frontiers | Frontiers in Language Sciences

TYPE Original Research
PUBLISHED 21 November 2024
DOI 10.3389/flang.2024.1458887

OPEN ACCESS

EDITED BY
Jon Andoni Dunabeitia,
Nebrija University, Spain

REVIEWED BY
Jose Luis Tapia,
University of Valencia, Spain
Carolina Gattei,
Universidad Torcuato Di Tella, Argentina

*CORRESPONDENCE
Josu Goikoetxea
✉ josu.goikoetxea@ehu.eus

RECEIVED 10 August 2024
ACCEPTED 28 October 2024
PUBLISHED 21 November 2024

CITATION
Goikoetxea J, San Martin I and Arantzeta M (2024) Bridging Natural Language Processing and psycholinguistics: computationally grounded semantic similarity datasets for Basque and Spanish.
Front. Lang. Sci. 3:1458887.
doi: 10.3389/flang.2024.1458887

Bridging Natural Language Processing and psycholinguistics: computationally grounded semantic similarity datasets for Basque and Spanish

Josu Goikoetxea[1]*, Itziar San Martin[2] and Miren Arantzeta[3]

[1]HiTZ Research Center, Bilbao School of Engineering (EHU/UPV), Bilbao, Spain, [2]The Bilingual Mind - Micaela Portilla Research Center, Basque Language and Communication (EHU/UPV), Vitoria-Gasteiz, Spain, [3]The Bilingual Mind - Micaela Portilla Research Center, Linguistics and Basque Studies (EHU/UPV), Vitoria-Gasteiz, Spain

Introduction: Semantic relations are crucial in various cognitive processes, highlighting the need to understand concept interactions and how such relations are represented in the brain. Psycholinguistics research requires computationally grounded datasets that include word similarity measures controlled for the variables that play a significant role in lexical processing. This work presents a dataset for noun pairs in Basque and European Spanish based on two well-known Natural Language Processing resources: text corpora and knowledge bases.

Methods: The dataset creation consisted of three steps, (1) computing four key psycholinguistic features for each noun; concreteness, frequency, semantic, and phonological neighborhood density; (2) pairing nouns across these four variables; (3) for each noun pair, assigning three types of word similarity measurements, computed out of text, Wordnet and hybrid embeddings.

Results: A dataset of noun pairs in Basque and Spanish involving three types of word similarity measurements, along with four lexical features for each of the nouns in the pair, namely, word frequency, concreteness, and semantic and phonological neighbors. The selection of the nouns for each pair was controlled by the mentioned variables, which play a significant role in lexical processing. The dataset includes three similarity measurements, based on their embedding computation: semantic relatedness from text-based embeddings, pure similarity from Wordnet-based embeddings and both categorical and associative relations from hybrid embeddings.

Discussion: The present work covers an existent gap in Basque and Spanish in terms of the lack of datasets that include both word similarity and detailed lexical properties, which provides a more useful resource for psycholinguistics research in those languages.

KEYWORDS

WordNet, text, psycholinguistic features, word similarity, embeddings, nouns

1 Introduction

Semantic similarity is a measure of distance between items determined by the closeness of their meanings. It represents a type of relationship grounded in shared characteristics between concepts. For instance, "cat" has more marked semantic similarity with "tiger" than with "rhinoceros." In contrast, semantic relatedness, frequently used interchangeably with semantic similarity, denotes any relation between concepts, not necessarily taxonomical. For example, "cat" is related to "tiger," "milk," and "veterinary." In short, semantic similarity offers a metric of categorical semantic relations, whereas semantic relatedness is closer to depicting associative relations. For now, we will use

these two concepts interchangeably and show how the metrics used in the present dataset constitute independent proxies for different types of semantic relations.

Collins and Loftus (1975) proposed a spreading activation theory of semantic processing, suggesting that our mental lexicon is organized into a network of concepts linked by semantic relations. This network allows for the co-activation of related concepts based on shared properties. For example, concepts within the same category (e.g., vehicles or colors) are more closely interlinked than those with fewer shared properties. They explain that the degree of relatedness between concepts is determined by the number and strength of the connections between them. This implies that concepts with many common properties will activate one another more readily, facilitating access to related information during cognitive tasks.

Psycholinguistic research has made significant efforts to highlight the important role that semantic relations play in various cognitive processes by helping us understand how concepts interact and influence each other. For instance, studies show that memory recall — the ability to remember specific information, such as word pairs — is strongly modulated by the semantic relations between the items (Kenett et al., 2017; Kowialiewski and Majerus, 2020). This suggests that related word pairs are more easily retrieved than unrelated ones, highlighting the impact of semantic connections on memory performance. Additionally, semantic relations affect mental imagery, as individuals often visualize concepts based on their interconnectedness. For example, when asked to imagine a "bird," a person might also visualize "nest" or "feather" due to the semantic relations they share. Research indicates that mental imagery plays a significant role in cognitive processes, facilitating not only recall but also problem-solving and comprehension by providing a visual representation that aids in organizing and manipulating information (Kosslyn et al., 2006). Semantic relations similarly influence language comprehension; the ability to predict the next word in a sentence is closely tied to the relations between the elements involved. Stronger semantic relations enhance the co-activation of related concepts, which promotes anticipatory processing and increases predictability (Federmeier and Kutas, 1999; Federmeier, 2007). For example, in the sentence "The cat sat on the...," the word "mat" is more predictable than "cloud" due to the stronger semantic association with "cat."

Taxonomic and associative relations significantly influence how the brain processes semantic information, affecting processing in distinct ways. Sass et al. (2009) showed that while both taxonomic (e.g., "cat" and "dog") and associative (e.g., "cat" and "wool") relations facilitate faster lexical decisions, thematic relations elicit stronger priming effects and engage distinct neural networks compared to taxonomic ones. Similarly, evidence from aphasia studies supports a neural dissociation between these relations. Schwartz et al. (2011) found that taxonomic errors in post-stroke aphasia localized to the left anterior temporal lobe, whereas thematic errors localized to the left temporoparietal junction. These findings suggest that the brain processes taxonomic and thematic knowledge through distinct neural pathways, reinforcing the idea that these semantic relations are represented differently in the human brain.

Studying semantic relations has become a significant focus in both Natural Language Processing (NLP) and psycholinguistics. Computational tools have been developed to gauge the strength of semantic connections between concepts, facilitating various cognitive experiments. Two important resources in NLP are text corpora and knowledge bases (KBs), both of which are valuable in investigating semantic similarity. In the case of KBs, they provide a quantitative and structured framework for representing the meanings of words, enhancing our understanding of semantic relations.

This work presents two automatically calculated semantic similarity datasets restricted to nouns, to provide useful material to build up psycholinguistic experiments; one in Basque and the other in Spanish. The similarity measurements of noun pairs were computed using vectors or embeddings built up in text corpora and KB. To capture different nuances of semantic similarity between noun pairs, in this work we employed three types of word representations; text embeddings, KB-based embeddings, and hybrid embeddings. The selection of noun pairs was conducted by controlling for length, word frequency, concreteness, and the number of semantic and phonological neighbors due to the widely proven effect of these variables in language processing. Controlling for these linguistic features allows leveraging crucial lexical properties within pairs of nouns beyond semantic similarity, broadening the usability of the dataset and the interpretation of the results.

Additionally, the unique linguistic properties of Basque suggest the development of materials that may be of key interest for psycholinguistic research. In particular, Basque (a language isolate) is an agglutinative language, as opposed to the fusional nature of Spanish. This implies that in Basque noun formation strategies, the morphological processes of derivation and compounding are both productive and involve clear addition and stacking of morphemes, creating longer but highly transparent words in meaning (Hualde and De Urbina, 2011). In contrast, Spanish derivation and compounding processes result in shorter and less transparent units. Additionally, unlike in Spanish, Basque compounds are a very frequent strategy for word creation, with compounds that have their semantic head either first or last in the compound. This provides an excellent ground for investigating whether lexical decomposition takes place in lexical access and whether the position of the semantic head in the compound has any effect on parsing. Acknowledging these distinct processes of lexical construction and the different linguistic typologies of these languages will enhance our understanding of how semantic relations operate across diverse linguistic contexts, ultimately enriching the field of psycholinguistics.

The paper is structured as follows: first, a summary of the related work of the current research; second, motivation of the inclusion of the linguistic features in the creation of the dataset; third, material and methods utilized in the paper; fourth, results obtained in several stages of the dataset creation; fifth, conclusion.

2 Related work

Psycholinguistics and NLP have been closely linked for a long time. In semantic similarity computation, two methods have been mainly used; KB-based and embedding-based. The former treats KBs as graphs and exploits their complete structural information, so that similarity measurements are based on the KB taxonomy. The latter encodes word meaning in a numeric vector

or embedding in a Euclidean space following the Distributional Hypothesis (Harris, 1954). Thus, embeddings latently encode the semantic (and also syntactic) features, and therefore, offer comparable representations for words.

Embedding measurements have revolutionized the field of NLP, increasing model accuracy significantly in several tasks such as named entity recognition (Pennington et al., 2014; Lample, 2016), text classification (Zhang et al., 2015), coreference resolution (Clark and Manning, 2016) or machine translation (Lample et al., 2017). Furthermore, they have surpassed KB-based methods in the word similarity task, proving to be robust resources (Lastra-Díaz et al., 2019).

In recent years, several works have given meaningful insights into the correspondence between word-embedding features and human cognition. Despite the misconceptions in understanding word embedding from a cognitive perspective (Günther et al., 2019), they remain the most meaningful proxy for human semantic representations. Thus, the performance of embeddings in the semantic similarity tasks has gained attention lately in combination with a variety of experimental methods, such as semantic priming (Ettinger and Linzen, 2016; Auguste et al., 2017; Hollenstein et al., 2019; Farhy and Veríssimo, 2019; Chersoni et al., 2021), brain imaging (Jain and Huth, 2018; Rodrigues et al., 2018; Toneva and Wehbe, 2019; Djokic et al., 2020; Zhang et al., 2024), eye-tracking (Luke and Christianson, 2018; Hollenstein et al., 2019; Kun et al., 2023; Zhang et al., 2024).

Mandera et al. (2017) proved that word vectors successfully explain semantic priming data, stating that they equal or outperform human-rated association datasets or feature norms. Salicchi et al. (2021) showed a strong correlation between similarity measurements of contextual and non-contextual embeddings and eye-tracking data collected from participants reading two English corpora. Hayes and Henderson (2021) explored the relationship between the visual scene and attention, grounding the semantic scene representation with word embeddings and evidencing the strong relationship between the semantic similarity of a scene region and the gaze-fixation pattern of the viewers.

Other psycholinguistic research has gauged the semantic distance of words via KB-based methods, as the latter provides a means of quantifying the relationships between words and concepts within the semantic structure of a language. Kenett et al. (2017) created a Hebrew KB, quantifying semantic distance as the path length between word pairs by counting the number of steps for traversing from one word to another in the KB. They stated that a distance of 4 was the turning point for the performance in the semantic relatedness judgement task and subsequent recall from memory. Similarly, Benedek et al. (2017) proved that a KB based on semantic relatedness judgements was feasible and valuable to test the associative and executive accounts of creativity. Likewise, Jelodar et al. (2010) modeled brain activation patterns measured with fMRI using semantic feature vectors of concrete nouns based on Wordnet (Miller, 1995). It outperformed previous models.

