

# HSE at LSCDiscovery in Spanish: Clustering and Profiling for Lexical Semantic Change Discovery

Kseniia Kashleva Alexander Shein Elizaveta Tukhtina Svetlana Vydrina\*

HSE University

## Abstract

This paper describes the methods used for lexical semantic change discovery in Spanish. We tried the method based on BERT embeddings with clustering, the method based on grammatical profiles and the grammatical profiles method enhanced with permutation tests. BERT embeddings with clustering turned out to show the best results for both graded and binary semantic change detection outperforming the baseline.

## 1 Introduction

Lexical semantic change detection (LSCD) aims to identify whether the words change their meaning over time, or not. LSCD is usually divided into two subtasks: graded change discovery and binary change detection. Graded LSCD is a subtask of ranking the intersection of (content-word) vocabularies according to their degree of change between a diachronic corpus pair C1 and C2 (Kurtyigit et al., 2021). Binary LSCD is a subtask of identifying whether a target word lost or gained senses from the 1st set of its usage to the second, or not (Schlechtweg et al., 2020).

Previous shared tasks on lexical semantic change detection (LSCD) were developed for English, German, Latin, and Swedish (Schlechtweg et al., 2020), Italian (Basile et al., 2020), and Russian (Kutuzov and Pivovarova, 2021). This one was in Spanish (D. Zamora-Reina et al., 2022). Spanish is a fusional Romance language of the Indo-European language family with rich morphology and a lot of national varieties. So far, LSCD in shared tasks were developed for three Romance languages, three German languages, and one Slavic language. Only two of them are analytical (English and Swedish), while others are fusional.

In this shared task we tested several methods. For graded change discovery we used BERT embeddings with clustering (Montariol et al., 2021).

\*Equal contribution, the authors listed alphabetically.

For binary change detection we used 3 methods. The first one was word embeddings again. Two others were grammatical profiling (Kutuzov et al., 2021), and grammatical profiling combined with permutation tests (Liu et al., 2021).

Though grammatical profiles by themselves yield worse performance than embedding-based method, they could be significantly improved by applying of additional significance tests.

## 2 Methods

### 2.1 BERT embeddings method

For this method we used a base version of BERT with 12 attention layers and a hidden layer size of 768. The exact pre-trained model was the one for Spanish<sup>1</sup> (Devlin et al., 2019). All parameters were set to the default as in the Transformers library ver. 4.14.1 (Wolf et al., 2020).

The method consisted of several steps. First, we split the corpora into train and test sets. The train/test ratio was 90/10. We used the lemmatized version of the corpora in this method. Then we took the pre-trained BERT model for Spanish and ran a fine-tuning process on the train set of the corpora using the test set for evaluation. The code we used for fine-tuning is provided as one of the examples in the Transformers library repository<sup>2</sup>.

After fine-tuning the model we extracted the embeddings for the target words from the full corpora provided. The embeddings were extracted separately for two time periods. To generate a final embedding for each target word, the embeddings from all 12 attention layers of the BERT model were summarized. The embeddings for all entries of every target word were extracted this way.

As a result, we obtained two matrices for every

<sup>1</sup>We used the following model: <https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased>

<sup>2</sup><https://github.com/huggingface/transformers/tree/main/examples/pytorch/language-modeling>

target word. One matrix represented one time period. The dimension of the resulting matrix was  $N \times 768$ , where  $N$  is the number of occurrences of the target word in the corpus of particular time period.

The final step was clustering. We ran a  $k$ -means clustering algorithm on the rows of the resulting matrices. It should be noted that we also attempted to use the affinity propagation algorithm, but it proved unfeasible at this point, as the number of target words and the number of their embeddings was too large for the affinity propagation approach. So the final decision was to resort to the  $k$ -means algorithm which is much faster. The number of clusters was set as a hyperparameter which we tuned at the development phase. The development phase demonstrated that the results were the best when the number of clusters equaled to a multiple of 7 with the larger numbers showing better results. In order to find a balance between the clustering time and the results we decided that the number of clusters should be 28.

The resulting clusters presumably represented some gradations of word meanings. In order to calculate the graded change between the sets of clusters from two time periods, we used the average of the cosine distances between all pairs of the cluster centroids. The binary change was calculated by clustering the resulting graded changes into two clusters: the words that fall into the cluster with higher centroid value were considered as changed. The other words were considered as unchanged.

To detect binary gain/loss we took the cluster centroids calculated on one of the previous steps. Those centroids were clustered once again, but this time we used the affinity propagation method that determined the number of clusters automatically. The result clusters presumably represented the basic meanings of target words. After that we compared the number of resulting clusters for both time periods. If the number of clusters in the first period was larger than that in the second period, we assumed that this word lost a sense. If not, we assumed the word gained a sense.

As for the optional COMPARE task, our submission was identical to that for the main Graded task. We did not use any other method for that.

## 2.2 Grammatical profiling

All language aspects are strongly interconnected. It means that semantic changes may be tied with

grammatical changes. Diachronically, it can be observed through lexicalization and grammaticalization in particular. In Spanish, the modern usage of the verb *andar* 'to go' can be a good example of grammaticalization:

*De que Blasillo **ande** al escuela me e holgado mucho* (16th c.).

