Distributional Semantics for Diachronic Search

Pablo Gamallo[a], Iván Rodríguez-Torres[a], Marcos Garcia[b]

[a]Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), University of Santiago de Compostela, Galiza
[b]LyS Group, University of A Coruña, Galiza

Abstract

This article describes a system aimed at searching for word similarity over different time periods. The strategy is based on distributional models obtained from a chronologically structured language resource, namely Google Books Syntactic Ngrams. The models were created using dependency-based contexts and a strategy for reducing the vector space, which consists of selecting the more informative and relevant word contexts. A quantitative evaluation of the distributional models was performed. The linguistic data are stored in a NoSQL DB, which is provided with a Web interface allowing linguists to analyze the meaning change of Spanish words in written texts across time.

Keywords: natural language processing, diachronic semantics, distributional semantics, language change

1. Introduction

From a synchronic perspective, the relation between the form and the meaning of any word is permanent. Language speakers cannot change or modify such a relation because in that case, language communication would not be possible. However the form-meaning relationship is not immutable from a diachronic or historical perspective. Words are being given new meanings or senses and losing other ones across time. For instance, the word *plastic*, which derives from the Latin word *plasticus* and the Greek one *plastikos*, appears in English as early as the 1600's, used as an adjective to relate to something that could be easily molded or shaped. However, in the 20th century, this word was provided with a new nominal sense referring to a new material consisting of synthetic organic compounds that are malleable and can be molded into solid objects. The two senses coexist as separate (even if somehow related) meanings for the same word form.

Frequency-based methods (such as Google Trends or Google Books Ngram Viewer search engines) are useful for finding words whose use significantly increases or decreases during a specific period of time, but they do not allow language experts to detect semantic changes, such as those mentioned above.

Taking the above into account, our objective is to describe a methodology based on distributional semantics and natural language processing to design and build a system which can be used by linguists and other researchers in the Humanities to identify, analyse and explain the semantic changes undergone by words in a specific period of time. Distributional semantics is a theory of meaning which is computationally implementable. Word meaning is represented as a vector containing the set of contexts in which the word occurs. So, the notion of word meaning is derived from the distribution of the word's contexts and the term distributional semantics is often used to describe the vector space models representing the meaning of words [1].

In our framework, we built a distributional language model on yearly corpora, learned from Spanish Google Books Ngrams, so as to obtain word vectors for each year from 1900 to 2009. Then, we compared the pairwise similarity of word vectors for each year and inserted this information in a non structured database. Finally, we implemented a web interface to provide researchers with a tool, called Diachronic Explorer, to search for semantic changes. Diachronic Explorer offers different advanced searching techniques and different visualizations of the results. In this article, we will describe the methodology used to build this distributional-based diachronic tool.

The main contribution of our proposal is the design of a distributional-based system to deal with semantic changes across historical Spanish texts. Even though similar distributional approaches exist for English, our proposal is the first piece of work to use this technique for Spanish. Although the Spanish n-grams are based on much smaller corpora than the English counterpart, the experiments described later in the paper demonstrate that the Spanish corpus is still large enough to build meaningful distributional models for each year since 1900. In addition, we have provided the Hispanic community with a useful tool to convey out linguistic research. The system is open source[1] and includes a web interface[2]. To the best of our knowledge, this is the first unrestricted web search tool to explore diachronic evolution of meaning for any language.

The rest of the article is organized as follows. This introduduction is followed by an overview of related work in Section 2. In Section 3 we describe the strategy we use to construct distributional semantic models. Then, in Section 4 we outline our specific proposal to design a system for searching for semantic changes across time. Next, in Section 5, we evaluate the distributional models by measuring their performance in word similarity tasks. And finally, we address the main conclusions and possible future work in Section 6.

2. Related Work

In recent years, we have observed an increasing number of quantitative-based works on diachronic linguistics and semantic/grammatical change, in addition to more theoretical studies. This is motivated by the growing number of new digitized historical corpora as well as the availability of natural language processing tools.

Liberman et al. [2] used quantitative methods for analysing grammatical change. They showed the relationship between frequency and morphological change in the past tense of English verb forms. For this purpose, the authors analyzed the irregular verbs of the last 1200 years. This quantitative-based diachronic analysis tried to demonstrate that the change from an irregular verb form to the regular one (*-ed*) is related with frequency: an irregular verb used

[1] `https://github.com/citiususc/explorador-diacronico`

[2] `https://tec.citius.usc.es/explorador-diacronico/index.php`

100 times less than another one will start to be used with the regular form 10 times faster.

Sagi et al. [3] analyze the distribution of words over time in order to identify specific types of meaning change, focusing on the widening and narrowing of meaning. They use the notion of *density* within a vector space to find semantic changes. The higher the density, the lesser the probability of semantic change. It is worth noting that this evluation is more qualitative than quantitative, concerning just a few English examples.

