Characterizing Linguistic Shifts in Croatian News via Diachronic Word Embeddings

David Dukić **Ana Barić** **Marko Čuljak** **Josip Jukić** **Martin Tutek**
TakeLab, Faculty of Electrical Engineering and Computing, University of Zagreb
{name.surname}@fer.hr

Abstract

Measuring how semantics of words change over time improves our understanding of how cultures and perspectives change. Diachronic word embeddings help us quantify this shift, although previous studies leveraged substantial temporally annotated corpora. In this work, we use a corpus of 9.5 million Croatian news articles spanning the past 25 years and quantify semantic change using skip-gram word embeddings trained on five-year periods. Our analysis finds that word embeddings capture linguistic shifts of terms pertaining to major topics in this timespan (COVID-19, Croatia joining the European Union, technological advancements). We also find evidence that embeddings from post-2020 encode increased positivity in sentiment analysis tasks, contrasting studies reporting a decline in mental health over the same period.[1]

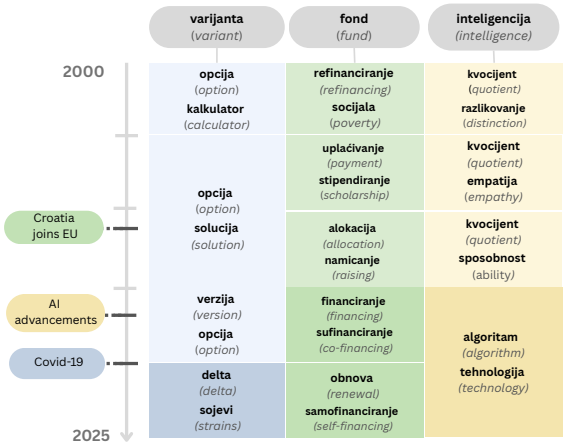

Figure 1: Linguistic shifts in Croatian news outlets over 25 years, driven by three major events: EU membership in 2013, technological progress in 2017, and COVID-19 in 2020.

1 Introduction

The progress of culture and technology is reflected in language, which adapts to incorporate novel meanings into existing words or by entirely changing their semantics. Such changes exhibit systematic regularities with respect to word frequency and polysemy (Bréal, 1904; Ullman, 1962), and can be detected by studies on distributed word representations (Hamilton et al., 2016b). Studies of diachronic word embeddings have detected known changes in word meaning in English-language books spanning multiple centuries. However, such analyses are limited to languages historically abundant in text corpora, as learning high-quality distributed word representations requires diverse contexts. In our work, we rely on a Croatian online news corpus containing articles from the last 25 years (Dukić et al., 2024). We investigate whether major topics in this period are reflected in word semantics and evaluate the practical implications of semantic shift on the use case of sentiment analysis.

We split the corpus into five periods of equal duration, train distributed word representations (Mikolov et al., 2013) for each period, and verify their quality. Next, we select three major topics that likely influenced the meaning of Croatian words during these periods and semi-automatically curate a list of related words for each topic. We show that these words undergo strong linguistic shifts (Hamilton et al., 2016a), acquiring new meanings and demonstrating the rapid impact of narrative on distributional semantics (see Figure 1).

To evaluate whether linguistic shifts affect word representations in practice, we first align word embeddings from different periods, then transfer such aligned embeddings onto a model based on embeddings from another period and observe the change in average predicted sentiment intensity. We find that embeddings from later periods are *more positive* despite studies showing that mental health has been negatively affected (Rozanov et al., 2019; Cullen et al., 2020). In short, our contributions are

[1] https://github.com/dd1497/cro-diachronic-emb

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 108–115

as follows: (1) We train diachronic word embeddings on a corpus of Croatian news articles, which we make available for further studies;[2] (2) We show that corpora spanning short timespans accurately reflect major topics through linguistic shifts of associated words; (3) We find that the sentiment of word embeddings trained on news corpora becomes more positive in recent periods.