'Since Blasillo has been **going** to school, I have been very happy.'

– *¿Y eso es todo el problema? — **Ándale**, exactamente eso.* (21th c.)

'And that's the whole problem? **Yes, yes** (lit. walk to it), that's exactly it.' (Company Company, 2008)

So here we can see that this verb changed its meaning while changing its form.

The idea of grammatical profiling is that semantic change can be discovered through significant changes in the distribution of morphosyntactic categories. This method is described in (Kutuzov et al., 2021) in detail, so here we explain only the main points. To get grammatical profiles, the frequency of morphological and syntactic categories for each target word were counted in both corpora, that were in advance tagged and parsed with UDPipe (Straka and Straková, 2017)<sup>3</sup>. Then, for each target word and for both morphological and syntactic dictionaries, a list of features was created by taking the union of keys in the corresponding dictionaries for the two time bins. After that, feature vectors  $\vec{x}_1$  and  $\vec{x}_2$  were made. Each dimension of these vectors represented a grammatical category and the value it took was the frequency of that category in the corresponding time period (Kutuzov et al., 2021). Then, the cosine distance  $\cos(\vec{x}_1; \vec{x}_2)$  between the vectors were calculated to estimate the change in the grammatical profiles of the target word<sup>4</sup>. These distances can be used for graded change discovery. For binary detection, the top  $n$  target words were classified in the ranking as 'changed' (1) and others as 'stable' (0).

## 2.3 Grammatical profiling enhanced with permutation-based statistical tests

Earlier statistical significance tests were applied to semantic change detection methods based on contextual word embeddings (Liu et al., 2021). Permutation-based statistical testing can be applied when data is limited. We used permutation tests to improve the results obtained with grammatical

<sup>3</sup>We used the following model: spanish-gsd-ud-2.5-191206.udpipe

<sup>4</sup>The code is available at <https://github.com/glmario/semchange-profiling>

profiling, as the aim of the permutation test is to discover whether the observed test statistic (i.e. the cosine distance) is significantly different from zero (Liu et al., 2021). Permutation tests reassigned group labels (time periods) to all observations by sampling without replacement.

For binary change detection we calculated the default distance between grammar profiles. Then, we took sentence indices from the first and the second corpus for every target word and permute them by randomly splitting them between two time periods. If the number of possible permutations were less than 1000 we used all permutations. Then we calculated cosine distance between grammar profiles generated after shuffling. So, we have 2 sets of distances: the original cosine distance between grammar profiles and the permuted cosine distances between grammar profiles.

Let us assume, there were 5 permutations, so we got 5 distances, e.g., 0.1, 0.7, 0.4, 0.15, and 0.2, and the original cosine distance was 0.3. We took only those permuted cosine distances that were larger than the default cosine distance. In this example, these are 0.7 and 0.4 (two values). So, we divided the number of these larger permuted distances by the number of permutations. In this example, this is 2/5. This result is a p-value (Liu et al., 2021).

If the number of permutations were greater than 1000, the procedure was the same, but we corrected the p-value for every digit capacity, i.e., we took the first significance threshold as 0.05 and step-by-step reduced it till 0.005 (Liu et al., 2021). In other words, we first randomly selected 1000 permutations and computed p-value. If this was larger 0.05, we stopped the procedure, otherwise took more permutations for more precise estimations.

As a result, we had the cosine distance between grammar profiles and the p-value for every target word. For binary change detection we sorted these values both by the distance and the p-value and labeled top  $n$  target words as changed.

### 3 Results

The submission results are presented in Table 1.

Clustering turned out to be the best one among all our methods. In graded change discovery it was proved to be better than both baselines and took the 3rd place in the leaderboard.

Grammatical profiling demonstrated the worst results among three methods we used (see Table 1).

| Graded         |              |              |              |
|----------------|--------------|--------------|--------------|
|                | COMPARE      | Spearman     |              |
| Clusters       | 0.558        | 0.553        |              |
| Baseline       | 0.561        | 0.543        |              |
| <i>Grammar</i> | —            | 0.390        |              |
| Binary         |              |              |              |
|                | Precision    | Recall       | F1           |
| Clusters       | 0.567        | <b>0.607</b> | <b>0.586</b> |
| Grammar        | 0.714        | 0.357        | 0.476        |
| Stats          | <b>0.750</b> | 0.429        | 0.545        |
| Baseline       | 0.846        | 0.393        | 0.537        |
| Gain           |              |              |              |
|                | Precision    | Recall       | F1           |
| Clusters       | 0.192        | 0.357        | 0.250        |
| Baseline       | 0.400        | 0.143        | 0.211        |
| Loss           |              |              |              |
|                | Precision    | Recall       | F1           |
| Clusters       | 0.421        | 0.320        | 0.364        |
| Baseline       | 0            | 0            | 0            |

Table 1: Submission results: *Clusters* means embedding clustering method, *Grammar* means grammatical profiles and *Stats* means grammatical profiles combined with a permutation test. Grammatical profiling for graded discovery was made after the competition.

