/prendre/tomar*, while the status of *receive/recevoir/recibir* is more ambivalent. In passive voice, where the preceding noun cues prediction of the following verb, lower LVC surprisal suggests partial sensitivity to associations between predicative nominal bases and their conventional LVs. Variation across verbs is greater in passive than active configurations in all three languages: *make* shows the lowest surprisal values, suggesting that it is the most strongly expected light verb in the tested passive configurations, whereas the corresponding pattern is less clear for French and Spanish; conversely, *have* and its translation equivalents consistently show the highest surprisal values. Figure 2 compares the behavior of five models for English (cf. Appendix B.2 for French and Spanish). Overall, the models show consistent behavior, with only minor deviations; see, e.g., gemma-2-2b for *have*-constructions.

4.2 Contextual Embedding Comparison

We compare contextual representations at the object position in LVC and full-verb constructions. For each sentence, we average the hidden states of the object noun phrase, yielding one object-level representation per instance, and test whether these representations separate by construction type. Since the surface verb and sentence template are controlled, such separation indicates that the object representations encode information distinguishing LVC from full-verb contexts, although correlated lexical-semantic properties of the nouns may also contribute. We assess this by reducing the dimensionality of the object representations and applying clustering analyses, both to the layer-wise representations and to Δ_l , defined as the change from layer $l - 1$ to layer l . We evaluate whether the induced clusters align with the LVC and full-verb conditions, i.e., whether the surprisal contrast is also reflected in contextual representation space. Figure 3 illustrates this pattern for English: full object representations show clear light/full-verb separability across most models and layers, though its strength and development vary by model family. Layer-wise changes are less stable, suggesting that the distinction is more consistently reflected in accumulated

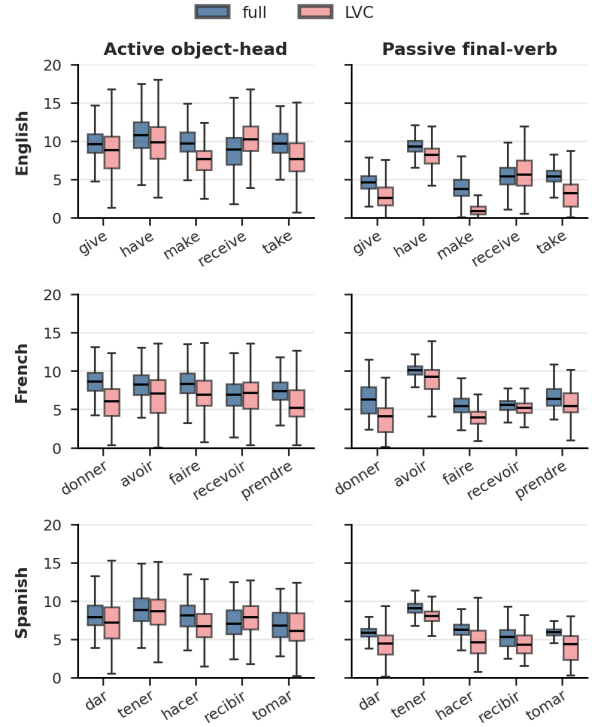


Figure 1: Qwen2.5-3B surprisal (y-axis, in nits) for LVCs and full-verb controls in English, French, and Spanish. The results are computed from 47,600 English, 41,310 French, and 45,080 Spanish sentences per voice. For active sentences, surprisal is measured at the sentence-final object; for passive sentences, at the sentence-final verb.

representations than in individual-layer updates. The multilingual comparison shows further cross-linguistic variation, with some models exhibiting more stable separation across English, French, and Spanish than others; complete results are reported in Appendix B.2. Figure 4 provides an example, showing a PCA projection of gemma-3-270m representations in which the two construction types form clearly distinct regions. Together with the surprisal results, these findings show that the light/full-verb contrast is reflected both in model predictions and in contextual representation geometry.

We further examine whether the observed separation could be explained by lexical properties of the bases themselves, specifically concreteness. To this end, we assess its contribution in our English dataset using human-rated concreteness scores on a 1–5 scale, where 5 indicates maximum concreteness (Brysbaert et al., 2014). We found that full-verb objects are, on average, more concrete than LVC bases (mean = 4.82, SD = 0.213 vs. mean = 3.00, SD = 0.914). Concreteness thus possibly contributes to the observed separation; however,

