Cross-linguistic Geometry of Adjective Representations in Multilingual Transformers: Semantic Class, Gradability, and Positional Effects

Tancredi Monterosso
University of Verona / Free University of Bozen
tancredi.monterosso@univr.it

Abstract

In this study, we examine whether multilingual contextual embeddings encode properties of adjectives that are theoretically relevant to formal analyses of nominal modification. Using Universal Dependencies corpora for Arabic, English, and Italian, we extract contextualized adjective embeddings from the multilingual XLM-RoBERTa model and analyze them with respect to (i) semantic classes, (ii) the distinction between relational and descriptive adjectives, (iii) the distinction between gradable and non-gradable adjectives, and (iv) prenominal versus postnominal position in Italian. Our results indicate that adjective representations are organized in a shared multilingual space, but that this space is not best accounted for by a rigidly aligned universal hierarchy of semantic classes. Rather, the most salient organizing dimensions correspond to broader semantic-syntactic contrasts, in particular the relational/descriptive opposition, gradability, and, in the case of Italian, position-conditioned variation.

1 Introduction

Adjectives and adjectival modification have long been central topics in generative linguistics, where they provide a productive domain for investigating the internal organization of nominal modification. A substantial body of work has examined the structure of the noun phrase (NP) and has sought to account for cross-linguistic variation in the ordering and linearization of adjectives, numerals, demonstratives, and possessives (e.g., Cinque 2010 or Fassi Fehri, 1999 for Arabic). Adjective placement, in particular, varies considerably across languages: English predominantly exhibits adjective–noun (A–N) order, Arabic predominantly exhibits noun–adjective (N–A) order, and Italian productively allows both prenominal and postnominal adjectives. From a generative perspective, adjectives have been analyzed either as specifiers of functional projections within DP, with their ordering constrained by hierarchically organized semantic classes, building on distinctions such as those proposed by Dixon (1982) (e.g., Cinque 2010), or as adjectival modifiers merged freely as adjuncts, in which relative order is shaped by factors beyond a fixed cartographic hierarchy and in which externalization processes operating outside narrow syntax play a crucial role (e.g., Svenonius 2008).

Concurrently, recent work on large language models (LLMs) and multilingual contextual encoders has made it possible to investigate the internal representations learned by neural networks in much greater detail than before (Chang et al., 2022). Rather than treating such models as opaque "black boxes," current research increasingly probes whether their vector spaces encode linguistically meaningful distinctions. This line of work is relevant to generative linguistics for at least two reasons. First, it bears on long-standing theoretical questions concerning the nature of linguistic knowledge and the role of data-driven acquisition (Chesi, 2024). Second, it offers new empirical tools for evaluating whether neural representations capture distinctions that are central to formal analyses of syntax and semantics. As argued, for instance, by Kennedy (2025), neural networks may provide a useful testing ground for both generative and non-generative hypotheses about linguistic structure.

In this study, we investigate whether multilingual contextual embeddings encode theoretically relevant properties of adjectives. Using Universal Dependencies corpora for Arabic, English, and Italian, we extract contextualized adjective embeddings from the multilingual XLM-RoBERTa model. We then analyze these representations with respect to semantic classes, the relational/descriptive distinction, the gradable/non-gradable distinction, and prenominal versus postnominal adjectival position in Italian.

Our results suggest that adjective representations are organized in a shared multilingual space; how-

Third Workshop on the Bridges and Gaps between Formal and Computational Linguistics, pages 22–30

ever, this space is not best characterized by a rigidly aligned universal semantic hierarchy. Rather, the strongest organizing dimensions correspond to broader semantic-syntactic contrasts, especially relationality, gradability, and position-conditioned variation. In this respect, the findings suggest that approaches deriving adjectival order primarily from a fixed cartographic hierarchy (e.g. in Cinque, 2010) may overgenerate the observed patterns. Instead, the evidence points toward a model in which adjectival representation is shaped by factors other than semantic hierarchy.

More specifically, the Italian prenominal/postnominal contrast supports the hypothesis that adjectives may be associated with two distinct positions of external merge. On this view, surface order is not reducible to a single hierarchical template, but reflects the interaction between merge position and externalization. The neural evidence therefore suggests that externalization plays a significant role in the syntactic ordering of adjectives and should be treated as a central component of their formal analysis.

