The Corpora They Are a-Changing: a Case Study in Italian Newspapers

Pierpaolo Basile🌿 **Annalina Caputo**🍀 **, Tommaso Caselli**🌷 **,**
Pierluigi Cassotti🌿 **, Rossella Varvara**🌱
🌿Dept. of Computer Science, University of Bari
🍀 ADAPT Centre School of Computing, Dublin City University
🌷 CLCG, University of Groningen
🌱 Dept. of Computer Science, University of Turin
{pierpaolo.basile,pierluigi.cassotti}@uniba.it
annalina.caputo@dcu.ie, t.caselli@rug.nl
rossella.varvara@unito.it

Abstract

The use of automatic methods for the study of lexical semantic change (LSC) has led to the creation of evaluation benchmarks. Benchmark datasets, however, are intimately tied to the corpus used for their creation questioning their reliability as well as the robustness of automatic methods. This contribution investigates these aspects showing the impact of unforeseen social and cultural dimensions. We also identify a set of additional issues (OCR quality, named entities) that impact the performance of the automatic methods, especially when used to discover LSC.

1 Introduction

Natural languages are *de facto* living entities always subject to change and evolution. The diachronic dimension of natural language has played a pivotal role in the history of Linguistics. Understanding and explaining why a community of speakers "speak" as they do is of primary importance to access one's cultural heritage and perspectives on the world.

In recent years, the Natural Language Processing (NLP) community has developed an interest in historical linguistics, and in particular in the study of lexical semantics change (LSC). Previous work has investigated LSC using different approaches, including statistical tests over time period (Popescu and Strapparava, 2013), supervised methods (Mihalcea and Nastase, 2012), count-based distributional approaches (Gulordava and Baroni, 2011), sense-based methods (Kim et al., 2014; Mitra et al., 2014; Frermann and Lapata, 2016), and neural language models (Hamilton et al., 2016a,b; Schlechtweg et al., 2018; Orlikowski et al., 2018; Brandl and Lassner, 2019; Gonen et al., 2020; Giulianelli et al., 2020; Schlechtweg et al., 2020). This has been possible thanks to two factors: increased availability of machine-readable texts covering different periods and increased processing capabilities. The use of computational models for studying LSC is not free from problems, however, as highlighted by Hengchen et al. (2021).

Almost every computational model for LSC is grounded on the Distributional Hypothesis of meaning according to which "the meaning of a word is its use" (Wittgenstein, 2010) and the "difference in meaning correlates with difference in distribution" (Harris, 1954). Distributional models are powerful, yet they suffer from some limitations, namely: (i) they require large amount of text; (ii) they are sensitive to the type of texts and the distribution (i.e., frequency) of the lexical items; and (iii) they tend to conflate different types of information and variables such as semantics, social and topical information.

This contribution investigates two strictly connected aspects: the reliability of LSC benchmark data and the sensitivity of a state-of-the-art approach for LSC, grounded on the distributional hypothesis, when changing the source corpus. The results of our work will help to shed light on systems' robustness and stability by verifying whether methods tuned on one corpus can be directly applied to another.

Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021, pages 14–20
August 6, 2021. ©2021 Association for Computational Linguistics

2 Methodology

To test benchmark independence and models' robustness for LSC, we design a set of experiments using two source corpora, a common benchmark, and a common architecture for LSC detection.

The first corpus is the "L'Unità" corpus (Basile et al., 2020a). It covers a time span between 1945–2014 and it has been collected, pre-processed, and released for the DIACR-Ita (Diachronic Lexical Semantics in Italian) task (Basile et al., 2020b), a LSC change shared task for Italian. Texts were extracted from PDF files by using the Apache Tika library[1] and pre-processed with spaCy[2] for tokenization, PoS-tagging, lemmatization, named entity recognition and dependency parsing. The second corpus was obtained by crawling a publicly available digital archive of the Italian newspaper "La Stampa". The corpus covers a shorter time period (1945–2005) and it was pre-processed using the same tools and pipeline of "L'Unità". Each corpus is split into two sub-corpora, C_1 and C_2, covering different time periods. Table 1 summarises the basic statistics of corpora and the time periods of each sub-corpus.

Corpus	Subcorpus	Tokens
L'Unità	C_1 [1945 – 1970]	52,287,734
L'Unità	C_2 [1990 – 2014]	196,539,403
La Stampa	C_1 [1945 – 1970]	670,281,513
La Stampa	C_2 [1990 – 2005]	1,193,959,080

Table 1: Corpora statistics.