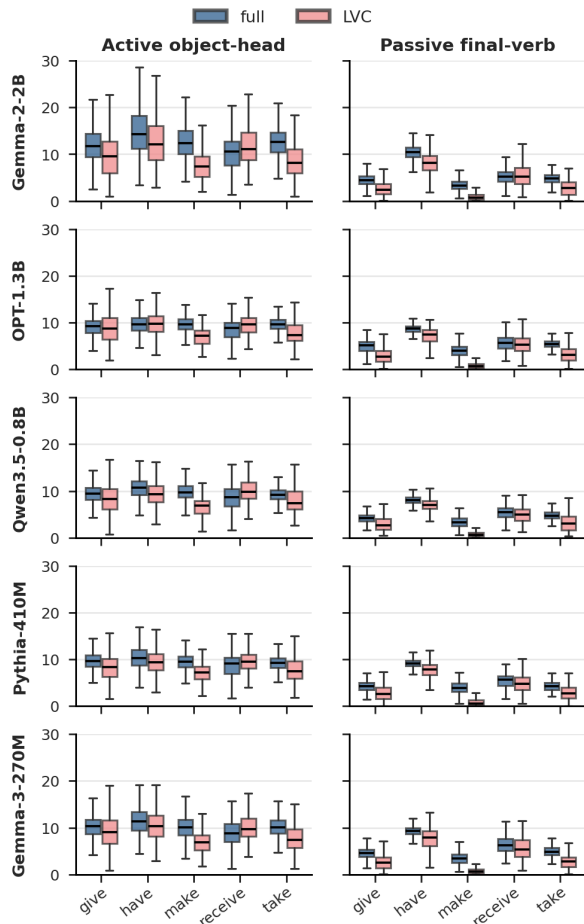


Figure 2: Surprisal (y-axis, in nits) for five language models on 47,600 active and 47,600 passive English sentences. For active sentences, surprisal is measured at the head of the sentence-final object; for passive sentences, it is measured at the sentence-final verb.

LVC bases span the full concreteness scale, and embedding-based clustering achieves perfect separation in some layers, beyond what concreteness alone predicts. We therefore interpret the clustering as reflecting semantic differences between the conditions, of which concreteness is an important component, rather than construction type alone.

5 Conclusions

Our dataset provides a controlled probing instrument for investigating differences between verbs used in LVCs and in full-verb constructions in English, French, and Spanish. It is designed to support further probing experiments, including extensions to additional verbs, broader contextual settings, other languages, and alternative analytical or experimental approaches including psycholinguistic testing. The obtained surprisal scores suggest that the six tested models are sensitive to LVCs.

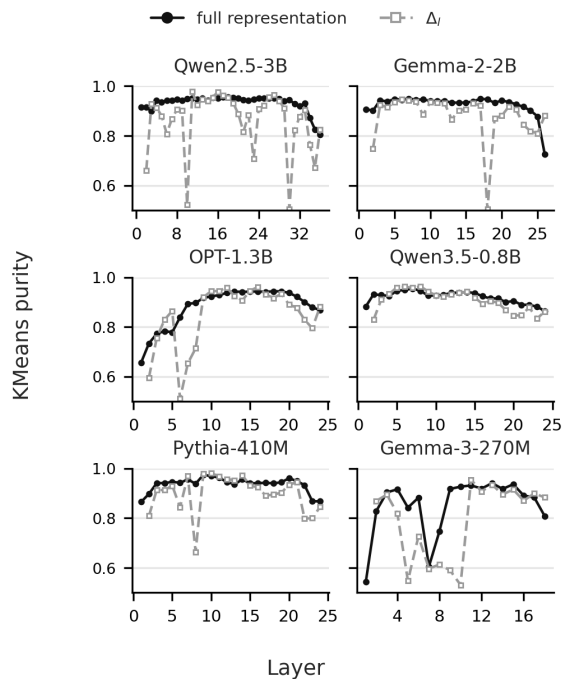


Figure 3: KMeans purity for full object-head hidden states and Δ_l , the change in the object-head representation from layer $l - 1$ to layer l for English across all models. Clustering done over 47,600 active sentences.

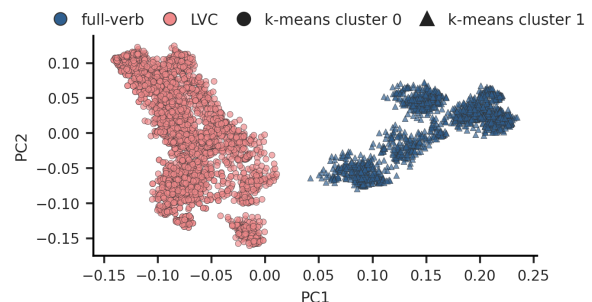


Figure 4: PCA projection of object-level representations from layer 15 of gemma-3-270m for 5,000 active *receive* sentences. Object-token embeddings are averaged per sentence. PC1/PC2 explain 26.6%/9.0% of the variance; KMeans uses the first 50 PCs (96.7% retained variance).

LVC-compatible nominal bases are generally less surprising after LVs in active sentences, and predicative nominal bases make their associated LVs more predictable in passive sentences. However, this varies across verbs, which implies the importance of the degree of conventionality of the considered LVCs. The contextual embedding comparison experiment shows that at some point of processing the two object types are clearly separated, suggesting that the models capture at least some of their differential properties.

Limitations

While this database provides controlled minimal pairs, it focuses on a limited set of verbs and constructions, potentially restricting generalization across broader verb classes. Additionally, sentence contexts are semi-controlled, which may reduce ecological validity. Finally, our analyses focus on three languages only, leaving a more in-depth cross-linguistic variation in LVCs unexplored. However, we view these choices as enabling precise comparisons, designed this dataset and methods to support future extensions across verbs, contexts, and languages.

Ethical Statement

This work introduces a dataset of constructed sentences designed for research on language processing in LLMs. All items have been automatically generated starting from human-designed materials and do not contain offensive or biased content.

AI support was used for coding and writing assistance, but not for research.

Acknowledgments

We are grateful to Beatriz Fisas, Igor Kuzmin and Marco Baroni for their valuable scientific input and insightful discussions. This work has been funded by the European Research Council (ERC) under the contract No.101201497, granted to the third author. The work reflects the authors' view only, and the funding agency is not responsible for any use that may be made of the information it contains.