2 Data and Extraction Pipeline

2.1 Datasets and Vector Processing

We extract adjective tokens from Universal Dependencies (UD) corpora (Nivre et al., 2020) for three languages: Arabic, English, and Italian. Specifically, we use UD_Arabic-PADT for Arabic, UD_English-GUM for English, and UD_Italian-ISDT for Italian. These languages were selected because they instantiate distinct typological patterns of adjective placement: English is predominantly adjective–noun (A–N), Arabic predominantly noun–adjective (N–A), and Italian allows both prenominal and postnominal adjectives. This typological contrast makes it possible to test whether positional variation within a single language (Italian) aligns with broader cross-linguistic differences in adjectival ordering.

To control the syntactic environment as tightly as possible, we restrict extraction to adjective tokens satisfying the following criteria: UPOS = ADJ and DEPREL = amod. This restriction ensures that we focus exclusively on adjectives in noun-internal modification, excluding predicative uses and other non-attributive contexts. We also require a minimum of five occurrences per adjective lemma in order to ensure that averaged contextual embeddings reflect recurring distributional

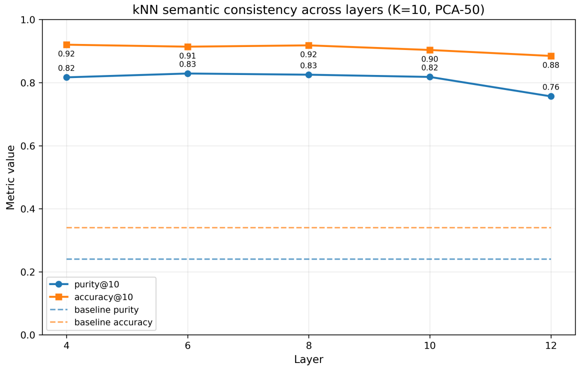

Figure 1: kNN semantic consistency across layers.

properties rather than noise arising from a small number of attested contexts. For each lemma, we aggregate all contextual occurrences and compute an averaged contextual representation (Apidianaki, 2022). Embeddings are extracted from the multilingual XLM-RoBERTa model, primarily from layer 8. This choice is motivated by a layer-wise analysis: a sweep across layers 4, 6, 8, 10, and 12 shows that semantic-class structure in local neighborhoods is most stable in the middle layers, a pattern also reported in prior work (Chang et al., 2022). In particular, kNN-based semantic consistency (which is discussed in greater detail below) remains high from layers 4 to 10 and decreases at layer 12. We therefore select layer 8 as a representative mid-layer setting rather than averaging across layers. This procedure resulted in a dataset of 1,598 adjective embeddings after applying the minimum frequency threshold of five occurrences per lemma.

Following extraction, embeddings are post-processed using mean-centering and row-wise L2 normalization, a standard procedure in representation analysis (Mahajan et al., 2025). This step reduces anisotropy and improves the interpretability of cosine similarity (Rajaee & Pilehvar, 2022). Several analyses are furthermore conducted in a PCA-reduced space (50 dimensions), which preserves a substantial portion of the original variance while providing a more stable basis for neighborhood-based diagnostics than lower-dimensional projections. In particular, we compute purity@10, that is, the average proportion of the ten nearest neighbors that share the same label as the target item, and accuracy@10, that is, the proportion of items whose label is correctly recovered by majority vote over their ten nearest neighbors. Together, these measures quantify the extent to which local neighborhoods in the embedding space are homogeneous and class-consistent. To capture the relative geometry among semantic classes, we also perform Representational Similarity Analysis (RSA) (Abnar et al., 2019) using class-level centroids for each language. We additionally report a set of standard clustering evaluation measures in the Discussion section. Specifically, we compute the Silhouette score, the Calinski–Harabasz index, the Dunn index, and the Adjusted Rand Index (ARI). These measures are reported both for the original embedding space and for the PCA50-reduced space, allowing us to evaluate whether the observed clustering structure is preserved under dimensionality reduction and to verify that the main results are not an artifact of the PCA transformation.

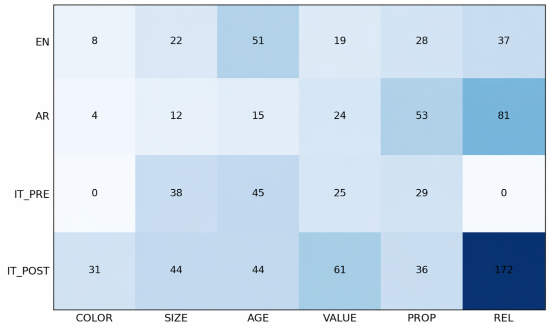

Figure 2: Adjective class coverage divided by language and semantic class. Italian is divided in prenominal adjective (IT_PRE) and post-nominal adjective (IT_POST).