2 Related Work

Various studies explore word embeddings as a diachronic tool (Hamilton et al., 2016a,b; Schlechtweg et al., 2019; Fišer and Ljubešić, 2019; Kurtyigit et al., 2021; Schlechtweg et al., 2024, *inter alia*). By leveraging methods from distributional semantics, which encode individual words in vector spaces based on co-occurrence (Mikolov et al., 2013), researchers study how global and local neighborhoods of individual words change over time (Hamilton et al., 2016b). There is a variety of causes driving semantic shift, with two major ones being *linguistic shift*, where words take on a new meaning while retaining previous ones, and *cultural shift*, where technological progress completely alters the way a word is used (Hamilton et al., 2016a). In our work, we follow the methodology used by Hamilton et al. (2016b), apply it to a corpus of Croatian newswire texts, and extend the setup to evaluate practical effects of *linguistic shift* on major topics and sentiment analysis.

The majority of diachronic embedding studies explore corpora spanning several centuries, grounded in books (Hamilton et al., 2016b,a; Schlechtweg et al., 2019; Kurtyigit et al., 2021). Due to the lack of such corpora of sufficient scale in Croatian, we leverage a recently introduced dataset of Croatian newswire corpora (Dukić et al., 2024), which covers a shorter period of 25 years. Despite the narrower timeframe, we hypothesize that the corpus sufficiently captures the diachronic shift in word meaning, which we experimentally verify in this work.

3 Methodology

3.1 Diachronic Word Embeddings

Dataset. We train word embeddings on the TakeLab Retriever corpus of Croatian newswire articles (Dukić et al., 2024). The corpus consists of 9,450,929 articles crawled from 33 Croatian news outlets across 25 years (2000–2024) and contains around 3.7 billion words (see Table 1 for more details). We use spaCy `hr_core_news_lg` (Honnibal et al., 2020) to sentencize, tokenize, and tag parts of speech in the corpus. As the Croatian language is highly inflectional, we lemmatize the corpus with the lexicon-based MOLEX lemmatizer (Šnajder et al., 2008) and differentiate between homonyms with part-of-speech tags obtained using the Croatian spaCy tagger applied to raw words from articles. We split the corpus into 5 five-year periods.

Period	#Words	#Unique words
1 (2000–2004)	53,062,322	589,769
2 (2005–2009)	158,028,732	1,191,784
3 (2010–2014)	551,701,502	2,583,363
4 (2015–2019)	1,170,882,497	3,462,601
5 (2020–2024)	1,753,495,356	3,975,631
Total	3,687,170,409	11,803,148

Table 1: The number of words and unique words per 5 five-year periods in the Croatian online news corpus.

Method. We use the skip-gram with negative sampling (SGNS) method from Word2Vec (Mikolov et al., 2013) to train our word embeddings. We use the GENSIM implementation of SGNS to train the embeddings (Řehůřek and Sojka, 2010). We list the hyperparameter values and hardware details in Appendix C.

3.2 Embedding Quality

We validate the quality of the learned embeddings on two word similarity corpora for Croatian: CroSemRel450 (Janković et al., 2011) and CroSYN (Šnajder et al., 2013). CroSemRel450 contains human-annotated pairs of words rated for semantic relatedness, while CroSYN is a synonym choice dataset comprising one correct synonym and three unrelated options for each target word.

3.3 Topical Linguistic Shift

We hypothesize that diachronic embeddings over periods can reveal significant topical linguistic shifts. To unveil these shifts, we curate words pertaining to three major topics relevant globally and/or to Croatia: the COVID-19 crisis, Croatia joining the *European Union (EU)*, and *technological progress*. We expect COVID-19 to produce the highest shift in the fifth period, joining the EU in the second, third, and fourth periods (as Croatia entered the EU in 2013), and technological progress in the fourth and fifth periods (digitalization after