Recent years have seen increased interest in the relationship between semantic relations and the human mind, particularly concerning cognitive aging. Advances in understanding the linguistic and semantic changes throughout the adult lifespan have fundamentally shaped the mental lexicon — the repository of lexical and conceptual representations. A crucial element of this research is the use of quantitative methods to measure semantic relations, providing valuable insights into language processing dynamics. Ongoing debates address the sources of linguistic and semantic changes, specifically the roles of environmental exposure and cognitive mechanisms involved in learning, representation, and retrieval (see Wulff et al., 2019). Wulff et al. (2022) demonstrate that life experiences influence the structure and flexibility of semantic knowledge, revealing that individual and age differences in semantic networks impact cognitive processes. Their findings indicate that older adults possess lexical networks with smaller average degrees and longer path lengths compared to younger adults, leading to greater variability in lexical representations. These studies underscore the dynamic nature of semantic relations throughout the lifespan and their implications for cognitive functioning.

Broderick et al. (2021) reveal that older adults employ different predictive mechanisms when processing semantic information compared to younger adults. Their use of semantic dissimilarity models based on embeddings shows that, while both age groups rely on context-based lexical predictions, older adults exhibit reduced pre-activation of semantic features, affecting their ability to comprehend natural speech effectively. Additionally, Cosgrove et al. (2021) examine age-related differences in the flexibility of semantic memory networks through percolation analysis, finding that older adults' networks are less adaptable, which may contribute to declines in language production despite a rich store of semantic knowledge. Collectively, these studies highlight the value of quantitative methods in elucidating the complexities of semantic relations and their relevance to cognitive processes across different age groups.

There has also been substantial progress in developing psycholinguistic datasets for English that include both word similarity measures and detailed lexical properties. For instance, the English Lexicon Project (Balota et al., 2007) offers comprehensive lexical and psycholinguistic data, including word frequency, orthographic and phonological neighborhood sizes, and lexical decision times. Similarly, SUBTLEX-US (Brysbaert and New, 2009) provides word frequency data based on U.S. film subtitles, making it a reliable source for spoken language frequency counts. The CELEX Lexical Database (Baayen, 1995) further enriches lexical studies with detailed morphological, phonological, and frequency information. Other notable English resources include the Bristol Norms (Stadthagen-Gonzalez and Davis, 2006), which offer concreteness, imageability, and familiarity ratings. Additionally, the British Lexicon Project (Keuleers et al., 2012) provides valuable lexical decision data for British English. Several word similarity datasets in English have been extended to other languages, including RG65 (Rubenstein and Goodenough, 1965), WordSIm353 (Finkelstein et al., 2001), Simlex999 (Hill et al., 2015),[1] MTURK287 (Radinsky et al., 2011), Rarewords (Luong et al., 2013), and MEN (Bruni et al., 2014).

Despite the increasing interest in psycholinguistic datasets, there remains a significant gap when it comes to resources available for Basque and Spanish. For Spanish, EsPal dataset (Duchon et al., 2013) provides extensive lexical information such as word frequency, orthographic and phonological neighborhood size, and

1 This dataset also includes concreteness measurements along with the similarity values.

concreteness ratings but it does not focus specifically on word similarity. Additionally, SUBTLEX-ESP (Cuetos et al., 2011) offers word frequency counts based on Spanish subtitles, providing a valuable resource for more colloquial language use. There are also three-word similarity datasets in Spanish, namely, SimLex-999 (Etcheverry and Wonsever, 2016), Wordsim353 (Hassan and Mihalcea, 2009) and RG65 (Camacho Collados et al., 2015), but these lack essential lexical information. For Basque, there only available word similarity datasets, RG65, and WordSim353 (Goikoetxea et al., 2018), are similarly limited in scope. To date, no dataset exists in either Spanish or Basque that integrates both word similarity measures and crucial lexical properties. To address these gaps, our study introduces computationally grounded word similarity datasets for both Basque and Spanish that integrate not only word similarity measures but also crucial lexical properties such as frequency, concreteness, and neighborhood size. This combined approach provides a more comprehensive resource for psycholinguistic research in these languages.

3 Linguistic features included in the dataset

Although the core of this dataset is word similarity calculation, the dataset is controlled by four additional features: concreteness, frequency, semantic neighborhood density and phonological neighborhood density. The following sections summarize the scientific evidence highlighting the significant role of these variables in psycholinguistics and NLP. A somewhat deeper comprehension of the connections between various linguistic features and word similarity from a cognitive viewpoint is needed. This dataset may lead to gaining insight into ongoing studies.

3.1 Concreteness

Concreteness is a term to refer to the degree to which a word denotes a tangible thing. This measurement was introduced by Paivio (1971), and has been proven to play a role in several aspects, such as working memory (Mate et al., 2012), embodied cognition (Barsalou, 1999; Fischer and Zwaan, 2008; Hauk et al., 2004), and neural representations on word processing (Wang et al., 2010). The literature suggests that concrete words show thicker links to associated semantic information and involve visual imagery processes. Accordingly, they elicit faster response times and larger N400 and N700 electrophysiological signals (Schwanenflugel, 2010; Wang et al., 2010). Remarkably, when contextual information and mental imageability are controlled, response times become faster for abstract words. However, the neural correlates do not change, suggesting that concrete words involve more significant semantic processing during meaning activation (Barber et al., 2013).

Due to the relevance of concreteness in psycholinguistic research, several works on NLP have computationally predicted concreteness values using word embeddings (Ljubešić et al., 2018; Charbonnier and Wartena, 2019; Incitti and Snidaro, 2021), which have been proven to be useful in tasks like metaphor detection (Tsvetkov et al., 2014; Alnafesah et al., 2020) and sentiment analysis (Rothe et al., 2016; Long et al., 2019). In Long et al. (2019), for example, the authors propose a cognition-grounded attention model in sentiment analysis, considering concreteness for leveraging the model's attention mechanism. Further, concreteness seems to be gaining more attention in the NLP field, as some concreteness norms rated by humans have been published in various languages lately, the English (Brysbaert et al., 2014b) and Dutch (Brysbaert et al., 2014a) ones being quite extensive, and the Croatian (Ćoso et al., 2019), Russian (Solovyev et al., 2022), French (Bonin et al., 2018) and Spanish (Guasch et al., 2016) ones rather more reduced.

3.2 Word frequency

Word frequency is a critical factor in both psycholinguistics and NLP. In psycholinguistics, it is established that the frequency with which a word occurs in language significantly influences how individuals process and recall information. This effect has been extensively studied concerning lexical access and recall tasks, which are essential for understanding memory usage. Higher word frequency leads to faster reaction times and increased accuracy in tasks requiring word recognition and recall (Balota and Chumbley, 1984; Balota et al., 2004; Brysbaert and New, 2009; MacLeod and Kampe, 1996; Gregg, 1976; Kinsbourne and George, 1974).

Online studies have shown that words with higher frequencies elicit shorter gaze fixation durations during reading, indicating that readers can process these words more efficiently (Joseph et al., 2013; Raney and Rayner, 1995). Raney and Rayner (1995) found that participants made shorter fixations on high-frequency words compared to low-frequency words during both the first and second readings of text passages. The consistent effect across both readings indicates that even when encountering familiar text, the inherent frequency of the words continues to influence cognitive processing.

Moreover, this phenomenon extends beyond reading. Both Dahan et al. (2001) and Magnuson et al. (2007) employed the visual world paradigm to explore the impact of word frequency on language processing. Dahan et al. (2001) demonstrated that higher-frequency names resulted in shorter fixation latencies on referent pictures, indicating that participants identified and responded to objects associated with high-frequency words more quickly. In contrast, Magnuson et al. (2007) revealed that early and continuous effects of frequency facilitate the activation of high-frequency words while inhibiting competition from phonologically similar low-frequency words. Together, these studies underscore the crucial role of word frequency in guiding attention and processing efficiency across both spoken language comprehension and reading.

This aligns with electrophysiological data, where earlier and greater neural responses are recorded for high-frequency words, reflecting their ease of access in the mental lexicon (Strijkers et al., 2010). This indicates that word frequency not only affects memory retrieval but also influences the real-time processing of language. For a comprehensive review of these effects, see Brysbaert et al. (2018).

From a computational perspective, word frequency is also a fundamental feature in Text Meaning, and Information Retrieval tasks, among others. Language structure provides information about how important a word is in a text or corpus by measuring its occurrences. In NLP, it can be used in a wide variety of tasks, such as to determine the most frequent words in a language (Spink et al., 2001), to identify rare words (Dave et al., 2003), to

synthesize information (Haghighi and Vanderwende, 2009) or to answer questions automatically (Koehn and Knowles, 2017).

3.3 Semantic neighborhood density

The semantic structure of the lexicon has always played an essential role in psycholinguistics and NLP, as the organization of knowledge is a pivotal aspect of studying meaning. Semantic neighborhood density[2] (SND) of a word correlates with several cognitive processes and their respective brain activations but remains a hitherto poorly explored field.

Different semantic neighborhood size measures have been used in lexico-semantic research (e.g., metrics based on feature semantics, co-occurrence, and categorical relations), showing diverging effects of SND in lexical processing. For instance, when SND is measured considering semantic associations between words (or word co-occurrence is taken as a proxy of it), words with large semantic neighborhoods generate faster responses than those with sparse semantic neighborhoods in lexical decision tasks (Yates et al., 2003; Buchanan et al., 2001; Locker et al., 2003). Conversely, when featural semantic information is taken as a base for calculating the density, words with sparse neighborhoods are related with faster recognition and naming (Rabovsky et al., 2016; Reilly and Desai, 2017). In line with these findings, Abdel Rahman and Melinger (2007) shows that associative relations facilitate word processing, whereas categorical connections between words create interference. Duñabeitia et al. (2008) analyze the influence of the number of associates of a word in four different visual word recognition tasks, showing that the words with higher amount of associates were processed faster. Consequently, different metrics of SND tap different constructs that need to be understood as complementary.

From an NLP perspective, KBs such as WordNet (Miller, 1995), FrameNet (Baker et al., 1998), BabelNet (Navigli and Ponzetto, 2010) or even Wikipedia provide a structured and quantifiable framework for words. These resources capture various semantic relations (e.g., synonymy, hypernymy, and meronymy) that strongly correlate with our mental lexicon, particularly in terms of semantic relatedness (Rogers and McClelland, 2004; Spivey, 2008; Boden, 2008; Jones et al., 2015). Specifically, this correlation refers to how these semantic relations inform our understanding of word meanings and relationships in human cognition.

While SND appears relevant to cognitive processes, the extent to which SND specifically impacts model performance in NLP has not been extensively explored. This lack of research highlights a significant gap in the literature regarding the implications of SND for NLP applications, suggesting that further investigation could yield valuable insights into the relationship between SND and human language processing.

3.4 Phonological neighborhood density

As with the SND, Phonological Neighborhood Density (PND) of words affects lexical processing, which indicates a need to control for such variables when designing a study. A word's PND refers to the number of words in the lexicon that can be formed by substituting a single phoneme of the target word. Likewise, orthographic neighborhood density (OND) is defined by the number of words that can be formed by replacing a single letter of the target word (Colheart et al.'s N metric) (Coltheart et al., 1977). These two measures imply considerable differences in opaque languages, such as English, Arabic, or French, because phonemes do not present a one-to-one mapping into graphemes. Some studies on opaque languages have argued that neighborhood effects reflect phonological rather than orthographic similarity (Mulatti et al., 2006; Yates et al., 2004); when orthographic similarity is controlled, phonological similarity still affects lexical decision times. In contrast, PND and OND measures can be used interchangeably in shallow languages such as Basque and Spanish, as both languages exhibit a direct correspondence between phonemes and graphemes, with a few exceptions. Consequently, we adopted OND as a measure of phonological neighborhood density in our study, using orthographic neighborhood data to effectively estimate phonological neighborhood density.

