LSCDiscovery: A shared task on semantic change discovery and detection in Spanish

Frank D. Zamora-Reina[1], **Felipe Bravo-Marquez**[1], **Dominik Schlechtweg**[2]
[1]Department of Computer Science, University of Chile, IMFD & CENIA
[2]Institute for Natural Language Processing, University of Stuttgart
fzamora@dcc.uchile.cl, fbravo@dcc.uchile.cl,
schlecdk@ims.uni-stuttgart.de

Abstract

We present the first shared task on semantic change discovery and detection in Spanish and create the first dataset of Spanish words manually annotated for semantic change using the DURel framework (Schlechtweg et al., 2018). The task is divided in two phases: 1) Graded Change Discovery, and 2) Binary Change Detection. In addition to introducing a new language the main novelty with respect to the previous tasks consists in predicting and evaluating changes for all vocabulary words in the corpus. Six teams participated in phase 1 and seven teams in phase 2 of the shared task, and the best system obtained a Spearman rank correlation of 0.735 for phase 1 and an F1 score of 0.716 for phase 2. We describe the systems developed by the competing teams, highlighting the techniques that were particularly useful and discuss the limits of these approaches.

1 Introduction

Lexical Semantic Change Detection (LSCD) is the task of detecting words which have changed their meaning over time in a diachronic corpus of text (Schlechtweg et al., 2020), usually an unsupervised task. In recent years, several LSCD shared tasks have been organized (Schlechtweg et al., 2020; Basile et al., 2020; Kutuzov and Pivovarova, 2021). These tasks have contributed to a better understanding of LSCD, but have also had their shortcomings: (i) they have used mainly small pre-selected sets of target words creating an unrealistic evaluation scenario for the application of computational models in historical semantics and lexicography where researchers typically aim to cover the full vocabulary of a language (Kurtyigit et al., 2021), (ii) different formalizations of the LSCD task have been proposed including binary classification and ranking tasks (Schlechtweg et al., 2018; Schlechtweg and Schulte im Walde, 2020; Schlechtweg, 2022) and these have been employed inconsistently, and (iii) none of them have focused on Spanish, despite the fact that there are more than 450 million native speakers of this language.

We tackle these shortcomings by organizing a shared task on Spanish diachronic data with a more realistic evaluation scenario requiring participants to provide Lexical Semantic Change (LSC) predictions for the full corpus vocabulary (Discovery). Additionally, we cover previous scenarios by asking participants to predict LSC only in the limited sample of annotated target words (Detection). By offering a range of additional optional tasks (defined on the same annotated data) participants are able to evaluate and compare models on various formalizations of the LSCD task. In order to derive gold LSC labels for target words, we annotate and publish the largest existing data set of semantic proximity judgments covering 100 words with approximately 62k judgments from 12 human native speakers.[1]

2 Related Work

The detection of lexical semantic changes is of great interest in research areas such as historical semantics, lexicography, linguistics and NLP. For a comprehensive review of the literature on the area we refer the reader to the recent surveys (Tahmasebi et al., 2021; Kutuzov et al., 2018; Hengchen et al., 2021). In previous years several shared tasks have been organized: SemEval-2020 Task 1 (Schlechtweg et al., 2020) for English, German, Latin, and Swedish, DIACR-Ita for Italian (Basile et al., 2020), and RuShiftEval for Russian (Kutuzov and Pivovarova, 2021).[2] All shared tasks applied an evaluation setup where LSC was measured between pairs of time periods.

SemEval used a total of 156 target words for all languages with no development/test split. Ap-

[1]The data set is available at https://zenodo.org/record/6300104.

[2]There was also a student shared task on German data (Ahmad et al., 2020).

proximately half of these were drawn from etymological dictionaries or research literature, while the other half was drawn from the corpus vocabularies by selecting lemmas with similar POS and frequency as the first half of target words. Target word occurrences in sentences (usages) were combined into pairs and these were annotated for their semantic proximity (Schlechtweg et al., 2021). Target words were excluded if they had a high number of undecidable use pairs or were annotated too sparsely. Sense clusters were inferred from the annotation. From the clusters a binary (sense loss/gain vs. none) and a graded (Jensen-Shannon distance between cluster distributions) change score were derived and used to evaluate participants on a corresponding binary classification and ranking task.

DIACR-Ita used a total of 18 target words with no development/test split. All of these were drawn from an etymological dictionary. Target word usages were annotated with word sense definitions. Words with a high number of OCR errors and annotator disagreements were excluded. From the annotation Binary Change scores similar to SemEval were derived and used to evaluate participants on a binary classification task.

RuShiftEval used a total of 111 target words (all nouns) split into 12 for development and 99 for testing. These were selected in a similar procedure to SemEval: approximately half of these were drawn from etymological dictionaries, research literature or "invented" by the authors, while the other half was drawn from the corpus vocabularies by selecting lemmas with similar POS and frequency as the first half of target words. Target word usages from different time periods were combined into usage pairs and annotated for semantic proximity. From these the DURel COMPARE score (see Subsection 3.3 for more details) (Schlechtweg et al., 2018) was derived, which can be seen as an approximation of SemEval's Graded Change score (Schlechtweg, 2022). Participants were evaluated in a ranking task on the COMPARE scores.

As we can see, target words in previous shared tasks have been strongly preselected and systems have been evaluated on different tasks. They have also yielded (seemingly) contradictory results: while type-based model architectures have dominated in SemEval and DIACR-Ita, token-based architectures have dominated in RuShiftEval. In all tasks clustering-based models have shown rather low performance.

3 Task description

Our task was designed in two phases:

1. Graded Change Discovery, and
2. Binary Change Detection.

Note that *discovery* introduces additional difficulties for models as compared to the more simple semantic change *detection*, e.g. because a large number of predictions is required and the target words are not preselected, balanced or cleaned (cf. Kurtyigit et al., 2021). Yet, discovery is an important task, with applications such as lexicography where dictionary makers aim to cover the full vocabulary of a language.

3.1 Phase 1: Graded Change Discovery

Similar to Kurtyigit et al. (2021), we define the task of **Graded Change Discovery** as follows:

> Given a diachronic corpus pair C_1 and C_2, rank the intersection of their (content-word) vocabularies according to their degree of change between C_1 and C_2.

