

What type is it? Exploring the polysemy of *tipo* with contextualized embeddings

Caroline A. Williams & Patrícia Amaral*

Abstract. Across languages, nouns of scientific classification (e.g. *type*, *species*, *kind*) tend to acquire new meanings over time (Brems & Davidse 2010; Mihatsch et al. 2023). This paper analyzes the polysemy of *tipo* ‘type’ in Spanish, which can have the original taxonomic meaning and new meanings like *ad hoc* categorization and “general extender”. The boundaries between meanings remain hard to determine, and this challenge is not unique to Spanish. In this paper, we adopt a computational approach using a pretrained transformer. This paper asks: (i) Can contextualized embeddings capture the polysemy of *tipo* (*de*), and (ii) How do resulting sense clusters compare to meaning categories proposed in prior research? Our results show that computational methods shed light on the polysemy of *tipo* and provide ways to empirically test existing accounts of polysemy.

Keywords. taxonomic nouns; polysemy; categorization; word embeddings; Spanish; Romance languages

1. Introduction. Across languages, nouns denoting categories in scientific classification (e.g. *type*, *species*, *kind*, *sort*) tend to acquire new meanings beyond their original taxonomic function in binominal structures “a type of N” (Brems & Davidse 2010; Mihatsch 2023; Kolyaseva & Kisiel 2023). While these nouns originate in learned vocabulary (science), in different languages they show various degrees of integration into everyday language: some nouns may be restricted to technical domains, while others also occur in colloquial speech. Synchronic data reflect diachronic changes (involving extensions of meaning and syntactic change), showing similar patterns across Indo-European languages, see Mihatsch et al. (2023).

This paper analyzes the polysemy of *tipo* ‘type’ in Spanish. This noun is a reflex of the Latin noun *typus* ‘figure, image’, which was borrowed from Greek *typos*, meaning ‘mark, shape’. In contemporary Spanish, the noun *tipo* can have the original taxonomic meaning and has also developed new meanings. Among the senses listed by the Dictionary of the *Real Academia Española* are: ‘model, exemplar’, ‘symbol of something’, ‘class, category’, ‘physical aspect of a person, bearing’, ‘unusual person, character’, ‘printing type’, ‘guy, person’, and ‘financial rate’.¹

Previous studies of *tipo* and its Romance counterparts (e.g., Italian *tipo*) have relied on qualitative analyses of corpus data to propose categories for its different senses. These studies tend to list the range of syntactic constructions in which the noun participates and hypothesize relations between their meanings. However, the criteria for differentiating among senses are often not spelled out (e.g., Mihatsch 2016, 2023). A common problem is the lack of a measure of similarity between senses. The boundaries between meanings remain hard to determine, and this challenge is not unique to Spanish.

In this paper, we adopt a computational approach using a pretrained transformer. This paper asks: (i) How do contextualized embeddings capture the semantic and syntactic range of con-

* The authors would like to thank M.G. Puxeddu, Luke Gessler, and the members of Patrícia Amaral’s Semantics Lab for their help. Authors: Caroline Amodeo Williams, Indiana University (caramode@iu.edu) & Patrícia Amaral, Indiana University (pamaral@iu.edu).

¹ The 15 senses listed and their respective order can be found here: <https://dle.rae.es/tipo> (from the 23rd edition of the DRAE, 2014).

structions with *tipo*? (ii) How do resulting sense clusters compare to meaning categories proposed in prior research? (iii) Can contextualized embeddings capture relations between senses that may lead to attested patterns of semantic change?

While the polysemy of *tipo* and its cross-linguistic counterparts offer specific challenges to the semanticist, this is by no means a rare phenomenon. Polysemy is rarely a problem in language use, but it raises difficult questions both in semantic theory and in semantic applications. Semanticists and lexicographers often disagree on the number of senses of a word, their groupings, and the relations among them. Hence, our broader goal is to explore the potential of computational methods to address these difficulties and model polysemy in naturally-occurring language.

This paper is structured as follows. Section 2 presents previous studies on Spanish *tipo* and some of the questions raised by the sense classifications they propose. Section 3 introduces the methods used in our study. In section 4 we detail the results of our experiments, whose implications are discussed in section 5. Section 6 provides concluding remarks.

2. Previous studies on “type” nouns and the meanings of *tipo*. Recent work has identified tendencies in semantic shifts of taxonomic nouns within the Indo-European languages (see, e.g., Mihatsch et al. 2023). Across languages, these nouns gain new meanings, and often come to express gradient category membership (Lakoff 1975; Rosch 1973). Here we focus on Spanish *tipo* and on some of the questions that have been investigated with respect to its cross-linguistic counterparts within Romance.

Mihatsch (2016) proposes eight different meanings or uses of taxonomic nouns, among which *tipo* is examined:² (i) Taxonomic (see Example 1), (ii) Quantifying (see Example 2), (iii) Approximative or hedge (see Example 3), (iv) Determiner-like (in questions) (see Example 4), (v) Modifier (attributive function, similar to an attributive adjective) (see Example 5), (vi) ‘Phoric’ (see Example 6), (vii) List completer (see Example 7), and (viii) Identifying (see Example 8).³

(1) *circulan los dos tipos de partículas*
circulate-PRS.3PL the two types of particles
‘Both types of particles/particles of two types circulate.’ Mihatsch (2016: 140, ex. 2)

(2) *Rechazan todo tipo de juramento*
reject-PRS.3PL every type of vow
‘(They) reject all types of vows.’ Mihatsch (2016: 142, ex. 8)

² Mihatsch notes that some of these meanings are associated with particular agreement properties. She also notes that in what she calls the Approximative use, it is unclear whether *tipo* retains its categorial properties of a noun. However, the syntactic and semantic criteria for each class are not clearly laid out: while the author presents the uses listed here as “non-nominal uses” (Mihatsch 2023:2), the examples provided display nominal features, like modification of *tipo* by determiners, quantifiers, and adjectives, a crucial issue not addressed by Mihatsch. For this reason, in this study, we decided to include all the occurrences of *tipo* in our corpus and to include part-of-speech (POS) information in the analysis in order to better capture the properties of the data.