The influence of phonologically related words has been explored in different studies of phonological neighborhood density, and it has shown relevant effects in task-dependent language processing. In particular, PND seems to exert competitive effects on word recognition tasks, but facilitatory effects in production tasks (Dell and Gordon, 2011; Gahl et al., 2012). In spoken word recognition, the acoustic stimulus activates potential candidates that are phonologically similar. Thus, the larger the phonological neighborhood of a word, the harder it is to recognize the target word (Luce and Pisoni, 1998). Paradoxically, in word production, dense phonological neighbors seem to ease production. For example, (Vitevitch, 2002) found facilitatory effects of neighborhood density in picture naming in a study that controlled other key factors such as frequency, neighborhood frequency, familiarity and phonotactic probability. Although most studies have attested facilitatory effects in production, such effects are less consistent than in recognition tasks.

4 Materials and methods

This section is dedicated to describing the materials and methods used to create the computationally grounded Basque and Spanish datasets. Section 4.1 summarizes the text- and knowledge-based corpora and the three types of word representations that form the foundation of the datasets, while Section 4.2 outlines the procedure for creating the datasets.

4.1 Materials

Two primary NLP sources of semantic information have been employed to create the Basque and Spanish semantic similarity datasets based on embeddings; KBs and text corpora. The former, the multilingual version of Wordnet (Miller, 1995) being used for its reliability as a KB in NLP and the abundance of associated libraries and tools. Table 1 summarizes the corpora and their derived embeddings which constitute the core of the resources in this work:

2 Following the terminology in this field, we will refer to the size of the neighborhood when using the term "density," both for semantic and phonological neighborhoods.

TABLE 1 Acronyms along with the description of the main resources used in this work.

	Acronym	Description
Corpora	txt	Text corpora
	kb	Wordnet-based pseudo-corpora
embeddings	FT_{txt}	`fasText` over text corpora
	FT_{kb}	`fasText` over wordnet-based pseudo-corpora
	FT_{hyb}	Combination of previous two embeddings

The first group includes two types of corpora, KB and text corpora (*kb* and *txt*, respectively). The second group describes the types of embeddings, namely `fastText` over text corpora, over wordnet pseudo-corpora and the meta-embeddings, which combine the first two types (TF_{txt}, TF_{kb} and FT_{hyb}, respectively).

The next sections detail the characteristics of the resource mentioned in Table 1, namely, Wordnet, text and KB-based corpora and the three types of word embeddings.

4.1.1 Wordnets and NLTK toolkit

Wordnet (Miller, 1995) is an English lexical database organized by concept and meaning. Specifically, lexical forms of nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms called synsets, each expressing a language-independent concept. Furthermore, each synset is linked to several synsets via semantic relations such as hypernymy, hyponymy, meronymy, and antonymy, thus creating a semantic network.

The semantic structure of Wordnet gives us a robust framework for the present work. First, the nature of the semantic information it conveys around words and concepts allows for two of the aforementioned features to be assigned to each word of the dataset; SND and concreteness. Second, the semantic structure of the WordNet grants coding in two of the three types of embeddings employed in this work, namely wordnet and hybrid embeddings (see Section 4.1.3).

We have used the Open Multilingual Wordnet (OMW)[3] which is linked to the Princeton WordNet 3.0 (PWN).[4] OMW extends the former English Wordnet to 34 languages, including Basque and Spanish, and uses the PWN synset structure thus mapping the lexicalizations of all languages into the same semantic network. This extension method based on English Wordnet's structure is called *expand-approach*.[5] We have used Python NLTK toolkit[6] to extract concreteness and SND features from Wordnet synsets (see Section 4.2.1), for both Basque and Spanish.

Alongside the KB Wordnet, we have used language corpora as a complementary semantic information resource, which is introduced in the next section.

4.1.2 Corpora

In the following section, we describe the two types of corpora used during the construction of the dataset; the text corpora and the wordnet-based KB pseudo-corpora.

4.1.2.1 Text corpora

Text corpora are collections of vast amounts of texts, being usually annotated (e.g., British National Corpus)[7] or even syntactically parsed (e.g., Penn TreeBank Marcus et al., 1993). Plain text corpora with no structure were used to create this dataset. The use of plain text has allowed the greater use of text corpora in the computation of the relationships between words and their contexts via word occurrence. Thus, from the corpora, we directly calculated the similarity between words by using word embeddings. In the case of Basque, it was also been used for extracting word frequencies (see Section 4.2.1).

In Spanish, we have employed the readily available 2018 Wikipedia dump text corpora,[8] extracting the text from the dump using a script.[9] Since the size of Wikipedia is insufficient for building a large corpus in low-resourced languages such as Basque, web crawling was used to complement the corpora in this language, since it is an effective method for collecting texts to compensate for this deficiency (Leturia, 2012). Thus for Basque, we have employed the publicly available Euscrawl corpus[10] (Artetxe et al., 2022).

Both Basque and Spanish corpora were pre-processed with the standard procedure, that is, lowercase and tokenization. Token sparsity in Basque, as an agglutinative language, was avoided with an in-house stemmer. The final Spanish and Basque corpus comprises 608 and 288 million tokens, respectively.

4.1.2.2 Knowledge-based corpora

This kind of corpora is not as known as text corpora, but it has recently gained some attention in the NLP field (Perozzi et al., 2014; Xu et al., 2018). The knowledge-based corpora latently contain the semantic structure of a KB (in our case Basque or Spanish wordnets), and we have dubbed it as pseudo-corpora in the present work because it is not human-understandable. This pseudo-corpora represents concepts and lexicalizations of knowledge bases in a much more compact format than traditional methods. As explained in the following section, the pseudo-corpora were processed by a neural network model to encode the semantic structure of Basque or Spanish wordnets in a continuous vector space.

To do this, we applied the monolingual method for English introduced by Goikoetxea et al. (2015), but in the Basque and Spanish settings. This technique uses a Monte Carlo method to compute the PageRank algorithm (Avrachenkov et al., 2007). The algorithm considers the KB as an undirected graph comprised of concepts and links among concepts. It needs a dictionary which associates words with concepts, as well as a damping factor α that

3 https://compling.hss.ntu.edu.sg/omw/

4 https://wordnet.princeton.edu/

5 In this paper, we will refer to the original PWN and the one used in this work (Wordnet 3.0) as "Wordnet." In contrast, we will use "wordnet" for the rest which derive from PWM such as the Spanish and Basque ones.

6 https://www.nltk.org/howto/wordnet.html

7 https://www.english-corpora.org/bnc/

8 https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

9 `xml2textx` available on the same site. The content of tables and maths have been deleted.

10 https://ixa.ehu.eus/euscrawl/

TABLE 2 Wordnet sizes for English (EN), Spanish (ES), and Basque (EU).

	Lexicalizations	Synsets
EN	147,306	136,334
ES	53,039	55,814
EU	26,701	30,464

Number of lexicalizations and synsets in the middle and rightmost columns, respectively.

determines the continuity of the random walk and the maximum number of walks.

For creating every line of pseudo-corpus, the algorithm starts in a random concept and launches a random lexical form of the concept via the dictionary. Afterwards, it decides whether to jump to another concept[11] and to launch a random lexical form of the latter, or stop the walk and start over a new walk. Finally, if the number of walks reaches its maximum, the process ends. Note that the word is fed to a text file whenever the method launches a lexical form in the walk.

Each line of the following example shows a different walk of the monolingual algorithm in Wordnet 3.0, which is used by Goikoetxea et al. (2015). In this case, every walk has a different length, and each jump from concept to concept has launched a random lexicalization. It is worth mentioning that this pseudo-corpus is not human-readable, but every walk gathers semantically related words, so that implicitly it contains Wordnet's structure. The following example shows five random walks from an English Wordnet pseudo-corpus:

```
storyteller liar beguiler grifter dissimulation
revitalize strength delicate ethereal
paved patio terrace house living_room home
swimming dive
backlog fire re-afforest forest woods rainforest
```

As mentioned before, the monolingual version of this method is adapted to the Basque and Spanish setup, using the dictionaries of both languages from OMW; hence, aligned with Wordnet 3.0 semantic structure. This means that the former English lexical forms have been translated to the target language, but the semantic structure remains intact. While maintaining the original semantic structure when translating WordNet's English lexicon into Spanish and Basque facilitates consistency, it introduces biases that may overlook language-specific semantic associations; ideally, custom-built WordNets for these languages would be more accurate, but we account for this bias in our analysis.

The resulting size of wordnet is not the same for every language, since *expand-approach* wordnets do not have the same number of lexicalizations as the former English Wordnet 3.0. As shown in Table 2, Basque and Spanish wordnets' sizes are much smaller than the original one.

The difference in size of the semantic structure and the number of lexical forms in Basque and Spanish directly impacts the dimension of the wordnet-based corpora in both languages. To prevent saturation and redundant information as described by Goikoetxea et al. (2015, 2018), the sizes of the pseudo-corpora in both languages (see Section 4.1.2) were limited. Even though the wordnet-based corpus is smaller in Basque, they are big enough to encode their respective wordnets' semantic structure (see Section 4.1.3).

English Wordnet 3.0 with glosses consists of 147,306 lexical forms, and Goikoetxea et al. (2016) reached peak performance in word similarity task with 200 million random walks, which created an 1100 million token corpus. In this work, we kept the same proportion between the number of lexical forms and random walks as Goikoetxea et al. (2016) in order to ensure good performance. Thus 72 million and 36.3 million random walks were performed for Spanish and Basque respectively, resulting in a corpus of 406 million tokens for the former and 166 million for the latter.

4.1.3 Embeddings

Three types of static embeddings have been computed: text embeddings (FT_{txt}), wordnet-based embeddings (FT_{kb}), and hybrid embedding (FT_{hyb}). These word representations encompass the two sources of semantic information (text and KB embeddings) related to word similarity.

In order to compute the static embeddings, a neural network processes the whole text corpus by traversing the corpus word by word, computing the representations of words based on the Distributional Hypothesis. It calculates the representation of a given word based on the representations of all the neighbors found within a predefined window while traversing the corpus. Eventually, every token in the corpus ends up with a vector representation called embedding.

In this work, the neural-based model `fastText` (Bojanowski et al., 2017) was used because of its more robust performance comparing to the also non-contextual models `word2vec` (Mikolov et al., 2013) or `Glove` (Pennington et al., 2014). `fastText` implements a variant of the Distributional Hypothesis which, instead of exploiting the neighboring words' information to compute a representation of a given word as `word2vec` and `Glove`, exploits subword information, which enhances performance in handling rare words and morphological variations. Both text-based and wordnet-based embeddings were calculated using `fastText`.

In the case of Basque and Spanish text-based embeddings, we fed `fastText` with their respective text corpora separately. Regarding wordnet-based ones, we encoded the semantic structure of Basque and Spanish wordnets in a vector space following the method proposed by Goikoetxea et al. (2015), which comprises two steps. First, the creation of Basque and Spanish Wordnet pseudo-corpus, as explained in the previous section. Second, processing of Basque and Spanish pseudo-corpora with a neural-based model to obtain their respective wordnet embeddings. In the original proposal, Goikoetxea et al. (2015) used `word2vec` (Mikolov et al., 2013), but as mentioned before, in the present work `fastText` has been used with the same parameters as in Mikolov et al. (2018).