References

Marco Baroni, Emily Cheng, Iria de Dios-Flores, and Francesca Franzon. 2026. Tracing the complexity profiles of different linguistic phenomena through the intrinsic dimension of llm representations. *arXiv preprint arXiv:2601.03779*.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. *Is incoherence surprising? Targeted evaluation of coherence prediction from language models*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. *Pythia*:

A suite for analyzing large language models across training and scaling. Preprint, arXiv:2304.01373.

- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.
- Miriam Butt. 1995. *The Structure of Complex Predicates in Urdu*. CSLI Publications, Stanford.
- Miriam Butt. 2010. The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker, and Mark Harvey, editors, *Complex Predicates: Cross-Linguistic Perspectives on Event Structure*. Cambridge University Press, Cambridge.
- Wei-Te Chen and Martha Palmer. 2015. English light verb construction identification using lexical knowledge. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Carlos Ramisch van der Plas, Lonneke, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.
- Silvio Ricardo Cordeiro and Marie Candito. 2019. *Syntax-based identification of light-verb constructions*. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 97–104, Turku, Finland. Linköping University Electronic Press.
- Beatriz Fisas, Luis Espinosa Anke, Joan Codina-Filbá, and Leo Wanner. 2020. *CollFrEn: Rich bilingual English–French collocation resource*. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 1–12, online. Association for Computational Linguistics.
- Jens Fleischhauer, , and Anja Latrouite. 2025. *Light Verbs*. Language Sciences Press, Berlin.
- Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. *SyntaxGym: An online platform for targeted evaluation of language models*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.
- Gemma Team, Google DeepMind. 2024. *Gemma 2: Improving open language models at a practical size*. Preprint, arXiv:2408.00118.
- Gemma Team, Google DeepMind. 2025. *Gemma 3 technical report*. Preprint, arXiv:2503.19786.
- Adam Goodkind and Klinton Bicknell. 2018. *Predictive power of word surprisal for reading times is a linear function of language model quality*. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

- Stefan Th. Gries. 2013. [50-something years of work on collocations: What is or should be next ...](#) *International Journal of Corpus Linguistics*, 18(1):137–166.
- Jane Grimshaw and Armin Mester. 1988. Light verbs and θ -marking. *Linguistic Inquiry*, 19(2):205–232.
- Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744. Association for Computational Linguistics.
- Otto Jespersen. 1942. *A Modern English Grammar on Historical Principles, Part VI, Morphology*. Ejnar Munksgaard, Copenhagen.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. [MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs](#). *Transactions of the Association for Computational Linguistics*, 14:193–216.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). *Cognition*, 106(3):1126–1177.
- Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics.
- Igor A. Melčuk, André Clas, and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve.
- István Nagy, Veronika Vincze, and Richárd Farkas. 2020. Detecting light verb constructions across languages. *Natural Language Engineering*, 26(3):319–348.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Qwen Team. 2026. [Qwen3.5: Towards native multimodal agents](#).
- Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archana Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxo Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, and 7 others. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemizadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.
- Nathaniel J. Smith and Roger Levy. 2013. [The effect of word predictability on reading time is logarithmic](#). *Cognition*, 128(3):302–319.
- Anatol Stefanowitsch and Stefan Th Gries. 2003. Collocations: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243.
- Alba Táboas García and Leo Wanner. 2025. [Assessing the agreement competence of large language models](#). In *Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025)*, pages 36–53, Ljubljana, Slovenia. Association for Computational Linguistics.
- Kathleen Tan, Tong Ming Lim, Chi Wee Tan, and Wei Wei Chew. 2021. Automatic identification of light verb constructions: A review. *IEM Journal, Special Edition: International Conference on Digital Transformation and Applications*.
- Yee Fan Tan, Min-Yen Kan, and Hang Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 49–56, Sydney, Australia. Association for Computational Linguistics.
- Yuancheng Tu and Dan Roth. 2011. Learning English Light Verb Constructions: Contextual or Statistical. In *Proceedings of the Workshop on Multiword Expressions*, pages 31–39. ACL.
- Ashwini Vaidya, Sumeet Agarwal, and Martha Palmer. 2016. Linguistic features for Hindi light verb construction identification. In *Proceedings of the International Conference on Computational Linguistics*, pages 1320–1329, Osaka, Japan.
- Veronika Vincze, István Nagy, and János Zsibrita. 2013. Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing*, 10(2).
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Ethan Gotlieb Wilcox, Pranali Vani, and Roger Levy. 2021. A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 939–952. Association for Computational Linguistics.

Beilei Xiang and 1 others. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. [OPT: Open pre-trained transformer language models](#). *Preprint*, arXiv:2205.01068.

A Dataset specifications

In this section, we provide the details on the construction of the datasets used in the experiments. In addition to the datasets, we release the full sentence-generation framework and accompanying materials, designed to support large-scale controlled experimentation. The generation pipeline is rule-based and allows users to systematically manipulate a range of grammatical and lexical properties while preserving controlled sentence structure.