2.2 Semantic Classes

Since Dixon (1982), adjectives have often been associated with broad semantic classes argued to constrain their ordering within the noun phrase. Within the generative tradition, and particularly in the cartographic framework developed by Cinque (2010), nominal modification is modeled as a hierarchy of functional projections whose ordering reflects Dixon's typology of adjectival classes. Under this view, cross-linguistic variation arises through movement operations affecting the noun or larger nominal constituents. If this semantic hierarchy constitutes a primary organizing principle of adjectival structure, adjectives belonging to the same class should be expected to occupy coherent regions of the embedding space. The semantic-class analysis presented below therefore investigates whether this theoretical generalization is recoverable in multilingual contextual embeddings and reflected in the geometry of adjective representations. As a starting point, we first classify adjectives into broad semantic classes using a prototype-based method. The classes considered are COLOR, SIZE, AGE, VALUE, PROP (physical property), and REL (relational). We did not include SPEED in the final analysis because the number of available items was too small to support a reliable class-level evaluation. Semantic class labels are assigned to adjective lemmas in Arabic, English, and Italian using contextual embedding representations. Class centroids are constructed from manually defined prototype adjectives, and each lemma is classified on the basis of cosine similarity to these centroids, together with thresholds on score and margin. For the COLOR class, we additionally employ a closed lexical list in order to override centroid-based classification and ensure higher precision. While classes such as AGE, VALUE, PROP, and SIZE exhibit greater semantic variability and fuzzy boundaries, color adjectives are generally characterized by a small set of lexically identifiable items whose membership is largely uncontroversial across languages. For this reason, we used a manually curated lexical list for COLOR and allowed this classification to override the centroid-based procedure applied to the other classes. The purpose of this choice was to maximize precision and avoid the introduction of classification errors arising from prototype similarity. In other words, the departure from the general procedure reflects a methodological decision motivated by the exceptional lexical stability and semantic coherence of color terms, rather than a theoretical distinction between COLOR and the other semantic classes. This procedure resulted in a total of 879 adjective embeddings after filtering (EN = 165, AR = 189, IT_PRE = 137, IT_POST = 388). Their distribution across semantic classes is reported in Figure 2. For the purposes of the semantic-class visualizations, Italian adjectives occurring in both prenominal and postnominal positions were averaged and treated as a single lexical item, thereby preventing duplicate representations of the same lemma in the plots.

The resulting semantic-class structure shows that multilingual adjective vectors are organized more strongly by semantic value than by language identity. This finding is consistent with recent work on cross-linguistic representational spaces in multilingual transformers (Chang et al., 2022). PCA-based visualizations suggest that adjectives from different languages are substantially intermingled,

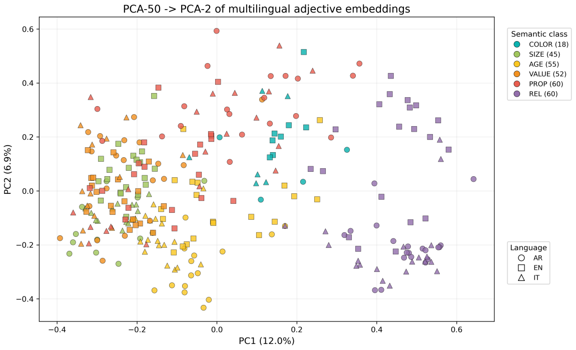

Figure 3: PCA of multilingual adjective embeddings divided by semantic class. The first two principal components explain 12.0% and 6.9% of the variance, respectively.