The participants were asked to rank the set of content words in the lemma vocabulary intersection of C_1 and C_2 according to their degree of semantic change between C_1 and C_2 where a higher rank means stronger change. The true degree of semantic change of a target word w was given by the Jensen-Shannon distance (Lin, 1991; Donoso and Sanchez, 2017) between w's word sense frequency distributions in C_1 and C_2 (cf. Schlechtweg et al., 2020). The two word sense frequency distributions were estimated via human annotation of word usage samples for w from C_1 and C_2 (see Subsection 4.4). Participants' predictions were *not* evaluated on the full set of target words, as this would be unfeasible to annotate, but on an (unpublished) random sample of words from the full set of target words. The predictions were scored against the ground truth via Spearman's rank-order correlation coefficient (Bolboaca and Jäntschi, 2006).

3.2 Phase 2: Binary Change Detection

Similar to Schlechtweg et al. (2020), we define the task of **Binary Change Detection** as follows:

Given a target word w and two sets of its usages U_1 and U_2, decide whether w lost or gained senses from U_1 to U_2, or not.

The participants were asked to classify a preselected set of content words into two classes, 0 for no change and 1 for change. The true binary labels of word *w* were inferred from *w*'s word sense frequency distributions in $C1$ and $C2$ (see Subsection 3.1). Participants' predictions were scored against the ground truth with the following metrics: F1 (main metric), Precision, and Recall. A crucial difference compared to Graded Change Discovery was that the public target words corresponded exactly to the hidden words on which we evaluated. Also, we published the usages sampled for annotation. Hence, participants could work with the exact annotated data, which was not possible in the first phase where participants could only work with the full corpora (from which the usages for annotation were sampled).

3.3 Optional tasks

Participants could submit predictions for several optional tasks:

Graded Change Detection was defined similar to Graded Discovery. The only difference was that the public target words corresponded exactly to the hidden words on which we evaluated. Participants were scored with Spearman correlation.

Sense Gain Detection was similar to Binary Change Detection. However, only words which gained (not lost) senses receive label 1. Participants were scored with F1, Precision and Recall.

Sense Loss Detection was similar to Binary Change Detection. However, only words which lost (not gained) senses received label 1. Participants were scored with F1, Precision and Recall.

COMPARE asked participants to predict the negated DURel COMPARE metric (Schlechtweg et al., 2018). This metric is defined as the average of human semantic proximity judgments of usage pairs for *w* between $C1$ and $C2$.[3] It can be seen as an approximation of JSD (Graded Change) (Schlechtweg, 2022). Participants were scored with Spearman correlation.

[3]Contrary to the original metric we first take the median of all annotator judgments for each usage pair and then average these values. For details see: https://github.com/Garrafao/WUGs.

Corpus	Time period	Tokens
Old corpus (C_1)	1810–1906	$\sim 13M$
Modern corpus (C_2)	1994–2020	$\sim 22M$

Table 1: Sizes of both corpora.

Participants' submission files only needed to include predictions corresponding to the obligatory tasks in order to get a valid submission. They did not see the leaderboard while the evaluation phases were running. Furthermore, participants only had three valid submissions for each evaluation phase.[4]

4 Data

In this section, we describe the corpora, the selection process of target words, the sampling of usages and their annotation. Moreover, we explain how the target words were presented to the participants considering the two phases of the shared task.

4.1 Corpora

We created two corpora covering disjoint time periods: 1810 to 1906 (old corpus, $C1$) and 1994 to 2020 (modern corpus, $C2$) (see Table 1). The former was created using different sources freely available from Project Gutenberg[5] and the latter using different sources available from the OPUS project[6] (Tiedemann, 2012). For the old corpus, all the sources collected were concatenated. As for the modern corpus, four datasets were used: Spanish portion of TED2013, Spanish portion of News-Commentary v16, Spanish portion of MultiUN and Spanish version of Europarl corpus. TED2013 was used in its entirety, while 50 snippets with 5000 lines each were extracted from the other datasets by cutting the corpora into snippets of the mentioned size and randomly choosing 50 of them.

Both corpora were parsed using spaCy (Honnibal et al., 2020).[7] Each corpus contains four

[4]We decided not to include the binary subtasks in phase 1, as the usage samples were not published which meant that participants needed to work with the full corpora instead of the samples on which the gold scores were inferred. We assumed that the sampling error between usages in the full corpora and our samples is much larger for Binary Change than for Graded Change (cf. Schlechtweg, 2022).

[5]https://www.gutenberg.org/browse/languages/es

[6]https://opus.nlpl.eu/

[7]Find details issues in Appendix A.

versions of the original dataset (raw, tokenized, lemmatized and POS-tagged).

4.2 Target words

4.2.1 Phase 1 (Graded Discovery)

Public target words was a list of 4385 words created in the following way: we first took the corpus vocabulary intersection from the lemmatized versions of both corpora. Then we removed words below a minimum frequency threshold of 40 for the old corpus and 73 for the modern corpus.[8] Then we removed all non-content words, i.e., we left only nouns, verbs, adjectives and adverbs. The final list of target words was published and participants were required to submit results for all 4385 words in the development and evaluation phase 1.

Hidden target words The large number of public target words was crucial to our task. However, it was not feasible to annotate all of them. Hence, we only annotated a subset of the public target words for semantic change. Participants' predictions for development and evaluation phase 1 were evaluated only on this subset of target words, which remained hidden from the participants. We selected the hidden target words in the following way: Initially, a list of 15 changing words was selected by scanning etymological dictionaries and consulting with a linguistic specialist to obtain words for changes from $C1$ to $C2$. Likewise, it was verified that these words were in both corpora. Additionally, a list of 85 words were randomly sampled from the public target words. The $85 + 15 = 100$ words were annotated as described in Section 4.4. Then, 20 words were excluded based on inter-annotator agreement.[9] The remaining set of 80 target words were split randomly into two groups, 20 words for the development set and 60 for the evaluation set (see Table 3). Uploaded submissions were scored against these 20/60 annotated words during development/evaluation phases.

4.2.2 Phase 2 (Binary Detection)

The target words corresponded to the 20/60 hidden words from Phase 1 for development/evaluation. There it was no distinction here between public and hidden target words. Participants also got access to the annotated usages (20+20 from each corpus). Uploaded submissions were scored against the 20/60 public annotated words.

[8] 40 was chosen by us for the old corpus and then we calculated 73 for the new corpus to reflect the same proportion of the frequency threshold to corpus size.

[9] We removed target words with agreements of less than 0.3 Krippendorf's α and less than 0.3 on a version of Krippendorf's α where expected disagreements were calculated from the full annotated data (instead of for each word separately). The latter measure is less sensitive to skewed judgment distributions for individual words.

↑	4: Identical
	3: Closely Related
	2: Distantly Related
	1: Unrelated

Table 2: DURel relatedness scale (Schlechtweg et al., 2018).