A.1 Light- and full-verb-object structures

The sentence databases were built around the mentioned five verbs per language, each paired in its light- and full-verb condition with an equal number of nominal objects, with the exception of French *faire*, which contains 35 objects in the full-verb condition. The number of objects in each condition is displayed in Table 3.

verb (Engl./Fr./Sp.)	number of objects
<i>make / faire / hacer</i>	20 / 60 / 50
<i>take / prendre / tomar</i>	20 / 10 / 5
<i>receive / recevoir / recibir</i>	20 / 20 / 5
<i>have / avoir / tener</i>	40 / 24 / 50
<i>give / donner / dar</i>	40 / 20 / 30

Table 3: The number of objects in each condition across languages.

Table 1 (Section 3) and Tables 4 and 5 below show contrastive examples of the use of these verbs in LVCs and in full-verb constructions for English, French, and Spanish.

A.2 Minimal sentence pairs

The light- and full-verb-nominal object structures from above are used to generate minimal sentence pairs. In all generated sentences, the target verb-object construction occurs in sentence-final position, facilitating evaluation with autoregressive language models. For each verb, collocational and

Verb	LVC use	Full-verb use
<i>faire</i>	<i>faire une promesse</i>	<i>faire un gâteau</i>
<i>prendre</i>	<i>prendre une décision</i>	<i>prendre une tasse</i>
<i>avoir</i>	<i>avoir un accident</i>	<i>avoir une voiture</i>
<i>donner</i>	<i>donner un ordre</i>	<i>donner un livre</i>
<i>recevoir</i>	<i>recevoir une demande</i>	<i>recevoir un fleur</i>

Table 4: Examples of contrastive LVC and full-verb – noun pairs for French.

Verb	LVC use	Full-verb use
<i>hacer</i>	<i>hacer una prueba</i>	<i>hacer una bufanda</i>
<i>tomar</i>	<i>tomar una decisión</i>	<i>tomar una botella</i>
<i>tener</i>	<i>tener miedo</i>	<i>tener un barco</i>
<i>dar</i>	<i>dar un consejo</i>	<i>dar un libro</i>
<i>recibir</i>	<i>recibir una visita</i>	<i>recibir un paquete</i>

Table 5: Examples of contrastive LVC and full-verb – noun pairs for Spanish.

full-verb object nouns were paired with temporal sentence contexts and optional adverbial modifiers. Sentences were generated through rule-based templates implementing grammatical constraints on tense, voice, determiner insertion (articles or possessives, where grammatical), adverb placement, and number agreement.

The active sentence database contains sentences with the personal pronoun *they* for English and *ils* for French as subject; for Spanish, the pronouns are dropped to reflect that Spanish is a pro-drop language. Passive forms were generated using manually specified past participles and singular/plural auxiliary agreement. Consider Tables 2 (Section 3), 6, and 7 for illustration of the examples of active and passive sentence structures in the dataset.

Cond.	Example
LVC _a	<i>Pendant la célébration, ils donneront un concert.</i>
Full _a	<i>Pendant la célébration, ils donneront un livre.</i>
LVC _p	<i>Pendant la célébration, un concert sera donné.</i>
Full _p	<i>Pendant la célébration, un livre sera donné.</i>

Table 6: Examples of matched active and passive sentence-level contrasts for French. ‘Cond.’: Condition; ‘LVC_a’: light-verb construction in active; ‘LVC_p’: light-verb construction in passive; ‘Full_a’: full-verb construction in active; ‘Full_p’: full-verb construction in passive.

The dataset includes 280 unique verb+object expressions (244 for French) distributed across collocational and concrete uses, 170 temporal specifications, adverbial modifiers (35 in English, 15 in Spanish and French), 3 tense specifications (English: past simple, future simple, past perfect – Spanish and French: future, present perfect, past perfect) and voice specifications (active, passive).

Cond.	Example
LVC _a	<i>Antes de la ceremonia, darán una respuesta.</i>
Full _a	<i>Antes de la ceremonia, darán una carta.</i>
LVC _p	<i>Antes de la ceremonia, una respuesta será dada.</i>
Full _p	<i>Antes de la ceremonia, una carta será dada.</i>

Table 7: Example of matched active and passive sentence-level contrasts for Spanish. ‘Cond.’: Condition; ‘LVC_a’: light-verb construction in active; ‘LVC_p’: light-verb construction in passive; ‘Full_a’: full-verb construction in active; ‘Full_p’: full-verb construction in passive.

Sentences were automatically inflected for tense and combined with temporal and adverbial contexts. The approximate size of the generated combinations corresponds to over 10,000,000 sentences for English, over 3,000,000 for French and Spanish.

B Complementary experimental material

B.1 Surprisal analysis

Surprisal differences between LVCs and full-verb controls were assessed separately for each language $\times$ language-model $\times$ voice combination using linear mixed-effects models. Each model included *LVC condition* and *verb* as fixed effects, with crossed random intercepts for *object* and *sentence context*: $condition + verb + (1 | object) + (1 | sentencecontext)$

Table 8 summarizes the results. Effects are reported as the estimated *LVC – full-verb difference in surprisal (nits)* with 95% confidence intervals; negative estimates indicate lower surprisal for LVCs. *p*-values were Holm-adjusted across the 36 comparisons.