while still exhibiting partial grouping by semantic class, especially in the case of relational adjectives. The kNN results further support this interpretation. Overall semantic-class purity is high (Purity@10 = 0.8498) relative to the chance baseline (0.2305). The three languages display comparable values: Arabic = 0.8529 (n = 191), English = 0.8275 (n = 167), and Italian = 0.8607 (n = 285). At the class level, all categories are well above baseline, though to different degrees. REL achieves the highest absolute purity (0.94), indicating that relational adjectives form a particularly coherent cluster in the embedding space. AGE (Purity@10 = 0.84) and PROP (Purity@10 = 0.83) also exhibit strong cohesion, whereas VALUE (Purity@10 = 0.67) is comparatively more diffuse, consistent with its broader and more context-dependent semantics. Interestingly, classes with lower baseline probabilities (i.e. classes that are relatively infrequent), such as COLOR (Purity@10 = 0.78) and SIZE (Purity@10 = 0.80), display especially large purity ratios, suggesting that relatively small classes can nonetheless form well-defined regions in the embedding space when sufficient contextual evidence is available. This pattern supports the view that semantic-class structure is not merely an artifact of frequency or class imbalance.

2.3 RSA of class geometry

We use RSA in the semantic-class analysis because simple centroid alignment is not sufficient to describe how classes are organized within each language. Figure 3 can show whether, for example, Italian AGE is close to English AGE, but it does not tell us whether the overall system of relations among AGE, VALUE, REL, PROP, and the other classes is similarly structured across languages. RSA is useful precisely because it shifts the focus from isolated class-to-class comparisons to the internal geometry of the whole semantic space. In practical terms, RSA starts from the class centroids computed for each language. For each pair of semantic classes within a language, we calculate a distance value, here based on cosine distance, and arrange all these pairwise values into a class-distance matrix. This matrix summarizes how close or distant the classes are from one another in that language. We then compare these matrices across languages, typically by correlating their off-diagonal values. If two languages have similar matrices, this means that their semantic classes are organized in a similar relative geometry, even if the absolute positions of the vectors are not identical. The RSA distance matrices show a broadly comparable class geometry across the three languages, but with some clear language-specific differences. In Arabic, the largest distances involve REL, especially in its relation to SIZE, and, to a lesser extent, COLOR and AGE. COLOR is also relatively distant from SIZE and AGE, whereas PROP tends to remain comparatively closer to the other classes overall. This suggests that the Arabic class space is structured around a strong separation between REL and part of the descriptive domain, while PROP occupies a more central position.

In English, the most pronounced distances are concentrated around COLOR and REL. In particular, COLOR shows very large distances from VALUE and SIZE, and REL is also strongly separated from VALUE and SIZE. By contrast, PROP and AGE remain relatively closer to the rest of the system. Compared with Arabic, English therefore shows a sharper polarization involving COLOR and REL as more isolated regions of the class space.

Italian is globally very similar to English. As in English, the largest distances involve COLOR and REL, especially in relation to VALUE and SIZE. REL and SIZE are also highly distant, and COLOR remains strongly separated from VALUE. At the same time, PROP and AGE again occupy more intermediate positions, with lower distances to the other classes. The Italian matrix therefore largely reproduces the English configuration, with only minor local differences in magnitude.

Taken together, the matrices indicate that English and Italian are more similar to each other in their class geometry, while Arabic shows a somewhat different pattern, especially in the distribution

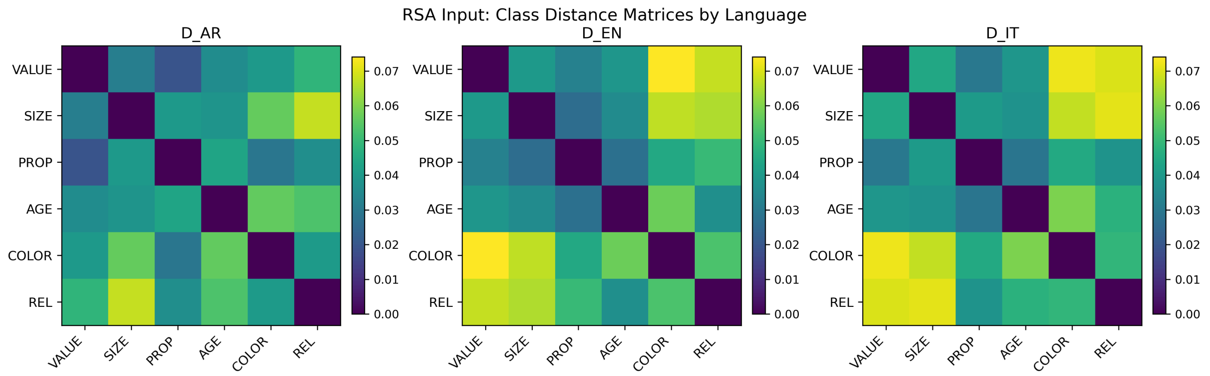

Figure 4: RSA of class distance matrices.

of distances involving REL. Thus, RSA is informative here because it reveals not only that the same classes are present across languages, but also how their relative organization differs in a measurable way from one language to another.