[2] https://www.takelab.fer.hr/retriever/cro-diachronic-emb.zip

entering EU and proliferation of AI in the fifth period). Finding no substantial shifts for verbs or adjectives, we focus on the change in nouns as they are more prone to linguistic shifts (Hamilton et al., 2016a). We measure the shift of each word using the cumulative shift score, based on the halved cosine distance (cos) over neighboring periods:

$$D_c = \sum_{i=1}^{4} \frac{1 - \cos(\mathbf{v}_i, \mathbf{v}_{i+1})}{2}.$$

For this analysis, we use Procrustes alignment (Schönemann, 1966) to align word embeddings across periods. We begin by recursively aligning pairs of embeddings, starting from the most recent, fifth period (2020-2024), and then moving toward the earlier ones. Let $\mathbf{E}_t$ denote the embedding matrix for period t, and let $\mathrm{PA}(\mathbf{A}, \mathbf{B})$ denote the Procrustes alignment of matrix $\mathbf{A}$ to $\mathbf{B}$. We use $\mathbf{E}_t^*$ to denote the aligned embeddings for period t. The alignment procedure can be written recursively as:

$$\mathbf{E}_t^* = \begin{cases} \mathbf{E}_5, & \text{if } t = 5, \\ \mathrm{PA}\big(\mathbf{E}_{t+1}^*, \mathbf{E}_t\big), & \text{if } t \in \{1, 2, 3, 4\}. \end{cases}$$

Further details are provided in Appendix A.

3.4 Sentiment Shift

Distributed word representations capture contextual cues helpful in determining the tone and sentiment of texts, serving as a more robust and effective alternative to lexicon-based and traditional machine learning approaches (Zhang et al., 2018; Al-Saqqa and Awajan, 2020; Wankhade et al., 2022). To quantify sentiment shifts in our corpus, we train a classifier C_i for each period t_i using embeddings E_i computed on the corpus from t_i. Each classifier predicts the sentiment label (positive, neutral, or negative) of a text sequence based on the average of the word embeddings within the sequence. Next, we compute the average sentiment of a classifier C_i on a test set using the word embeddings from E_i and denote this quantity by $\bar{s}_{i \leftarrow i}$. We repeat the same procedure for C_i with Procrustes-aligned embeddings from each other period E_j^*, $j \neq i$ to obtain quantities $\bar{s}_{i \leftarrow j}$. We hypothesize that using the embeddings from a period with an overall more positive (or negative) sentiment biases the classifier accordingly. Thus, we estimate the sentiment shift between periods t_i and t_j with $\bar{d}_{i \leftarrow j} = \bar{s}_{i \leftarrow j} - \bar{s}_{i \leftarrow i}$. We conduct the experiment on two Croatian news sentiment analysis datasets: STONE (Barić et al., 2023), comprising solely of news headlines, and 24sata (Pelicon et al., 2020), which focuses on full news articles.

To further validate the quality of word embeddings for sentiment drift, we also analyze the distribution of sentiment scores of news articles in each period. Specifically, we sample 25k unlabeled articles per period from the TakeLab Retriever corpus. To automatically assign sentiment labels, we train a transformer-based classifier using BERTić (Ljubešić and Lauc, 2021), on the STONE and 24sata datasets, respectively. Further details on the training procedure and hyperparameter settings can be found in Appendix B.

4 Results

4.1 Embedding Quality

We report the results of embedding quality evaluation in Table 2. We measure the Spearman correlation between embedding-based cosine similarity and human judgments on the word similarity dataset CroSemRel450. Additionally, we compute contrastive spread on the CroSYN dataset to evaluate how clearly word embeddings distinguish synonyms from unrelated words. Focusing on nouns, adjectives, and verbs, we calculate the contrastive spread as the difference between a word's cosine similarity to its synonym and its similarity to an unrelated word, where higher scores reflect stronger semantic discrimination. Overall, we find a moderate positive correlation of our estimated similarity with human judgments for word similarity across all periods. Both measurements indicate that embedding quality improves in later periods, highlighting the influence of data quantity on embedding quality. In contrast to similar embedding approaches for word similarity evaluation, our results are slightly worse albeit comparable ($\rho = 0.62$; Zuanovic et al. (2014)).