Recent works show that embeddings which combine semantic information from both text and KB (i.e., hybrid embeddings) have

11 This will depend on the parameter α, which is set to 0.85 as indicated by Goikoetxea et al. (2015).

TABLE 3 Embedding quantities in Basque (EU) and Spanish (ES) for three types of embeddings, namely, text (FT_{txt}), wordnet-based (FT_{wn}), and hybrid (FT_{hyb}) embeddings.

	EU	ES
FT_{txt}	472,166	931,101
FT_{wn}	26,135	52,347
FT_{hyb}	12,128	5,316

an overall higher performance in word similarity tasks. Lastra-Díaz et al. (2019) proved that hybrid embeddings outperformed most ontology-based measures and the rest of the word embedding models. Likewise, Goikoetxea et al. (2016) and García et al. (2020) proved that hybrid embeddings are more explicative of human perception of semantic distance than text or KB embeddings separately. In the present work, the method proposed by García et al. (2020) will be implemented in the computation of hybrid embeddings. In short, the latter's proposal consists of four steps:

1. To compute separate text and wordnet embeddings.
2. To map text embeddings space onto the wordnet one.
3. To estimate word embeddings for both spaces.
4. To combine equivalent text and wordnet embeddings.

In the creation of FT_{hyb} embeddings, García et al. (2020) were strictly followed, employing `vecmap` (Artetxe et al., 2018) for the mapping of text and wordnet embeddings.

Table 3 shows the embedding sizes for each language.

As expected, the text and wordnet-based Spanish embedding spaces are larger than those of Basque. However, for the hybrid embeddings, Basque's space is twice the size of Spanish's. This phenomenon is due to the small overlap between text and wordnet tokens in Spanish; many multiword expressions in Spanish wordnet do not have equivalents in text embeddings, making it impossible to create corresponding hybrid embeddings. In contrast, Basque shows a broader overlap between wordnet and text embeddings, as its WordNet contains fewer multiword expressions.

4.2 Data analysis

This work has sought to create a dataset of noun pairs, matched by length, that compiles information about the semantic similarity between them. As already stated, the semantic similarity was calculated using three types of embeddings; text, wordnet and hybrid embeddings. The linguistic features controlled in pairing the nouns were concreteness, word frequency, and semantic and phonological neighbor density.

Each feature was clustered via KNN classification (Cover and Hart, 1967) to match pairs of words. That is, each pair of nouns was set whenever there was a coincidence in the four feature clusters. Despite maximizing the number of potential noun pairs, we conducted a two-group clustering (low and high values) for concreteness, word frequency, semantic and phonological neighbor density (see Section 5.1).

Each line of the final dataset comprises the values of features of the noun pairs, their corresponding clusters and the three similarity measurements. Figure 1 shows the stages of the dataset creation procedure for both languages, which are described in the following points:

- KB-based corpus generation: word embeddings are computed feeding `fastText` with the aforementioned text and KB-based corpora.
- Embeddings generation: the final dataset will contain three types of word pair similarity measurements, computed out of their corresponding type of embeddings, namely, text-based (FT_{txt}), knowledge-based (FT_{wn}), and hybrid (FT_{hyb}) word representations or embeddings.
- Feature dictionaries generation: as mentioned previously, four extra-linguistic features are added to the similarity dataset. For each feature, a dictionary of nouns along with their feature values or values is generated, namely, word frequency dictionary ($DICT_{frq}$), phonological neighborhood density dictionary ($DICT_{pnd}$), semantic neighborhood density dictionary ($DICT_{snd}$), and concreteness dictionary ($DICT_{cnc}$).
- Dataset creation: the datasets in this study are comprised of noun pairs, along with their three types of word similarity measurements and the linguistic features associated with each noun in the pair. Thus the final step consists of finding all possible noun pairs that match two specific conditions (see Section 4.2.2), measuring their similarity values and adding a linguistic feature value for each noun in the pair.

The following sections seek to describe the details of the four stages mentioned in the above paragraphs to create the final dataset.

4.2.1 Computation of linguistic features

This section describes the details of the calculation of each linguistic feature before being L2-normalized and clustered in the final dataset.

4.2.1.1 Concreteness

Concreteness measurements were calculated automatically by exploiting Wordnet's taxonomy in Basque and Spanish. In order to do so, an algorithm proposed by Feng et al. (2011) was followed. Said algorithm predicts word concreteness via various lexical resources and features, namely, human ratings, lexical types, latent semantic analysis dimensions, word frequency and length, and hypernymy level. For the purposes of this work, of all these features, only the *Hypernymy Level* was considered for computing concreteness. The reason for this is that, with the exception of ratings and lexical type, which are not included in this research, the remaining features will be analyzed and treated separately here later.[12] The *Hypernymy Level* aspect scores the concreteness of a word following its hypernymy relations in Wordnet's hierarchical structure. For every noun in Basque and Spanish, the aforementioned method does the following in the corresponding wordnet:

12 Although latent semantic analysis is not used for word representations in the present word, more advanced techniques of static word embeddings are included (see Section 4.1.3).

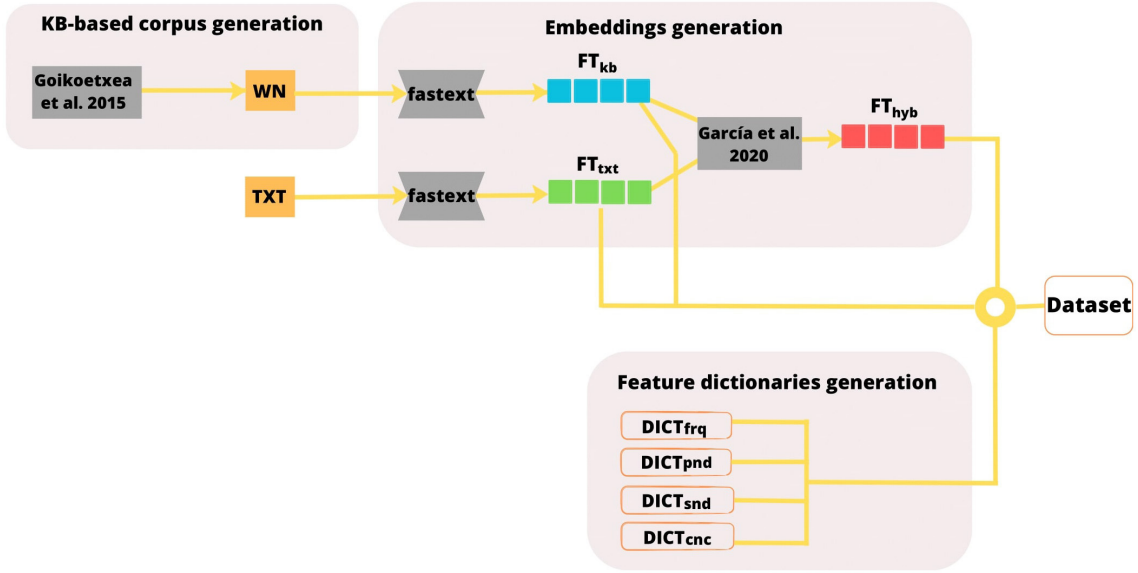

FIGURE 1
Overview of the dataset creation process. The figure outlines three main stages: generation of feature dictionaries, construction of a knowledge-based (KB-based) corpus followed by embedding generation, and the creation of the final dataset. First, four feature dictionaries are created: word frequency ($DICT_{frq}$), phonological neighborhood ($DICT_{pnd}$), semantic neighborhood density ($DICT_{snd}$), and concreteness ($DICT_{cnc}$). Next, the KB-based corpus is derived from WordNet (WN) as described by Goikoetxea et al. (2015). Three types of embeddings are then computed: text-based (FT_{txt}), KB-based (FT_{kb}), and hybrid (FT_{hyb}), following García et al. (2020). Finally, the feature dictionaries and embeddings are combined to generate the final dataset, which includes word similarity measurements and lexical property features. The same procedure is followed in Basque and Spanish languages.

- Check all of its synsets within a given noun.
- For each synset, the method counts every hypernym from the source synset until the topmost synset (i.e., Wordnet's root node *entity*), thus scoring the depth of the source synset in Wordnet's tree structure.
- Compute the concreteness of a given noun by averaging all the depths of its synsets.

In order to count the hypernyms in the Wordnet structure up to the *entity* node for each of the synsets in a noun, we employed synset-based semantic relation in the NLTK toolkit.

Note that the higher the depth, the more concrete the word is, and the lower the depth, the more abstract it will be. For example, the word *car* has five synsets in Wordnet 3.0. If we choose the most common synset for that noun,[13] its hypernymy path is the following:

car → *MotorVehicle* → *SelfPropelledVehicle*
→ *WheeledVehicle* → *Container* → *Instrumentality*
→ *Artifact* → *Whole* → *Object* → *PhysicalEntity* → **Entity**.

Thus, the depth of the path of the above synset of *car* is 10. The word *Wheeled Vehicle* has a depth of 7, meaning that it is less concrete than *car* and *ambulance* a depth of 11, therefore being more concrete than *car*. In order to obtain the final score of the noun, the method computes all the depths of the rest of the synsets to find the average of all of them.

4.2.1.2 Semantic neighborhood density

The framework proposed in this work for computing SND is also based on the semantic structure of Wordnet. The semantic neighbors of every noun are computed by counting all the semantic relations of its synsets. A procedure similar to the one described in Section 4.2.1 was used for every noun in Basque and Spanish:

- Check all of its synsets within a given noun.
- For each synset, count all its surrounding synsets linked by a Wordnet semantic relation. This approach departs from the source synset and checks all its semantic relations to find its neighboring synsets.
- Finally, compute the semantic neighbors of a noun averaging all the counts in its synsets.

Only first-degree semantic neighbors have been considered. Regarding the Wordnet semantic relations, we have taken every available synset-based semantic relation in the NLTK toolkit and discarded the more surface-level lexical relations like derivation and pertainymy.

Following the same example as in the previous section, the most common synset for the noun *car* has 31 hyponyms, 29 meronym parts and one hypernym. That is, 61 first-degree semantic neighbors.

13 The sense with the following definition in Wordnet: *a motor vehicle with four wheels; usually propelled by an internal combustion engine.*

4.2.1.3 Phonological neighborhood density

Phonological neighbors are pairs of words that differ in only one phonological segment, such as *cat* and *bat*. For the purpose of this work, phonological neighborhood size was calculated using the Levenstein distance method (Levenshtein, 1965). It has several advantages compared to other alternatives, such as the Hamming Distance or the Jaro-Winkler distance; it accounts for both substitution and insertion/deletion operations, so the distance is more accurately computed, independent of the length of the strings.

This work assumes that two words are phonological neighbors if their Levenstein distance equal to one. For example, in the case of the word *car*, the number of phonological neighbors that meet the condition mentioned above is 31, with first-degree neighboring words such as *cat*, *card*, *scar*, *ear*, *jar*, and *tar*. Multiword expressions and nouns with less than three characters in length have been excluded from the dataset.

4.2.1.4 Word frequency

Word frequency measures have been calculated using Zipf frequencies; a base-10 logarithm of its occurrences per billion words. Zipf frequency allows for a more comprehensive analysis than raw noun counts because it accounts for the relative importance of a word in the corpus.