The corpora present two major differences. First, as shown in Table 1, the number of tokens in "La Stampa" is consistently larger than "L'Unità". Second, the political and social orientations of the two newspapers are different. Historically, "L'Unità" has been the official newspaper of the Italian Communist Party and of its successors PDS/DS. "La Stampa" is the oldest newspaper in Italy, traditionally it has voiced centrist and liberal positions.

The only benchmark for Italian has been proposed in the context of DIACR-Ita. The dataset contains 18 target lemmas, 6 of which are instances of a LSC. The dataset was manually created using the "L'Unità" corpus, where a valid LSC corresponds to the acquisition of a new meaning by a target word in C_2.

As architecture for automatic LSC detection, we obtain comparable diachronic representations of word meanings by re-implementing the Word2Vec Skipgram model (Mikolov et al., 2013) with Orthogonal Procrustes (OP-SGNS) (Hamilton et al., 2016b). In particular, we adopted the implementation proposed by Kaiser et al. (2020), a state-of-the-art system that ranked 1^{st} both at DIACR-Ita and at SemEval 2020 Task 1: Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020). Model parameters are reported in Appendix A. Word embeddings were generated using lemmas to reduce sparseness and facilitate the evaluation against the benchmark.

3 Testing for Robustness and Independence

Testing for robustness and consistency for LSC is not trivial since it requires to distinguish between two strictly connected dimensions: (i) reliability of the benchmark (dataset dimension), and (ii) variations in data distributions (corpora dimension). The first dimension (dataset) is analysed by comparing on the DIACR-Ita benchmark the performances of the same model trained on the two corpora. The corpora dimension is investigated by manually inspecting the disagreements between the model predictions. All 18 target words in the benchmark satisfy a minimal frequency threshold of 10 both in C_1 and C_2 in "La Stampa", allowing us to reliably compare the results.

To study the reliability of the DIACR-Ita benchmark with respect to the underlying corpus, for each target word in the benchmark, we computed the cosine similarity of its embedding representation from each sub-corpus (C_1 and C_2). To account for the random initialisation of the OP-SGNS parameters, we ran 10 experiments with different initialisations and averaged the results. The system accuracy is computed as the fraction of correctly predicted words over the total number of words in the benchmark. A target word is deemed as an instance of LSC when its cosine similarity across the two time periods is below a given threshold λ^*.

Since the focus is on the reliability of the benchmark across corpora, and not the system performances, the threshold λ^* for each corpus is set up to the value that maximises the system performance on the corpus.

Using the optimal threshold, our implementation of OP-SGNS obtained an accuracy of $0.96 \pm .02$

[1] https://tika.apache.org/

[2] https://spacy.io/

when trained on “L’Unità” and $0.83 \pm .00$ when trained on “La Stampa”, a difference spawned by the incorrect classification of the words *ape* (LSC), *rampante* (LSC), and *brama* (stable).

To understand the role of the two corpora, we compare the target word similarities between C_1 and C_2 on the two corpora. Figure 1a and Figure 1b illustrate the similarities of the stable and LSC target words, respectively. Overall, the identification of LSC target words seems consistent among the two corpora, and lets us assume that the benchmark is reliable and the algorithm is robust.

We further analyse the system’s disagreements by manually exploring their occurrences in each corpus for every time period.[3] For the target *brama* (‘yearning’), “La Stampa” indicates a potential LSC. The manual inspection, however, has confirmed the annotation in the benchmark (i.e., a stable meaning) showing that the change is triggered by the presence of this word in band names in the C_2 portion of the corpus. *Ape* (‘bee’) is listed in the benchmark as an LCS, since in C_2 it refers not only to the insect, but also to a three-wheeled vehicle. Despite this new sense is present in the C_2 sub-corpus of “La Stampa”, the difference in similarity is above the threshold. Interestingly, in this corpus we observe the three-wheeled vehicle sense also in C_1, especially as part of paid advertisements. This points to a bias in the corpus (i.e, “L’Unità”) used to create the benchmark, namely the lack of (or extremely limited) presence of advertisements, which has obfuscated the occurrence of the three-wheeled vehicle sense and suggested *ape* as a good candidate for an LSC. *Ape* is interesting also for another reason: the discrepancy between when it was first on the market (1948) and its first attestation in the Sabatini Coletti dictionary (1983). Further related to the more commercial nature of the “La Stampa”newspaper is the higher difference in similarity with respect to the “L’Unità” for the word *rampante* (‘rampant’/‘high-flying’). In “La Stampa”, the word occurs also in C_1 as part of the book title “Il barone rampante”; this has mitigated the variation in context of usage with the occurrences of *rampante* in C_2.