2.4 REL vs. DESC

The semantic-class analysis reveals that relational adjectives form the most coherent category in the embedding space, consistently yielding higher purity values than the other classes. This observation is theoretically significant, since the distinction between relational and descriptive adjectives has long been recognized as a fundamental dimension of adjectival semantics and syntax. Relational adjectives are commonly associated with classificatory or non-intersective modification, whereas descriptive adjectives denote properties that can be directly attributed to the noun (ten Hacken, 2019). Because this opposition cuts across individual semantic classes and has been argued to play a role in adjectival ordering (e.g., Cinque 2010), it provides a natural candidate for a broader organizing principle of adjective representations. We therefore collapse the semantic classes into a binary contrast between relational (REL) and descriptive (DESC) adjectives in order to test whether this distinction explains the geometry of the multilingual embedding space more effectively than the finer-grained semantic categories considered above.

The binary distinction between relational and descriptive adjectives is strongly encoded in the embedding space across all three languages. Both purity@10 and majority-vote accuracy@10 reach near-ceiling values: accuracy exceeds 0.98 in all cases, while purity remains above 0.94. These results indicate that local neighborhoods are highly homogeneous with respect to the REL/DESC distinction and that nearest neighbors overwhelmingly preserve this binary opposition. More generally, they suggest that broad structural distinctions between classifying and descriptive modification play a central role in organizing adjective representations in multilingual transformers, even more so than fine-grained semantic classes.

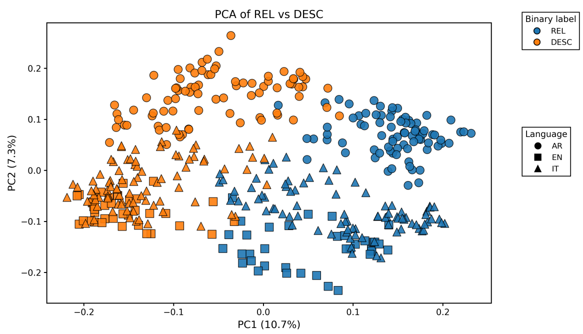

Figure 5: PCA of multilingual adjective embeddings divided into REL and DESC categories.

2.5 Gradable vs. Non-Gradable

While the REL/DESC distinction captures an important opposition between classificatory and descriptive modification, another property that has long been argued to play a central role in adjectival semantics is gradability. Since the work of Kennedy and McNally (2005), gradability has been treated as a fundamental dimension of variation among adjectives, distinguishing predicates that can participate in degree constructions and scalar interpretations from those that cannot. Beyond its semantic relevance, gradability has also been argued to interact with the syntactic behavior of adjectives, including their distribution and ordering within the nominal domain. This makes it a particularly interesting candidate for an organizing principle of multilingual adjective representa-

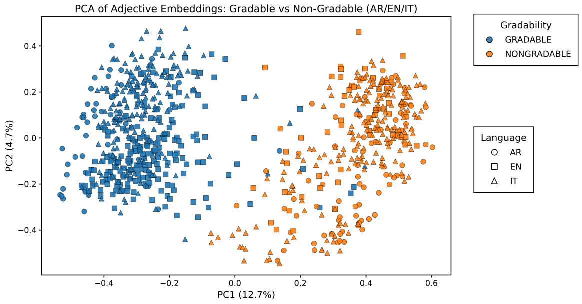

Figure 6: PCA of multilingual adjective embeddings divided into gradable and non-gradable categories.