Wijaya and Yeniterzi [4] use English Google Books Ngrams to generate yearly distributional models, in a very similar way to what we did. The difference is that their objective is to automatically find the period where word meaning changes. For this purpose, word vectors are classified in clusters. If two vectors of the same word in two consecutive years are found in different clusters, then one may infer that its meaning changed at that time. A similar work is described by Gulordava and Baroni [5], but they are focused on finding meaning change with regard to two specific time periods (1960-64 and 1995-99). As in the previously cited work, they claim that if a word has low similarity to itself in the two periods then its meaning could have changed. They found that 1.6% of words with low similarity may have undergone semantic change. Besides, they found a relation between semantic change and frequency: more frequent words tend to change in meaning. Jatowt [6] also uses Google Books Ngrams to build yearly distributional models. However, they are based on different strategies: unordered bag-of-words, ngrams, and LSA. Besides analyzing the evolution of word meaning through time, he also studies changes in polarity (sentiment orientation). With similar objectives, Kim et al. [7] train neural models from Google Books Ngrams and implement algorithms for change point detection. So they try to generalize the process of identifying meaning change by applying the method to whatever historical point instead of being focused on a specific time period as in the works cited above.

Finally, Hamilton [8] applies neural methods which are similar to those described in [7], but with the objective of defining a methodology for quantifying semantic change on the basis of four languages: English, French, German, and Chinese. The historical analysis across the four languages led to the proposal of two statistical "laws" that govern the evolution of word meaning. First, the *law of conformity* states that rates of semantic change scale with a negative power of word frequency, that is, frequent words have meanings that are more stable over time. Secondly, the *law of innovation* states that polysemous words have significantly higher rates of semantic change.

Besides the recent automatic approaches to the study of meaning change, this subject has been studied deeply since the XIX century in the field of Linguistics. We think that automatic strategies such as those introduced above should be accompanied by linguistic studies, carried out by experts with the help of high-level semantic tools. For this purpose, we built Diachronic Explorer, which may help linguists find meaning changes over large and chronologically organized text corpora. Unlike the work cited above, our system works with Spanish, uses syntactic ngrams (Sec: 3.3), and is based on a filtering strategy (Sec: 3.2) to build distributional models.

3. Distributional Semantic Models

3.1. Count-Based Models and Embeddings

The distributional hypothesis states that words appearing in similar contexts are semantically similar [9]. Distributional methods based on natural language processing and information extraction learn a semantic model in which words are represented as vectors of contexts from a large corpus. Those vectors are proxies for word meaning representations that can be measured and compared in order to acquire the semantic similarity between words.

Distributional methods differ mainly in the way the word space model is built: count-based approaches and neural-based embeddings are two of the most popular strategies. The former strategy collects context vectors and then reweighes them based on some association measure (e.g. log-likelihood, mutual information, PMI, etc). As word distributions give rise to sparse matrices, these count-based strategies make use of dimensionality reduction techniques such as singular value decomposition (SVD) or context filtering [10]. More recently, other strategies based on neural-network language modeling have been proposed to represent words [11]. These methods give weight to vectors so as to optimally predict the contexts in which the corresponding words tend to appear. Since similar words occur in similar contexts, the system learns to naturally assign similar vectors to similar words [12]. These efficient representations are known as *word embeddings* or *predictive models*.

There is some controversy about the performance of the different types of word space models when they are applied to specific natural language tasks. Even if word embeddings have gained popularity in recent years, some researchers showed that there are no significant differences between count-based models and embeddings when applied to tasks such as word similarity discovery [13, 14]. In addition, count-based sparse vectors can also be represented in an efficient way on the basis of hashing functions whose keys are word-context pairs and their values are non-zero scores [10].

3.2. A Count-Based Model with Context Filtering

The method we use to build the historical semantic track of Spanish words is a count-based approach with context filtering. This method allows us to build an *explicit model* of word meanings. Their performance and efficiency are comparable to predictive strategies [10, 14]. Also, explicit models are transparent for linguists and fully interpretable, unlike embeddings which transform contexts into opaque dimensions and weights. The full interpretation of contexts in an explicit model is a very strong motivation to use count-based approaches in digital humanities projects. Any linguist required to study and analyze the similarity between two words can easily check the more relevant contexts they share. This linguistically transparent information is hidden in word embeddings.

Given the Zipf-law distribution of words in a corpus, all word-context matrices are sparse. When storing and manipulating large sparse matrices, it is very useful to store only non-zero values in a hash table where keys are word-context pairs [10, 14]. Following Gamallo and Bordag [10], we claim that it is not necessary to reduce the dimensionality of the entire vector space. Instead of representing the whole word space model as a full matrix (with n^2 being required storage space), it could be represented in such a way that a vector uses only as much memory as non-zero entries are in it. Zero values are easily induced, or rather assumed, later by the algorithm used to compute vector similarity. An efficient storage mode for a sparse matrix is a hash table with a key-value representation. Keys are structured as a two-dimensional array, containing only row-column pairs with non-zero values. Like in a matrix structure, hashes also allow access to any arbitrary element in a constant amount of time by means of using a hash function that, given a key, computes the address of the value stored for that key.