4 Models into the Wild

We further extended the analysis to the whole common vocabulary of the two corpora to test the robustness of the computational model. In particular, we consider the vocabulary intersection V of the two sub-corpora, that consists of 48,681 lemmas. Then, we compute the two sets X and Y of cosine similarities for all the words in V. A first analysis was conducted to understand to which extent the rank order of the two sets X and Y are correlated. The Spearman Correlation between the two sets is 0.67 (p-value < 0.01), which indicates a positive correlation between the two rank orders, suggesting that the output of OP-SGNS is similar across the two corpora. The plots of the correlations are reported in Figure 2 in Appendix B.

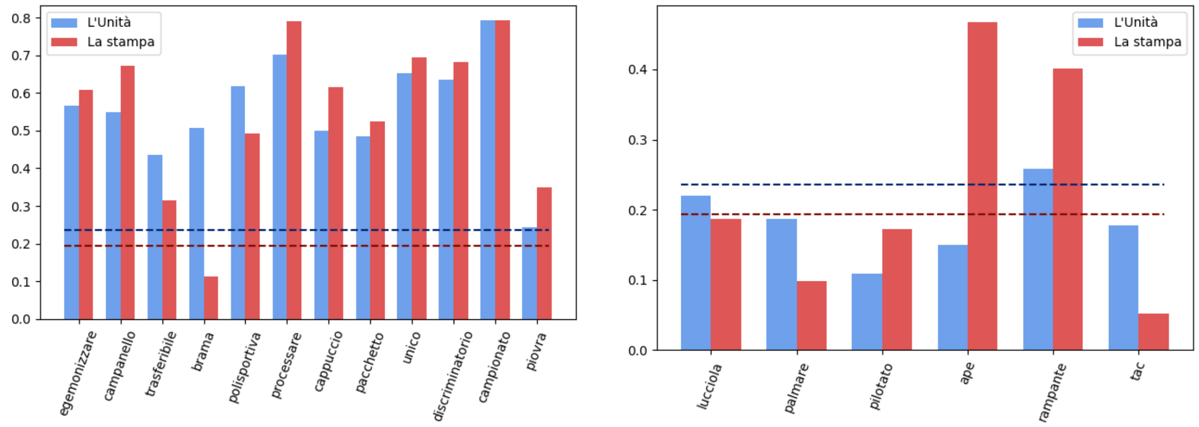
In this analysis, the optimal thresholds cannot be computed due to the lack of a gold-standard for the whole vocabulary intersection V. Potential LSC instances are identified by using as threshold the difference between the average of the cosine similarities (μ) and the standard deviation (σ) over the set V:

$$LCS(X) = \{t_i \in V \mid x_i < \mu(X) - \sigma(X)\}$$

Where t_i is the term associated with the ith similarity $x_i \in X$. Similarly, we compute the set $LCS(Y)$. The intersection of the two LCS sets consists of 2,283 lemmas. A quick inspection of the proposed LSCs immediately triggers observations concerning two aspects: (i) the well formedness of a lemma; and (ii) the presence of named entities (NEs). By well formedness, we refer to the lemma being an actual word attested in a reference dictionary of Italian. Indeed, some of the lemmas with the lowest similarity scores, e.g., *gaucha*, *bwa*, *bill*, *-anche*, do not appear to be well formed Italian words. Reasons for this are to be found in the quality of the digitized versions of the documents of the two corpora, the presence of foreign words (e.g., *frere*, French for ‘brother’), as well as tokenization errors of the pre-processing tool. We use the list of lemmas in the Sabatini Coletti dictionary to filter out all of these entries.

NEs appear to be an additional source of noise. Lemmas like *albertarelli*, *beraudo*, *napoleoni*, *armellini*, are all instances of NEs referring to people’s surnames. We automatically filter NEs in two steps: (i) for each word in a sequence tagged as NE by spaCy, we retrieve and store separately the corresponding lemma; (ii) every candidate LSC lemma is matched against the list generated in (i), greedily filtering all lemmas found to be part of a NE.

[3] We use NoSketch Engine `https://nlp.fi.muni.cz/trac/noske`.



(a) Stable targets.

(b) LSC targets.