³ Only some examples are translated in Mihatsch (2016), so we provide translations. Note that the full notation details of the corpora used by Mihatsch are not reproduced here.

- (3) fue justamente a ... tipo diez, diez y media de la mañana.
 be-PST.3SG precisely at ... type ten ten and half of the morning
 ‘(It) happened at about 10, 10:30am.’ Mihatsch (2016: 151, ex. 63)
- (4) preguntándole en una ocasión qué tipo de muerte elegiría,
 ask-GER-CL in one occasion which type of death choose-COND.3SG
 ‘asking her on one occasion which type of death (she) might choose,’ Mihatsch (2016: 148, ex. 37)
- (5) un tipo de paisaje muy optimizado
 a type of landscape very optimized
 ‘a very optimized type of landscape’ Mihatsch (2016: 149, ex. 43)
- (6) a mí ese tipo de personas así no me gustan
 to me that type of persons like.that not CL please-PRS.3PL
 ‘I don’t like that type of people (lit. like that).’ Mihatsch (2016: 153, ex. 73)
- (7) en la idea de comunidad planetaria, este tipo de cosas.
 in the idea of community planetary, this type of things
 ‘in the idea of a planetary community, that type of things.’ Mihatsch (2016: 154, ex. 77)
- (8) otro tipo de ciudad más eficiente,
 other type of city more efficient
 ‘another type of city, a more efficient one’ Mihatsch (2016: 154, ex. 81)

These categories, as well as the examples presented by the author, raise a number of questions. How are these meanings related? Are some of these meanings compositionally obtained, in which case no multiplication of meanings would be needed? For instance, in the case of “phoric” uses, in which the author identifies an anaphoric function, it is unclear whether the anaphoric meaning comes from *tipo* or from other elements in the sentence: notice that in Example 6, we can find a demonstrative determiner (*ese* ‘that’) and the anaphoric adverb *así* ‘that way’, which both anaphorically retrieve a contextually salient set of properties of people. Likewise, in the literature, examples that are labeled “list completer” may contain demonstrative determiners (as in Example 7) or universal quantifiers. So, one may wonder why these examples should be considered “list completers” rather than be classified under categories that they share properties with (quantifying, phoric). Overall, it is unclear in this proposal whether the categories above are based on formal properties of the examples, their discourse uses, or both. This raises the following question: Which criteria can we use to tease apart the different meanings of *tipo*?⁴

⁴ For research that tries to provide a stricter, clearer set of criteria based on syntactic and semantic factors for English nouns *type*, *sort*, and *kind*, see Denison (2002); De Smedt et al. (2007).

Beyond Spanish, there are some more general difficulties facing the analysis of taxonomic nouns. First, while these nouns originate in technical discourse about scientific classes, the taxonomic meaning ('kind') can occur with an *ad hoc* category in discourse (in the sense of Mauri & Sansó 2018). Distinctions among scientific categories are based on necessary and sufficient conditions, and are evidence-based. On the other hand, *ad hoc* categories are vague, subjective, and have a strong context-dependence. To further complicate things, it is almost impossible to distinguish the value of 'category' that is created on the spot in discourse from the expression of similarity ('like'). This can be seen in Examples 9 and 10 below, obtained from the CREA corpus:

- (9) Lippi... por la estructura de equipo que hace, es tipo Capello.
 Lippi... by the structure of team that do-PRS.3SG be-PRS.3SG type Capello
 'Lippi...for the structure of the game he promotes, is like Capello.'

In 9, the speaker is describing the strategy of the soccer coach Lippi and compares him with a famous coach, Capello. While there is no extant category of "coaches like Capello", the mention of this name is enough to evoke a number of characteristics known in communities that follow Italian soccer. The example continues with an enumeration of the properties that justify such categorization.⁵ The creation of *ad hoc* categories can be found in the co-occurrence of *tipo* with proper nouns but also with common nouns, as shown by Example 10:

- (10) en un despliegue visual más bien del tipo pavo real.
 in a display visual more well of-the type peacock
 'in a visual display more like a peacock'

Note that from a syntactic point of view, the creation of an *ad hoc* category is not associated with a specific structure. In 10, unlike 9, the noun *tipo* is preceded by a definite determiner. We believe that these examples raise interesting questions about the linguistic expression of the concept of similarity, an issue that we return to in Section 5.

A second issue raised by nouns like *tipo* is that it is difficult to determine whether the multiple meanings of *tipo* are better accounted for as instances of polysemy or homonymy. While the sense of 'financial rate' intuitively seems a good candidate for homonymy with respect to 'category' or 'person', it is not always straightforward to decide what counts as a related sense. This question is relevant beyond the analysis of *tipo*. As noted by Ravin and Leacock:

The distinction between polysemy and homonymy is important because it separates the principled from the accidental and poses the following questions: If different senses of polysemous words are systematically related, how do they derive from each other, and how should they be organized to reflect this regularity? (Ravin & Leacock 2002:2)

⁵ The entire example is: *Lippi me da la sensación, por la estructura de equipo que hace, es tipo Capello. Farragoso de estructura. Muy de juego de conjunto. Primero, estructura y luego, calidad.* 'Lippi gives me the impression that, for the structure of the game that he promotes, is like Capello. Very much [a supporter of] team play. [Creating a] confusing structure. First, the structure, and then, quality.' (our translation) The *Corpus de Referencia del Español Actual* (CREA) can be found at <https://corpus.rae.es/creanet.html>.