To reduce the number of keys in a hash table representing word-context co-occurrences, we apply a technique to filter out contexts by relevance. The compressing technique consists of computing an association measure between each word and their contexts (for instance, log-likelihood, mutual information, or PMI). Considering the experiments performed in [15], we use log-likelihood as an association measure [16]. Then, for each word, only the R (relevant) contexts with highest log-likelihood scores are kept in the hash table. The top R contexts are considered to be the most relevant and informative for each word. R is a global, arbitrarily defined constant whose usual values range from 10 to 1000 [17]. However, this value can be computed by selecting a proportion over the total number of dependency contexts. In our work, $R = \sqrt{\|C\|}$, where $\|C\|$ is the total number of different contexts in the corpus. In short, we keep the R most relevant contexts for each target word.

A filtered model is then based on selecting the most relevant context per target word. It is an explicit representation readable by humans. Methods based on dimensionality reduction and embeddings, by contrast, make the vector space more compact with dimensions that are not transparent in linguistic terms.

The filtering-based approach turned out to be as efficient as other strategies based on dimensionality reduction such as SVD [10].

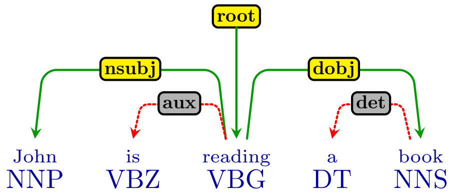
3.3. Syntax-Based Contexts

Another important element of distributional models is the type of context used to represent word distributions. The contexts of words can be based on their sequential order in the sentence (bag-of-words) or on their respective position in a syntactic parse tree. The different contextual words identified by syntactic-based techniques may be far apart from the target word in the sentence, yet close to it syntactically. Previous work comparing different types of contexts showed that syntax-based (usually dependencies) methods outperform bag-of-words strategies [18, 19, 20, 21], in particular when the objective is to compute semantic similarity between functional equivalent words, such as detection of co-hyponym/hypernym word relations (i.e. near synonymy).

The syntax-based context used in this work is derived from syntactic dependencies defined in [22]. Dependency parsing is a natural choice, as it emphasizes individual words and explicitly models the connections between them. Dependency parse trees contain binary relationships between words in the same sentence. More precisely, binary dependencies are extracted from ngrams ($n \leq 5$) annotated with POS tags and dependency relations. Content-words, which are meaning bearing elements, are distinguished from functional markers (or non-content words), which do not carry semantic meaning of their own, such as the auxiliary verb *have*. Only dependencies between content words are considered.

Figure 1 shows the syntactic analysis of a 5-gram, “John is reading a book”, and the two binary dependencies extracted from it. We use an arrow that points from the head word to the modifier word (e.g., $head => modifier$) to indicate a dependency binary relation. Non-content words (e.g. *is*, *a*) are syntactically analysed but they are not considered as semantically relevant binary dependencies.

As the utility of syntactic contexts of words for constructing vector-space models of word meaning is well established, we will use word dependencies to to build our distributional models.



reading => John

reading => book

Figure 1: A syntactic 5-gram which includes two dependency relations for three content words: *John*, *reading*, and *book*. The determiner and the auxiliary verb are non-content words (linked with dashed arrows) and therefore are not extracted as relevant dependency pairs.

3.4. Word Similarity

The process of measuring word similarity is based on identifying the contexts shared by two words. In Equation 1, the similarity Sim between two words, w_1 and w_2, is the Cosine coefficient which turned out to be one of the best measures in distributional semantics [10] :

$$\mathsf{Sim}(\mathsf{w}_1, \mathsf{w}_2) = \frac{\sum_{\mathsf{i}} \mathsf{A}(\mathsf{w}_1, \mathsf{c}_\mathsf{i})\mathsf{A}(\mathsf{w}_2, \mathsf{c}_\mathsf{i})}{\sqrt{\sum_{\mathsf{j}} (\mathsf{A}(\mathsf{w}_1, \mathsf{c}_\mathsf{j}))^2} \sqrt{\sum_{\mathsf{k}} (\mathsf{A}(\mathsf{w}_2, \mathsf{c}_\mathsf{k}))^2}} \tag{1}$$

where $A(w_1, c_i)$ is an association measure (e.g. loglikehood) between w_1 and syntactic context c_i.

4. The Diachronic Explorer

We built a system, The Diachronic Explorer, which relies on a set of distributional semantic models for each year from 1900 to 2009 for the Spanish language. This semantic resource is stored in a NoSQL database which feeds a web server used for searching for and visualizing the lexical changes of hundreds of thousands of Spanish words through time.

