DWUGs-IT: Extending and Standardizing Lexical Semantic Change Detection for Italian

Pierluigi Cassotti[1,*], Pierpaolo Basile[2] and Nina Tahmasebi[1]

[1] *University of Gothenburg, Department of Philosophy, Linguistics and Theory of Science, Gothenburg, Sweden*
[2] *University of Bari Aldo Moro, Department of Computer Science, via E. Orabona, 70125, Bari, Italy*

Abstract

Lexical Semantic Change Detection (LSCD) is the task of determining whether a word has undergone a change in meaning over time. There has been a marked increase in interest in this task, accompanied by a corresponding growth in the scientific community involved in developing computational approaches to semantic change. In recent years, a number of resources have been made available for the evaluation of LSC models in a number of languages, including English, Swedish, German, Latin, Russian and Chinese. DIACR-ITA is the only existing resource for LSCD in Italian. However, DIACR-ITA has a different format from that used for other languages. In this paper, we present DWUGs-IT, which extends the DIACR-ITA dataset with additional target words and usage-sense pair annotations and adapts it to the DURel format, including the first implementation of a LSCD graded task for Italian.

Keywords

Lexical Semantic Change, Sense-annotated corpora, Italian, Historical Linguistics

1. Introduction

As is the case with both society and culture, language is subject to change over time. Two key factors cause such linguistic change. Firstly, there are purely evolutionary and linguistic considerations driven by the need for more efficient communication [1]. One example of this is the use of abbreviations and acronyms, such as *LOL* (Laughing Out Loud), which have become commonplace on social media platforms. Secondly, changes in society and culture lead to changes in language. This can be seen, for example, in the adoption of a more inclusive language, as evidenced by grammatically gendered languages, including Italian and the introduction of ə to replace masculine and feminine endings [2].

Language may undergo alteration at various levels, including morphological, syntactic, and semantic. Semantic change concerns the alteration of the meaning of words over time. The study of semantic change is a prominent area of research in Historical Linguistics, with the aim of investigating the linguistic mechanisms that characterize the change and the causes that trigger it. For instance, Blank [3] provides a broad study on the characterization of semantic change, identifying a number of different types of change, including metaphor, metonymy, generalization, specialization, co-hyponym transfer and auto-antonym. The English word *bad*, for example, has acquired an auto-antonym meaning, i.e. a meaning that is the opposite of its original meaning. In addition to its original connotation of *poor quality* or *negative*, it has also acquired the opposite connotation of *good* or *cool*. The term *meat* has undergone a process of specialization in its meaning, whereby it has shifted from referring *to any kind of food in general* to exclusively denoting the *meat of animals consumed as food*.

While traditional linguistic methods are informative, they are often based on small, carefully curated samples. In contrast, linguistic analyses using computational models not only accelerate our understanding of language change but also provide broader and more detailed insights, thereby facilitating the study of vast corpora across a wider range of genres and time [4, 5].

From a computational perspective, two key challenges emerge in the study of semantic change: **the modelling of word meanings over time** and the **detection of change** [6, 7]. At the synchronic level, ignoring the temporal dimension with a focus on modern corpora, the Natural Language Processing community has made significant strides in modelling word meanings, with approaches such as Word Sense Disambiguation (WSD) [8] playing a pivotal role. Computational modelling of semantic change introduces a significant level of complexity, as it necessitates the handling of meanings that are either extinct or novel in comparison to existing lexicographic resources, such as WordNet, as well as dynamically changing meaning representations.

In recent years, great efforts have been made to advance the field of computational methods for Lexical Semantic Change Detection. With initiatives such as the Workshop on Computational Approaches to Historical Language Change [9] promoting research in this

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

[*] Corresponding author.

pierluigi.cassotti@gu.se (P. Cassotti); pierpaolo.basile@uniba.it (P. Basile); nina.tahmasebi@gu.se (N. Tahmasebi)
0000-0001-7824-7167 (P. Cassotti); 0000-0002-0545-1105 (P. Basile); 0000-0003-1688-1845 (N. Tahmasebi)

field or shared tasks such as SemEval 2020 Task 1 [10], RuShiftEval [11], DIACR-ITA [12], or LSCD Discovery [13] leading to the development of the first evaluation resources. DIACR-Ita, hosted in EVALITA 2020 [14], is the first shared task specifically created for the evaluation of models for Lexical Semantic Change in Italian. The majority of the evaluation resources follow a two-task approach: (1) a binary task, which requires the assignment of a word to either the *changed* or *stable* label, based on whether the word has undergone a change in meaning or not; and (2) a graded (ranking) task, which requires the sorting of words based on the extent of their change (over time). These labels are assigned on the basis of human-annotated data, typically in the form of a graded word-in-context task.