4.3 Word usages

All occurrences of the target words per corpus were extracted according to the lemma. Then, 20 usages were randomly sampled per target word from each corpus.

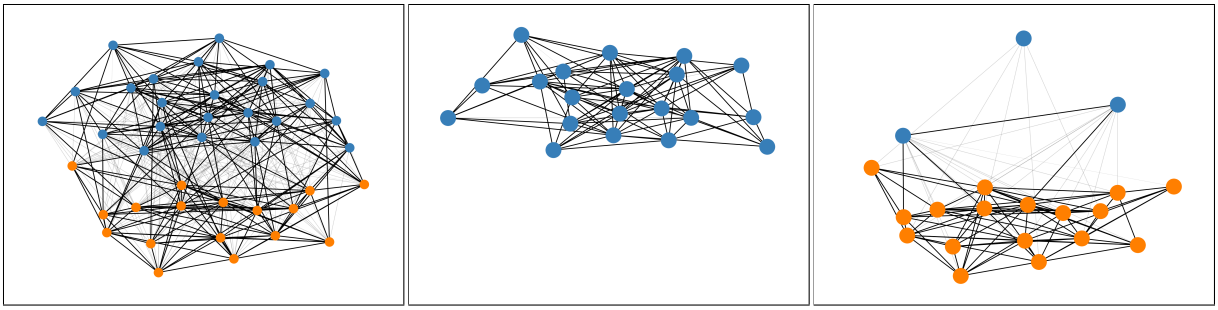
4.4 Annotation

We applied the SemEval procedure to annotate target word usages, as described in Schlechtweg et al. (2020, 2021). Annotators were asked to judge the semantic relatedness of pairs of word usages, such as the two usages of *servidor* in (1) and (2), on the scale in Table 2.

(1) Todo esto lo hago con mi iPhone; se va derecho al **servidor**, allí se hace el trabajo de archivo, clasificación y ensamble.
'*I do all this with my iPhone; it goes straight to the **server**, there the work of archiving, sorting and assembling is done.*'

(2) Llamó a grandes voces a sus **servidores**, y únicamente le contestó el eco en aquellas inmensas soledades, y se arrancó los cabellos y se mesó las barbas, presa de la más espantosa desesperación.
'*He called out to his **servants**, and only the echo in those immense solitudes answered him, and he pulled out his hair and ruffled his beard, prey to the most frightening desperation.*'

The annotated data of a word was represented in a Word Usage Graph (WUG), where vertices represented word usages, and weights on edges represented the (median) semantic relatedness judgment of a pair of usages such as (1) and (2). The final



$G, D = (23, 17)$ $G_1, D_1 = (20, 0)$ $G_2, D_1 = (3, 17)$

Figure 1: Word Usage Graph *servidor* (left), subgraphs for old corpus G_1 (middle) and for modern corpus G_2 (right). The colors correspond to the clusters. **black**/gray lines indicate **high**/low edge weights.

WUGs were clustered with correlation clustering (Bansal et al., 2004; Schlechtweg et al., 2020, 2021) (see Figure 1, left) and split into two subgraphs G_1 and G_2 representing nodes from subcorpora C_1 and C_2 respectively (middle and right). Clusters were then interpreted as word senses and changes in clusters over time as lexical semantic change.[10]

In contrast to Schlechtweg et al., we used the openly available DURel interface for annotation and visualization.[11] This also implied a change in sampling procedure, as the system implemented only random sampling of usage pairs (without SemEval-style optimization, i.e., sampling in rounds with connection of clusters). For each target word we sampled $|U_1| = |U_2| = 20$ usages (sentences) per subcorpus (C_1, C_2) and uploaded these to the DURel system, which presented usage pairs to annotators in randomized order. We recruited twelve Spanish native speakers (4 Chileans, 4 Colombians, 2 Cubans, 1 Spaniard and 1 Venezuelan). All had university level education, while seven had a background in linguistics of which two had one in historical linguistics. We monitored agreement between annotators during the annotation process and discussed some strong annotation disagreements with certain annotators. This led to the exclusion of one annotator early in the process who often completely inverted the annotation scale (e.g. judged 1 while agreeing that the two usages have identical meanings).

Similar to Schlechtweg et al. (2020), we ensured the robustness of the obtained clusterings by continuing the annotation of a target word until all clusters in its WUG were connected by at least one judgment.[12] For 16 words the annotation had to be stopped before this condition was met. We manually inspected the unconnected clusters of some words and concluded that missing connections did not lead to clustering errors.

We finally labeled a target word as Binary Change if it gained or lost a cluster over time. For instance, *servidor* in Figure 1 was labeled as change as it gained the orange cluster from C_1 to C_2. Consequently, *servidor* was also labeled as gaining a sense; but not as losing a sense, since the blue cluster persists. Graded Change was defined as the Jensen-Shannon distance between the normalized cluster frequency distributions D_1 and D_2 yielding a high value of 0.82 (ranges between 0.0 and 1.0) for *servidor*, as sense probabilities changed drastically. The negated COMPARE score was derived by averaging over all graph edges with nodes from different time periods and negating this value, yielding a high score of -1.97 (ranges between -4.0 and -1.0) for *servidor*.[13] Following Schlechtweg et al. (2020) we used k and n as lower frequency thresholds for the binary notions to avoid that small random fluctuations in sense frequencies caused by sampling variability or annotation error were misclassified as change. As proposed in Schlechtweg and Schulte im Walde (submitted) for comparability across sample sizes we set $k = 1 \leq 0.01 * |U_i| \leq 3$ and $n = 3 \leq 0.1 * |U_i| \leq 5$, where $|U_i|$ was the number of usages from the respective time period.[14]

[10] We used Schlechtweg et al. (2020, 2021)'s code provided at https://www.ims.uni-stuttgart.de/data/wugs.

[11] https://www.ims.uni-stuttgart.de/data/durel-tool.

[12] Note that this condition was more strict than Schlechtweg et al. (2020)'s where only connection of multi-clusters (clusters with more than one usage) was guaranteed. Their condition was always met in our data.

[13] Find a more detailed discussion of different change scores in Schlechtweg et al. (2020) and Schlechtweg (2022).

[14] That is, k was always between 1 and 3. There are three possible cases: $k = 1$ if $0.01 * |U_i| \leq 1$, $k = 0.01 * |U_i|$ if $1 < 0.01 * |U_i| < 3$, $k = 3$ if $0.01 * |U_i| \geq 3$. Similarly for

This resulted in $k = 1$ and $n = 3$ for all target words.