For Spanish, Python's wordfreq library was used to obtain Zipf frequencies because it provides reliable word frequency data derived from large, diverse corpora, including web texts, Wikipedia, and subtitles. Given the nature of the 10-base logarithm scale, Zipf values in this library are within the range of 0-8. However, it is important to note that Basque is not included in this library, thus in the case of Basque, Zipf frequencies were computed from the most extensive Basque public corpus, Euscrawl (Artetxe et al., 2022). Euscrawl corpus was pre-processed as described in Section 4.1.2, and raw word counts were calculated using the `fastext` model from the gemsim library in Python. Afterwards, those counts were converted into Zipf frequencies, limiting their maximum and minimum values between 0 and 8.

4.2.2 Creation of feature dictionaries

As mentioned in previous sections, each noun is accompanied by four features in the dataset. An independent dictionary was created for each feature (i.e., concreteness, frequency, SND, and PND) in Basque and Spanish, so as to assign the measurements of its feature along with its cluster number to each noun. Once the feature dictionaries were computed, the next step was to build the final dataset (see Section 4.2.4).

The lists of single-word nouns from Spanish and Basque wordnets were extracted via NLTK toolkit, constraining the candidates to 28,647 and 22,877 in each language, respectively.

Although the Spanish wordnet is nearly double the Basque one in size (see Table 2), the imbalance is less apparent in the size of the feature dictionaries because the number of available nouns in the NLTK toolkit is comparable across these two languages. To improve the KNN classification of the nouns across each feature, outliers were removed using the interquartile range, and L2-normalization has been applied to every raw measurement. Although the latter strategy diminishes the size of the datasets, we opted to apply it in order to reduce noisy samples and balance cluster sizes (see Section 5.1).

4.2.3 Word similarity measurement

In NLP, word similarity is commonly calculated by the cosine similarity of the angle between two-word embeddings, measuring the similarity of the words they represent. The cosine similarity is determined by computing the dot product of the two vectors and dividing it by the product of the Euclidean norms of the vectors. The cosine similarity is a measure ranging from 0 to 1, where 0 means the complete absence of similarity and 1 means complete similarity (i.e., synonyms). The similarity measurement is independent of the origins of the embeddings, in the way that text, wordnet and hybrid representations are processed in the same way.

In this paper, the term similarity has been used indistinctly, but the difference between pure similarity and relatedness has been widely recognized in cognitive sciences for a long time (Tversky, 1977). Pure similarity measures the degree to which two concepts share semantic features, while relatedness is the degree of association between two words. Regarding the semantic relations involved, pure similarity includes synonymy and hyponymy/hyperonymy. In contrast, relatedness encompasses the previous ones and a wider variety of relations, such as meronymy, functional associations and other unusual relations. For example, *wolf* and *dog* are taxonomically linked by hypernymy relations in the same semantic structure and share many features; thus, they have high similarity. In contrast, *wolf* and *moon* do not share any semantic features (low similarity) but are related by association; hence, they have a high relatedness.

Cosine similarity measurement does not distinguish between these two aspects but can be applied to different types of embeddings to measure distinctive semantic relations. In this work, we have chosen three types of embeddings, which perform differently in pure semantic similarity (Hill et al., 2015; Agirre et al., 2009; Rubenstein and Goodenough, 1965) and relatedness (Finkelstein et al., 2001; Bruni et al., 2014) measures. Broadly speaking, wordnet embeddings measure pure similarity relations more accurately, while text embeddings are more sensitive to compute relatedness, and hybrid embeddings have been proven to be more robust for capturing semantic relations in general (Goikoetxea et al., 2016, 2018; García et al., 2020). Results in Section 5.3 illustrate the distribution of cosine measurements for both languages across five different ranges, revealing a clear tendency toward lower values for all types of embeddings.

4.2.4 Creation of word pair matrix

The final step in this work consisted of creating the noun pair dataset out of the feature dictionaries described previously. There are two conditions for noun pairing:

- First, the two nouns composing the pair must figure in the three types of embeddings.

TABLE 4 Concreteness (CNC), frequency (FRQ), phonological neighborhood density (PND) and semantic neighborhood density (SND) dictionaries for Basque (EU) and Spanish (ES).

		EU	ES
CNC	size	19,660	14,771
	avg	8.41	8.38
	var	3.44	2.99
FRQ	size	14,380	12,146
	avg	6.39	2.95
	var	1.05	1.27
SND	size	18,044	15,534
	avg	2.47	2.18
	var	3.05	2.25
PND	size	14,671	14,608
	avg	1.75	1.38
	var	1.25	0.43

The upper line in each feature shows the size of a specific dictionary, the medium one shows the average (*avg*) value of each measurement and the last one the variance (*var*).

- Second, the two nouns composing the pair must have the same length and share the same cluster in all four linguistic features.

All the nouns that met the first condition were traversed to find all possible pairs of nouns that also share the four linguistic features, as stated in the second condition. Every time a pair was set, featural values in concreteness, frequency, SND, and PND, as well as their cluster identification,[14] were inserted in the dataset, along with the three types of similarity measurements. Altogether, each noun pair is followed by 19 columns which include the three types of similarity measurements, followed by both nouns' four features' cluster numbers and measurements.

5 Results

The present section aims to summarize the main results obtained during the whole dataset creation process, namely, feature dictionaries' sizes, embedding evaluation, similarity values' distribution and size of final similarity datasets.

5.1 Feature dictionaries' sizes

In this section, we show the dictionary sizes for each of the four features described in Section 4.2.1, along with the average value for each feature, in both languages. Table 4 shows the resulting feature dictionaries' size and the mean value for each feature.

Table 4 shows that the Basque feature dictionaries are slightly larger than the ones in Spanish, due to the higher amount of outliers in the Spanish dataset. Both languages exhibit similar average concreteness values, with Basque showing slightly higher variance. Basque nouns have a much higher average word frequency with lower variance, indicating more consistent and higher frequency usage. For semantic SND, Basque nouns show higher average values but also greater variability compared to Spanish. In terms of PND, Basque has a higher average and significantly larger variance, reflecting a broader range of phonological similarities. After conducting the two-sample t-test, it indicates a statistically significant difference in PND between Basque and Spanish, with a t-value of $t_{(8,212)} = 16.68$ and p-value of $p < 0.001$.

Special mention goes for frequency. On average, Basque nouns are more frequent than the Spanish ones.[15] A reasonable explanation for this difference may be the size of the corpora used to calculate word frequencies in each language. The Scrawl corpus is smaller than the one used for computing Spanish Zipf frequencies in the NLTK toolkit, and it has far fewer infrequent words. This may bias the Zipf frequency in favor of a higher average value in Basque.

Table 5 shows the token distribution of all previous dictionaries in KNN-based clusters. Features are clustered in two groups, differentiating between high and low values. The size of the dictionary of phonological neighbors features is far smaller than the rest. This is because nouns with 0 neighbors have been excluded from the KNN classification (14,671 in Basque and 12,085 in Spanish) because they tend to unbalance the cluster distribution, leaving the cluster with high-PND value with almost no content. Therefore, nouns with 0 neighbors have been extracted into a subgroup and treated as a separate cluster when creating the final dataset.

Overall, Table 5 reveals that the cluster distributions for PND and SND features in both languages tend to be more uneven compared to concreteness and frequency. Regarding concreteness, both languages exhibit a similar distribution, predominantly favoring lower values. There is a slight disparity in frequency measurements between Basque and Spanish: Basque shows a higher number of high-valued noun pairs, whereas Spanish exhibits nearly equal amount of pairs in both clusters. In terms of SND and PND distributions, while both languages maintain similar cluster ratios, they show opposite trends: Basque has more high-valued pairs compared to low-valued ones, whereas Spanish demonstrates the opposite pattern. Regarding the mean values and variances observed for each language, Basque shows higher mean values in all features, while Spanish seems more stable in terms of variance.

5.2 Evaluation of the quality of the embeddings in similarity task

As mentioned in Section 4.2.3, in order to verify the quality of the three types of embeddings, we tested them in a word similarity task, including pure similarity and also relatedness datasets. In Spanish, we operated with the pure similarity datasets RG65 (RG)

14 We used −1 number for 0 neighbor subgroup in phonological neighbor matching, as that group was not part of a KNN cluster. For the rest of the linguistic features, clustering was marked with 0 and 1 to indicate low or high-value clustering.

15 Note that the mean Zipf frequency value of the Basque words is 6, meaning that a word appears once per a thousand words, whereas the mean Zipf value in Spanish is 2, indicating that a word appears once per million words.

TABLE 5 Concreteness (CNC), frequency (FRQ), phonological neighborhood density (PND) and semantic neighborhood density (SND) high-valued and low-valued cluster data for Basque (EU) and Spanish (ES).

		EU		ES	
		Low	High	Low	High
CNC	size	11,373	8,287	9,692	6,179
	avg	7.13	10.16	7.27	10.09
	var	0.93	1.59	1.34	0.96
FRQ	size	5,646	8,734	6,103	6,043
	avg	5.33	7.08	2.01	3.91
	var	0.41	0.25	0.31	0.43
SEM	size	13,915	4,129	4,713	9,836
	avg	1.65	5.2	1.31	4.0
	var	0.55	1.77	0.25	1.51
PND	size	3,791	974	1,017	2,432
	avg	1.24	3.7	1	2.32
	var	0.19	0.62	0	0.22

The upper line in each feature shows the size of a specific dictionary, the medium one shows the average (*avg*) value of each measurement and the last one the variance (*var*).

TABLE 6 Spearman correlation results in word similarity task for RG, Wordsim353 (WS) and SimLex999 (SL) datasets in Basque (EU) and Spanish (ES).

		Dataset		
		RG	WS	SL
EU	Baseline	0.7705	0.657	—
	FT_{txt}	0.7786	0.7331	—
	FT_{kb}	0.8567	0.6588	—
	FT_{hyb}	**0.8655**	**0.7457**	—
ES	Baseline	**0.879**	0.578	0.3658
	FT_{txt}	0.8657	0.5728	0.287
	FT_{kb}	0.7284	0.5732	0.3993
	FT_{hyb}	0.8725	**0.6345**	**0.4057**

Text (FT_{txt}), wordnet-based (FT_{kb}) and hybrid embedding (FT_{hyb}) representations are compared to their baselines. The best results for each dataset and language are expressed in bold.

(Camacho Collados et al., 2015) and SimLex999 (SL) (Etcheverry and Wonsever, 2016), and the relatedness dataset Wordsim353 (WS) (Hassan and Mihalcea, 2009). In the case of Basque, we used the pure similarity RG dataset and the relatedness dataset WS created by Goikoetxea et al. (2018).

In both Spanish and Basque, FT_{txt}, FT_{kb} and FT_{hyb} representations were compared with the baseline, using publicly available text-based representations `fastText`.[16] All of our embeddings were computed using the same set of parameters as those described on the website from which the baseline fastText embeddings were obtained.

Spearman correlation is a widely used statistical method in word similarity tasks within NLP, as it assesses the strength and direction of the monotonic relationship between two variables; in this case, our embeddings and the human similarity judgments. The Spearman correlation values between the word similarity scores obtained from the embeddings and the human-annotated values in the gold standard datasets is presented in Table 6.