DIACR-Ita, however, diverges from the evaluation process employed in SemEval 2020 Task 1, RushiftEval and several other datasets that emerged subsequently. This results in a distinct configuration of the task and the released data. For example, DIACR-Ita only has a binary task but does not include a graded task. Moreover, only the target words with their gold truth labels were made available for the shared task, while the remaining data produced during the annotation process were not. In this paper,

1. we release DWUGs-IT [1], a new dataset for Lexical Semantic Change Detection for Italian, which:
    - **extends** the original DIACR-ITA with 12 new words;
    - provides **sense-annotated usages** with the respective sense labels
    - **standardizes** DIACR-ITA providing the data in the DURel format [15, 16, 17]
    - introduces the first LSC **graded task** for Italian
2. we **evaluate** DWUGs-IT using XL-LEXEME[18], the state-of-the-art model for Lexical Semantic Change Detection [19]

2. Related Work

DURel [15] is a framework for the annotation of Lexical Semantic Change across a pair of time periods or corpora. The annotation involves human labelling of pairs of sentences containing the target word. The sentences can be contemporary, i.e. originating from the same time period, or diachronic, denoting a divergence in time between the two periods under consideration. An annotator has to decide whether the meaning expressed by the word in the two sentences is Unrelated (1), Distantly Related (2), Closely Related (3) or Identical (4). The scale of semantic relatedness is derived from the cognitive model proposed by Blank [20] and corresponds to the values of Homonymy (1), Polysemy (2), Context Variance (3) and Identity (4).

The annotations are then presented in the form of a graph, specifically a Word Usage Graph (WUGs) or a **Diachronic Word Usage Graph (DWUGs)** [21] in cases where the usages originate from different time periods. In these graphs, the nodes correspond to the word uses and the edges correspond to the median of the annotations. The diachronic graph is then subjected to clustering in order to identify the senses. Before clustering, a new graph is created by binarizing the edges, where an edge between two uses is established if the score of the original edge weight is less than 2.5, or in other words if the average annotation for this pair of uses is less than 2.5. Since the graph typically exhibits considerable sparsity, which limits the applicability of conventional clustering algorithms, a variation of the correlation clustering algorithm [22] is typically used, as it is able to model this type of sparsely connected graph.

Once the (diachronic) clusters have been obtained, they can be considered to represent the senses. The distribution of the usages from different time periods in each cluster (sense) is then analyzed to obtain a change score. For instance, one can determine a graded change score by computing the Jensen-Shannon Distance (JSD) on the probability distributions of senses across various time periods. This is expressed as

$$\sqrt{\frac{D(P \,||\, M) + D(Q \,||\, M)}{2}}$$

where P and Q represent the probability distributions of clusters from different historical periods, D denotes the Kullback-Leibler divergence, and $M = \frac{(P+Q)}{2}$ [23, 24].

Furthermore, a binary label can be obtained, whereby words that have undergone a change in meaning over time are assigned a *changed* label (words that have gained/lost a sense), while words that have retained their meaning are labelled *stable*. The label is typically assigned by evaluating the frequency of senses in different time periods and establishing thresholds to distinguish stable and changed words.

Datasets based on DURel SemEval 2020 Task 1 [10] is the first initiative to standardize the evaluation of computational approaches to semantic change. SemEval 2020 Task 1 focuses on English, German, Swedish and Latin and proposes a common evaluation framework with two tasks: classifying target words as those whose meaning has changed or remained stable, and ranking words according to their degree of change. Special attention is given to Latin due to the lack of native speakers. Therefore, in the annotation of the Latin dataset, usage-sense

[1] DWUGs-IT is available on Zenodo https://zenodo.org/records/13941618.