Find an overview over the final set of WUGs in Table 3. We reached an inter-annotator agreement of Krippendorff's $\alpha = .53$ and Spearman's $\rho = .57$ which was comparable to previous studies (e.g. Schlechtweg et al., 2018; Rodina and Kutuzov, 2020; Kurtyigit et al., 2021; Baldissin et al., 2022).[15]

5 Systems

We now summarize the baseline systems as well as the systems and resources used by the participating teams.

5.1 Baselines

For both phases we use five baselines:

baseline1 *Skip-Gram with Negative Sampling + Orthogonal Procrustes + Cosine Distance (SGNS+OP+CD)* This approach learned vector representations for each word (type-based) in two input corpora with a shallow neural language model (Mikolov et al., 2013a,b).[16] These were then aligned using Orthogonal Procrustes (Hamilton et al., 2016). For phase 1, the method computed Graded Change as the cosine distance between old and modern vectors for all words in the vocabulary. This same value was used in the COMPARE subtask. In phase 2, binary predictions were computed by setting a threshold to the cosine distances, which was calculated as the sum between the mean and the standard deviation (std) of all these distances (Kaiser et al., 2020b). All words with values above the threshold were classified as *change*, and values below were classified as *no change*. This approach has shown high performance in several previous studies and shared tasks (Schlechtweg et al., 2019; Pömsl and Lyapin, 2020; Kaiser et al., 2020b; Pražák et al., 2020).

baseline2 *Normalized Log-Transformed Frequency Difference (FD)* For phase 1, this method calculated the frequency of each target word in each of the two corpora, normalized it by the logarithm of the total corpus frequency and then calculated absolute differences between these values as a measure of change. We submitted these values for the change graded and COMPARE subtasks. For phase 2, the method applied the same thresholding approach used in baseline1. For the sense loss subtask, it first verified that the target word presents change using the value of the change binary subtask. Then, if the differences were negative, the words were classified as loss $= 1$ and as loss $= 0$ otherwise. For sense gain the labeling is reversed.

baseline3 *Grammatical profiles* were generated from tagged and parsed corpora (Kutuzov et al., 2021). These profiles were essentially frequency vectors of various morphological and syntactic features (for example, *case* = *Nominative*, or *syntax role* = *subject*) for a given word in a given historical corpus. The cosine distance between the profile vectors of the same word for the two periods was used as an estimate of graded semantic change and COMPARE. Binary predictions were generated from ordered lists of graded scores for all target words by applying an offline change-point detection algorithm based on dynamic programming. The algorithm finds a point (a word) in an ordered list of scores, where the scores become significantly higher. This word and all words with score values above it were assigned the class "changed". This baseline did not produce predictions for the sense loss and sense gain subtasks.[17]

baseline4 *Minority class* This baseline produced predictions by labeling each word with the minority class label of the respective Binary Change score (change binary, loss, gain). This is label 1 (change) in all cases. It only applied to phase 2.

baseline5 *Random baseline* This baselines produced random predictions for all subtasks in both phases. For phase 1, we generated random values between 0 and 1 from a uniform distribution for all hidden target words and computed Spearman correlation with the gold scores. This process was repeated 100 times and we reported the average performance over all repetitions. For phase 2, we used a parallel procedure generating random labels $\in \{0, 1\}$ from a uniform distribution.[18]

n.

[15] We provide WUGs as Python NetworkX graphs, descriptive statistics, inferred clusterings, change values and interactive visualizations for all target words and the respective code at https://www.ims.uni-stuttgart.de/data/wugs (DWUG ES).

[16] As parameters we chose dim=100, window size=10, epochs=5, number of negative samples=5, subsampling threshold=0.001 (cf. Kaiser et al., 2020a).

[17] All results for baseline3 were computed and submitted by Andrey Kutuzov using the code at https://github.com/glnmario/semchange-profiling.

[18] Baseline3, baseline4 and baseline5 were added after the shared task finished.

Data set	n	N/V/A	|U|	AN	JUD	AV	KRI	SPR	UNC	LOSS	LSC_B	LSC_G
development	20	13/4/3	40	10	12k	40	.53	.59	0	.53	.55	.39
evaluation	60	30/14/16	40	12	38k	40	.58	.60	0	.45	.47	.37
discarded	20	8/6/6	40	12	12k	40	.27	.33	0	.52	.30	.18
full	100	51/24/25	40	12	62k	40	.53	.57	0	.48	.45	.34

Table 3: Overview target words. n = no. of target words, N/V/A = no. of nouns/verbs/adjectives+adverbs, $|U|$ = avg. no. of usages per word, AN = no. of annotators, JUD = total no. of judged usage pairs, AV = avg. no. of judgments per usage pair, KRI = Krippendorff's α, SPR = weighted mean of pairwise Spearman, UNC = avg. no. of uncompared multi-cluster combinations, LOSS = avg. of normalized clustering loss * 10, $LSC_{B/G}$ = mean binary/Graded Change score.

5.2 Participating systems

Below we present a summary of the methods developed by the participants:[19]

HSE *(Kashleva et al., 2022)* This team participated with two different methods. The first consisted of fine-tuning BERT (Devlin et al., 2019) on the lemmatized versions of the corpora in order to extract embeddings of the target words separately for each period, which are then clustered using K-means. Graded Change was estimated as the average cosine distance between all pairs of cluster centroids in the first and second periods. In order to estimate Binary Change, the Graded Change scores were thresholded by clustering them into two clusters.

The second method was based on grammatical profiles (Kutuzov et al., 2021). The frequency of morphological and syntactic categories for each target word in both corpora (parsed with UdPipe, Straka and Straková, 2017) were counted and used as features in two time-specific vectors. Graded Change was measured by the cosine distance between these vectors, while Binary Change was measured by thresholding the graded scores.

GlossReader *(Rachinskiy and Arefyev, 2022)* This system fine-tuned the XLM-R multilingual language model (Conneau et al., 2019) as part of a gloss-based Word Sense Disambiguation (WSD) system on a large English WSD dataset. It employed zero-shot cross-lingual transferability to build contextualized embeddings for Spanish data. The Graded Change score for each word was calculated as the Average Pairwise (Manhattan) Distance (APD) between the embeddings for (non-preprocessed) word usages in the old and new corpus. Binary changes were estimated by thresholding these scores. For the sense gain and sense loss subtasks the same predictions were reused.