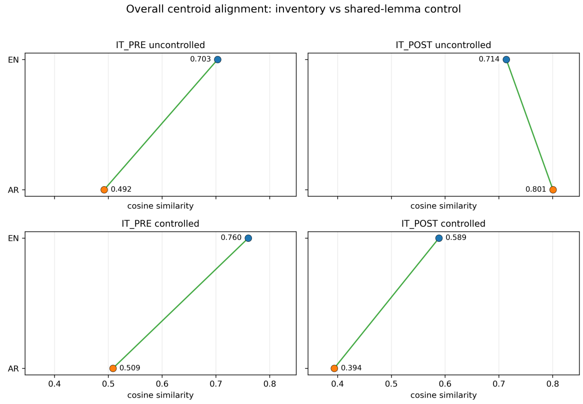

Figure 7: Overall centroid alignment between Italian (PRE/POST) and Arabic and English.

tions. We therefore investigate whether gradable and non-gradable adjectives form distinct regions of the embedding space and whether this distinction provides a stronger account of its geometry than the semantic-class and REL/DESC contrasts considered above. To derive gradability labels, we combine weak supervision from corpus-based degree-modifier evidence with a supervised classifier trained on contextual embeddings. Degree-attested adjectives are treated as positive examples, whereas relational adjectives are used as conservative negative examples. A PCA-50 plus logistic regression pipeline is evaluated by cross-validation and then fit on the full training set. The resulting model assigns each lemma a probability of being gradable; this probability is combined with thresholding and prior evidence to partition the lexicon into GRADABLE, NON_GRADABLE, and UNKNOWN sets. In this way, gradability is operationalized as a distributional property of the embedding space while preserving a conservative distinction between attested evidence and model-based generalization.

The gradable vs. non-gradable distinction is also strongly encoded in the embedding space. Across all languages, both purity@10 and accuracy@10 reach near-ceiling values (approximately 0.97–0.99), substantially exceeding their baselines. This indicates that local neighborhoods are highly homogeneous with respect to gradability and that nearest neighbors reliably preserve this distinction. The effect is robust across Arabic, English, and Italian, suggesting that gradability constitutes a stable and cross-linguistically consistent organizing dimension in multilingual contextual representations.

2.6 Positional Split in Italian

Italian provides a particularly informative internal comparison because adjectives can appear both as prenominal modifiers and postnominal modifiers. We therefore extract two separate Italian spaces: IT_PRE (prenominal adjective embeddings) and IT_POST (postnominal adjective embeddings). We compare these two spaces to English and Arabic at three levels: overall centroid alignment, class-centroid alignment, and RSA of class geometry. We also distinguish between uncontrolled analyses, which use all available lemmas, and shared-lemma controls, which restrict the comparison to lemmas attested in both Italian positions.

These analyses reveal a systematic positional effect in the embedding space. At the global centroid level, prenominal adjectives align more closely with English, while postnominal adjectives appear more similar to Arabic. However, once lexical inventory is controlled for by restricting the analysis to lemmas attested in both positions, the apparent Arabic alignment of postnominal adjectives is substantially reduced. This indicates that part of the effect is driven by distributional differences in the types of adjectives that occur in postnominal position. As shown in Figure 7, prenominal Italian exhibits consistent alignment with English in both uncontrolled and shared-lemma conditions. By contrast, postnominal Italian appears closer to Arabic only in the uncontrolled setting; this effect weakens or disappears under lexical control, suggesting that it is driven primarily by lexical inventory rather than position alone. Nevertheless, class-level centroid analyses show that the relative organization of semantic classes remains more English-like in IT_PRE and more Arabic-like in IT_POST, indicating that positional effects are not purely compositional.

3 Discussion

Taken together, the results suggest that multilingual transformer models encode adjectives in a way that is structurally meaningful, but not reducible to a simple one-dimensional universal hierarchy. Instead, the most stable dimensions across languages appear to be broad contrasts such as relational/descriptive adjectives and gradable/non-gradable adjectives, with the latter yielding the strongest results in Purity@10 and Accuracy@10. In addition to these measures, the clustering metrics further support this pattern, particularly the Adjusted Rand Index (ARI), which confirms the hierarchy observed in the previous analyses: the gradable/non-gradable distinction exhibits the strongest clustering structure, followed by the REL/DESC distinction and, finally, by semantic classes. Furthermore, the PCA reduction appears to function primarily as a denoising procedure, improving cluster compactness and separability without artificially creating the observed clustering structure.

From a generative perspective, these findings are compatible with approaches to the human faculty of language and the syntactic computational system in which adjectival modification is not reduced to a simple sequence of hierarchically ordered lexical classes, but is instead shaped by broader semantic-syntactic contrasts (e.g., Manzini 2025; Svenonius 2008). More specifically, our results suggest that semantic class alone is not the primary organizing principle in the multilingual embedding space, even though semantic classes remain recoverable and empirically useful. Rather, broader distinctions such as the relational/descriptive contrast and, even more strongly, gradability appear to structure the space more robustly.