However, the results indicate that it was significantly improved by applying a permutation test. It should also be noted that grammatical profiling with a permutation test demonstrated the best precision among all participants and was only outperformed by the baseline. We also applied grammatical profiling for graded change discovery after the competition. The result was worse than baseline (see Table 1).

The clustering method was our only method that was applied to the optional Gain/Loss task, however, it did not show good results. While this method surpassed the baseline numbers, it proved to be significantly inferior to the other methods participating in the task. We assume that it happened because we approached the Gain/Loss task as a separate task. The better approach might have been to somehow use the results we received on the main Binary task in order to calculate the gain/loss values.

There is another problem with the method that we can think of. The method assigned a gain/loss label for the word if the number of clusters in two time epochs differs even by one. Perhaps a better approach would have been to decrease the sensi-

| word          | change graded | change graded golden | change graded difference |
|---------------|---------------|----------------------|--------------------------|
| actitud       | 0.369         | 0.925                | 0.556                    |
| propiamente   | 0.473         | 0                    | 0.473                    |
| fallecimiento | 0.468         | 0                    | 0.468                    |
| viernes       | 0.447         | 0                    | 0.447                    |
| trato         | 0.490         | 0.051                | 0.439                    |
| distribuir    | 0.438         | 0                    | 0.438                    |
| banco         | 0.514         | 0.925                | 0.411                    |
| canal         | 0.607         | 1                    | 0.393                    |
| variedad      | 0.392         | 0                    | 0.392                    |
| socialista    | 0.391         | 0                    | 0.391                    |

Table 2: BERT-based predictions compared with the gold standard.

tivity of the method and to ignore the insignificant differences between the number of clusters.

## 4 Discussion

Table 2 presents the top 10 words with the largest difference between BERT-based predictions and the gold standard. Closer inspection shows that there are two error types. According to the standard, some words (*actitud*, *banco*) changed a lot, while our prediction for these words appeared to be much lower. Meanwhile, there were words that did not change, however, our model labeled them as changed (*propiamente*, *fallecimiento*, *viernes*, *distribuir*, *variedad*, *socialista*). Interestingly, that within the top 10 words, the model fell into errors on the side of changing more often.

Table 3 presents the top 10 words with the largest difference between grammatical profiling predictions and the gold standard. Our prediction for these words was much lower than the gold standard. Some incorrect predictions are the same with the incorrect predictions obtained with the BERT-based method (*actitud*, *canal*, *banco*). A likely explanation is that these words have a complicated semantic structure and more than one meaning.

## 5 Conclusion

Further studies need to be carried out in order to evaluate the combination of profiling with statistical significance testing for other languages. Although the BERT-based method demonstrated the best results, more detailed error analysis is still required.

| word       | change graded | change graded golden | change graded difference |
|------------|---------------|----------------------|--------------------------|
| marco      | 0.018         | 1                    | 0.982                    |
| prima      | 0.118         | 1                    | 0.882                    |
| actitud    | 0.115         | 0.925                | 0.810                    |
| indicativo | 0.202         | 1                    | 0.798                    |
| canal      | 0.240         | 1                    | 0.760                    |
| disco      | 0.167         | 0.915                | 0.748                    |
| pendiente  | 0.096         | 0.781                | 0.685                    |
| corriente  | 0.072         | 0.753                | 0.681                    |
| banco      | 0.246         | 0.925                | 0.678                    |
| cólera     | 0.098         | 0.741                | 0.643                    |

Table 3: Grammatical profiles predictions compared with the gold standard.

## Acknowledgements

This research was supported in part through computational resources of HPC facilities at HSE University (Kostenetskiy et al., 2021).

## References

- Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. *Diacrita @ evalita2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task*. In *EVALITA*.
- Concepción Company Company. 2008. *The directionality of grammaticalization in spanish*. *Journal of Historical Pragmatics*, 9(2):200–224.
- Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. *Lscdiscovery: A shared task on semantic change discovery and detection in spanish*. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- P. S. Kostenetskiy, R. A. Chulkevich, and V. I. Kozyrev. 2021. *HPC resources of the higher school of economics*. *Journal of Physics: Conference Series*, 1740(1):012050.

- Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. [Lexical semantic change discovery](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6985–6998, Online. Association for Computational Linguistics.
- Andrey Kutuzov and Lidia Pivovarova. 2021. [Three-part diachronic semantic change dataset for Russian](#). In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.
- Andrey Kutuzov, Lidia Pivovarova, and Mario Giulianelli. 2021. [Grammatical profiling for semantic change detection](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 423–434, Online. Association for Computational Linguistics.
- Yang Liu, Alan Medlar, and Dorota Glowacka. 2021. [Statistically significant detection of semantic shifts using contextual word embeddings](#). In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 104–113, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. [Scalable and interpretable semantic change detection](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [SemEval-2020 task 1: Unsupervised lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.
- Milan Straka and Jana Straková. 2017. [Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe](#). In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.