B.2 Cross-model comparison

Figures 5 and 6 contrast the surprisals of LVC and full-verb constructions in both active and passive voice configurations for English, French, and Spanish as obtained by all six models. In the active voice configuration, the surprisals for English are somewhat higher for all five verbs across the models, compared to the surprisals for their translation equivalents. This is visible in particular in Gemma2-2B for the verb *have*.

In the passive voice configuration, the surprisal values are more balanced across languages and models. It is interesting to note that despite its high frequency in general language corpora, for *have* and its translation equivalents the values are in this configuration in the average the highest. Further

research is needed to determine why this is the case.

Figure 7 contrasts the quality of the separation of the LVCs and full-verb construction hidden states. Qwen2.5-3B, gemma-2-2b, and Qwen3.5-0.8B show the most stable separation across layers, while the larger OPT-1.3b performs much worse for French and Spanish. At the same time, smaller models such as pythia-410m and gemma-3-270m can still reach very high purity in some layers. Multilingual models, especially the Qwen models, tend to show more consistent results across languages, but multilingual training is not required for this, as the mainly English-trained gemma-2-2b also shows very high separation in French and Spanish. Overall, this suggests that training data and training recipe, including techniques such as knowledge distillation, as well as model generation and architecture, may matter as much as model size or multilingual exposure. The results from gemma-2-2b are specially relevant, since its strong cross-lingual separation may partly reflect its extensive training and distillation from a larger model despite its mainly English-oriented training data. Since these factors vary together across model families, however, we cannot isolate the effect of any one of them from this comparison alone.

Figure 8 illustrates PCA projections of object-level contextual representations in Gemma-3-270m for the English, French, and Spanish verbs that yield the lowest and highest cluster purity. Despite the parallels observed across the three languages in terms of surprisal values, the representation-level analysis reveals interesting contrasts. In English, the weakest separation is observed for *have*, whereas its French equivalent, *avoir*, yields the strongest separation. By contrast, French and Spanish coincide in identifying *prendre/tomar* as the verbs with the weakest separation.

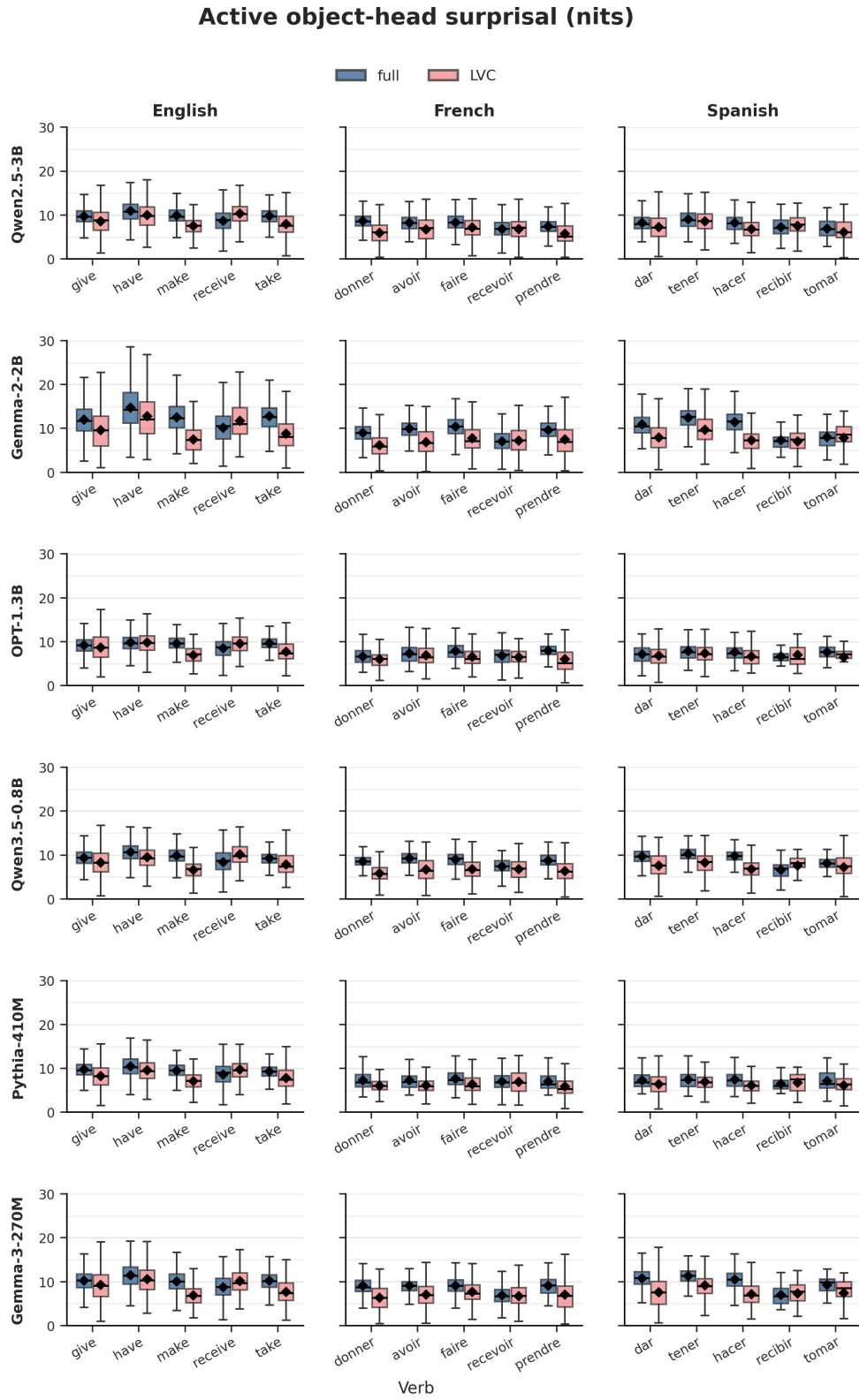


Figure 5: Surprisal (y-axis, in nits) in the six language models for the active voice sentences in all three languages.