The relational/descriptive opposition reflects a distinction that has long been recognized in the literature on adjective semantics, including contrasts related to subsective and intersective interpretation. In broad terms, descriptive adjectives tend to pattern with the former, while relational adjectives tend to pattern with the latter. As argued by Cinque (2010), this distinction is relevant to adjective ordering and should therefore be taken seriously in any formal account of nominal modification. Our results support this view by showing that the REL/DESC contrast is encoded more strongly than fine-grained semantic classes in the multilingual representation space.

Gradability, which yields the strongest results in our study, has likewise been identified as a central property of adjectival meaning, especially since Kennedy & McNally (2005), where it is treated as a key dimension of semantic variation across natural languages. The fact that the multilingual encoder

Label Set	Lang.	Dimensionality	Silh.	CH	Dunn	ARI
Semantic classes	AR	768D	0.099	9.627	0.229	0.750
Semantic classes	AR	PCA50	0.186	13.779	0.259	0.420
Semantic classes	EN	768D	0.106	6.922	0.241	0.593
Semantic classes	EN	PCA50	0.156	10.090	0.211	0.571
Semantic classes	IT	768D	0.102	10.114	0.166	0.287
Semantic classes	IT	PCA50	0.124	14.666	0.159	0.423
REL/DESC	AR	768D	0.192	24.964	0.215	0.937
REL/DESC	AR	PCA50	0.270	37.302	0.376	0.937
REL/DESC	EN	768D	0.106	11.711	0.298	0.474
REL/DESC	EN	PCA50	0.153	16.547	0.270	0.596
REL/DESC	IT	768D	0.134	24.311	0.166	0.492
REL/DESC	IT	PCA50	0.193	35.004	0.154	0.805
Gradability	AR	768D	0.346	28.812	0.393	0.966
Gradability	AR	PCA50	0.355	33.561	0.527	0.966
Gradability	EN	768D	0.131	12.103	0.296	0.741
Gradability	EN	PCA50	0.163	16.462	0.220	0.718
Gradability	IT	768D	0.145	27.760	0.105	0.778
Gradability	IT	PCA50	0.193	39.890	0.047	0.767

Table 1: Clustering metrics for adjective representations in the original and PCA50 embedding spaces. Silh. = Silhouette score; CH = Calinski–Harabasz score; Dunn = Dunn index; ARI = Adjusted Rand Index.

learns to separate and cluster adjective embeddings so effectively on the basis of gradability suggests that this feature plays a major role in the organization of adjective meaning. Moreover, because these embeddings are extracted from an intermediate transformer layer that is plausibly sensitive to both semantic and syntactic information (Jawahar et al., 2019), the results also suggest that gradability may be relevant not only to semantic interpretation, but also to the factors that shape adjective ordering. In this sense, gradability emerges as a plausible candidate for a feature that contributes to the computational organization of adjectival modification, rather than as a merely lexical or descriptive property.

The Italian PRE/POST contrast is particularly informative from a theoretical perspective. Prenominal Italian aligns more closely with English, whereas postnominal Italian shows a stronger affinity with Arabic, especially at the level of class geometry. At the same time, the shared-lemma control demonstrates that this pattern cannot be reduced to a uniform positional shift affecting the same lexical items across contexts; lexical distribution plays a substantial role. This suggests that adjective position in Italian may correspond to more than one type of modification relation, with prenominal adjectives aligning more closely with the English prenominal system and postnominal adjectives only partially resembling the Arabic postnominal system.

More broadly, these findings are consistent with the view that adjective position is not determined solely by linearization through a fixed functional hierarchy, but also reflects differences in the semantic-syntactic organization of modification (Manzini 2025). In this respect, our results are compatible with Manzini's proposal that adjectival modifiers may be externally merged in two distinct structural positions within the nominal domain: at the edge of the NP phase in the case of gradable adjectives, and in the complement domain of nP in the case of non-gradable adjectives. Although our findings on their own do not provide conclusive evidence for such a derivational account, they are consistent with an analysis in which positional alternations in Italian reflect structurally differentiated modification sites rather than a single uniform ordering mechanism.