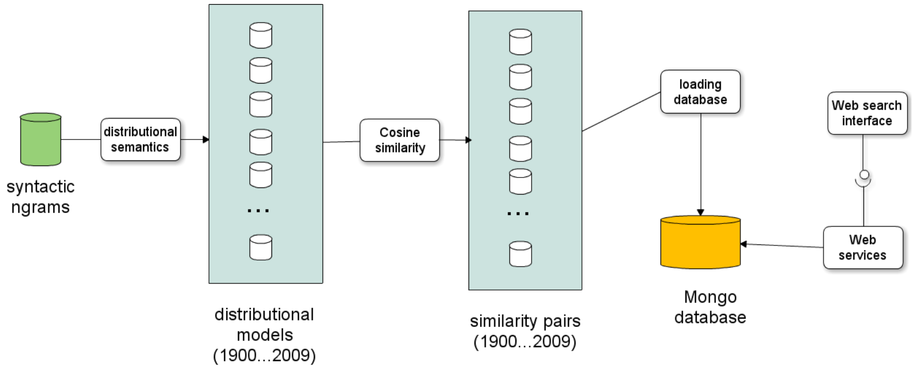


Figure 2: Architecture of Diachronic Explorer

4.1. Architecture

Figure 2 shows the architecture of The Diachronic Explorer. As input, it takes a large corpus of dependency-based ngrams to build 110 distributional models, one per year since 1900 to 2009. Then, Cosine similarity is computed for all words in each model in order to generate CSV files consisting of pairs of similar words. Each word is associated with its N most similar words (where N = 20), according to the Cosine coefficient. These files are stored in MongoDB to be accessed by web services which generate different types of word searching.

4.2. Yearly Models from Syntactic-Ngrams

To build the distributional semantic models through time, we made use of the dataset based on the Google Books Ngrams Corpus, which is described in detail in [23]. The whole project contains over 500 billion words (45B in Spanish). The majority of the content was published after 1900. More precisely, we used the Spanish syntactic-ngrams from the 2012-07-01 Version [3]. The method used to extract syntactic ngrams was described in [22]. Syntactic ngrams were extracted following the strategy defined above in subsection 3.3. Each syntactic ngram is accompanied by a corpus-level occurrence count, as well as a time-series of

[3] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

counts over the years. The temporal dimension allows inspection of how the meaning of a word evolves over time by looking at the contexts the word appears in within different time periods.

We transformed the syntactic ngrams into distributional 'word-context' matrices by using the filtering-based strategy that selects for relevant contexts (Section 3). A matrix was generated for each year, where each word is represented as a context vector. The final distributional-based resource consists of 110 matrices (one per year from 1900 to 2009). The size of the whole resource is 2,8G, with 25M per matrix on average. Then, the Cosine similarity between word vectors was calculated and, for each word, the 20 most similar ones were selected by year[4]. In total, a data structure with more than 300 million pairs of similar words was generated, giving rise to 110 CSV files, one per year.

4.3. Data Storage

We stored data in two ways. First, we stored the distributional models and the similarity pairs in temporary CSV files. These temporary files are generated from offline processes which can be executed periodically as the linguistic input (ngrams) is updated or enlarged with new language sources. Secondly, these files are imported into a database to make data access easier and more efficient.

The database system that we chose was MongoDB, which is a NoSQL DB. This type of database system fits well with the task for two main reasons:

- Our data do not follow a relational model, so we do not need some of the features that make relational databases powerful, for example, reference keys or table organization.[5]

- NoSQL databases scale really well. Unlike relational databases they im-

[4] We performed experiments with different values of N (from 1 to 20). Performance grows in a significant way from 1 to 5, but the curve tends to stabilize at N=20.

[5] We decided that our data should not be stored based on a relational model as the test using an unstructured NoSQL database engine has proven to be quicker when applied to the same queries.

plement automatic sharding, which enables us to share and distribute data among servers in the most efficient way possible. So, as the data increase it will be easy to add a new instance without any additional engineering.

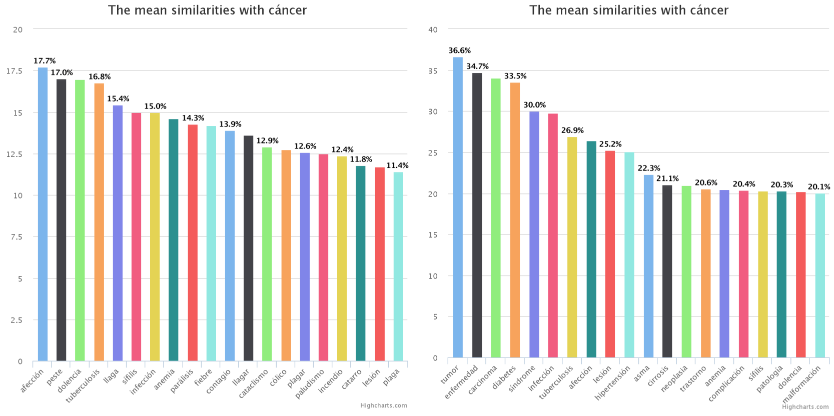
Among the different types of NoSQL databases, we chose MongoDB because it is document oriented, which fits very well with our data structure. Besides, its user and developer community is very strong and active, so it is easy to find support from them.