Period	Similarity (↑)	Contrastive spread (↑)		
		Noun	Adjective	Verb
1 (2000–2004)	0.49†	0.08	0.07	0.05
2 (2005–2009)	0.49†	0.14	0.10	0.09
3 (2010–2014)	0.52†	0.21	0.18	0.15
4 (2015–2019)	0.51†	0.26	0.23	0.19
5 (2020–2024)	0.51†	0.27	0.25	0.21
All (2000–2024)	0.52†	0.32	0.27	0.23

Table 2: Intrinsic embedding evaluation: word similarity († = $p < 0.001$, Spearman correlation) and contrastive spread by period and part of speech.

4.2 Topical Linguistic Shift

We provide a summary of words exhibiting most prominent shifts in Table 3. We show that neighboring words of top-shifting words inside a topic can pinpoint the period when words acquire new meanings. We provide complete results of the top-picked shifting words inside each topic: COVID-19, EU, and technology in Table 5 in Appendix A.

COVID-19. The COVID-19 crisis, which began in 2020, is reflected in the semantic shifts of words that were previously topically neutral, such as *maska (mask)* and *varijanta (variant)*. The word *maska* changes from referring to a clothing item to an instrument for reducing viral transmission. The noun *varijanta* changes its dominant meaning during the fifth wave from an option or possibility to characterizing different strains (variants) of the coronavirus. The word *pandemija (pandemic)* changed a lot during the 25 year period due to its connection to diverse diseases (from Ebola to flu and finally COVID-19). However, it was always used in the context of infectious diseases.

EU. The evolution of EU-related terminology mirrors Croatia's path through three periods: considering EU membership, preparing for admission, and utilizing the benefits of being a member state. The word *integracija (integration)* changes from emphasizing bureaucratic *harmonization* (2000–2004) to entering the *union* (2013) and practical implementation and *Europeanization* by 2020–2024. *Komisija (commission)* increasingly associates with legislative bodies such as the *council*, *ombudsman*, and *parliament*, reflecting the importance of legal procedures for Croatia's admission into the EU. Finally, *fond (fund)* shifts from associating with financial terms such as *quotation* and *portfolio* to *sufinanciranje (co-financing)* and *obnova (renewal)* in the last two periods, reflecting usage of EU funds.

Technology. Technological advancements are also reflected in linguistic shifts. *Vjerodajnica (credential)* evolves from diplomatic words (*delegation*, *telegram*) to digital identifiers (*password*, *document*), signalling the transition into the digital era. *Inteligencija (intelligence)* changes from abstract cognitive attributes (*quotient*, *erudition*) to AI concepts (*algorithms*, *automation*), reflecting the post-2010 AI revolution. Finally, *privola (consent)* shifts from legal, in-person authorization to digital mechanisms such as *kolačić (cookie)* and *pohrana (data storage)*.

4.3 Sentiment Shift

We report results of sentiment shift on STONE and 24sata datasets in Figure 2. We observe that transferring aligned embeddings from later periods into earlier periods increases average predicted sentiment, while the opposite holds when transferring embeddings from earlier periods to later. Additionally, we observe a similar trend regarding the increased share of positive words in more recent periods using a SentiLex lexicon for Croatian (Glavaš et al., 2012).

We further investigate the increase in news positivity, through the distribution of sentiment labels for both news headlines and full articles across different time periods Figure 3. We find that in general, the amount of articles labeled as positive increases at the expense of neutral ones. The proportion of negative labels also slightly increased over time, particularly in news headlines. These results corroborate the findings of sentiment shift, indicating an increase of positivity in news in recent periods.

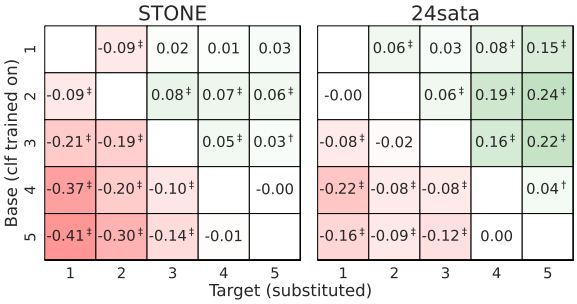

Figure 2: Sentiment shift between periods. Each cell (i,j) contains the value $\bar{d}_{i \leftarrow j}$. We compute statistical significance levels of the quantities being greater than zero using 10-fold cross validation. We denote $p < 0.05$ with † and $p < 0.01$ with ‡.