pairs are considered rather than usage-usage pairs, and the annotator is asked to decide how related the considered usage is to a particular sense, using the DURel scale from Unrelated to Identical. RuShiftEval [11] aimed to detect semantic shifts in Russian across pre-Soviet, Soviet, and post-Soviet periods. The dataset included 111 Russian nouns, with participants ranking them by their degree of change (using the COMPARE measure [15], an approximation of the JSD). The task focused on ranking changes, with evaluations based on Spearman rank correlations. LSC Discovery [13] focused on detecting and discovering semantic changes in Spanish. It is divided into Graded Change Discovery and Binary Change Detection. The task required evaluations for all vocabulary words in the corpus, covering periods from 1810-1906 and 1994-2020. NorDiaChange [25] studied diachronic semantic change in Norwegian. The dataset included 80 nouns reflecting significant historical periods, such as pre- and post-war events and technological advances. ZhShiftEval [26, 27] assessed semantic change in Chinese over 50 years, focusing on the period around Reform and Opening Up. The dataset used texts from the People's Daily and included 20 words chosen for their frequency and noted changes.

3. DIACR-ITA

The DIACR-ITA annotation was conducted on word usages extracted from *L'Unità* corpus [28]. *L'Unità* corpus comprises a collection of Italian texts extracted from the newspaper L'Unità. In order to evaluate semantic change, the corpus has been divided into two sub-corpora, covering the period from 1948 to 1970 and the period from 1990 to 2014, respectively. A time window of 20 years between the sub-corpora ensures sufficient distance between the two periods, allowing for the tracking of potentially more pronounced semantic changes. The sub-corpora statistics are presented in Table 1.

The selection of target words was based on the information provided in the Sabatini-Coletti dictionary of the Italian language, which records the year of the first occurrence of a word's sense. The initial step involved the extraction of a list of words from Sabatini-Coletti for which the dictionary reported a semantic change, i.e. the introduction of at least one new sense after 1970. Moreover, an examination of the set of words was conducted to ensure that the sampled words appeared at least 10 times in each of the two periods and that the occurrences of these words were not significantly affected by OCR errors. Consequently, 26 target words were identified. For each target word, up to 50 occurrences from each of the two sub-corpora were extracted.

The senses of each word were classified into two groups: the senses recorded in the Sabatini-Coletti dictionary for the period 1948-1970 (Group 1) and the new senses introduced after 1970 (Group 2). The annotators were required to determine whether the sense of each word usage belonged to Group 1, Group 2, or to another category if the word sense did not align with either group (Other). Additionally, the annotator may indicate a preference of *Cannot decide* for the uses in which they were uncertain. Five annotators fluent in Italian annotated DIACR-ITA. Each sentence was annotated by two annotators. The disagreement cases were resolved by the two annotators involved, analyzing the disagreement and deciding on an unambiguous label.

Each target word was labelled as *stable* or *changed*. A word was considered *changed* if there was at least one instance of Group 2 among the extracted usages from the period between 1990 and 2014 and no instances of Group 2 among the extracted usages from the period between 1948 and 1970. The final dataset consisted of 18 words, of which 6 were changed and 12 were stable.

Corpus	Period	#Tokens
L'Unità	1948-1970	52,287,734
L'Unità	1990-2014	196,539,403

Table 1
Sub-corpora statistics.

4. DWUGs-IT

DWUGs-IT builds on the DIACR-ITA dataset, adapting it to the DURel format and adding eight new words. It also provides the usage-sense annotated pairs that were not initially released, as summarized in Table 2. For each target word, we format the annotated usages following the WUG style, including the time period of the usage and the word's position in the sentence. Similarly, we format and release the annotated sense labels in a way similar to DWUG LA [29].

Unlike the traditional WUG approach, where sense preference is not explicitly marked, in DIACR-ITA, annotators clearly indicate their preference for one sense over others. For example, in the usage of the word *api* (Italian for *bees*), in the sentence "Dalle api un dolce dono" ("From bees, a sweet gift"), the annotators choose the sense *insect* while discarding the alternative sense *means of transport*. For each use-sense pair not selected by annotators, a rating of 1 (*Unrelated*) is assigned, while matched pairs receive a rating of 4 (*Identical*), in line with the DURel scale.