4.4. Web Interface

The Diachronic Explorer provides a web interface that offers five different ways of making a diachronic search:

Simple: The user enters a target word and selects a specific period of time (e.g. from 1920 to 1939). Then, the system returns the 20 most similar terms (on average) within the selected time period. The user can select a word cloud image to visualize the output. In Figure 3, we show the similar words to *cáncer* (*cancer*) returned with a simple search in two different periods: 1900 (a) and 2009 (b). Notice that this word is similar to common diseases at the beginning of the 20th century —such as *peste* (*plague*) and *tuberculosis*—, while nowadays it is more similar to more technical words, such as *tumor* (*tumour*) or *carcinoma*.

Pairs: The user can carry out the same search but focused on a pair of words. In this case, the system returns the similarity score obtained for the two words during a selected period of time. Figure 4 shows the pair search of two words: *ordenador* (*computer* in European Spanish) and *computador* (*computer* in American Spanish). Even if the search was extended to the whole period (1900-2009), the two words started to be similar in the mid-sixties and became fully similar quickly in the seventies with the emergence of computer science. Notice that the two words are not comparable before the sixties because there are no occurrences of *computador* in the Google Books Ngram corpus before that time.



(a) (b)

Figure 3: Similar words to *cáncer* (*cancer*) using a simple search in 1900 (a) and 2009 (b)

Track record: As in the simple search, the user inserts a word and a time period. However, this new type of search returns the specific similarity scores for each year of the period, instead of the similarity average. In addition, the system allows the user to select any word from the set of all words (and not just the top 20) semantically related to the target one in the selected time period. Figure 5 shows the similarity record between *plástico* (*plastic*) and two other words *vidrio* (*glass*) and *musical* (*musical*), during the period 1950-2009. The similarity score with the material noun *vidrio* begins to grow in the sixties at the same time as this synthetic material gains in popularity. Besides, its relation of similiarity with *musical* decreases because its use as an adjective begins to be much less common at that time.

Transitive: In this case, the system returns the 20 most similar words after having computed a new similarity metric: *transitivity similarity*. This kind of search is based on transitivity property. To explain the metric, let us suppose we are provided with a tree similarity structure at three levels. The root, i.e. the search word, is at the first level and its 20 most

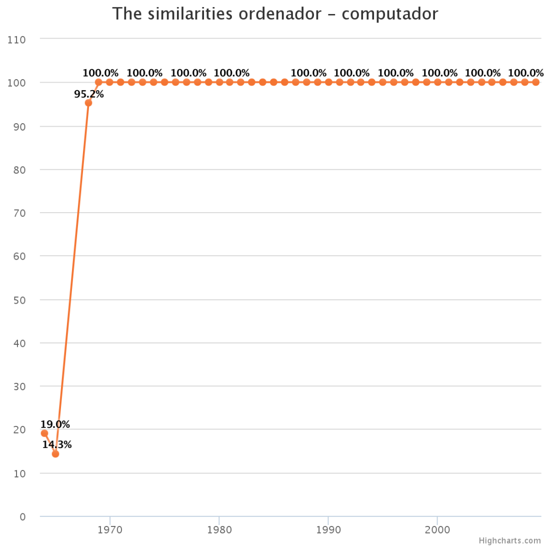


Figure 4: Comparing how similar two full synonyms are, *ordenador* (*computer* in European Spanish) and *computador* (*computer* in American Spanish), using the pair search between 1900 and 2009.

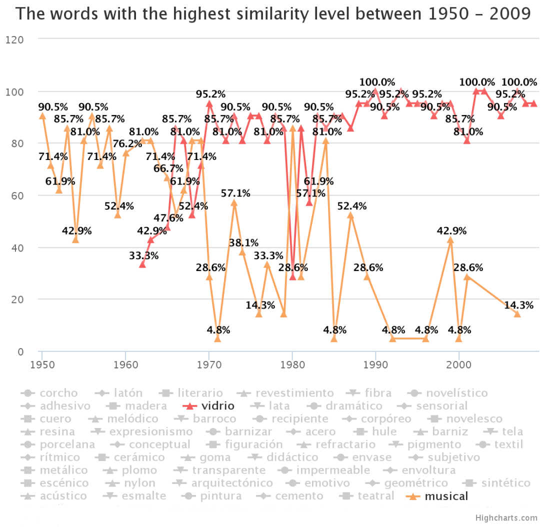


Figure 5: Historical record (1950-2009) of the similarity between *plástico* (*plastic*) and two words *vidrio* (*glass*, red line) and *musical* (*musical*, orange line).

similar words are at the second one. The third level is also considered and consists of the words that are similar to those from the second level. By considering the transitivity property of similarity, we compute a new similarity score between the root word (first level) and those found at the third level. This is done by adding the direct similarity scores and by normalizing the results. The final values range between 0 and 100.

Cloud: The fifth search option is the classic cloud. It allows the user to generate a wordcloud with the most similar words related to the search word. As is usual in this kind of representation, there are different word sizes. Word sizes are determined by the similarity level. So the bigger a word is the strongest its correlation with the search word, while the small ones are those whose correlation level is lower.