As shown in this section, determining the number of senses of *tipo* and the relations between them is not an easy task. In the following, we address these questions empirically and adopt an approach informed by distributional principles.

3. Methods. Data analysis followed a multi-step procedure that involved preparing the corpus, extracting contextual embeddings, part-of-speech (POS) processing, applying clustering algorithms to the data, and visualizing the results. The following sections detail this process.

3.1. **CORPORA USED.** The data included two sections of spoken, transcribed Spanish from the *Large Spanish Corpus* (Cañete 2019) from Hugging Face. The two sections, TEDx and Europarl, correspond to varying registers: the former involves more informal speech and the latter semi-formal parliamentary notes. From these, all sentences containing *tipo(s) (de)* were extracted, resulting in an initial dataset of approximately 17,000 instances across TEDx ($n = 2,584$) and Europarl ($n = 14,561$). To balance the two genres, 2,550 sentences were randomly extracted from each. After cleaning the data from pre-processing artifacts (e.g., incorrect subword tokenization), the final dataset consisted of 5,075 instances of *tipo(s) (de)*.

3.2. **CONTEXTUALIZED EMBEDDINGS: BETO.** For each instance of *tipo(s) (de)* in the dataset ($n = 5,075$), we extracted contextual embeddings using $\text{BETO}_{\text{BASE-CASED}}$ (BETO), a pretrained Spanish neural language model (Cañete et al. 2023). For multi-word spans, such as *tipos de*, single vectors were extracted by averaging the individual vectors into a combined representation (i.e., mean-pooling).

To provide some context, a word embedding is a vector representation of a word’s meaning as a point in high-dimensional semantic space. With roots in distributional semantics (cf., Harris 1954; Firth 1957), embeddings capture how a word’s meaning can be inferred from its collocational patterns and range of syntactic environments. Embeddings fall into two categories: *static* and *contextual*, and prior work has shown that embeddings can be used to measure word similarity and nuanced usage differences.

In such work, similarity between embeddings is typically quantified using cosine similarity ($\cos \text{sim}$), the measure of the angle between vectors in high-dimensional space whose value lies between -1 and 1 ($[-1, 1]$). Intuitively, vectors that point in similar directions represent more similar meanings ($\cos \text{sim} \approx 1$), while those more widely divergent indicate greater semantic differences ($\cos \text{sim} \approx 0$ or even -1 if vectors have opposite directions). We adopt this approach here, using cosine similarity to quantify the relationship between individual contextualized embeddings and average-sense representations.

Though *static* embeddings are efficient and readily available, they are not well suited for the present task. Static embeddings assign each word a single, fixed representation, regardless of contextual variation. Essentially, each word form in a corpus is encoded as a single vector corresponding to that word as a *type* across all contexts in which it occurs. While this is useful for identifying a word’s top collocates or for approximating its average representation (e.g., in studies of semantic change, see Amaral et al. 2023; Hamilton et al. 2016), contextually sensitive differences in meaning are obscured.

Transformer-based language models, such as BETO, address this limitation—that all instances of a word form share the same, single vector—by using surrounding context, word order, and other syntactic and probabilistic cues to generate a contextualized representation (i.e. embedding). The result is an individual embedding for each *token* of the target word. This is the approach we pur-

sue here.

To briefly illustrate, consider the word *vino* in Spanish in Example 11. From the surrounding context, the separate lexemes *vino*₁, ‘came’, and *vino*₂, ‘wine’, can be identified. While static embeddings would create one vector representation for both lexemes, contextual embeddings instead encode syntactic and semantic information for each instance, or *token*, of *vino*. This allows homonyms, and more importantly polysemous words, to be represented distinctly by individual contextualized embeddings for every token in the given corpus.

- (11) a. El verdadero premio vino de la Sociedad... (TEDx)
The true prize come-3SG.PST from the Society
‘The true prize came from the Society.’
- b. La cuestión de producir vino de calidad... (Europarl)
The question of producing wine-NOM of quality...
‘The question of producing quality wine’

Prior work demonstrates that contextualized embeddings robustly capture syntactic and semantic distinctions. Experiments with transformer language models show an ability to capture polysemy, such as the multiple senses of *break*, a verb occurring in a wide range of syntactic constructions (Petersen & Potts 2023) as well as the many meanings of the preposition *over* (Fonteyn 2021). In addition, these models provide a comprehensive representation of syntax, as shown through probing model architectures (Tenney et al. 2019) and by causal probing work on specific constructions, such as fillergap dependencies (Boguraev et al. 2025) and subject-verb agreement (Aryaman 2024). Though what exactly is happening “under the hood” with language models, like transformers, is a focus of experimental work, studies affirm that contextual embeddings are capable of encoding gradient syntactic and semantic differences.

3.2.1. PART OF SPEECH INFORMATION. Due to the relative novelty of using contextual embeddings for descriptive linguistic work, alongside the inability to “view” BETO’s internal processes, we added part-of-speech information (i.e., POS tags) to our analysis. This served as an interpretable validation of the syntactic information encoded by the contextual embeddings, enabled a clearer analysis, and served as a diagnostic for clustering quality: any improvements with POS tags would indicate that the embeddings did not adequately capture syntactic information, whereas no improvement would suggest the embeddings were already sufficient. Beyond providing an external measure that the contextualized embeddings capture gradient semantic and syntactic distinctions, including POS tags also provides a way to more accurately account for the syntactic properties of the constructions in which *tipo* appears.