The discussion below describes the findings extracted from Table 6:

- FT_{txt} embeddings: FT_{txt} results perform similarly to the baseline ones in both languages, with two exceptions. One, the Basque FT_{txt} result in the WS dataset is higher than the baseline, likely due to the greater corpus size Euscrawl. Second, the Spanish FT_{txt} in the SL dataset is lower than the baseline. The only plausible explanation for this result in the Spanish SL may lie in the differences in the pre-processing of the corpus; the baseline text corpus was tokenized with Europarl pre-processing tools,[17] whereas the NLTK tokenizer was used in the present work.
- FT_{kb} embeddings: this type of embeddings being more suited to pure similarity datasets (RG and SL) than relatedness ones (WS), FT_{kb} results are higher than the baseline in the Basque RG and the Spanish SL, but not in the Spanish RG. This underperformance in Spanish must be interpreted with caution, as the RG dataset is small (64 pairs) and it has low statistical power.

The most noticeable result, though, is in the relatedness WS dataset, in which FT_{kb} performs at baseline. Note that the baseline txt-based embeddings are supposed to have a better performance in WS due to their superior capacity for capturing relatedness relations, so we expect them to obtain better results than the FT_{kb} measurement. However, the incorporation of gloss relations (which are relatedness relations) when creating the wordnet-based corpora may have enhanced the capability of FT_{kb} embeddings to measure relatedness.

- FT_{hyb} embeddings: comparing the results of FT_{txt}, FT_{kb}, and FT_{hyb} embeddings, we find that the combination of the first two into the latter enhances their performance in similarity tasks for both languages across all datasets. FT_{hyb} also outperforms the baseline in all datasets, with only exception of RG in Spanish, likely due to its small size, as mentioned above.

In general, the three types of embeddings in this work have performed as expected with FT_{hyb} embeddings offering the best overall results. The quality of the embeddings in the Basque language has proven to be the best to date. Note that the excellence of the embeddings is critical in this work, as they are used to create the three semantic similarity measurements that constitute the core features of the dataset. Overall, all embeddings created in this work were deemed suitable for inclusion in the final dataset.

16 https://fasttext.cc/docs/en/crawl-vectors.html

17 https://www.statmt.org/europarl/

TABLE 7 Percentages of word similarity values in the Basque and Spanish dataset across five ranges and three types of embeddings; text (FT_{txt}), wordnet (FT_{kb}), and hybrid (FT_{hyb}).

	EU			ES		
	FT_{hyb}	FT_{kb}	FT_{txt}	FT_{hyb}	FT_{kb}	FT_{txt}
0.0–0.2	89.1	98.44	84.9	93.52	98.68	89.13
0.2–0.4	10.3	1.48	13.97	6.09	1.24	10.12
0.4–0.6	0.54	0.066	1.045	0.36	0.06	0.71
0.6–0.8	0.025	0.009	0,048	0.015	0.007	0.03
0.8–1.0	0.00135	0.0015	0.0018	0.001	0.0013	0.0014

5.3 Similarity measurement distribution

We also analyzed the similarity measurement values' distribution across the tree types of embeddings. Table 7 shows the percentages of semantic similarity values for all possible noun pairs classified along five different ranks. Only nouns with the three types of embeddings that match all four linguistic features are considered. The percentages presented in Table 7 show a bias toward lower-ranked similarity values for all types of embeddings in both Basque and Spanish. This is even more evident in the wordnet embeddings.

This phenomenon was already pointed out in a cross-lingual setting by Lample and Conneau (2019). The authors observed that `fasText` based embeddings' cosine similarity mean value was considerably lower than that achieved by XLM (Lample and Conneau, 2019), a large cross-lingual language model. The authors suggested that the distinctive magnitude of the mean cosine similarity based on `fasText` and XLM relies on the vector space of the language model, which is more compact in XLM due to its training in a sentence encoder. This phenomenon only shows that the non-contextual embeddings like `fasText` or `word2vec` are sparser than the language model ones. As the results in Table 6 demonstrate, the different organization of the lexicon in the vector spaces does not affect the performance of the embeddings in the word similarity task.

5.4 Final datasets' sizes

Finally, we indicated the number of nouns which fulfill the conditions defined in Section 4.2.4, that is, the size of the final similarity datasets for Basque and Spanish. The first condition defined in Section 4.2.4 filters the pairs of nouns without the three types of embeddings. The Spanish wordnet contains a high number of multiword expressions, and the multiword expressions are discharged from the dataset (see Section 4.2.1). The suppression of the latter leads to a lower overlap of text and wordnet embeddings and, therefore, a smaller amount of hybrid embeddings (see Table 3). The latter phenomenon limits severely the potential noun pairs for Spanish, as both nouns have to have all three types of embeddings. As a result, we obtained a Spanish dataset of 136,886 noun pairs and a Basque one with 996,514. Thus, the substantially higher number of pairs in Basque is only attributed to the different amounts of available nouns for each language, as Basque and English clusters seem to be similarly distributed thus not influencing the matching for creating noun pairs.

6 Conclusions

Psycholinguistic evidence supports the idea that the overlap of semantic features across concepts is crucial in the computation of semantic relations, and by extension, in semantic processing. A comprehensive work on NLP has shown that embeddings or vectors are a sensitive method for artificially computing similarities across concepts by accounting for a large number of semantic properties, which may be applied to different types of linguistic resources to uncover distinctive semantic relations. This work has aimed to bridge computational linguistics and psycholinguistics to automatically build a dataset of vectorised word similarity measures in Basque and Spanish.

We have presented a computationally grounded dataset that encompasses different aspects of the semantic information present in both text corpora and knowledge bases. On the one hand, each dataset includes three similarity measurements based on their corresponding embedding computations, namely, text-based, wordnet-based and hybrid embeddings. These measurements encode different subtleties of meaning. Text-based embeddings are computed out of word co-occurrence in large natural language corpora; being prone to better measure relatedness relations. Wordnet-based embeddings encode the semantic structure of Wordnet, so that they are considered a measurement of pure similarity relations. Finally, hybrid embeddings combine both text-based and wordnet-based embeddings, binding categorical and associative relations. In addition, all the materials were controlled for several linguistic features (concreteness, frequency, and semantic and phonemic neighbor size) to adjust the dataset to various research interests and requirements.

It is important to acknowledge the potential influence of the specific characteristics of the corpora used in this study. Wikipedia, used for Spanish, tends to represent a more formal and encyclopedic register, which may not fully capture everyday language use, informal expressions, or dialectal variations. In contrast, EusCrawl, used for Basque, includes a wider variety of web content, providing greater domain diversity and a mix of formal and informal registers. While our focus was on producing a computationally grounded dataset based on widely available and robust corpora, alternative sources such as social media or forums might result in different distributions of word pair similarities, especially in terms of capturing more colloquial or diverse language use. Future work could explore these sources to further evaluate their impact on word similarity measures.

Orthographic neighborhoods enhance visual lexical processing by reflecting how words are recognized and retrieved in written contexts, while phonological neighborhoods support auditory recognition and retrieval during spoken language processing. In our study, we used orthographic neighborhood density data as a proxy for phonological neighborhood density, which is appropriate in transparent languages like Basque and Spanish. This approach is effective due to the close correspondence between phonemes and graphemes in these languages, allowing for the analysis of sound

relationships based on the number of orthographic neighbors a specific lexical item possesses. However, this method may not apply to opaque languages, where the mapping between phonemes and graphemes is less direct, and thus the relationships between orthographic and phonological neighborhoods may not be as tightly linked. In such cases, orthographic neighborhoods alone may not accurately reflect phonological relationships, complicating lexical processing and recognition. Therefore, future research should consider employing a combination of both phonological and orthographic measures to gain a more comprehensive understanding of lexical processing across different languages. By integrating these approaches, researchers can effectively account for the interaction between auditory and visual representations in language processing, thereby increasing the relevance of findings across a wider array of linguistic contexts.

A crucial aspect of our dataset is the distinction between semantic similarity and relatedness measures, which has significant implications for psycholinguistic research. This database allows researchers to carefully control semantic similarity and relatedness between concepts, either separately or together. It facilitates studies on the semantic organization in both taxonomic and associative relations through semantic relation judgment tasks, as well as sentence prediction experiments, where the relationships between prime and target words can be precisely controlled. Additionally, it supports vocabulary learning and generalization efforts based on shared semantic features across lexical terms.

The potential applications of this dataset are extensive, spanning multiple areas within psycholinguistics and beyond. It is valuable for researching semantic processing across various contexts throughout the lifespan, addressing key questions related to language acquisition, semantic memory retrieval, and organization in aging populations. Furthermore, it contributes to bilingualism studies by providing insights into how semantic relations are processed in diverse linguistic contexts and supports cross-linguistic comparisons to explore how different languages encode semantic information under specific circumstances. Overall, this versatile resource enables thorough control of important variables while quantifying the degree of semantic overlap, enhancing our understanding of cognitive processes.

The presented dataset can be applied to several NLP tasks. It could be relevant for word sense disambiguation, where lexical features like concreteness and frequency can help distinguish between multiple word meanings (Navigli, 2009) and also in machine translation where the dataset can aid in refining meaning alignment across languages, particularly for the included Basque and Spanish word pairs (Vaswani, 2017). Additionally, sentiment analysis can utilize the dataset to capture subtle differences in word meaning, which is critical for modeling context-dependent emotional content (Poria et al., 2020).

In relation to the latter task, we also intend to extend the features of the dataset with the two main emotional dimensions of words, namely, valence and arousal (Citron et al., 2013). Recent works (Buades-Sitjar et al., 2021; Planchuelo et al., 2022) have shown that the strength of word associations is correlated with both valence and arousal dimensions, indicating that emotionally charged words tend to be more related. Further, large language models (LLMs) have been used to capture semantic relationships and emotional dimensions of language by encoding rich contextual information (Devlin et al., 2018; Radford et al., 2019). While our dataset focuses on psycholinguistic features, it could complement LLMs in tasks requiring disentanglement of context, sentiment, and word similarity. These psycholinguistic measures could contribute to fine-tuning LLMs for sentiment analysis, providing a structured understanding of noun processing and improving context-rich semantic evaluations.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://data.mendeley.com/datasets/6xr2rp8gvh/4.

Author contributions

JG: Writing – original draft, Writing – review & editing, Data curation, Methodology, Resources. IS: Writing – review & editing. MA: Funding acquisition, Methodology, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Leonardo grant of the BBVA Foundation for Researchers and Cultural Creators 2021 (FP2157) (JG, MA, and IS), Consolidated Group funding from the Basque Government (IT1439/22-GIC21/132) (MA and IS), and the grant RYC2021-033222-I funded by the Ministry of Science and Innovation/State Research Agency/10.13039/501100011033 and by the European Union NextGenerationEU/Recovery, Transformation and Resilience Plan of Spain (MA).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abdel Rahman, R., and Melinger, A. (2007). When bees hamper the production of honey: lexical interference from associates in speech production. *J. Exper. Psychol.* 33:604. doi: 10.1037/0278-7393.33.3.604

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). "A study on similarity and relatedness using distributional and WordNet-based approaches," in *Proceedings of HLT-NAACL*, 19–27. doi: 10.3115/1620754.1620758

Alnafesah, G., Madabushi, H. T., and Lee, M. (2020). "Augmenting neural metaphor detection with concreteness," in *Proceedings of the Second Workshop on Figurative Language Processing*, 204–210. doi: 10.18653/v1/2020.figlang-1.28

Artetxe, M., Aldabe, I., Agerri, R., Perez-de Vi naspre, O., and Soroa, A. (2022). Does corpus quality really matter for low-resource languages? *arXiv preprint arXiv:2203.08111*.