UAlberta *(Teodorescu et al., 2022)* This team applied different methods to the two subtasks. For Graded Change Discovery, they followed the design of CIRCE (Pömsl and Lyapin, 2020) and computed distances based on both static (type-based) and contextual (token-based) embeddings, with their relative weights tuned on the development set. For static embeddings, they used SGNS+OP+Euclidean Distance on the lemmatized versions of the corpora. For contextual embeddings, the XLM-R model was trained on the combined corpus (tokenized) to predict masked instances of the target words and Graded Change was measured using Euclidean APD. For Binary Change Detection, they framed the task as a WSD problem, creating sense frequency distributions for target words in the old and modern corpus with an end-to-end WSD system (Orlando et al., 2021). It was assumed that the word semantics has changed if: (1) a sense is observed in the modern corpus but not in the old corpus (or vice versa), or (2) the relative change for any sense exceeds a tuned threshold.

CoToHiLi *(Sabina Uban et al., 2022)* This team proposed a type-based embedding model combined with hand-crafted linguistic features. The system computed several features for every target word based on embedding distances between time periods and linguistic hand-crafted features, which were then weighted into an ensemble model to predict the final score. First, the system obtained word embeddings separately on the two corpora (tokenized) with the Continuos Bag-of-Words (CBOW) model (Mikolov et al., 2013a,b), which were then

[19] The descriptions are based on the system description papers submitted by the participating teams, with the exception of Rombek who did not provide a paper but gave us a brief description by e-mail.

aligned to obtain a common embedding space. The alignment algorithms used were: supervised alignment using a seed word dictionary and a linear mapping method, a semi-supervised algorithm and unsupervised alignment based on adversarial training (Artetxe et al., 2016, 2017, 2018a,b). Finally, cosine distance between embeddings of the same word in different corpora was used as an indicator of graded semantic change. For the binary task, the system used thresholding the graded scores.

DeepMistake *(Homskiy and Arefyev, 2022)* This team employed a Word-in-Context (WiC) model, i.e., a model designed to determine if a particular word has the same meaning in two given contexts. In essence, they attempted to directly apply a model trained on a related task to our problem. The WiC model was initially trained by fine-tuning the XLM-R language model on the Multilingual and Cross-lingual Word-in-Context (MLC-WiC) dataset (Martelli et al., 2021). Subsequently, it was further fine-tuned on the provided annotations for the development set in this shared task and on the Spanish portion of the multi-language XL-WSD dataset (Pasini et al., 2021). Graded Change was measured similarly to APD by averaging same-sense probabilities between embeddings for usages (no preprocessing) from different time periods. For the change binary subtask, the authors applied thresholding to the Graded Change scores, for the sense gain and sense loss subtasks the same predictions were reused.

They also experimented with clustering by representing word usages and their same-sense probabilities in a weighted undirected graph, which was then clustered with Correlation Clustering. Graded Change was measured with JSD, while Binary Change was measured with the Binary Change score definition from Section 4.4.

BOS *(Kudisov and Arefyev, 2022)* The system described by this team was based on generating lexical substitutes that describe old and new senses of a given word. These were generated using the XLM-R masked language model. For polysemous words, lexical substitutes depended on the meaning expressed in a particular context. For each target word, usages were sampled from both corpora, lemmatized and used to generate lexical substitutes. Next, two sets of vectors were built for old and new usages where each usage is represented by a vector of the probabilities of its substitutes. For Graded Change the Cosine APD between old and new vectors was computed, while for Binary Change a threshold was applied to this score. The authors also proposed three different approaches based on pairwise distances for the sense gain and loss subtasks.

Rombek This system adapted ideas from the Word Sense Induction (WSI) task. Lexical substitutes were generated in the same way as with the BOS system (see above) and arranged in a matrix. Agglomerative clustering was then applied to each target word to obtain clusters with candidate senses. JSD was applied between clusters to obtain Graded Change estimates. Thresholding was applied to produce binary predictions.[20]

5.3 Summary

Most systems were based on three main components: (i) a semantic representation of words or word usages as vectors, (ii) an aggregation method over vectors, and (iii) a change measure. Type-based systems usually employed an additional alignment step over semantic representations. Also, the preprocessing of data was crucial for the performance of contextualized embeddings (Laicher et al., 2021).

Preprocessing Some teams only used the tokenized version of the shared task dataset (CoTo-HiLi, UAlberta), while other teams only used the lemmatized version (UAlberta, BOS, HSE). One team varied the preprocessings with systems (UAlberta): lemmatization for type-based embeddings and tokenization, lemmatization and POS-tagging for the WSD system. Two teams did not use any sort of preprocessing (GlossReader, DeepMistake), while two teams used substitution with dynamic patterns (e.g. *<mask> (y [target])*, *[target] (por ejemplo <mask>)*) for their lexical substitution models (BOS, Rombek).

Semantic representations Most systems used token-based contextualized embeddings such as BERT (HSE) and XLM-R (DeepMistake, GlossReader, Rombek, UAlberta, BOS). Some teams further fine-tuned these embeddings on Language Modeling, WSD or WSI/WiC tasks. One team (DeepMistake) fine-tuned on the semantic proximity judgments from the published development data. Only three teams used type-based semantic rep-

[20] This team did not submit a paper to the shared task.

resentations including SGNS (UAlberta), CBOW (CoToHiLi) and Grammatical Profiling (HSE).

Vector aggregation Participating teams used different approaches to aggregate vectors into more abstract semantic representations. A common strategy was to model the COMPARE score by computing Average Pairwise Distances (APD) between vectors from different time periods (DeepMistake, GlossReader, UAlberta, BOS). This strategy has shown to perform well in various previous studies and shared tasks (Kutuzov and Giulianelli, 2020; Laicher et al., 2021; Kurtyigit et al., 2021; Arefyev et al., 2021). Another strategy was to cluster the vectors (HSE, Rombek, DeepMistake). Clustering algorithms used are: Agglomerative Clustering (Rombek), K-means (HSE) and Correlation Clustering (DeepMistake). One system used a WSD system to assign cluster labels (UAlberta).

Change Measure For Graded Change most teams using contextualized embeddings directly relied on APD scores as described above. They used different distance measures such as: Cosine (BOS), Euclidean (UAlberta) and Manhattan (GlossReader) distances. One team averaged same-sense probabilities (DeepMistake). The teams relying on clustering mostly used the JSD to measure Graded Change (Rombek, DeepMistake). One team instead used cosine distance between cluster centroids (HSE). The teams relying on type-based representations used either Cosine (CoToHiLi, HSE) or Euclidean distance (UAlberta). For Binary Change most teams relied on thresholding the graded predictions (DeepMistake, GlossReader, Rombek, HSE, CoToHiLi, BOS). This strategy has shown high performance in several previous studies and shared tasks (Schlechtweg et al., 2020; Kaiser et al., 2020b; Kurtyigit et al., 2021). Two teams using a clustering approach measured Binary Change by applying exactly the definition from the annotation process (DeepMistake) or a similar definition (UAlberta).