Using spaCy’s Spanish model `es_core_news_lg` (Honnibal & Montani 2017), POS tags were assigned to the three words preceding and following each instance of *tipo* (i.e., window size=7). This threshold was set to balance efficiency with being able to characterize syntactic patterns described in previous literature on *tipo* (*de*) (Mihatsch et al. 2023). After retrieving POS information for each sentence, the tags were one-hot-encoded (OHE) and concatenated to their corresponding contextual embedding. In addition, detailed morphological features were extracted for each sentence using spaCy for later analysis. Ultimately, POS tags did not affect clustering metrics (described below). However, they did improve how interpretable the results were. We take this as further evidence for the robustness of contextual embeddings.

The final data set consisted of the 5,075 sentences containing *tipo(s) (de)* and their vectors: the contextual embeddings for each *tipo(s) (de)* token, and the POS tags as OHE vectors.

3.2.2. DATA ANALYSIS. We analyzed the *tipo(s) (de)* vectors ($n = 5,075$) using K-means clustering, an unsupervised machine learning algorithm that groups vectors based on similarity. The algorithm was run across multiple cluster values ($K=2-21$) and the final cluster parameter was selected based on (i) cluster silhouette scores⁶ and (ii) qualitative analysis of the cluster outputs. The silhouette score, which ranges from $[-1, 1]$, measures how well each data point (i.e., vector) fits within its assigned cluster compared to other clusters. This provides a measure of cluster distinctiveness and a way to evaluate how much different tokens and clusters overlap.

The final cluster structure was examined through multiple visualizations and methods that organize clusters into a hierarchy to show how the senses in each cluster relate to each other. Initial inspection relied on t-SNE scatterplot graphs to explore emergent patterns in the data (van der Maaten & Hinton 2008), and second-level analysis consisted of comparing cosine similarities and distances between the average values of each cluster (i.e., cluster centroids). While several studies on English use contextualized embeddings for descriptive analysis and to test the range of meanings they capture (Fonteyn 2021; Petersen & Potts 2023), to our knowledge, no comparable work exists for Spanish. Likewise, a growing body of research uses contextualized embeddings to study semantic change (cf. Periti & Tahmasebi 2024; Liu et al. 2025); however, we are not aware of studies that apply these methods as a way to quantitatively model possible instances of polysemy versus homonymy. In response to this gap, the present study extends prior work on contextualized embeddings by introducing dendrogram- and graph-based methods to quantify relationships among average-sense representations. In this secondary analysis, each cluster’s centroid served as an average representation of its sense (cf. Fonteyn 2021), which we examined as a semantic network using hierarchical dendrogram- and graph-based methods.

Although network science is more widely recognized in social sciences, computer science, neuroscience, and related fields, semantic network analysis can trace its roots back to Quillian (1967) and Sowa (1992). Network science has varied interpretations, especially in linguistics (cf., de Barros Pereira et al. 2022). As an analytical tool, it generally refers to a set of points (i.e., *nodes* or *vertices*) that are connected pairwise by lines (i.e., *edges*) that quantify the relationship between nodes (Newman 2010). These structures are also classified into *communities*, or groups of nodes that are highly interconnected.

Using the cluster centroids as the average senses for *tipo*, the dendrogram is constructed with the `SciPy` Python library (Virtanen et al. 2020) using cosine distance, while the sense graph is implemented with the `NetworkX` Python package (Hagberg et al. 2008), using greedy modularity optimization and cosine similarity. The sense graph was systematically tested by altering the similarity thresholds that determined the communities of closely related meanings. Analysis of cluster centroids using these hierarchical methods highlights the most prominent senses that emerge from the data, as well as the degree of similarity among them. Both the dendrogram and the communities in the sense graph show groupings of most-similar senses, but the sense graph has the added advantage of visualizing the complex relationships among more than two clusters at a given time.

Finally, the POS tags and detailed morphological features extracted from `spaCy` were used to

⁶ We also inspected the silhouette scores using the elbow method, but there was no distinct or “sharp elbow” in our data, which may further indicate the need to analyze a larger number of K clusters to disentangle different senses.

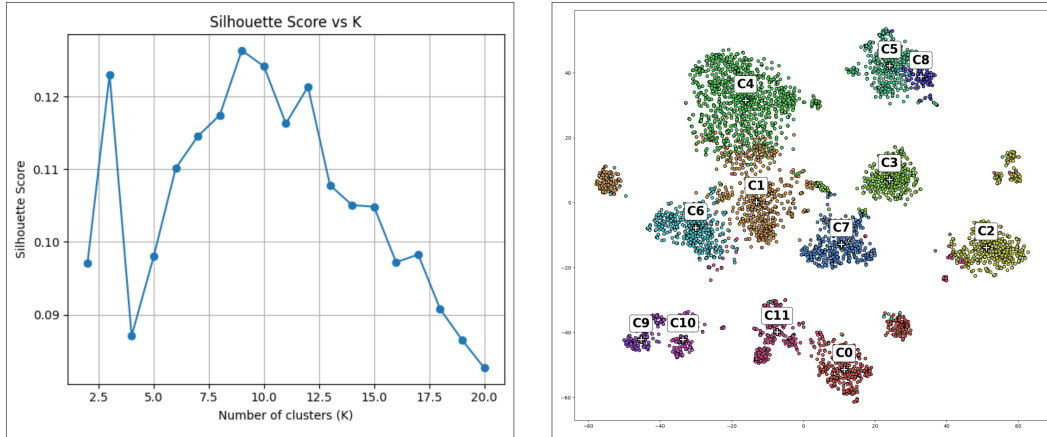


Figure 1. [Left] Silhouette score to measure cluster distinctiveness by cluster count (K : [2,21]) in the K-Means model. $k = 3, 9, 10$ result in best separation (≈ 0.13), but within-cluster results contain syntactically and semantically divergent uses. [Right] K-means clusters ($k = 12$) of *tipo* visualized with t-SNE.

analyze syntactic environment and collocational variation within and across clusters.