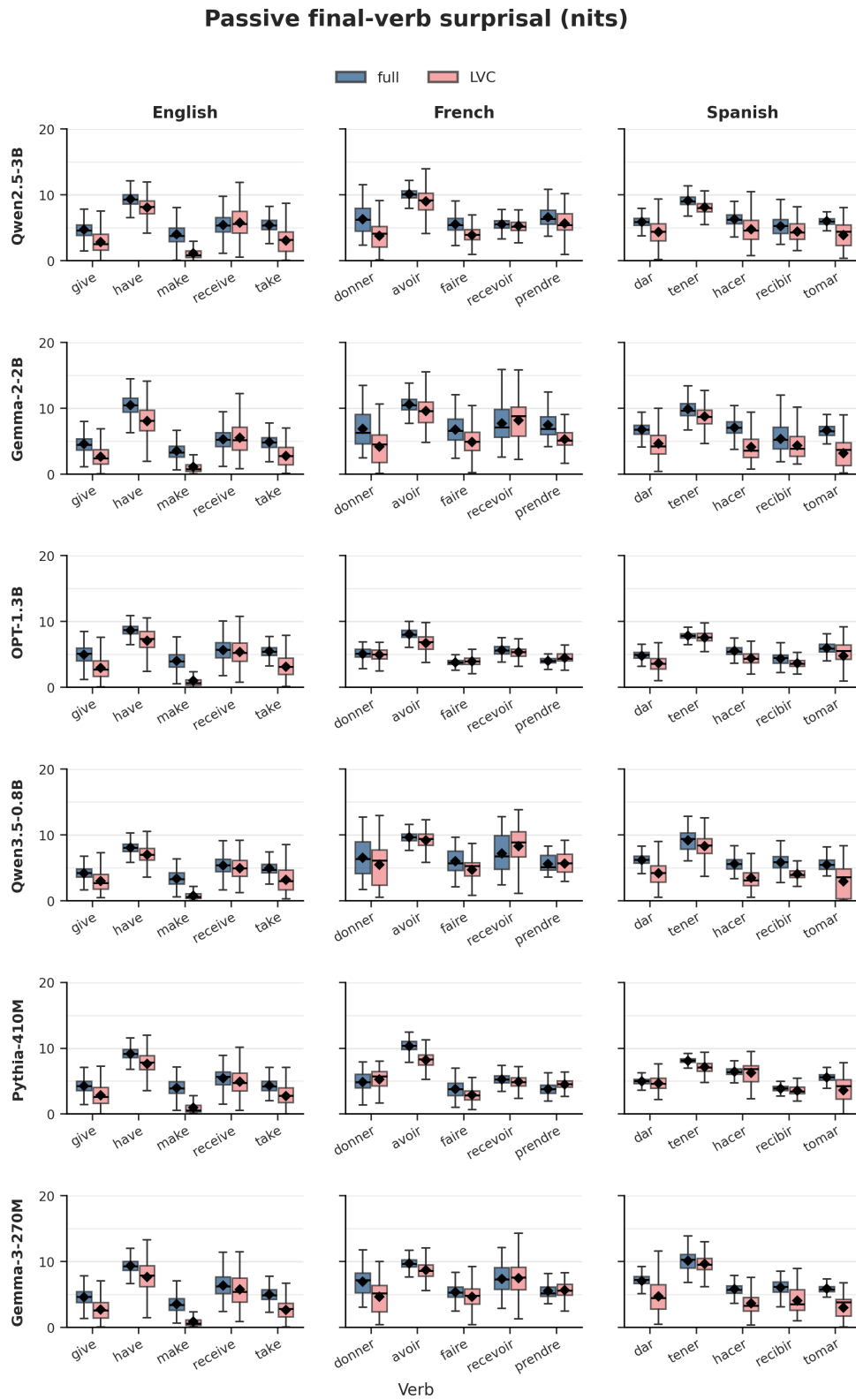


Figure 6: Surprisal (y-axis, in nits) in the six language models for the passive voice sentences in all three languages.