Artetxe, M., Labaka, G., and Agirre, E. (2018). "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 789–798. doi: 10.18653/v1/P18-1073

Auguste, J., Rey, A., and Favre, B. (2017). "Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks," in *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, 21–26. doi: 10.18653/v1/W17-5304

Avrachenkov, K., Litvak, N., Nemirovsky, D., and Osipova, N. (2007). Monte carlo methods in pagerank computation: when one iteration is sufficient. *SIAM J. Numer. Anal.* 45, 890–904. doi: 10.1137/050643799

Baayen, R. (1995). *The Celex Lexical Database (release 2)*. Linguistic Data Consortium, University of Pennsylvania.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). "The berkeley framenet project," in *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. doi: 10.3115/980451.980860

Balota, D. A., and Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *J. Exper. Psychol.* 10:340. doi: 10.1037//0096-1523.10.3.340

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition of single-syllable words. *J. Exper. Psychol.* 133:283. doi: 10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., et al. (2007). The english lexicon project. *Behav. Res. Methods* 39, 445–459. doi: 10.3758/BF03193014

Barber, H. A., Otten, L. J., Kousta, S.-T., and Vigliocco, G. (2013). Concreteness in word processing: Erp and behavioral effects in a lexical decision task. *Brain Lang.* 125, 47–53. doi: 10.1016/j.bandl.2013.01.005

Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660. doi: 10.1017/S0140525X99002149

Benedek, M., Kenett, Y. N., Umdasch, K., Anaki, D., Faust, M., and Neubauer, A. C. (2017). How semantic memory structure and intelligence contribute to creative thought: a network science approach. *Think. Reason.* 23, 158–183. doi: 10.1080/13546783.2016.1278034

Boden, M. A. (2008). *Mind as Machine: A History of Cognitive Science*. Oxford: Oxford University Press.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5, 135–146. doi: 10.1162/tacl_a_00051

Bonin, P., Méot, A., and Bugaiska, A. (2018). Concreteness norms for 1,659 french words: relationships with other psycholinguistic variables and word recognition times. *Behav. Res. Methods* 50, 2366–2387. doi: 10.3758/s13428-018-1014-y

Broderick, M. P., Di Liberto, G. M., Anderson, A. J., Rofes, A., and Lalor, E. C. (2021). Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. *Sci. Rep.* 11:4963. doi: 10.1038/s41598-021-84597-9

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *JAIR* 49, 1–47. doi: 10.1613/jair.4135

Brysbaert, M., Mandera, P., and Keuleers, E. (2018). The word frequency effect in word processing: an updated review. *Curr. Dir. Psychol. Sci.* 27, 45–50. doi: 10.1177/0963721417727521

Brysbaert, M., and New, B. (2009). Moving beyond kučera and francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behav. Res. Methods* 41, 977–990. doi: 10.3758/BRM.41.4.977

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., and Storms, G. (2014a). Norms of age of acquisition and concreteness for 30,000 dutch words. *Acta Psychol.* 150, 80–84. doi: 10.1016/j.actpsy.2014.04.010

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014b). Concreteness ratings for 40 thousand generally known english word lemmas. *Behav. Res. Methods* 46, 904–911. doi: 10.3758/s13428-013-0403-5

Buades-Sitjar, F., Planchuelo Fernández, C., and Dunabeitia Landaburu, J. A. (2021). *Valence, arousal and concreteness mediate word association*. Psicothema.

Buchanan, L., Westbury, C., and Burgess, C. (2001). Characterizing semantic space: neighborhood effects in word recognition. *Psychon. Bull. Rev.* 8, 531–544. doi: 10.3758/BF03196189

Camacho Collados, J., Pilehvar, M. T., and Navigli, R. (2015). *A framework for the construction of monolingual and cross-lingual word similarity datasets*. Association for Computational Linguistics. doi: 10.3115/v1/P15-2001

Charbonnier, J., and Wartena, C. (2019). "Predicting word concreteness and imagery," in *Proceedings of the 13th International Conference on Computational Semantics-Long Papers* (Association for Computational Linguistics), 176–187. doi: 10.18653/v1/W19-0415

Chersoni, E., Santus, E., Huang, C.-R., Lenci, A., et al. (2021). Decoding word embeddings with brain-based semantic features. *Comput. Ling.* 47, 663–698. doi: 10.1162/coli_a_00412

Citron, F. M., Weekes, B. S., and Ferstl, E. C. (2013). Effects of valence and arousal on written word recognition: time course and erp correlates. *Neurosci. Lett.* 533, 90–95. doi: 10.1016/j.neulet.2012.10.054

Clark, K., and Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*. doi: 10.18653/v1/D16-1245

Collins, A., and Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychol. Rev.* 82, 407–428. doi: 10.1037//0033-295X.82.6.407

Coltheart, M., Davelaar, E., Jonasson, J., and Besner, D. (1977). Access to the internal lexicon. *Attent. Perfor.* 6, 535–555. doi: 10.4324/9781003309734-29

Cosgrove, A. L., Kenett, Y. N., Beaty, R. E., and Diaz, M. T. (2021). Quantifying flexibility in thought: the resiliency of semantic networks differs across the lifespan. *Cognition* 211:104631. doi: 10.1016/j.cognition.2021.104631

Ćoso, B., Guasch, M., Ferré, P., and Hinojosa, J. A. (2019). Affective and concreteness norms for 3,022 croatian words. *Quart. J. Exper. Psychol.* 72, 2302–2312. doi: 10.1177/1747021819834226

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Cuetos, F., Glez-Nosti, M., Barbon, A., and Brysbaert, M. (2011). Subtlex-esp: frecuencias de las palabras espanolas basadas en los subtitulos de las peliculas. *Psicológica* 32, 133–144.

Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cogn. Psychol.* 42, 317–367. doi: 10.1006/cogp.2001.0750

Dave, K., Lawrence, S., and Pennock, D. M. (2003). "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th International Conference on World Wide Web*, 519–528. doi: 10.1145/775152.775226

Dell, G., and Gordon, J. (2011). "Neighbors in the lexicon: Friends or foes?" in *Phonetics and phonology in Language Comprehension and Production*, eds. N. O. Schiller and A. S. Meyer (Berlin: Walter de Gruyter), 9–38. doi: 10.1515/9783110895094.9

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Djokic, V. G., Maillard, J., Bulat, L., and Shutova, E. (2020). Decoding brain activity associated with literal and metaphoric sentence comprehension using distributional semantic models. *Trans. Assoc. Comput. Ling.* 8, 231–246. doi: 10.1162/tacl_a_00307

Du nabeitia, J. A., Avilés, A., and Carreiras, M. (2008). Noa's ark: influence of the number of associates in visual word recognition. *Psychon. Bull. Rev.* 15, 1072–1077. doi: 10.3758/PBR.15.6.1072

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., and Carreiras, M. (2013). Espal: one-stop shopping for spanish word properties. *Behav. Res. Methods* 45, 1246–1258. doi: 10.3758/s13428-013-0326-1

Etcheverry, M., and Wonsever, D. (2016). "Spanish word vectors from wikipedia," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3681–3685.

Ettinger, A., and Linzen, T. (2016). "Evaluating vector space models using human semantic priming results," in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 72–77. doi: 10.18653/v1/W16-2513

Farhy, Y., and Veríssimo, J. (2019). Semantic effects in morphological priming: the case of hebrew stems. *Lang. Speech* 62, 737–750. doi: 10.1177/0023830918811863

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology* 44, 491–505. doi: 10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., and Kutas, M. (1999). A rose by any other name: long-term memory structure and sentence processing. *J. Mem. Lang.* 41, 469–495. doi: 10.1006/jmla.1999.2660

Feng, S., Cai, Z., Crossley, S., and McNamara, D. S. (2011). "Simulating human ratings on word concreteness," in *Twenty-Fourth International FLAIRS Conference*.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., et al. (2001). "Placing search in context: the concept revisited," in *Proceedings of the 10th international conference on World Wide Web*, 406–414. doi: 10.1145/371920.372094

Fischer, M. H., and Zwaan, R. A. (2008). Embodied language: a review of the role of the motor system in language comprehension. *Quart. J. Exper. Psychol.* 61, 82–850. doi: 10.1080/17470210701623605

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *J. Memory Lang.* 66, 789–806. doi: 10.1016/j.jml.2011.11.006

García, I., Agerri, R., and Rigau, G. (2020). A common semantic space for monolingual and cross-lingual meta-embeddings. *arXiv preprint arXiv:2001.06381*.

Goikoetxea, J., Agirre, E., and Soroa, A. (2016). Single or multiple? Combining word representations independently learned from text and wordnet," in *Thirtieth AAAI Conference on Artificial Intelligence. doi: 10.1609/aaai.v30i1.10321*

Goikoetxea, J., Soroa, A., and Agirre, E. (2018). Bilingual embeddings with random walks over multilingual wordnets. *Knowl.-Based Syst.* 150, 218–230. doi: 10.1016/j.knosys.2018.03.017

Goikoetxea, J., Soroa, A., Agirre, E., and Donostia, B. C. (2015). "Random walks and neural network language models on knowledge bases," in *Proceedings of HLT-NAACL*, 1434–1439. doi: 10.3115/v1/N15-1165

Gregg, V. (1976). "Word frequency, recognition and recall," in *Recall and recognition*, ed. J. Brown (New York: John Wiley & Sons).

Guasch, M., Ferré, P., and Fraga, I. (2016). Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behav. Res. Methods* 48, 1358–1369. doi: 10.3758/s13428-015-0684-y

Günther, F., Rinaldi, L., and Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: a discussion of common misconceptions. *Persp. Psychol. Sci.* 14, 1006–1033. doi: 10.1177/1745691619861372

Haghighi, A., and Vanderwende, L. (2009). "Exploring content models for multi-document summarization," in *Proceedings of Human Language Technologies: the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 362–370. doi: 10.3115/1620754.1620807

Harris, Z. S. (1954). Distributional structure. *Word.* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Hassan, S., and Mihalcea, R. (2009). "Cross-lingual semantic relatedness using encyclopedic knowledge," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1192–1201. doi: 10.3115/1699648.1699665

Hauk, O., Johnsrude, I., and Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron* 41, 301–307. doi: 10.1016/S0896-6273(03)00838-9

Hayes, T. R., and Henderson, J. M. (2021). Looking for semantic similarity: what a vector-space model of semantics can tell us about attention in real-world scenes. *Psychol. Sci.* 32, 1262–1270. doi: 10.1177/0956797621994768

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: evaluating semantic models with (genuine) similarity estimation. *Comput. Ling.* 41, 665–695. doi: 10.1162/COLI_a_00237

Hollenstein, N., de la Torre, A., Langer, N., and Zhang, C. (2019). Cognival: a framework for cognitive word embedding evaluation. *arXiv preprint arXiv:1909.09001*. doi: 10.18653/v1/K19-1050

Hualde, J. I., and De Urbina, J. O. (2011). *A grammar of Basque, volume 26*. Berlin: Walter de Gruyter.

Incitti, F., and Snidaro, L. (2021). "Fusing contextual word embeddings for concreteness estimation," in *2021 IEEE 24th International Conference on Information Fusion (FUSION)* (IEEE), 1–8.. doi: 10.23919/FUSION49465.2021.9626843

Jain, S., and Huth, A. (2018). "Incorporating context into language encoding models for fMRI," in *Advances in Neural Information Processing Systems*, 31. doi: 10.1101/327601

Jelodar, A. B., Alizadeh, M., and Khadivi, S. (2010). "Wordnet based features for predicting brain activity associated with meanings of nouns," in *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, 18–26.