4. Results. Before reviewing the clustering results in detail, it is worth mentioning how the added POS information for each use of *tipo* affected cluster quality. Comparisons between models with and without POS tags showed minimal differences in silhouette scores and in the location of optimal or near-optimal values of k . While including POS tags did not substantially improve global clustering metrics, they were useful during linguistic evaluation by facilitating the qualitative interpretation of clusters. They primarily allowed for more systematic inspection of syntactic and morphological patterns. For this reason, despite not improving evaluative cluster metrics, POS-tag information was retained.

The K-means clustering models were fit for values of k ranging from 2 to 21 and evaluated using silhouette scores (Fig 1). The most prominent peak occurred at $k = 3$, which aligns with intuitive and lexicographic expectations about broad functional distinctions in the data (e.g., *taxonomic vs financial vs person*). However, the resulting clusters remained internally heterogeneous and included multiple senses and syntactic environments. At the same time, the silhouette line graph has several local optima ($k = 9, 10, 12$); this suggests that the data reveal multiple levels of meaning in a notable way. These patterns are consistent with a distinction between coarse-grained and finer-grained analyses that reflect related but distinct senses. In particular, the former may approximate notions of homonymy while the latter may capture polysemy.

To investigate these finer-grained sense distinctions, we inspected cluster outputs at the secondary local optima. We ultimately selected $k = 12$ to balance its relatively high silhouette score ($= 0.12$) and greater syntactic coherence within clusters shown by consistent POS tags and collocations. The panel on the right of Figure 1 shows the output of the clustering algorithm as a scatterplot visualized with t-SNE. Each cluster consists of instances of *tipo* that are most similar to one another; together, these clusters can be interpreted as average senses for *tipo* in this data. Some senses, such as ‘exchange, financial rate’ (clusters 9 and 10) are clearly separated, while

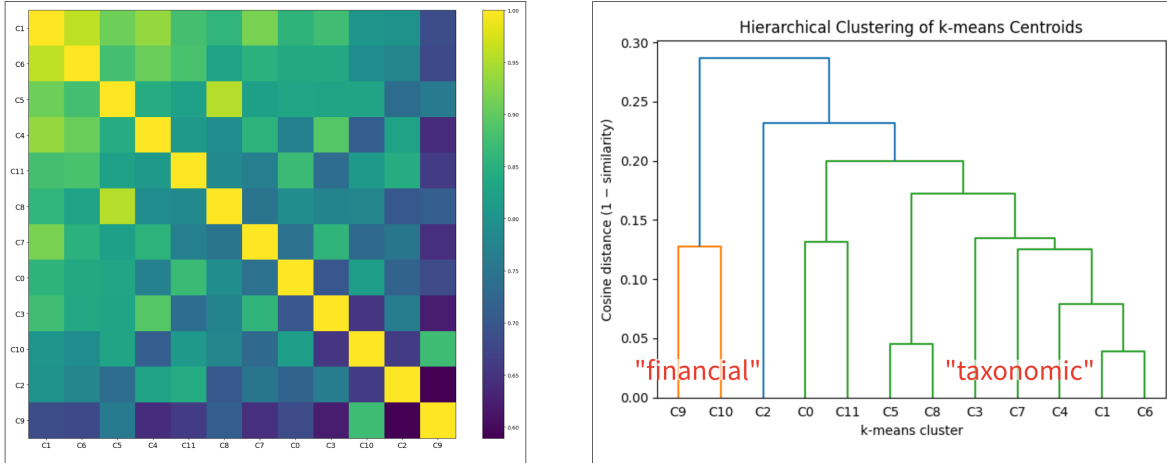


Figure 2. [Left] Cluster centroid cosine similarity matrix (ordered); lighter color is greater similarity. [Right] Dendrogram clustering of *tipo* based on cosine distance. Higher branches show weaker similarity, and lower branches show more intra-cluster similarity.

other clusters show substantial overlap. This is evident with clusters 1, 6, 4, 7, and 3, all containing the taxonomic value of *tipo*. We also see a grouping of clusters 0 and 11; the latter has some overlap with cluster 2. A manual inspection of a sample of the examples in each cluster shows that the differences between 1, 6, 4, and 3 pertain to the presence of determiners (definite vs. indefinite), quantifiers, and numerals that modify *tipo*. In cluster 0 we find instances of the sense ‘person, guy’. In clusters 11 and 2 we find examples of *tipo* expressing *ad hoc* categorization and similarity (‘like’), as well as instances in which *tipo* is modified by a demonstrative and occurs in a post-nominal PP, i.e., Det N [of this type], with anaphoric reference. We return to these in Section 5.

Despite the contextualized embeddings’ success in capturing a wide range of senses in the dataset, visualization alone does not sufficiently characterize the relationship among these meanings. Specifically, this representation captures salient, large-scale patterns in the use of *tipo*, but it does not provide a fine-grained view of similarity between senses, which is central to an account of polysemy. This limitation is tied in part to the dimensionality reduction behind visualization methods, like t-SNE, which involves projecting high-dimensional embeddings into two dimensions by prioritizing the most salient, characteristic structures in the data. Although this method preserves broad patterns in how vectors relate to one another, it inevitably distorts the original vector space, blurring nuanced distinctions represented in the full embeddings (originally greater than 700 dimensions). As such, visualization is well-suited to data exploration but not for measuring semantic relatedness.