6 Results

The results shown in Tables 4, 5 and 6 correspond to the best submissions per subtask.[21]

Graded Change Discovery As shown in Table 4, **GlossReader** and **DeepMistake** obtained first and second place in the main task of evaluation phase 1, while **HSE** came third.[22] These were the only teams that managed to outperform baseline1 (SGNS+OP+CD) and baseline3 (Grammatical Profiles). The three winning systems were based on fine-tuned versions of contextualized embeddings with average vector aggregation (GlossReader, DeepMistake) or clustering (HSE). Interestingly, the top two systems did not model the JSD between cluster distributions (as done on the annotation to derive gold scores), but instead model the COMPARE score (with APD). We discuss this observation further in Subsection 6.1.

COMPARE Discovery GlossReader and DeepMistake also reached the first and second place on the COMPARE task in evaluation phase 1. This is not surprising, because they actually modeled the COMPARE score with APD. Consequently, also the correlation was considerably higher than with Graded Change (e.g. $\rho = 0.842$ vs. 0.735). Baseline1 took the third place.

Binary Change Detection For Phase 2 (Tables 5 and 6), again **GlossReader** performed best, this time followed by **UAlberta** and **Rombek**. Interestingly, with the exception of GlossReader the systems used in Phase 1 did not obtain a good performance in Phase 2. However, participants managed to outperform all baselines with the exception of HSE not outperforming baseline4 (minority class). Two out of the winning systems used thresholding (GlossReader, Rombek), i.e., they modeled the COMPARE score or the JSD and then thresholded these scores to obtain Binary Change predictions. From these teams only UAlberta inferred sense clusters. Hence, here we saw again what we saw for phase 1: the top-performing teams were often not modeling the annotation procedure.

Sense Gain/Loss Detection The top performance for sense gain ($\mathrm{F1} = 0.591$) was clearly lower than for Binary Change, while for loss the top performance ($\mathrm{F1} = 0.688$) approaches the one for Binary Change. The best results for sense gain were obtained by **DeepMistake**, followed by **BOS** and **GlossReader**. In the sense loss subtask, **GlossReader** obtained the best performance, followed by **Rombek** and **BOS**. GlossReader and DeepMistake submitted the same results to both subtasks as for Binary Change Detection implicitly assuming

[21] In the case of HSE who used two different systems, the displayed results correspond to the token-based system.

[22] Since not not all users reported a team name on Codalab, some leaderboard entries are filled with usernames.

Task		Change graded	COMPARE
#	Team name	SPR	SPR
1	GlossReader	**0.735 (1)**	0.842 (1)
2	DeepMistake	**0.702 (2)**	0.829 (2)
3	HSE	**0.553 (3)**	0.558 (4)
4	baseline1	0.543 (4)	0.561 (3)
5	baseline3	0.508 (5)	0.459 (5)
6	Rombek	0.497 (6)	0.456 (6)
7	CoToHiLi	0.282 (7)	–
8	baseline2	0.092 (8)	0.088 (7)
9	baseline5	0.064 (9)	-0.072 (8)
10	BOS	-0.125 (10)	-0.129 (9)

Table 4: Summary of system performance in phase 1. Teams are ranked according to SPR score for the Graded Change subtask in decreasing order. The values corresponding to the three best systems are highlighted in bold type.

Task		Change binary			Change graded	COMPARE
#	Team name	F1	P	R	SPR	SPR
1	GlossReader	**0.716 (1)**	0.615 (3)	0.857 (3)	0.735 (1)	0.842 (1)
2	UAlberta	**0.709 (2)**	0.549 (7)	1.000 (1)	–	–
3	Rombek	**0.687 (3)**	0.590 (4)	0.821 (4)	0.535 (5)	0.546 (5)
4	BOS	0.658 (4)	0.510 (8)	0.929 (2)	0.209 (8)	0.163 (7)
5	DeepMistake	0.655 (5)	0.633 (2)	0.679 (6)	0.676 (2)	0.821 (2)
6	CoToHiLi	0.636 (6)	0.553 (6)	0.750 (5)	0.282 (7)	–
7	baseline4	0.636 (6)	0.467 (11)	1.0 (1)	–	–
8	HSE	0.586 (7)	0.567 (5)	0.607 (7)	0.553 (3)	0.558 (4)
9	baseline3	0.548 (8)	0.500 (9)	0.607 (7)	0.373 (6)	0.423 (6)
10	baseline1	0.537 (9)	0.846 (1)	0.393 (9)	0.543 (4)	0.561 (3)
11	baseline5	0.508 (10)	0.484 (10)	0.536 (8)	0.064 (10)	-0.072 (9)
12	baseline2	0.222 (11)	0.500 (9)	0.143 (10)	0.092 (9)	0.088 (8)

Table 5: Summary of the results of Phase 2 for substasks Graded Change, COMPARE and Binary Change. Teams are ranked according to F1 score for subtask Change binary in decreasing order. The values corresponding to the three best systems are highlighted in bold type.

that gain and loss always occur together. In this way, they mostly outperformed Rombek and BOS who tried a more principled approach.

Graded Change/COMPARE Detection The top performance for these tasks was the same in evaluation phase 1 and 2 ($\rho = 0.735$ and 0.842). Some teams had the same results in both phases (GlossReader, HSE, CoToHiLi) and thus likely submitted the same predictions. Two teams improved their results (Rombek, BOS), while one team had lower results (DeepMistake). We are unsure about the impact of the published target words and their usages on these results, as teams did not consistently report whether they used this information in phase 2.

6.1 Discussion

The Graded Change Discovery subtask was solved with a rather high performance by the winning team ($\rho = 0.735$). This is comparable to the top performance in SemEval ($\rho = 0.725$ for DE) obtained with type-based embeddings. The COMPARE Discovery subtask was solved with even higher performance ($\rho = 0.842$). This is comparable to the top performance in RuShiftEval ($\rho = 0.822$). However, the results in our shared task were obtained under harder conditions, i.e., for a large number of uncleaned target words (Discovery).[23] This suggests that, as far as Graded Change is concerned, LSCD

[23]We assume that the performance of participating systems obtained on the hidden target words generalizes roughly to the full set of public target words as the sample was taken largely random.