Since human annotators already provide the sense labels, we do not cluster usages automatically (as is typically done in the WUG approach), but directly assign the annotated meanings. All subsequent calculations, such as

Lemma	Group 1	Group 2	Other
ultima	Che viene dopo tutti gli altri in una serie numerica, in una classifica, in una graduatoria o in una successione spaziale o temporale	Nel l. fam., l'ultima cosa; la novità, la notizia più recente: la sai l'ultima?	
emulare	Prendere qlcu. a modello, imitarne meriti e virtù: e. i genitori, le imprese di uno scalatore	ambito informatico	
affido		Affidamento di un minore	✔
bombetta	S1. Cappello maschile di feltro rigido a cupola con tese corte leggermente rialzate ai lati	S2. Fialetta puzzolente che i ragazzi lanciano per divertimento per strada o in ambienti chiusi	✔
cantieristica	maschile - Di cantiere, relativo ai cantieri: il settore c. oppurre con riferimento al cantiere	Attività di costruzione, riparazione navale	
fondista	Giornalista che scrive l'articolo di fondo su un quotidiano - Atleta	Nel gergo della finanza, sottoscrittore di fondi di investimento	✔
portatile	Che può essere trasportato agevolmente da una persona: televisore p.	Piccolo computer facilmente trasportabile, funzionante anche a batteria e quindi utilizzabile in viaggio - telefono portatile	
impegnativa	agg. che richiede impegno	Dichiarazione con cui si assume un impegno; in partic. nel l. burocr., documento con cui un ente mutualistico si impegna a coprire, nella misura prevista dalla legge, le spese sanitarie di un suo iscritto: fare l'i. per le analisi	

Table 2
Newly introduced words together with the senses of Group 1 (1948-1970), Group 2 which involves senses introduced after 1970, and an indication of the presence of other senses not listed in Group 1 and Group 2.

change scores and related statistics, follow the standard WUG methodology.

5. Evaluation

XL-LEXEME has been tested on different languages before but has never been evaluated on Italian. In this section, we evaluate XL-LEXEME on the new DWUGs-IT dataset using the traditional evaluation pipeline for the DWUGs [19, 30]. We assess the ability to derive a reliable change score (Graded Change Detection) and evaluate the possibility of clustering the XL-LEXEME vectors to automatically induce target word senses, which are then compared to the DWUGs-IT annotations via the Adjusted Rand Index and the Purity measure.

5.1. Model

XL-LEXEME, built on XLM-RoBERTa large [31], is trained for the Word-in-Context (WiC) task [32], which determines if a word has the same meaning in two sentences. Using a Siamese architecture [33], it creates word vectors. The loss function adjusts weights via cosine distance, aligning vectors for the same meanings and separating them for different meanings. To calculate the change score, a classic approach is to use the Average Pairwise Distance between the vectors computed over the two different periods:

$$\mathrm{LSC}(s_w^{t_0}, s_w^{t_1}) = \frac{1}{N \cdot M} \sum_{i=0}^{N} \sum_{j=0}^{M} \delta(s_{w,i}^{t_0}, s_{w,j}^{t_1}) \quad (1)$$

where δ is the cosine distance and s_w^t is the set of sentences containing the word w at time t. For the Word Sense Induction step, we cluster the vectors into senses using Agglomerative Clustering [2] with a cosine threshold of 0.5 and Average Linkage, which merges clusters with a similarity greater than 0.5.

5.2. Metrics

We test the ability of XL-LEXEME in ranking words according to their change scores (Graded Change Detection) using Spearman Correlation. Cluster quality is assessed using the Adjusted Rand Index (ARI) [34], which is defined as follows:

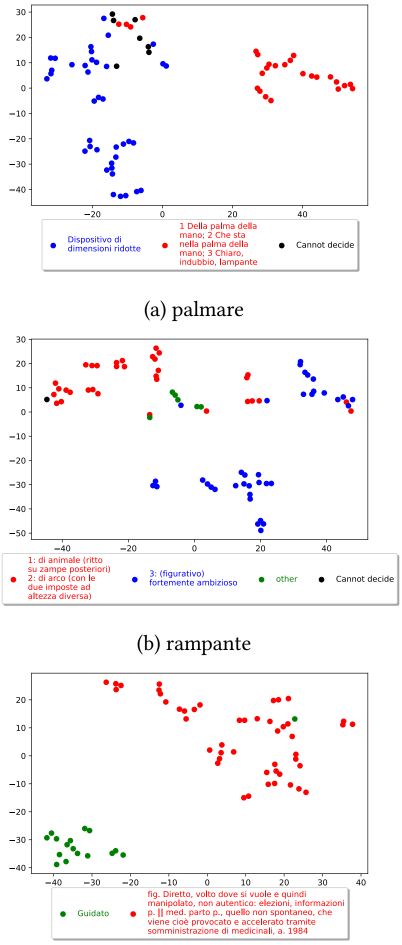
$$ARI = \frac{RI - Expected_{RI}}{max(RI) - Expeted_{RI}}$$