We quantitatively capture distinctions between clusters by comparing average sense representations (i.e., centroids) using cosine similarity and distance, with the latter providing the complementary notion of dissimilarity. Figure 2 shows two representations of centroid distinctiveness. On the left, an ordered similarity matrix confirms that some senses are more closely related than others (e.g., *taxonomic* uses in clusters 1 and 6). Given the high cosine similarity values between all clusters (> 0.6 , of a possible range of $[-1, 1]$, Fig 2), this suggests that the representations and contextualized embeddings for each sense are also very similar. However, the similarity

matrix does not explicitly encode a global structure of the *tipo* senses. Therefore, we applied hierarchical clustering to the centroids to create a dendrogram/tree-based representation, as shown in the right panel of Fig 2. Here, the higher branches indicate weaker similarity, while the lower branches indicate closer meanings, as well as which ones may be subtypes of others. The dendrogram once again aligns with intuition: *financial* senses are the most distinct and branch highest, while prototypical *taxonomic* uses continue to cluster more closely towards the right side with consistently lower branches. The clusters in the middle of the tree include the *person, guy* sense (cluster 0) and uses expressing *ad hoc* categorization and similarity (cluster 11).

The relational structure among the emerging sense categories is further examined by analyzing the cluster centroids as a semantic network. Here, the senses of *tipo* are modeled as a semantic, graph-based representation with nodes as the cluster centroids (i.e., average senses), the edges depicting the cosine similarity between nodes, and the colors indicating communities of more closely related senses (Fig 3). At varying thresholds of cosine similarity (≥ 0.82), the nodes are connected with different numbers of edges, showing at which thresholds the senses overlap or diverge from each other. This representation illustrates, once again, that the sense represented by clusters 9 and 10 (‘financial rate’) diverges from the other senses at lower similarity thresholds. In addition, the senses in lilac and green are more closely connected than those in yellow (clusters 2, 11, 0). While clusters 2, 11, and 0 are still connected to the uses associated with taxonomic sense (represented in lilac and green), the similarity measures represented indicate semantic shifts, particularly to the meaning ‘a class of person’ (including character properties and physical appearance) and a broad notion of similarity.

This graph highlights a modular structure of *tipo*, offering a more interpretable, data-driven representation of its meanings and how they relate to one another. Taken together, these results suggest that most uses of *tipo* do not fall neatly into discrete sense categories, but rather exhibit a gradient, relational structure, particularly among taxonomic uses which anchor a “sense network” for *tipo*. To our knowledge, this type of analysis—modeling homonymy and polysemy as a relational structure using contextualized embeddings for a single word—remains largely unexplored.

5. Discussion. Our results show that contextualized embeddings identify meaningful sense structure without predefined categories (as it is an unsupervised method). Interestingly, the combination of contextualized embeddings with part of speech information recovers distinctions proposed in previous literature, but not in the same configuration as in those studies. For instance, in our results the taxonomic sense (cluster 1) remains central and well-connected with other senses at different cosine similarity thresholds in the nodal representation (Fig. 3). This indicates that we have a core taxonomic meaning, which is not just one among multiple senses. Rather, other senses (e.g., quantifying, phoric, identifying, list completer) are instead compositionally obtained from the taxonomic meaning and the meaning of determiners, quantifiers, and adjectives that modify *tipo*—and our methods—are sensitive to such distributional patterns. In fact, if we look at individual examples within these clusters, we see factors in the different uses identified: the presence of (in)definiteness, quantification, and restrictive vs. non-restrictive modifiers.

In addition, our results show that examples based on similarity pattern together, and are well-connected. If we look at the examples that appear in clusters 2 and 11, we find both instances containing demonstrative determiners with anaphoric uses (e.g., *incentivos de este tipo* ‘incentives of this type’) and instances of *ad hoc* categories (e.g., *somos del tipo que se disculpa* ‘we are of the apologetic type’). Such examples show that reference to a common type is a means of

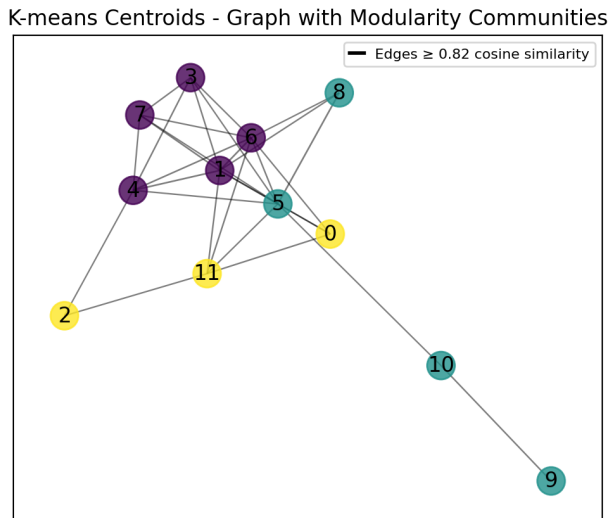


Figure 3. Sense graph of the average sense representations for *tipo*. Nodes represent average cluster senses, edges represent cosine similarity (≥ 0.82), and colors represent communities of broader senses.

expressing similarity (vs. dissimilarity), which is crucial for categorization. We believe that the closeness in sense structure between clusters 2 and 11 corresponds to the semantic domain of similatives, i.e. “meanings that lie at the intersection of the semantic dimensions of similarity and demonstration” (van der Auwera & Sahoo 2020). The category of similatives was originally identified by Haspelmath and Buchholz for adverbials expressing similarity, as in *Fatmir sings like a nightingale* (Haspelmath & Buchholz 1998:277-278). However, the notion was later expanded to cover a range of words that defy strict categorization as determiners, quantifiers or pronouns, like the reflexes of Latin TALIS (French *tel*, Portuguese *tal*, Spanish *tal*) and English *such*.⁷ The sense graph captures the relatedness of such senses, and how they diverge from the original taxonomic meaning. A detailed exploration of the connection of *tipo* with anaphoricity, demonstration, and similarity, revealed here, is left for future work.