Task		Sense gain			Sense loss		
#	Team name	F1	P	R	F1	P	R
1	GlossReader	**0.511 (3)**	0.333 (5)	0.929 (2)	**0.688 (1)**	0.564 (2)	0.880 (2)
2	DeepMistake	**0.591 (1)**	0.433 (1)	0.929 (2)	0.582 (5)	0.533 (3)	0.640 (4)
3	HSE	0.250 (8)	0.192 (9)	0.357 (5)	0.364 (7)	0.421 (5)	0.320 (5)
4	baseline1	–	–	–	–	–	–
5	Rombek	0.50 (4)	0.409 (2)	0.643 (4)	**0.681 (2)**	0.727 (1)	0.640 (4)
6	baseline3	–	–	–	–	–	–
7	BOS	**0.520 (2)**	0.361 (4)	0.929 (2)	**0.610 (3)**	0.529 (4)	0.720 (3)
8	baseline2	0.211 (9)	0.400 (3)	0.143 (6)	0 (8)	0 (8)	0 (7)
9	UAlberta	0 (10)	0 (10)	0 (7)	0 (8)	0 (8)	0 (7)
10	CoToHiLi	0.462 (5)	0.316 (6)	0.857 (3)	0 (8)	0 (8)	0 (7)
11	baseline4	0.378 (6)	0.23 (8)	1.0 (1)	0.588 (4)	0.416 (6)	1.0 (1)
12	baseline5	0.333 (7)	0.313 (7)	0.357 (5)	0.367 (6)	0.375 (7)	0.36 (6)

Table 6: Summary of the results of Phase 2 for subtasks Sense loss and Sense gain. The values corresponding to the three best systems are highlighted in bold type.

systems are applicable to solve real-world problems and may be useful in historical semantics or lexicography. However, the more relevant task for these fields is Binary Change Detection/Discovery (Schlechtweg and Schulte im Walde, 2020). The results for Binary Change Detection were lower ($F1 = 0.716$), but still clearly higher than the best baseline (0.636). Results in SemEval were mixed, but mostly not higher than $F1 = 0.7$ (DE), while results in DIACR-Ita were high with an accuracy of 0.94, which was, however, obtained with a different metric and on a very small and strongly preselected set of target words. A future challenge will thus be to improve performance on the binary task.

Our shared task was clearly dominated by token-based systems. Out of seven participants only two used a (standalone) type-based system which also performed much worse than the winning teams (CoToHiLi, HSE).[24] Also, our type-based baseline1 was clearly outperformed by a number of token-based systems (three in phase 1 and six in phase 2). This confirms the tendency observed in RuShiftEval where token-based systems outperformed type-based ones on LSCD. Before that, in SemEval and DIACR-Ita the type-based systems had dominated. Potential reasons for this switch are the understanding of biases in contextualized embeddings (Laicher et al., 2021), their optimization through fine-tuning (Arefyev et al., 2021; Arefyev and Bykov, 2021) and the optimization of vector aggregation methods (Kutuzov and Giulianelli, 2020; Laicher et al., 2021; Arefyev et al., 2021).

In our task, we saw clustering methods amongst the best-performing systems (HSE, UAlberta) for the first time. This is an important development, because the current top-performing system (GlossReader), as well as many other systems not relying on clustering, did not model the target word annotation procedure (cf. Subsection 4.4). Instead, it exploited correlations between the COMPARE score and JSD as well as Binary Change. These scores are known to correlate strongly in current LSCD datasets (Schlechtweg, 2022), including ours. The correlation between gold (negated) COMPARE and JSD scores in our dataset is 0.92, while it is 0.69 for gold (negated) COMPARE and Binary Change. This means that modeling the COMPARE score is a good predictor for Graded as well as Binary Change. However, this also means that, the current best-performing systems have a clear upper bound on their potential to solve LSCD tasks (where this upper bound is higher for Graded than for Binary Change). Hence, if we want to break through this upper bound in the future, we need to develop or improve other system types possibly relying on clustering to model the annotation procedure.[25]

In order to see how far the current approach of thresholding COMPARE/JSD/graded scores carries, we compared performance of the top three systems in evaluation phase 1 across binarization thresholds in Figure 2. As we can see, the three

[24] The result reported by the HSE team in the leaderboard corresponds to the first method described in Section 5.2.

[25] Homskiy and Arefyev (2022) had promising results with applying the clustering framework used in the annotated data and semantic proximity graphs derived from fine-tuned contextualized embeddings.

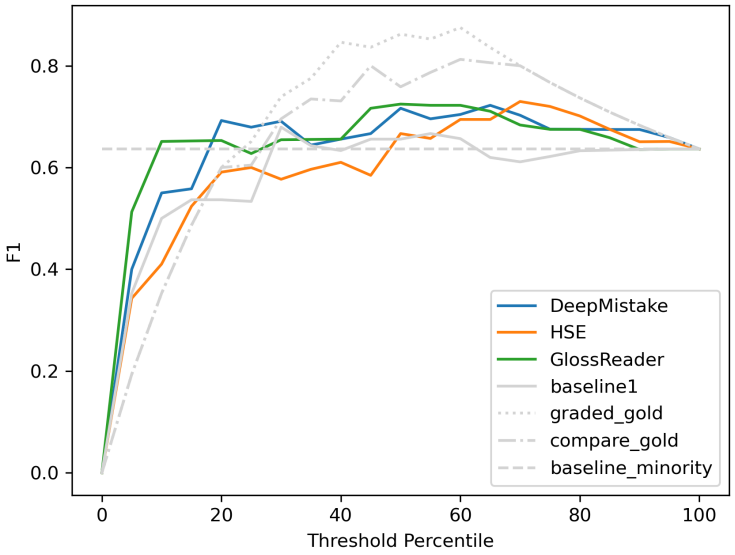


Figure 2: F1 scores over binarization thresholds based on percentiles on submitted Graded Change predictions for top four teams in evaluation phase 1.

systems had a similar maximum performance of roughly $F1 = 0.72$ around a binarization threshold of $50 - 70$ %.[26] At 100 % they all converged to the minority class baseline (all target words labeled as 1). The upper bound on this approach was given by the maximum performance of the gold JSD (graded_gold) and the gold COMPARE score (compare_gold). These upper bounds were 0.88 and 0.81 respectively. This means that perfectly modeling the COMPARE or even the JSD score can reach high but never perfect performance on Binary Change.

7 Conclusion

We conducted the first shared task on semantic change discovery and detection in Spanish. We manually annotated 100 Spanish words for semantic change between two corpora, an old one covering the period between 1810 and 1906, and a modern one covering the years between 1994 and 2020. The discovery part of our task imposed several computational challenges for participants, as it required calculating semantic change scores for all words in the vocabulary.