5. Evaluation

In order to check whether the distributional-based similarity scores generated using our models are semantically coherent, we evaluated the results from a sample of models using standard test datasets. The sample consists of selecting the similarity scores obtained from one distributional model every 10 years from 1900 to 2009. In total, we built a sample containing similarity pairs of 12 different models to be evaluated. Concerning the test data, as there are not many resources available for evaluation purposes in Spanish, we used the only three datasets that, as far as we know, are freely available:

- A translation into Spanish of WordSim-353 [24]. The original dataset contains 353 word pairs, each one associated with an average of 13 to 16 human judgments according to its similarity.
- Another translation into Spanish of the standard RG-65 dataset provided by [25]. The original RG dataset consists of 65 pairs of words collected by [26], who had them judged by 51 human subjects on a scale from 0.0 to 4.0.

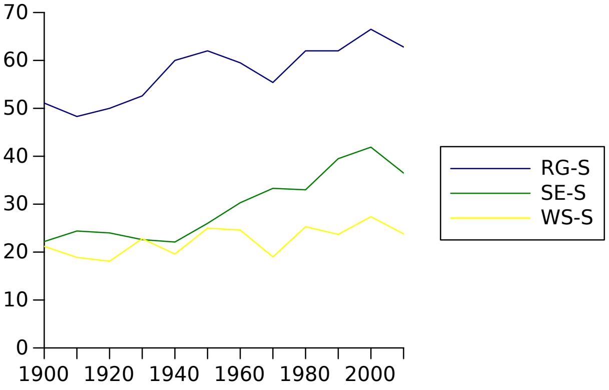


Figure 6: Spearman correlation (y-axis) of 12 sample models (from 1900 to 2009) on the three datasets: RG45-Spanish (RG-S), SemEval-Task2-Spanish (WS-S), and WordSim353-Spanish (WS-S).

- The Spanish dataset used in SemEval-2017 Task 2 on multilingual words and multiword similarity [27]. The original dataset consists of 500 pairs but we only used those pairs that do not contain multiwords, 378 in total.

The test datasets were constructed by asking humans to rate the degree of semantic similarity between two words on a numerical scale. The performance of a computational system was measured in terms of correlation (Spearman) between the values assigned by humans to the word pairs and the similarity coefficient assigned by our strategy.

As Figure 6 depicted, the quality of our yearly models does not vary too much across time. Even if the earlier models tend to be poorer, the difference with regard to the latest ones is not very significant. So, the semantic coherence of all models remains somehow stable through time. Notice the symmetry of the three tests: models behave in a similar way in the three datasets even if they are completely different. This makes clear that the test datasets are well balanced and representative.

Table 2 compares the Spearman correlation obtained by our best model (year 2000) and another comparable model we trained using Word2Vec[6]. To train this model, we used a corpus of the same size (about 6 million words) as the n-grams required to build the year-2000 model. The corpus was extracted from the Spanish Billion Words Corpus and Embeddings project[7]. Table 2 also compares the results of these two models with the highest scores reached so far on those datasets in Spanish: [28], [29] and SemEval-2017 Task 2 [27]. However, comparison between our models and state-of-the-art systems is not fair. First, the score obtained by Agirre et al. and reported in [29] was not obtained based on the same RG-Spanish dataset but on a different bilingual English-Spanish list based on the original RG dataset. And secondly, our models are very small matrices if compared with those derived from the large text corpora used in [28], [29] and SemEval-2017 Task 2 [27]. Our models are very limited in size because each one was built from texts belonging to a specific year. In spite of this size limitation, the behaviour of our yearly models can be seen as acceptable since they are not very far from the best systems relying on very large corpora. In particular, our best model is just 3 points below (0.27 *vs* 0.30) with regard to the experiment described in [28]. In addition, it is worth noting that our year-2000 model derived from a medium-size corpus clearly performs better than the neural-based embeddings trained on a corpus of a similar size.

In order to check whether our syntax-based models outperform word embeddings when both are trained on medium-size corpora, we tested them on a different task: categorization. Given a set of nominal concepts, the objective of the task is to group them into natural categories (e.g., cars and motorcycles should go into the vehicle class, dogs and elephants into the mammal class). To perform clustering and evaluation, we used CLUTO toolkit[8]. Performance is evaluated in terms of *entropy* and *purity*. Small entropy values and large purity

[6] `code.google.com/p/word2vec/`

[7] `http://crscardellino.me/SBWCE`

[8] `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download`

System	*RG-S* (ρ)	*SE-S* (ρ)	*WS* (ρ)
Model-2000	0.67	0.42	0.27
Word2Vec-6M	0.33	0.29	0.10
Agirre 2009	0.83	-	-
SemEval-17 (best system)	-	0.72	-
Etcheverry 2015	-	-	0.30

Table 1: Spearman correlation ρ between three Spanish datasets and the rating obtained by our best yearly model (Model-2000), the word embeddings trained from a 6M Spanish corpus (Word2Vec-6M), and the state-of-the-art scores in those three datasets. The datasets are the following: RG45-Spanish (RG-S), SemEval-Task2-Spanish (WS-S), and WordSim353-Spanish (WS-S).

values indicate good clustering solutions.