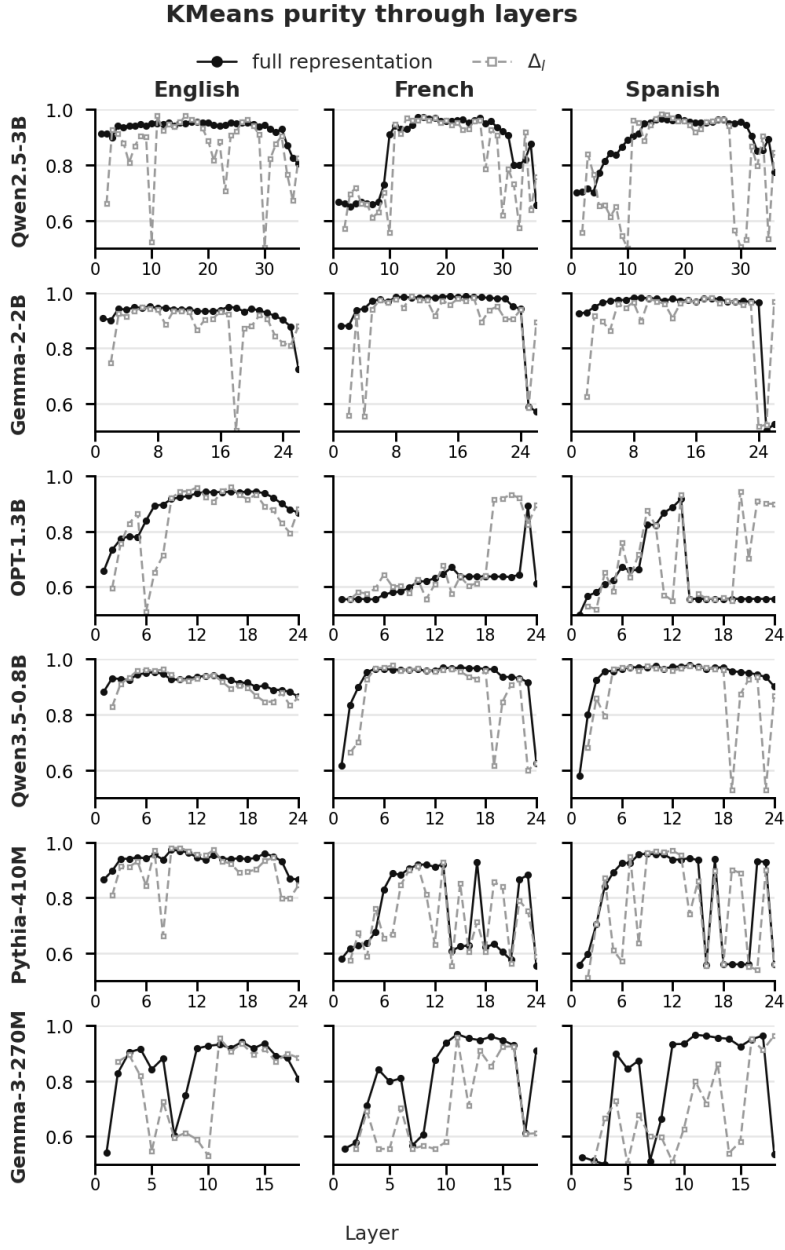


Figure 7: KMeans purity for full object-head hidden states and Δ_l , the change in the object-head representation from layer $l - 1$ to layer l for all languages across all models.