RI stands for the Rand Index, which measures the number of pair agreements within the data – that is, pairs of instances that are correctly placed in the same cluster. The $Expetcted_{RI}$ is the expected number of such agreements by chance, calculated based on the distribution of the clusters, while the $max(RI)$ is the maximum possible value of RI, which occurs when all pairs are classified perfectly. We use Purity in addition to ARI to capture cluster homogeneity and provide clearer insight about how mixed the clusters are in terms of class labels, i.e.

$$\text{Purity} = \frac{1}{N} \sum_k \max_j |c_k \cap t_j|$$

where N is the total number of instances, c_k denotes cluster k, and t_j represents class j.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering



(a) palmare

(b) rampante

(c) pilotato

Figure 1: t-SNE visualization of XL-LEXEME embeddings with respect to the annotated clusters for changed words palmare, rampante, and pilotato.

5.3. Results

We begin to discuss qualitative results. Figure 1 illustrates the t-SNE visualization of XL-LEXEME embeddings for the usages of the words *palmare*, *rampante*, and *pilotato*. For *palmare* (Figure 1a), the senses are well separated except for some instances of the sense *relating to the palm, clear, evident* that are placed closer to the *PDA device* meaning, for example:

> *sono state n levate di le impronte palmari che saranno inviate al1' archivio generale segnaletico di Roma.* (en. *The palm prints have been removed and will be sent to the general sign archive of Rome.*)

For the word *rampante* (Figure 1b), the annotators identified an additional meaning (Other) that refers to a named entity, i.e., *Il barone rampante* written by Italo Calvino. The instances of *Il barone rampante* fall in the middle of the cluster of the *rearing* and *ambitious* meanings. Interestingly, the only instance annotated as *Cannot decide* falls in the *rearing* cluster:

> *Uno rampante » non ci aia ancora nulla da fare, comunque i tecnici....supremazia di le Ferrari.* (en. *A rampant » there is still nothing to be done, in any case the technicians.... supremacy of the Ferrari.*)

This instance is ambiguous since the subject of *rampante* is missing in the sentence. However, interestingly, XL-LEXEME assumes it to have the *rearing* meaning, probably due to the presence of the word *Ferrari*, referring to the Ferrari logo. Figure 1c shows how the embeddings of the usages of *pilotato* are perfectly split according to the sense labels. However, one instance of the meaning *driven* falls in the cluster of the *manipulated* instances, which can be considered ambiguous and open to interpretation:

> *Twingo Easy offre la grande comodità di un cambio con frizione pilotata, ovvero: non c' è più il pedale della frizione.* (en. *Twingo Easy offers the great convenience of a gearbox with a piloted clutch, that is: there is no longer a clutch pedal.*)

The quantitative results of XL-LEXEME are reported in Table 3. Compared to LSCD benchmarks in other languages, XL-LEXEME shows similar results for the GCD score (ranging from 0.567 in NO to 0.851 in RU) and the ARI score (ranging from 0.249 in SV to 0.400 in ES). It also performs slightly better using the purity measure (ranging from 0.766 in SV to 0.836 in ZH). These results likely stem from the properties of the dataset that includes several monosemous words, but also from the process that has been used for DWUGs-IT where senses are modeled explicitly. Purity measures the extent to which clusters contain a single class. With many monosemous words, achieving high purity is easier since these words inherently belong to one sense group. ARI, on the other hand, evaluates the similarity between the clustering results and the ground truth, accounting for both the clustering quality and the number of clusters. In DWUGs-IT, most groups of word senses have just one meaning. But sometimes, a group of words can have several meanings, and how often each meaning is used can change over time. For example, the word *palmare* has three meanings in its Group 1: i) related to the palm of the hand, ii) something that fits in your hand, and iii) something that is obvious or clear. Over time, some of these meanings might be used more or less often. However, because all three meanings are grouped together, DWUGs-IT does not take into account how the use of each of those meanings changes over time. This broad categorization of senses

Graded Change Detection (Spearman Correlation)	0.51
Adjusted Rand Index (ARI)	0.28
Purity	0.89

Table 3
XL-LEXEME Results

can impact the performance of XL-LEXEME, which analyzes meanings at a more detailed level. Additionally, XL-LEXEME has been tested on different languages before but has never been evaluated on Italian. DWUGs-IT models senses explicitly, whereas previous datasets inferred senses automatically by comparing pairs of usages. This automatic inference process is similar to the approach XL-LEXEME uses, potentially making it better suited for datasets without explicit sense modelling.