We translated two datasets into Spanish and removed multiwords. The first dataset is the ESSLLI 2008 Distributional Semantic Workshop shared-task set (*esslli*), which contains 44 concepts to be clustered into 6 categories[9] The second one is the Battig (*battig*) test set introduced by Baroni et al. [30], which includes 83 concepts from 10 categories. The results in terms of both entropy and purity show that our yearly model clearly outperforms the embeddings acquired from a corpus with similar size. Moreover, our model is close to the performance of the embeddings trained on a much bigger corpus. The results of this last experiment seems to confirm the previous conclusion: transparent and syntax-based models perform better that word embeddings when both are acquired from medium size corpora.

6. Conclusions

In this article, we have described a system, the Diachronic Explorer, that allows linguists to analyze the meaning evolution (change or stability) of Spanish words in written language across time. The system is based on distributional

[9] `http://wordspace.collocations.de/doku.php/workshop:esslli:task`

System	*esslli*		*batting*	
	entropy	purity	entropy	purity
Model-2000	0.380	0.700	0.151	0.817
Word2Vec-6M	0.545	0.588	0.518	0.480
Word2Vec-1B	0.308	0.725	0.053	0.958

Table 2: Comparing the performance (entropy and purity) in categorization between our best yearly model (Model-2000) and the word embeddings trained from a 6M Spanish corpus (Word2Vec-6M). The performance of word embeddings trained on a much bigger corpus with 1 billion words (Word2Vec-1B) is also considered. The two data sets of the experiment are *esslli* and *batting*.

models obtained from a chronologically structured language resource, namely Google Books Syntactic Ngrams. The models were created using dependency-based contexts and a strategy for reducing the vector space, which consists of selecting the more informative and relevant word contexts.

As far as we know, our system is the first attempt to build diachronic distributional models for the Spanish language. Besides, it uses NoSQL storage technology to scale easily as new data is processed, and provides an interface enabling useful types of word searches across time. This is an open source project that is freely available[10].

A future interesting direction of research could involve the use of the system infrastructure for structuring other types of language variety, namely diastratic or diatopic variation. Distributional models can be built from text corpora organized, not only by diachronic information, but also by social and dialectal features. For instance, we could adapt the system to search for meaning changes across different diatopic varieties: Spanish from Argentina, Mexico, Spain, Bolivia, and so on. The structure of our system is generic enough to deal with any type of variety, not only that derived from the historical axis.

[10] https://github.com/citiususc/explorador-diacronico

Acknowledgments

This work has received financial support from a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, TelePares (MINECO, ref:FFI2014-51978-C2-1-R), *Juan de la Cierva formación* Grant (Ref: FJCI-2014-22853), the Consellería de Cultura, Educación e Ordenación Universitaria (acreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

[1] S. Clark, Vector Space Models of Lexical Meaning, John Wiley & Sons, Ltd, 2015, pp. 493–522. `doi:10.1002/9781118882139.ch16`.
URL `http://dx.doi.org/10.1002/9781118882139.ch16`

[2] E. Lieberman, J.-B. Michel, J. Jackson, T. Tang, M. A. Nowak, Quantifying the evolutionary dynamics of language, Nature 449 (7163) (2007) 713–716. `doi:10.1038/nature06137`.

[3] E. Sagi, S. Kaufmann, B. Clark, Semantic density analysis: Comparing word meaning across time and phonetic space, in: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics, Association for Computational Linguistics, Athens, Greece, 2009, pp. 104–111.
URL `http://www.aclweb.org/anthology/W09-0214`

[4] D. T. Wijaya, R. Yeniterzi, Understanding semantic change of words over centuries, in: Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web, DETECT '11, ACM, New York, NY, USA, 2011, pp. 35–40. `doi:10.1145/2064448.2064475`.
URL `http://doi.acm.org/10.1145/2064448.2064475`

[5] K. Gulordava, M. Baroni, A distributional similarity approach to the detection of semantic change in the google books ngram corpus, in: Proceedings

of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, GEMS '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 67–71.
URL http://dl.acm.org/citation.cfm?id=2140490.2140498

[6] A. Jatowt, K. Duh, A framework for analyzing semantic change of words across time, in: Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '14, IEEE Press, Piscataway, NJ, USA, 2014, pp. 229–238.
URL http://dl.acm.org/citation.cfm?id=2740769.2740809

[7] Y. Kim, Y. Chiu, K. Hanaki, D. Hegde, S. Petrov, Temporal analysis of language through neural language models, CoRR abs/1405.3515.
URL http://arxiv.org/abs/1405.3515

[8] W. L. Hamilton, J. Leskovec, D. Jurafsky, Diachronic word embeddings reveal statistical laws of semantic change, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, 2016, pp. 1489–1501.
URL http://aclweb.org/anthology/P/P16/P16-1141.pdf

[9] J. Firth, A synopsis of linguistic theory 1930-1955, Studies in linguistic analysis (1957) 1–32.