As pointed out above, one of the difficulties in studies of polysemy, and taxonomic nouns in particular, is how to decide between polysemy and homonymy. Our results suggest that the ‘financial rate’ meaning is an example of homonymy: it separates at a lower threshold than other meanings (see clusters 9 and 10). We take this as an indication that the meaning ‘financial rate’ corresponds to a different lexeme. Hence, information from the nodal representation provides a data-driven way to distinguish instances of polysemy (‘category’, ‘person’) from homonymy (‘financial rate’), matching native speakers’ intuitions. Note that this provides a principled way to distinguish among meanings. Comparatively, the DRAE lists ‘financial rate’ as the tenth meaning of *tipo*, with no clear reason being indicated for this decision.

6. Conclusion. In this paper, we explored the potential of contextualized embeddings to examine the range of meanings of *tipo*. This data-driven method has allowed us to test existing propos-

⁷ For a recent volume highlighting the connection between similatives and taxonomic nouns, see Kisiel et al. (2025).

als and identify replicable criteria to analyze the uses of *tipo*. Importantly, our approach was not supervised, so we included all uses of *tipo* as represented in the corpus. This is unlike previous approaches (Mihatsch 2016, 2023), who exclude upfront some of the values of *tipo*, e.g. ‘financial rate’, and argue that only non-nominal uses are considered. We believe that such an approach is problematic: the distribution of *tipo* in examples classified by Mihatsch e.g., as Quantifying or Identifying, reveals the maintenance of nominal properties of *tipo*, like the co-occurrence with determiners or modification by adjectives. Furthermore, using sense relatedness to make determinations about polysemy requires considering all the senses.

The clustering obtained, together with the methods that target sense similarity, yielded a sense network, revealing a structure akin to a radial category (cf. Lakoff 1987). We take these results as a first step; the nodal representation we leveraged provides a measure of similarity and addresses questions regarding number of senses and their relationships, pressing issues in studies of polysemy. In this paper we do not further develop this model to specify the types of relations (e.g., metaphor, inference, cf. Jurafsky 1996) that bind the different senses; this is left for future work.

This methodology can contribute to study other instances of polysemy that raise similar problems. While our goal here is purely synchronic, from our results one can also hypothesize a historical path based on the nodal representations presented. The connections among senses thus identified may provide a blueprint for an empirically-based semantic map of *tipo* that may capture tendencies in semantic shifts.

References

- Amaral, Patrícia, Hai Hu & Sandra Kübler. 2023. Tracing semantic change with distributional methods: The contexts of *algo*. *Diachronica* 40(2). 153–194. <https://doi.org/10.1075/dia.21012.ama>.
- Aryaman, Dan; Potts Christopher, Arora; Jurafsky. 2024. Causalgym: Benchmarking causal interpretability methods on linguistic tasks. <https://arxiv.org/abs/2402.12560>.
- de Barros Pereira, Hernane Borges, Marcos Grilo, Inácio de Sousa Fadigas, Cleônidas Tavares de Souza Junior, Marcelo do Vale Cunha, Renata Souza Freitas Dantas Barreto, Júlia Carvalho Andrade & Trazíbulo Henrique. 2022. Systematic review of the semantic network definitions. *Expert Systems with Applications* 210. 118455.
- Boguraev, Sasha, Christopher Potts & Kyle Mahowald. 2025. Causal interventions reveal shared structure across English filler–gap constructions. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose & Violet Peng (eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing*, 25021–25042. Suzhou, China: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.1271>.
- Brems, Lieselotte & Kristin Davidse. 2010. The Grammaticalisation of Nominal Type Noun Constructions with *kind/sort of*: Chronology and paths of change. *English Studies* 91(2). 180–202. <https://doi.org/10.1080/00138380903355023>.
- Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang & Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data. <https://arxiv.org/abs/2308.02976>.
- Cañete, José. 2019. Compilation of large spanish unannotated corpora. <https://doi.org/10.5281/zenodo.3247731>.