4 Conclusion

We have presented a cross-linguistic study of adjective embeddings in Arabic, English, and Italian using multilingual transformer representations extracted from Universal Dependencies corpora. Our analyses show that adjective embeddings form a multilingual space structured by semantic classes, but even more strongly by the broader distinctions between relational and descriptive adjectives and between gradable and non-gradable adjectives, with the latter being the strongest shaping factor. We further show that adjective position in Italian conditions cross-linguistic alignment in systematic ways, with prenominal contexts behaving more similarly to English and postnominal contexts exhibiting a more Arabic-like class geometry, although this latter effect is partly driven by lexical inventory.

The present study has a number of limitations. First, the analysis is based on contextual embeddings derived from a single multilingual encoder and a restricted set of UD corpora, so the results may partly reflect model-specific and corpus-specific properties rather than general characteristics of multilingual representations. Second, our semantic-class labels are obtained through a prototype-based procedure operating on the same embedding space that is subsequently analyzed. Consequently, the semantic-class and, to a lesser extent, the REL/DESC results cannot be regarded as entirely independent of the labeling process itself and should therefore be interpreted primarily as measures of the internal coherence of the induced categories rather than as fully independent validations of their existence. Importantly, however, the strongest effects reported in this study concern gradability, whose labels are derived through a substantially different procedure combining corpus-based evidence from degree modification with weakly supervised classification. Unlike semantic classes and REL/DESC, gradability is therefore not directly induced from the same embedding-space geometry used in the subsequent analyses. The fact that gradability consistently emerges as the strongest organizing dimension across kNN, clustering, and classification measures suggests that its role cannot be reduced to an artifact of the labeling procedure. More generally, the geometric patterns observed in the embedding space should not be interpreted as direct evidence for a particular syntactic derivation, but rather as indirect representational tendencies that may inform formal analyses of adjectival modification and ordering. Future work should therefore test the robustness of these findings across additional models, corpora, and annotation procedures, and should relate the present vector-space evidence more directly to controlled syntactic data. A further limitation is that some semantic classes, particularly COLOR and SIZE, are represented by relatively few items. Consequently, class-specific results for these categories are less stable and should be interpreted with caution. Overall, these findings contribute to our understanding of what multilingual encoders represent about adjectives and adjectival modification. More broadly, they suggest that vector-space diagnostics can provide useful evidence for formal analyses of linear ordering and modification, particularly if the human syntactic computational system is viewed as an efficient computational system (Kennedy, 2025).

References

Abnar, S., Beinborn, L., Choenni, R., Zuidema, W. (2019). Blackbox Meets Blackbox: Representational Similarity & Stability Analysis of Neural Language Models and Brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 191–203. Florence: Association for Computational Linguistics.

Apidianaki, M. (2022). From Word Types to Tokens and Back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, pp. 1–59.

Chang, T. A., Tu, Z., Bergen, B. K. (2022). The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 119–136. Abu Dhabi: Association for Computational Linguistics.

Cinque, G. (2010). *The syntax of adjectives: A comparative study*. Cambridge, MA: MIT Press.

Chesi, C. (2024). Is it the end of (generative) linguistics as we know it?. *Italian Journal of Linguistics*, 37.1, pp. 3-44.

Dixon, R. M. W. (1982). *Where have all the adjectives gone? And other essays in semantics and syntax*. Berlin: Mouton.

Fassi Fehri, A. (1999). Arabic Modifying Adjectives and DP Structures. *Studia Linguistica, 53*.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657. Florence: Association for Computational Linguistics.

Kennedy, C., McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2), pp. 345–381.

Kennedy, M. (2025). Evidence of generative syntax in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pp. 377–396. Vienna, Austria: Association for Computational Linguistics.

Manzini, M. R. (2025). Left–right asymmetries in Italian adjectives: Partial order and phases. *Syntactic Theory and Research, 1*, pp. 1–39.

Nivre, J., de Marneffe, M. C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4034–4043. Marseille, France: European Language Resources Association.

Rajaee, S., & Pilehvar, M. T. (2022). An isotropy analysis in the multilingual BERT embedding space. In *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics.

Svenonius, P. (2008). The position of adjectives and other phrasal modifiers in the decomposition of DP. In L. McNally & C. Kennedy (Eds.), *Adjectives and adverbs: Syntax, semantics, and discourse* (pp. 16–42). Oxford: Oxford University Press.

ten Hacken, P. (2025). Relational adjectives between syntax and morphology. *SKASE Journal of Translation and Interpretation, 19(1)*, pp. 77–92. Vienna, Austria: Association for Computational Linguistics.