We received predictions from six teams in phase 1 and seven teams in phase 2. Participants applied systems using static and contextualized word embeddings in combination with various fine-tuning procedures, vector aggregation methods and change measures. Graded Change Discovery was solved with high performance while Binary Change Detection still remains far from being solved. The most successful method winning both main tasks is a system fine-tuning contextualized multilingual XML-R embeddings on WSD data, aggregating vectors into cross-corpus pairs and measuring change as the average of their distances, or a binarization of these values. However, we showed that this approach has a clear upper bound which will not allow to solve the tasks completely reliably in the future. Another interesting result from our task was that clustering approaches are amongst the winning teams for the first time.

We hope that this shared task will help pave the way for future research in the discovery and detection of semantic lexical changes for the Spanish language, and that our data can be used in the future for the proposal of novel ideas and techniques.

8 Acknowledgements

This work was supported by ANID FONDECYT grant 11200290, U-Inicia VID Project UI-004/20, ANID -Millennium Science Initiative Program - Code ICN17_002, the National Center for Artifi-

[26] Interestingly, HSE here obtained maximum performance amongst all systems (0.73), much higher than their submission in evaluation phase 2. A similar observation holds for our baseline1. This shows how crucial threshold selection is in this approach.

cial Intelligence CENIA FB210017, Basal ANID, and SemRel Group (DFG Grants SCHU 2580/1 and SCHU 2580/2). Dominik Schlechtweg has been funded by the project 'Towards Computational Lexical Semantic Change Detection' supported by the Swedish Research Council (2019–2022; contract 2018-01184) and by the research program 'Change is Key!' supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

Adnan Ahmad, Kiflom Desta, Fabian Lang, and Dominik Schlechtweg. 2020. Shared task: Lexical semantic change detection in german. *CoRR*, abs/2001.07786.

Nikolay Arefyev and Dmitrii Bykov. 2021. An interpretable approach to lexical semantic change detection with lexical substitution. volume 2021-June, pages 31–46. ABBYY PRODUCTION LLC.

Nikolay Arefyev, Maksim Fedoseev, Vitaly Protasov, Daniil Homskiy, Adis Davletov, and Alexander Panchenko. 2021. Deepmistake: Which senses are hard to distinguish for a word-in-context model. volume 2021-June, pages 16–30.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018b. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 282–291, Brussels, Belgium. Association for Computational Linguistics.

Gioia Baldissin, Dominik Schlechtweg, and Sabine Schulte im Walde. 2022. DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine Learning*, 56(1-3):89–113.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Sorana-Daniela Bolboaca and Lorentz Jäntschi. 2006. Pearson versus spearman, kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, 5(9):179–200.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Gonzalo Donoso and David Sanchez. 2017. Dialectometric analysis of language variation in twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 16–25, Valencia, Spain.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. 2016. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas.

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. Challenges for Computational Lexical Semantic Change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, volume Language Variation, chapter 11. Language Science Press, Berlin.

Daniil Homskiy and Nikolay Arefyev. 2022. Deepmistake at lscdiscovery: Can a multilingual word-in-context model replace human annotators? In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Jens Kaiser, Dominik Schlechtweg, Sean Papay, and Sabine Schulte im Walde. 2020a. IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Jens Kaiser, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020b. OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still rocks Semantic Change Detection. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org. Winning Submission!

Kseniia Kashleva, Alexander Shein, Elizaveta Tukhtina, and Svetlana Vydrina. 2022. Hse at lscdiscovery in spanish: Clustering and profiling for lexical semantic change discovery. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

Artem Kudisov and Nikolay Arefyev. 2022. Bos at lscdiscovery: Lexical substitution for interpretable lexical semantic change detection. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

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)*, Online. Association for Computational Linguistics.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021. Rushifteval: a shared task on semantic shift detection for russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*.

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.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37:145–151.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

Martin Pömsl and Roman Lyapin. 2020. CIRCE at SemEval-2020 Task 1: Ensembling Context-Free and Context-Dependent Word Representations. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Ondřej Pražák, Pavel Přibákň, Stephen Taylor, and Jakub Sido. 2020. UWB at SemEval-2020 Task 1: Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Maxim Rachinskiy and Nikolay Arefyev. 2022. Glossreader at lscdiscovery: Train to select a proper gloss in english – discover lexical semantic change in spanish. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*. Association for Computational Linguistics.

Ana Sabina Uban, Alina Maria Cristea, Anca Daniela Dinu, Simona Georgescu, and Laurentiu Zoicas. 2022. Cotohili at lscdiscovery: the role of linguistic features in predicting semantic change. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

Dominik Schlechtweg. 2022. *Human and Computational Measurement of Lexical Semantic Change*. Ph.D. thesis, University of Stuttgart, Stuttgart, Germany.

Dominik Schlechtweg, Anna Hätty, Marco del Tredici, and Sabine Schulte im Walde. 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. 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 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Dominik Schlechtweg and Sabine Schulte im Walde. 2020. Simulating Lexical Semantic Change from Sense-Annotated Data. In *The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13)*.

Dominik Schlechtweg and Sabine Schulte im Walde. submitted. Clustering Word Usage Graphs: A Flexible Framework to Measure Changes in Contextual Word Meaning.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association 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.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection.

Daniela Teodorescu, Spencer McIntosh von der Ohe, and Grzegorz Kondrak. 2022. Ualberta at lscdiscovery: Lexical semantic change detection via word sense disambiguation. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Appendix

A Lemmatization

Manual inspection showed that spaCy sometimes yielded erroneous lemmatization. This happened more frequently for sentences in the old corpus and for tokens at the beginning of sentences as shown in the example below:

Example:
"Decidióse ésta por Teresa la expósita, y así se vio a la vagamunda tomar bajo su amparo a la pobre desheredada como ella."
Lemmatization:
Decidióse este por Teresa el expósita , y así él ver a el vagamunda tomar bajo su amparo a el pobre desheredado como él .

As can be seen, the lemma of the word *Decidióse* was not found, nor was the word converted to lowercase. SpaCy version 3.1.1 with es_core_news_md (3.1.0) was used.

B Target indices of annotated usages

In the first version of the extracted word usages which were uploaded to the DURel interface for annotation there were frequent errors for the target word indices. As a result, the wrong target words

were marked in these usages. However, annotators were instructed to search for the correct target words and to judge these instead. We corrected the indices for the data provided to participants during the shared task. However, we later noticed that some indices included punctuation immediately following the target word as shown below:

Example
lemma: sexo
context: 136. Los apellidos de familia no varían de terminación para los diferentes **sexos;** y así se dice «don Pablo Herrera», «doña Juana Hurtado», «doña Isabel Donoso». 137 (b).
indexes_target_token: 75:81

After the shared task we uploaded a data version with corrected indices.