- De Smedt, Liesbeth, Lieselotte Brems & Kristin Davidse. 2007. NP-internal functions and extended uses of the type nouns *kind*, *sort*, and *type*: towards a comprehensive, corpus-based description. In Roberta Facchinetti (ed.), *Corpus linguistics 25 years on*, Rodopi.
- Denison, David. 2002. History of the sort of construction family. Paper presented at the Second International Conference on Construction Grammar (Helsinki, 6-8 September 2002).
- Firth, J. R. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis*, Philological Society. Reprinted in Palmer, F. (ed.), *Selected Papers of J. R. Firth*, Longman, Harlow, 1968.
- Fonteyn, Lauren. 2021. Varying abstractions: a conceptual vs. distributional view on prepositional polysemy. *Glossa: a journal of general linguistics* 6(1). 90. <https://doi.org/10.5334/gjgl.1323>.
- Hagberg, Aric A., Daniel A. Schult & Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught & Jarrod Millman (eds.), *Proceedings of the 7th python in science conference*, 11–15. Pasadena, CA USA.
- Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In Katrin Erk & Noah A. Smith (eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 1489–1501. Berlin, Germany: Association for Computational Linguistics. <https://doi.org/10.18653/v1/P16-1141>.
- Harris, Zellig S. 1954. Distributional structure. *Word* 10(2–3). 146–162.
- Haspelmath, Martin & Oda Buchholz. 1998. Equative and similitive constructions in the languages of Europe. In Johan van der Auwera (ed.), *Adverbial Constructions in the Languages of Europe*, 277–334. Berlin, New York: De Gruyter Mouton. <https://doi.org/doi:10.1515/9783110802610.277>.
- Honnibal, Matthew & Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Jurafsky, Daniel. 1996. Universal tendencies in the semantics of the diminutive. *Language* 72(3). 533–578.
- Kisiel, Anna, Hlne Vassiliadou, Valentina Benigni, Beatrice Bernasconi, Lieselotte Brems & Dejan Stosic. 2025. Introduction. In *Similatives. semantic sources, pathways, and types of usage*, vol. 48, 179–188. John Benjamins. <https://doi.org/https://doi.org/10.1075/li.48.2>. Special Issue of *Linguisticae Investigationes*.
- Kolyaseva, Alena & Anna Kisiel. 2023. Taxonomic nouns in Slavic: An overview. In Wiltrud Mihatsch, Inga Hennecke, Anna Kisiel, Alena Kolyaseva, Kristin Davidse & Lieselotte Brems (eds.), *Type Noun Constructions in Slavic, Germanic and Romance Languages*, 455–500. Berlin, Boston: De Gruyter Mouton. <https://doi.org/doi:10.1515/9783110701104-012>.
- Lakoff, George. 1975. *Hedges: A study in meaning criteria and the logic of fuzzy concepts* 221–271. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-010-1756-5_9.
- Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.
- Liu, Yifan, Gelila Tilahun, Xinxiang Gao, Qianfeng Wen & Michael Gervers. 2025. A comparative study of static and contextual embeddings for analyzing semantic changes in medieval Latin charters. In Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan & Lasitha Uyangodage (eds.), *Pro-*

- ceedings of the first workshop on language models for low-resource languages, 182–192. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- van der Maaten, Laurens & Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9. 2579–2605.
- Mauri, Caterina & Andrea Sansó. 2018. Linguistic strategies for ad hoc categorization: theoretical assessment and cross-linguistic variation. *Folia Linguistica* 52(s39-s1). 1–35. <https://doi.org/doi:10.1515/flih-2018-0001>.
- Mihatsch, Wiltrud. 2016. Type-noun binominals in four Romance languages. *Language Sciences* 53. 136–159. <https://doi.org/https://doi.org/10.1016/j.langsci.2015.05.009>.
- Mihatsch, Wiltrud. 2023. General introduction: Taxonomic nouns and their derived functions in Germanic, Romance and Slavic languages. In Wiltrud Mihatsch, Inga Hennecke, Anna Kisiel, Alena Kolyaseva, Kristin Davidse & Lieselotte Brems (eds.), *Type Noun Constructions in Slavic, Germanic and Romance Languages*, 1–52. De Gruyter Mouton.
- Mihatsch, Wiltrud, Inga Hennecke, Anna Kisiel, Alena Kolyaseva, Kristin Davidse & Lieselotte Brems (eds.). 2023. *Type Noun Constructions in Slavic, Germanic and Romance Languages*. Berlin, Boston: De Gruyter Mouton. <https://doi.org/doi:10.1515/9783110701104>.
- Newman, Mark E.J. 2010. *Networks: An introduction*. Oxford University Press.
- Periti, Francesco & Nina Tahmasebi. 2024. Towards a complete solution to lexical semantic change: an extension to multiple time periods and diachronic word sense induction. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, David Alfter, Francesco Periti, Pierluigi Cassotti & Netta Huebscher (eds.), *Proceedings of the 5th workshop on computational approaches to historical language change*, 108–119. Bangkok, Thailand: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.lchange-1.10>.
- Petersen, Erika & Christopher Potts. 2023. Lexical semantics with large language models: A case study of English “break”. In Andreas Vlachos & Isabelle Augenstein (eds.), *Findings of the association for computational linguistics: Eacl 2023*, 490–511. Dubrovnik, Croatia: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-eacl.36>.
- Quillian, M. Ross. 1967. Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science* 12(5). 410–430. <https://doi.org/https://doi.org/10.1002/bs.3830120511>.
- Ravin, Yael & Claudia Leacock. 2002. Polysemy: An overview. In Yael Ravin & Claudia Leacock (eds.), *Polysemy: Theoretical and computational approaches*, 1–29. Oxford University Press.
- Rosch, Eleanor H. 1973. On the internal structure of perceptual and semantic categories. In Timothy E. Moore (ed.), *Cognitive development and acquisition of language*, 111–144. San Diego: Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-505850-6.50010-4>.
- Sowa, John F. 1992. Semantic network. In Stuart C. Shapiro (ed.), *Encyclopedia of artificial intelligence*, Wiley 2nd edn.
- Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das & Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. <https://arxiv.org/abs/1905.06316>.
- van der Auwera, Johan & Kalyanamalini Sahoo. 2020. Such similatives:

a cross-linguistic reconnaissance. *Language Sciences* 81. 101217.

<https://doi.org/https://doi.org/10.1016/j.langsci.2018.12.002>.

Virtanen, Pauli et al. 2020. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods* 17. 261–272.