Jones, M. N., Willits, J., Dennis, S., and Jones, M. (2015). Models of semantic memory. *Oxford Handb. Mathem. Comput. Psychol.* 1, 232–254. doi: 10.1093/oxfordhb/9780199957996.013.11

Joseph, H., Nation, K., and Liversedge, S. (2013). Using eye movements to investigate word frequency effects in children's sentence reading. Sch. Psychol. Rev. 42, 207–222. doi: 10.1080/02796015.2013.12087485

Kenett, Y. N., Levi, E., Anaki, D., and Faust, M. (2017). The semantic distance task: quantifying semantic distance with semantic network path length. *J. Exper. Psychol.* 43:1470. doi: 10.1037/xlm0000391

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012). The british lexicon project: lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behav. Res. Methods* 44, 287–304. doi: 10.3758/s13428-011-0118-4

Kinsbourne, M., and George, J. (1974). The mechanism of the word-frequency effect on recognition memory. *J. Verbal Lear. Verbal Behav.* 13, 63–69. doi: 10.1016/S0022-5371(74)80031-9

Koehn, P., and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Kosslyn, S. M., Thompson, W. L., and Ganis, G. (2006). *The Case for Mental Imagery*. New York: Oxford Psychology Series. Oxford University Press. doi: 10.1093/acprof:oso/9780195179088.001.0001

Kowialiewski, B., and Majerus, S. (2020). The varying nature of semantic effects in working memory. *Cognition* 202:104278. doi: 10.1016/j.cognition.2020.104278

Kun, S., Qiuying, W., and Xiaofei, L. (2023). An interpretable measure of semantic similarity for predicting eye movements in reading. *Psychon. Bull. Rev.* 30, 1227–1242. doi: 10.3758/s13423-022-02240-8

Lample, G. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Lample, G., and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lastra-Díaz, J. J., Goikoetxea, J., Taieb, M. A. H., García-Serrano, A., Aouicha, M. B., and Agirre, E. (2019). A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. *Eng. Appl. Artif. Intell.* 85, 645–665. doi: 10.1016/j.engappai.2019.07.010

Leturia, I. (2012). "Evaluating different methods for automatically collecting large general corpora for basque from the web," in *Proceedings of Coling 2012*, 1553–1570.

Levenshtein, V. (1965). Binary codes capable of correcting spurious insertions and deletion of ones. *Probl. Inf. Transm.* 1, 8–17.

Ljubešić, N., Fišer, D., and Peti-Stantić, A. (2018). Predicting concreteness and imageability of words within and across languages via word embeddings. *arXiv preprint arXiv:1807.02903*.

Locker, L., Simpson, G., and Yates, M. (2003). Semantic neighbourhood effects on the recognition of ambiguous words. *Memory Cogn.* 31, 505-515. doi: 10.3758/BF03196092

Long, Y., Xiang, R., Lu, Q., Huang, C.-R., and Li, M. (2019). "Improving attention model based on cognition grounded data for sentiment analysis. *IEEE Trans. Affect. Comput.* 12, 900–912. doi: 10.1109/TAFFC.2019.2903056

Luce, P., and Pisoni, D. (1998). Recognizing spoken words: the neighborhood activation model. *Ear Hear.* 19, 1–36. doi: 10.1097/00003446-199802000-00001

Luke, S. G., and Christianson, K. (2018). The provo corpus: a large eye-tracking corpus with predictability norms. *Behav. Res. Methods* 50, 826–833. doi: 10.3758/s13428-017-0908-4

Luong, M.-T., Socher, R., and Manning, C. D. (2013). "Better word representations with recursive neural networks for morphology," in *Proceedings of the Seventeenth Conference on Computational Natural Language learning*, 104–113.

MacLeod, C. M., and Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *J. Exper. Psychol.* 22:132. doi: 10.1037//0278-7393.22.1.132

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., and Aslin, R. N. (2007). The effects of word frequency, cohort density, and neighborhood density on eye movements during visual scene analysis. *J. Exper. Psychol.* 33, 1125–1138. doi: 10.1080/03640210709336987

Mandera, P., Keuleers, E., and Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. *J. Mem. Lang.* 92, 57–78. doi: 10.1016/j.jml.2016.04.001

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the penn treebank. *Comput. Ling.* 19, 313–330. doi: 10.21236/ADA273556

Mate, J., Allen, R. J., and Baqués, J. (2012). What you say matters: exploring visual-verbal interactions in visual working memory. *Q. J. Exp. Psychol.* 65, 395–400. doi: 10.1080/17470218.2011.644798

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Proceedings of Advances in Neural Information Processing Systems*, 3111–3119.

Miller, G. A. (1995). Wordnet: a lexical database for English. *Commun. ACM* 38, 39–41. doi: 10.1145/219717.219748

Mulatti, C., Reynolds, M., and Besner, D. (2006). Neighborhood effects in reading aloud: new findings and new challenges for computational models. *J. Exper. Psychol.* 32, 799–810. doi: 10.1037/0096-1523.32.4.799

Navigli, R. (2009). Word sense disambiguation: a survey. *ACM Comput. Surv.* 41, 1–69. doi: 10.1145/1459352.1459355

Navigli, R., and Ponzetto, S. P. (2010). "Babelnet: building a very large multilingual semantic network," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 216–225.

Paivio, A. (1971). *Imagery and Verbal Processes*. New York: Psychology Press.

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation," in *Proceedings of EMNLP*, 1532–1543. doi: 10.3115/v1/D14-1162

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710. doi: 10.1145/2623330.2623732

Planchuelo, C., Buades-Sitjar, F., Hinojosa, J. A., and Duñabeitia, J. A. (2022). The nature of word associations in sentence contexts. *Exper. Psychol.* 69:547. doi: 10.1027/1618-3169/a000547

Poria, S., Hazarika, D., Majumder, N., and Mihalcea, R. (2020). Beneath the tip of the iceberg: current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.* 14, 108–132. doi: 10.1109/TAFFC.2020.3038167

Rabovsky, M., Schad, D., and Abdel Rahman, R. (2016). Language production is facilitated by semantic richness but inhibited by semantic density: evidence from picture naming. *Cognition* 146, 240–244. doi: 10.1016/j.cognition.2015.09.016

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog* 1:9.

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). "A word at a time: computing word relatedness using temporal semantic analysis," in *Proceedings of the 20th International Conference on World wide web*, 337–346. doi: 10.1145/1963405.1963455

Raney, G. E., and Rayner, K. (1995). Word frequency effects and eye movements during two readings of a text. *Canad. J. Exper. Psychol.* 49, 151–172. doi: 10.1037/1196-1961.49.2.151

Reilly, M., and Desai, R. (2017). Effects of semantic neighborhood density in abstract and concrete words. *Cognition* 169, 46–53. doi: 10.1016/j.cognition.2017.08.004

Rodrigues, J., Branco, R., Silva, J., Saedi, C., and Branco, A. (2018). "Predicting brain activation with wordnet embeddings," in *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, 1–5. doi: 10.18653/v1/W18-2801

Rogers, T. T., and McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. London: MIT Press. doi: 10.7551/mitpress/6161.001.0001

Rothe, S., Ebert, S., and Schütze, H. (2016). Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Rubenstein, H., and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Commun. ACM* 8, 627–633. doi: 10.1145/365628.365657

Salicchi, L., Lenci, A., and Chersoni, E. (2021). "Looking for a role for word embeddings in eye-tracking features prediction: does semantic similarity help?" in *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, 87–92.

Sass, K., Sachs, O., Krach, S., and Kircher, T. (2009). Taxonomic and thematic categories: neural correlates of categorization in an auditory-to-visual priming task using fMRI. *Brain Res.* 1270, 78–87. doi: 10.1016/j.brainres.2009.03.013

Schwanenflugel, P. (2010). Why are abstract concepts hard to understand? The psychology of word meanings. *Hum. Brain Mapp.* 31, 1459–1468.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., et al. (2011). Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* 108, 8520–8524. doi: 10.1073/pnas.1014935108

Solovyev, V., Yu, V., Andreeva, M., and Zaikin, A. (2022). Russian dictionary with concreteness/abstractness indices. *Russian J. Ling.* 26, 515–549. doi: 10.22363/2687-0088-29475

Spink, A., Wolfram, D., Jansen, M. B., and Saracevic, T. (2001). Searching the web: the public and their queries. *J. Am. Soc. Inf. Sci. Technol.* 52, 226–234. doi: 10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.3.CO;2-I

Spivey, M. (2008). *The Continuity of Mind*. Oxford: Oxford University Press.

Stadthagen-Gonzalez, H., and Davis, C. J. (2006). The bristol norms for age of acquisition, imageability, and familiarity. *Behav. Res. Methods* 38, 598–605. doi: 10.3758/BF03193891

Strijkers, K., Costa, A., and Thierry, G. (2010). Tracking lexical access in speech production: electrophysiological correlates of word frequency and cognate effects. *Cerebral Cortex* 20, 912–928. doi: 10.1093/cercor/bhp153

Toneva, M., and Wehbe, L. (2019). "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)," in *Advances in Neural Information Processing Systems*, 32.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). "Metaphor detection with cross-lingual model transfer," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 248–258. doi: 10.3115/v1/P14-1024

Tversky, A. (1977). Features of similarity. *Psychol. Rev.* 84, 327. doi: 10.1037//0033-295X.84.4.327

Vaswani, A. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*.

Vitevitch, M. (2002). The influence of phonological similarity neighbors on speech production. *J. Exper. Psychol.* 28, 735–747. doi: 10.1037//0278-7393.28.4.735

Wang, J., Conder, J., Blitzer, D., and Shinkareva, S. (2010). Neural representation of abstract and concrete concepts: a meta-analysis of neuroimaging studies. *Hum. Brain Mapp.* 31, 1459–1468. doi: 10.1002/hbm.20950

Wulff, D. U., De Deyne, S., Jones, M. N., and Mata, R. (2019). New perspectives on the aging lexicon. *Trends Cogn. Sci.* 23, 686–698. doi: 10.1016/j.tics.2019.05.003

Wulff, D. U., O'Brien, E. J., and Heller, T. (2022). Structural differences in the semantic networks of younger and older adults. *J. Mem. Lang.* 120:104250. doi: 10.1038/s41598-022-11698-4

Xu, K., Wu, L., Wang, Z., Feng, Y., Witbrock, M., and Sheinin, V. (2018). Graph2seq: graph to sequence learning with attention-based neural networks. *arXiv preprint arXiv:1804.00823*.

Yates, M., Locker, L., and Simpson, G. (2004). The influence of phonological neighborhood on visual word perception. *Psychon. Bull. Rev.* 11, 452–457. doi: 10.3758/BF03196594

Yates, M., Locker, L., and Simpson, G. B. (2003). Semantic and phonological influences on the processing of words and pseudohomophones. *Memory Cogn.* 31, 856–866. doi: 10.3758/BF03196440

Zhang, X., Zhao, J., and LeCun, Y. (2015). "Character-level convolutional networks for text classification," in *Advances in Neural Information Processing Systems*, 28.

Zhang, Y., Zhang, X., Li, C., Wang, S., and Zong, C. (2024). Mulcogbench: a multi-modal cognitive benchmark dataset for evaluating Chinese and English computational language models. *arXiv preprint arXiv:2403.01116*.