[10] P. Gamallo, S. Bordag, Is singular value decomposition useful for word simalirity extraction, Language Resources and Evaluation 45 (2) (2011) 95–119.

[11] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, 2013, pp. 746–751.

[12] M. Baroni, R. Bernardi, R. Zamparelli, Frege in space: A program for compositional distributional semantics, LiLT 9 (2014) 241–346.

[13] O. Levy, Y. Goldberg, Linguistic regularities in sparse and explicit word representations, in: Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014, 2014, pp. 171–180.

[14] P. Gamallo, Comparing explicit and predictive distributional semantic models endowed with syntactic contexts, Language Resources and Evaluation First online: 13 May 2016.

[15] S. Bordag, A Comparison of Co-occurrence and Similarity Measures as Simulations of Context, in: 9th CICLing, 2008, pp. 52–63.

[16] T. Dunning, Accurate methods for the statistics of surprise and coincidence, Computational Linguistics 19 (1) (1993) 61–74.

[17] M. Padró, M. Idiart, A. Villavicencio, C. Ramisch, Nothing like good old frequency: Studying context filters for distributional thesauri, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, 2014, pp. 419–424.

[18] S. Padó, M. Lapata, Dependency-Based Construction of Semantic Space Models, Computational Linguistics 33 (2) (2007) 161–199.

[19] Y. Peirsman, K. Heylen, D. Speelman, Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts, in: CoSMO Workshop, Roskilde, Denmark, 2007, pp. 9–16.

[20] P. Gamallo, Comparing window and syntax based strategies for semantic extraction, in: PROPOR-2008, Lecture Notes in Computer Science, Springer-Verlag, 2008, pp. 41–50.

[21] O. Levy, Y. Goldberg, Dependency-based word embeddings, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, 2014, pp. 302–308.

[22] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, S. Petrov, Syntactic annotations for the google books ngram corpus, in: Proceedings of the ACL 2012 System Demonstrations, ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 169–174.
URL http://dl.acm.org/citation.cfm?id=2390470.2390499

[23] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, E. L. Aiden, Quantitative analysis of culture using millions of digitized books, Science 331 (6014) (2011) 176–182. arXiv:http://science.sciencemag.org/content/331/6014/176.full.pdf, doi:10.1126/science.1199644.
URL http://science.sciencemag.org/content/331/6014/176

[24] R. M. Samer Hassan, Cross-lingual semantic relatedness using encyclopedic knowledge, in: Proceedings of the conference on Empirical Methods in Natural Language Processing, Singapore, 2009.

[25] J. Camacho-Collados, M. T. Pilehvar, R. Navigli, A framework for the construction of monolingual and cross-lingual word similarity datasets, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers, 2015, pp. 1–7.
URL http://aclweb.org/anthology/P/P15/P15-2001.pdf

[26] H. Rubenstein, J. B. Goodenough, Contextual correlates of synonymy, Commun. ACM 8 (10) (1965) 627–633. doi:10.1145/365628.365657.
URL http://doi.acm.org/10.1145/365628.365657

[27] J. Camacho-Collados, M. Pilehvar, N. Collier, R. Navigli, Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity, in: Proceedings of SemEval, Vancouver, Canada, 2017.

[28] M. Etcheverry, D. Wonsever, Spanish word vectors from wikipedia, in: N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France, 2016, pp. 3681–3685.

[29] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, A. Soroa, A study on similarity and relatedness using distributional and wordnet-based approaches, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 19–27.
URL `http://dl.acm.org/citation.cfm?id=1620754.1620758`

[30] M. Baroni, B. Murphy, E. Barbu, M. Poesio, Strudel: A corpus-based semantic model based on properties and types, Cognitive Science 34 (2) (2010) 222–254. `doi:10.1111/j.1551-6709.2009.01068.x`.
URL `http://dx.doi.org/10.1111/j.1551-6709.2009.01068.x`

Short Bios

Pablo Gamallo is associate professor on Linguistics at University de Santiago de Compostela, research member of the CiTIUS center, and coordinator of LinguaKit, an open source project aimed at developing a multilingual suite of NLP tools. He is currently involved in NLP areas such as PoS tagging, dependency based parsing, relation extraction, sentiment analysis, bilingual extraction from comparable corpora, and so on.

Iván Rodríguez-Torres got his degree on Computer Engeniering at the Santiago de Compostela University in 2015. He is working as a research assistant in the CiTIUS and is finishing his master degree on Artificial Intelligence Research at the Meléndez Pelayo University.

Marcos Garcia obtained a PhD in Linguistics at the University of Santiago de Compostela in 2014, and works as a postdoctoral researcher in Natural Language Processing at the University of Corunha since 2016.