6. Conclusion

This paper presents DWUGs-IT, an extension and standardization of the Lexical Semantic Change Detection (LSCD) task for Italian, based on the existing DIACR-ITA dataset. The dataset is expanded with additional target words and its format is aligned with that of the resources used for other languages. This involves the introduction of the first graded task for Italian. The standardized dataset and the evaluation framework we provide can serve as a foundation for future research in LSCD for Italian. By aligning the Italian dataset with those of other languages, we facilitate cross-linguistic comparisons and contribute to the broader understanding of semantic change mechanisms. In addition, we provide a first evaluation of the state-of-the-art LSCD model, XL-LEXEME, for Italian and both show its effectiveness as well as set a baseline for future work.

Acknowledgments

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021). The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.
We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.
We would also like to thank Tommaso Caselli, Annalina Caputo and Rossella Varvara, who contributed to the development of the DIACR-ITA resource, and Dominik Schlechtweg for valuable feedback on a preliminary draft of this work.

References

[1] J. R. Firth, A synopsis of linguistic theory 1930-55., Studies in linguistic analysis 1952-59 (1957) 1–32.

[2] P. Cassotti, A. Iovine, P. Basile, M. de Gemmis, G. Semeraro, Emerging trends in gender-specific occupational titles in italian newspapers, in: E. Fersini, M. Passarotti, V. Patti (Eds.), Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022, volume 3033 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: https://ceur-ws.org/Vol-3033/paper52.pdf.

[3] A. Blank, Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen, volume 285, Walter de Gruyter, 2012.

[4] P. Cassotti, S. D. Pascale, N. Tahmasebi, Using synchronic definitions and semantic relations to classify semantic change types, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 4539–4553. URL: https://doi.org/10.18653/v1/2024.acl-long.249. doi:10.18653/V1/2024.ACL-LONG.249.

[5] F. Periti, P. Cassotti, H. Dubossarsky, N. Tahmasebi, Analyzing semantic change through lexical replacements, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 4495–4510. URL: https://doi.org/10.18653/v1/2024.acl-long.246. doi:10.18653/V1/2024.ACL-LONG.246.

[6] N. Tahmasebi, L. Borin, A. Jatowt, Survey of computational approaches to lexical semantic change detection, Computational approaches to semantic change 6 (2021).

[7] S. Montanelli, F. Periti, A Survey on Contextualised Semantic Shift Detection, arXiv preprint arXiv:2304.01666 (2023).

[8] R. Navigli, Word Sense Disambiguation: A Survey, ACM Comput. Surv. 41 (2009). URL: https://doi.org/10.1145/1459352.1459355. doi:10.1145/1459352.1459355.

[9] N. Tahmasebi, S. Montariol, H. Dubossarsky, A. Kutuzov, S. Hengchen, D. Alfter, F. Periti, P. Cassotti (Eds.), Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, Association for Computational Linguistics, Singapore, 2023. URL: https://aclanthology.org/2023.lchange-1.0.

[10] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Du-

bossarsky, N. Tahmasebi, SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1–23. URL: https://www.aclweb.org/anthology/2020.semeval-1.1/.

[11] A. Kutuzov, L. Pivovarova, RuShiftEval: A Shared Task on Semantic Shift Detection for Russian, in: Proc. of the International Conference on Computational Linguistics and Intellectual Technologies (Dialogue), 20, Redkollegija sbornika, (online), 2021.

[12] P. Basile, A. Caputo, T. Caselli, P. Cassotti, R. Varvara, Diacr-ita @ EVALITA2020: overview of the EVALITA2020 diachronic lexical semantics (diacr-ita) task, in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2765/paper158.pdf.

[13] F. D. Zamora-Reina, F. Bravo-Marquez, D. Schlechtweg, LSCDiscovery: A Shared Task on Semantic Change Discovery and Detection in Spanish, in: Proc. of the Workshop on Computational Approaches to Historical Language Change (LChange), Association for Computational Linguistics (ACL), Dublin, Ireland, 2022, pp. 149–164.

[14] V. Basile, D. Croce, M. Di Maro, L. C. Passaro, Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), CEUR.org, Online, 2020.

[15] D. Schlechtweg, S. S. im Walde, S. Eckmann, Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change, in: M. A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 169–174. URL: https://doi.org/10.18653/v1/n18-2027. doi:10.18653/v1/n18-2027.