Figure 1: LSC change scores computed using cosine similarity on both “L’Unità”and “La Stampa”corpus. The dashed lines indicate the λ^* thresholds, computed respectively on “L’Unità”and “La Stampa”corpus. Similarities below the thresholds trigger an LSC.

After the filtering, only 232 lemmas remain. We sample 50 lemmas (approx. 20%), for a manual inspection. For each lemma, we collected its definitions and the associated year(s) of first attestation from the Sabatini Coletti. Then we manually explored the context of occurrence of each lemma in each time period for each corpus. The manual validation followed a similar approach to the creation of DIACR-Ita gold standard: a lemma is considered to be undergone an LSC only if the definition(s) of the sense are attested in C_2 and not in C_1. The analysis was conducted by only one annotator, who is one of the authors of this paper. By simply using the date of first attestation in the dictionary, 37 lemmas do not qualify as having undergone LSC between the two time periods. Of the remaining 13 lemmas: three have no date of first attestation; five lemmas have a date of first attestation after 1970 (i.e. these lemmas were not used before); and five lemmas present new senses. However, when considering only those lemmas with a new sense attested after 1970, this list reduces to two lemmas.

The manual exploration of the contexts of occurrence in both corpora of the 50 lemmas showed that only four of them (8% of the total sample) can be considered correct examples of an LSC. Two of them, *palmare* (‘obvious’/‘palmar’/‘hand-held computer’) and *patteggiare* (‘negotiate’/‘plea’), are also attested in the Sabatini Coletti. The remaining two, *handicappare* (‘to handicap’) and *orgasmo* (‘orgasm’), indicate a change of use rather than an actual change of meaning. In particular, *handicappare*, and namely its participial form, was used during the 80s/90s to refer to people with disabilities, extending the initial meaning in C_1 of “to assign an handicap to a team”. The use of the word with this meaning is now derogatory and it is not attested in the dictionary. On the other hand, *orgasmo* was used in C_1 in its figurative meaning of great or extreme anxiety, e.g. “nell’orgasmo del momento” (‘in the excitement of the moment’). On the other hand, in C_2 is used with reference to sex and sexuality. Three additional lemmas are signalled as lexical changes: *pula*, *doc*, and *tac*. However, they are officially attested as different lemmas in the Sabatini Coletti, thus implying homonymy. All remaining entries are false positives being either NEs or OCR errors. For the NEs, these are cases where the NE also corresponds to a lemma in the reference dictionary. A good example of this is *borsellino*. In C_1, both corpora present context of use with the dictionary meaning of “a small purse”. However, in C_2, the contexts of use refer to the judge Paolo Borsellino[4], killed in a terrorist attack by the Mafia.

NEs introduce additional challenges while constructing a benchmark for LSC, especially when they are homonyms with common nouns. A viable solution to this problem would be to detect and disregard from the corpus those entities that are homonyms of common nouns. This also calls for the development of more robust systems for NE detection: besides our efforts at filtering NEs, lots of them have remained as potential targets of LSC.

[4] https://en.wikipedia.org/wiki/Paolo_Borsellino

5 Conclusion and Future Work

This contribution has tested the reliability of the DIACR-Ita benchmark for LSC when the underlying corpus used to train and detect LSCs varies. Furthermore, it has scrutinised the robustness of the LSCs, detected by a common algorithm, across different corpora.

Although preliminary, our results indicate that: (i) social and cultural dimensions must be carefully considered when creating LSC benchmark since potential positive examples may be biased; (ii) current approaches to unsupervised LSC are sensitive to the used corpora; (iii) quality of the data (i.e., OCR rendering) and the presence of NEs, especially homonyms with common nouns, are major sources of errors when such automatic methods are applied to actively discover cases of LSC. Strictly connected to this latter aspect is the hiatus between the results of the algorithm against the benchmark and its use "in the wild". This calls for the development of different and more realistic evaluation protocols for unsupervised LSC and research programmes to address the availability of high quality, distributable diachronic corpora.

Besides these limitations, the use of LSC methods on sources with clear differences along social, political, and cultural dimensions could promote a cross-fertilisation of disciplines.

As future work, we plan to extend our analysis to both other corpora and languages, as well to other lexical change detection algorithms, in order to confirm the validity of our findings.