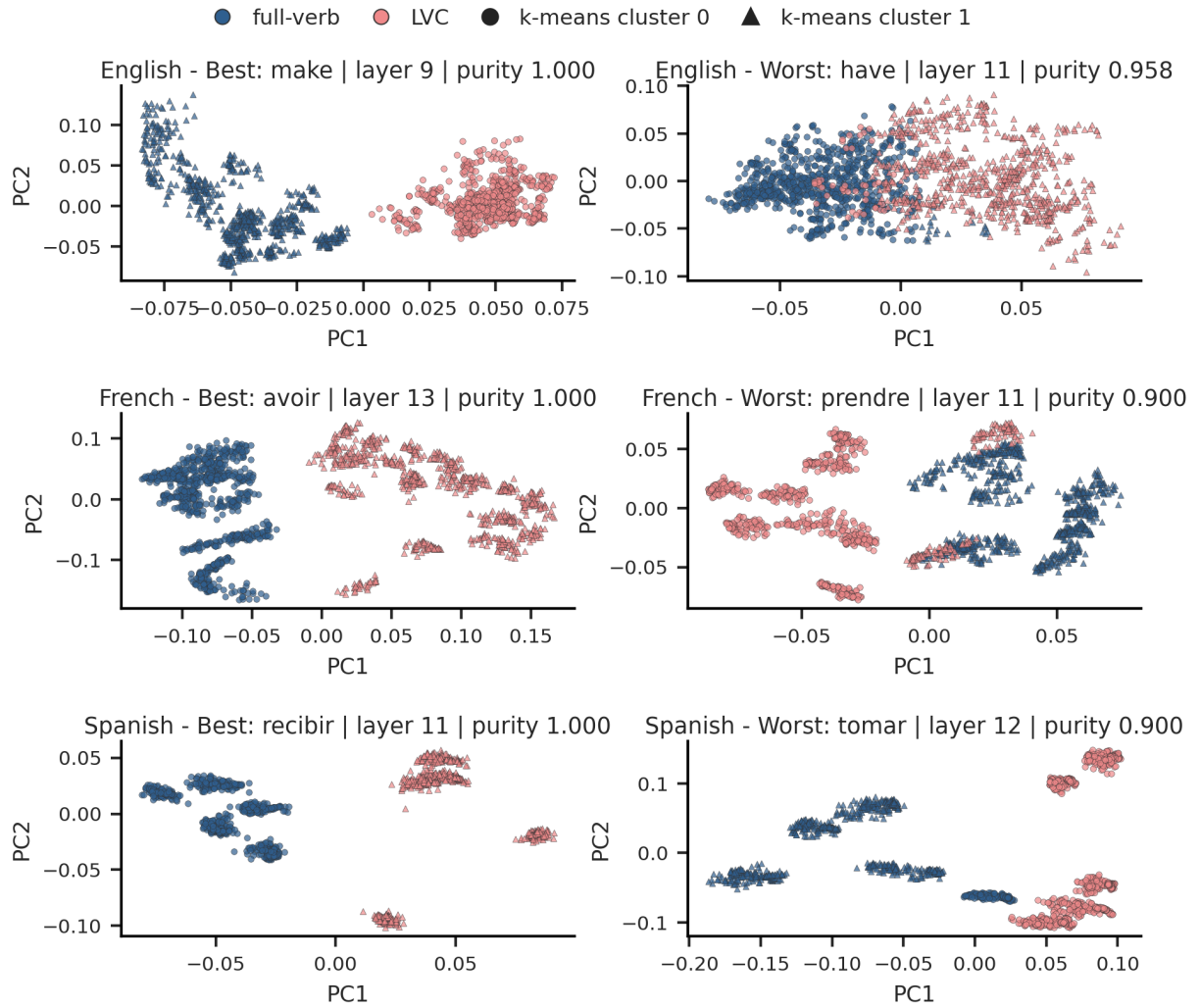


Figure 8: PCA projection of object-level contextual representations from specific layers of gemma-3-270m, computed over 1,610 active sentences. On the representation, each point corresponds to one sentence, with the object representation obtained by averaging the embeddings of all object tokens. KMeans clustering was performed on the first 50 PCA dimensions.

Language	Model	Voice	Δ surprisal	95% CI	Holm p
<i>English</i>					
	Qwen2.5-3B	Active	-0.974	[-1.475, -0.472]	.0013
	Qwen2.5-3B	Passive	-1.617	[-1.905, -1.329]	< .001
	Gemma-2-2B	Active	-2.293	[-3.174, -1.411]	< .001
	Gemma-2-2B	Passive	-1.863	[-2.173, -1.552]	< .001
	OPT-1.3B	Active	-0.634	[-1.104, -0.164]	.0326
	OPT-1.3B	Passive	-1.861	[-2.153, -1.569]	< .001
	Qwen3.5-0.8B	Active	-1.040	[-1.567, -0.512]	.0012
	Qwen3.5-0.8B	Passive	-1.331	[-1.585, -1.078]	< .001
	Pythia-410M	Active	-1.052	[-1.539, -0.565]	< .001
	Pythia-410M	Passive	-1.576	[-1.912, -1.240]	< .001
	Gemma-3-270M	Active	-1.179	[-1.789, -0.568]	.0013
	Gemma-3-270M	Passive	-1.832	[-2.158, -1.505]	< .001
<i>French</i>					
	Qwen2.5-3B	Active	-1.337	[-1.887, -0.786]	< .001
	Qwen2.5-3B	Passive	-1.401	[-1.765, -1.037]	< .001
	Gemma-2-2B	Active	-2.232	[-2.884, -1.581]	< .001
	Gemma-2-2B	Passive	-1.505	[-2.060, -0.950]	< .001
	OPT-1.3B	Active	-0.897	[-1.403, -0.392]	.0030
	OPT-1.3B	Passive	-0.281	[-0.498, -0.064]	.0326
	Qwen3.5-0.8B	Active	-2.050	[-2.534, -1.566]	< .001
	Qwen3.5-0.8B	Passive	-0.582	[-1.112, -0.052]	.0326
	Pythia-410M	Active	-0.997	[-1.514, -0.480]	.0013
	Pythia-410M	Passive	-0.728	[-1.017, -0.438]	< .001
	Gemma-3-270M	Active	-1.571	[-2.149, -0.993]	< .001
	Gemma-3-270M	Passive	-0.841	[-1.270, -0.413]	.0012
<i>Spanish</i>					
	Qwen2.5-3B	Active	-0.885	[-1.384, -0.387]	.0030
	Qwen2.5-3B	Passive	-1.334	[-1.593, -1.076]	< .001
	Gemma-2-2B	Active	-3.104	[-3.705, -2.503]	< .001
	Gemma-2-2B	Passive	-2.062	[-2.403, -1.721]	< .001
	OPT-1.3B	Active	-0.632	[-1.114, -0.150]	.0326
	OPT-1.3B	Passive	-0.800	[-0.998, -0.602]	< .001
	Qwen3.5-0.8B	Active	-2.177	[-2.627, -1.726]	< .001
	Qwen3.5-0.8B	Passive	-1.643	[-1.959, -1.327]	< .001
	Pythia-410M	Active	-0.911	[-1.341, -0.480]	< .001
	Pythia-410M	Passive	-0.548	[-0.768, -0.329]	< .001
	Gemma-3-270M	Active	-2.671	[-3.220, -2.121]	< .001
	Gemma-3-270M	Passive	-1.591	[-1.920, -1.261]	< .001

Table 8: Mixed-effects model estimates of the surprisal difference between LVCs and full-verb controls across languages, language models, and voices. Δ surprisal denotes the estimated LVC minus full-control difference in nits; negative values therefore indicate lower surprisal for LVCs. Confidence intervals are 95% CIs. p -values are Holm-adjusted across the 36 comparisons.