[16] D. Schlechtweg, S. M. Virk, P. Sander, E. Sköldberg, L. T. Linke, T. Zhang, N. Tahmasebi, J. Kuhn, S. S. im Walde, The durel annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change, in: N. Aletras, O. D. Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - System Demonstrations, St. Julians, Malta, March 17-22, 2024, Association for Computational Linguistics, 2024, pp. 137–149. URL: https://aclanthology.org/2024.eacl-demo.15.

[17] P. Sander, S. Hengchen, W. Zhao, X. Ma, E. Sköldberg, S. Virk, D. Schlechtweg, The durel annotation tool, in: Book of Abstracts of the Workshop Large Language Models and Lexicography, 8 October 2024 Cavtat, Croatia (ed. Simon Krek), 2024.

[18] P. Cassotti, L. Siciliani, M. DeGemmis, G. Semeraro, P. Basile, XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1577–1585. URL: https://aclanthology.org/2023.acl-short.135. doi:10.18653/v1/2023.acl-short.135.

[19] F. Periti, N. Tahmasebi, A systematic comparison of contextualized word embeddings for lexical semantic change, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 4262–4282. URL: https://aclanthology.org/2024.naacl-long.240.

[20] A. Blank, Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change, Historical semantics and cognition (1999).

[21] D. Schlechtweg, N. Tahmasebi, S. Hengchen, H. Dubossarsky, B. McGillivray, DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages, in: Annual Conference of the North American Chapter of the Association for Computational Linguistics, (NAACL-HLT 2021), Association for Computational Linguistics, Mexico City, Mexico, 2021.

[22] N. Bansal, A. Blum, S. Chawla, Correlation clustering, Machine Learning 56 (2004) 89–113. doi:10.1023/B:MACH.0000033116.57574.95.

[23] J. Lin, Divergence measures based on the shannon entropy, IEEE Transactions on Information Theory 37 (1991) 145–151.

[24] G. Donoso, D. Sanchez, Dialectometric analysis of language variation in twitter, in: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, Valencia, Spain, 2017, pp. 16–25.

[25] A. Kutuzov, S. Touileb, P. Mæhlum, T. R. Enstad, A. Wittemann, Nordiachange: Diachronic semantic change dataset for norwegian, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022, European Language Resources Association, 2022, pp. 2563–2572. URL: https://aclanthology.org/2022.lrec-1.274.

[26] J. Chen, E. Chersoni, C.-r. Huang, Lexicon of changes: Towards the evaluation of diachronic semantic shift in Chinese, in: N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, L. Borin (Eds.), Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 113–118. URL: https://aclanthology.org/2022.lchange-1.11. doi:10.18653/v1/2022.lchange-1.11.

[27] J. Chen, E. Chersoni, D. Schlechtweg, J. Prokic, C.-R. Huang, Chiwug: Diachronic word usage graphs for chinese (2023). URL: https://doi.org/10.5281/zenodo.10023263. doi:10.5281/zenodo.10023263.

[28] P. Basile, A. Caputo, T. Caselli, P. Cassotti, R. Varvara, A diachronic italian corpus based on "l'unità", in: J. Monti, F. Dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2769/paper_44.pdf.

[29] B. McGillivray, D. Schlechtweg, H. Dubossarsky, N. Tahmasebi, S. Hengchen, Dwug la: Diachronic word usage graphs for latin (2021). URL: https://doi.org/10.5281/zenodo.5255228. doi:10.5281/zenodo.5255228.

[30] D. Schlechtweg, F. D. Zamora-Reina, F. Bravo-Marquez, N. Arefyev, Sense through time: diachronic word sense annotations for word sense induction and lexical semantic change detection, Language Resources and Evaluation (2024). URL: http://dx.doi.org/10.1007/s10579-024-09771-7. doi:10.1007/s10579-024-09771-7.

[31] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[32] M. T. Pilehvar, J. Camacho-Collados, WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 1267–1273. URL: https://doi.org/10.18653/v1/n19-1128. doi:10.18653/v1/n19-1128.

[33] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[34] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, A. Bouras, A survey of clustering algorithms for big data: Taxonomy and empirical analysis, IEEE transactions on emerging topics in computing 2 (2014) 267–279. URL: https://ieeexplore.ieee.org/iel7/6245516/6939750/06832486.pdf.