Acknowledgements

Part of this research was conducted with the financial support of Science Foundation Ireland (Grant Agreement No. 13/RC/2106) at the ADAPT SFI Research Centre at Dublin City University; the EVALITA4ELG project, funded by ELG (European Language Grid) Pilot Projects Open Call 1 (Grant Agreement No. 825627 – H2020, ICT 2018-2020 FSTP). The ADAPT Centre for Digital Media Technology is funded by SFI through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF, Grant No. 13/RC/2106_P2).

References

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020a. A diachronic italian corpus based on "l'unità". In *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.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020b. Diacr-ita @ EVALITA2020: overview of the EVALITA2020 diachronic lexical semantics (diacr-ita) task. In *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.

Stephanie Brandl and David Lassner. 2019. Times are changing: Investigating the pace of language change in diachronic word embeddings. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 146–150, Florence, Italy. Association for Computational Linguistics.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. ACL.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings* *of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. *arXiv preprint arXiv:2101.07668*.

Jens Kaiser, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020. Op-ims@ diacr-ita: Back to the roots: Sgns+ op+ cd still rocks semantic change detection. *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org*.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 259–263, Jeju Island, Korea. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029, Baltimore, Maryland. Association for Computational Linguistics.

Matthias Orlikowski, Matthias Hartung, and Philipp Cimiano. 2018. Learning diachronic analogies to analyze concept change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11, Santa Fe, New Mexico. Association for Computational Linguistics.

Octavian Popescu and Carlo Strapparava. 2013. Behind the times: Detecting epoch changes using large corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 347–355, Nagoya, Japan. Asian Federation of Natural Language Processing.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

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, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Ludwig Wittgenstein. 2010. *Philosophical investigations*. John Wiley & Sons.

A OP-SGNS Parameters

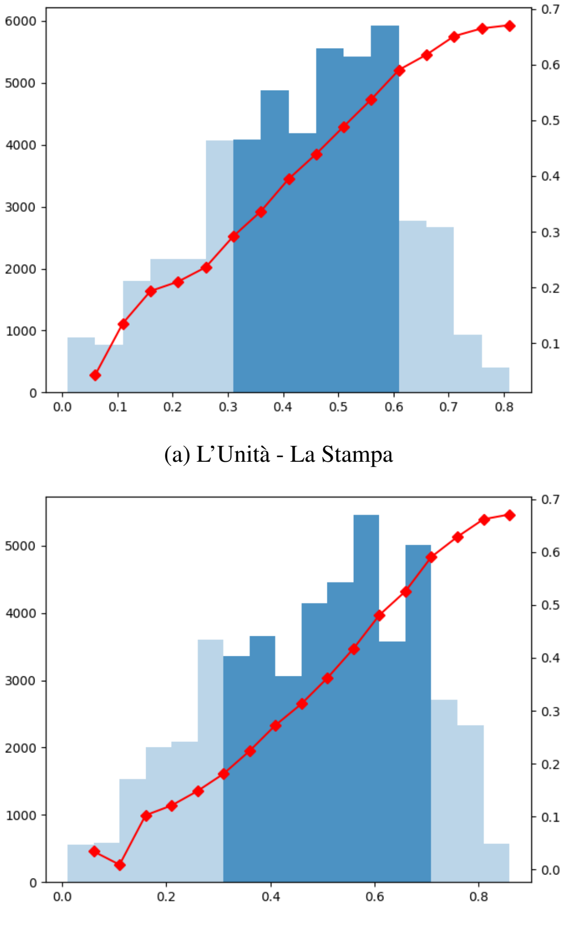
Parameter	Value
learning rate	0.025
min. frequency	10
downsampling rate	0.001
training epochs	5
negative sampling	5
context window	5
vector dimension	300

Table 2: OP-SGNS Parameters for the creation of the word embeddings.

The initial learning rate is set to 0.025, with a negative sampling of 5 and a context window size fixed to 5.

B Cosine similarities: Spearman Correlations

Figure 2 shows the plots of the Spearman correlations between the two sets of ranked similarities computed over the two sub-corpora, C_1 and C_2, of "L'Unità" and "La Stampa", respectively. The cosine similarities are binned in bin of size 0.05 in the interval $[0.0, 0.9]$. The background histogram reports the binned cosine similarity distribution for "L'Unità" (Figure (a)) and "La Stampa" (Figure (b)). The foreground red plot shows the corresponding Spearman correlation values when computed against the "La Stampa" (Figure (a)) and "L'Unità" (Figure (b)), respectively.



(a) L'Unità - La Stampa

(b) La Stampa - L'Unità

Figure 2: Correlation plots.