We hypothesize that increased positivity in news may be driven by one of several phenomena observed in media communication. Increased positivity could be the a reaction to general negativity, influenced by the decline of mental health in the general population (Rozanov et al., 2019; Cullen et al., 2020). The increase in positivity could also be attributed to online news covering more diverse, less serious topics, or the increase in satirical or comedic articles. Another potential factor is the increased polarization of media discourse, where news content is becoming more extreme in its use of emotionally charged language to elicit reactions from readers (Rozado et al., 2022). Nonetheless,

Topic	Top shift	Top five noun neighbors				
		2000–2004	**2005–2009**	**2010–2014**	**2015–2019**	**2020–2024**
COVID-19	varijanta (variant) $D_c = 0.53$	opcija (option) kalkulator (calculator) mogućnost (possibility) solucija (solution) opipavanje (palpation)	opcija (option) solucija (solution) alternativa (alternative) mogućnost (possibility) verzija (version)	opcija (option) solucija (solution) verzija (version) inačica (version) alternativa (alternative)	verzija (version) opcija (option) solucija (solution) alternativa (alternative) vrsta (type, kind)	delta (delta) sojevi (strains) mutacija (mutation) podvrsta (subtype) virus (virus)
EU	fond (fund) $D_c = 0.32$	portfelj (portfolio) kotacija (quotation) benefit (benefit) refinanciranje (refinancing) socijala (poverty)	alokacija (allocation) benefit (benefit) tranša (tranche) uplaćivanje (payment) stipendiranje (scholarship)	alokacija (allocation) namicanje (raising) kapital (capital) dividenda (dividend) banka (bank)	financiranje (financing) sufinanciranje (co-financing) alokacija (allocation) novac (money) proračun (budget)	alokacija (allocation) ulaganje (investment) sufinanciranje (co-financing) obnova (renewal) samofinanciranje (self-financing)
Tech	inteligencija (intelligence) $D_c = 0.51$	kvocijent (quotient) razlikovanje (distinction) instinkt (instinct) jasnoća (clarity) evolucija (evolution)	kvocijent (quotient) empatija (empathy) nadarenost (giftedness) opažanje (perception) habitus (habitus)	kvocijent (quotient) sposobnost (ability) upućenost (familiarity) racionalnost (rationality) erudicija (erudition)	algoritam (algorithm) tehnologija (technology) automatizacija (automation) kvocijent (quotient) robotika (robotics)	tehnologija (technology) algoritam (algorithm) automatizacija (automation) učenje (learning) robotika (robotics)

Table 3: Topical linguistic shift with respect to three topics: COVID-19, *European Union (EU)*, and *Technology (Tech)*. We pick one top shift noun word per topic based on the cumulative shift score (second column). For each of the picked words, we show the top five nearest noun neighbors over five periods. Translations are in parentheses.

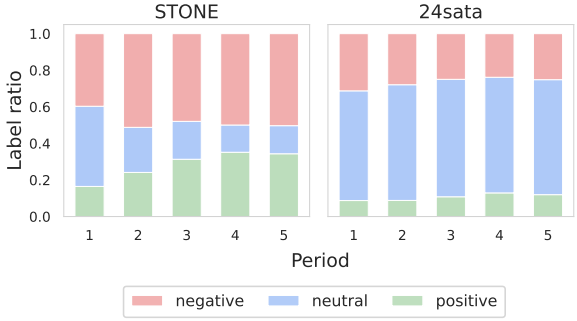

Figure 3: Change of predicted sentiment ratios when using classifiers trained on STONE and 24sata to categorize a sample of articles from Retriever. The trend of increased news polarization is more evident when using classifiers trained on STONE, but the same is evident for 24sata.

we believe that this phenomenon, in which sentiment expressed in news articles contrasts broader negativity, warrants further study as it may affect the quality of models trained on corpora from different time periods.

5 Conclusion

We apply diachronic word embedding analysis to Croatian, a language scarce in historical corpora. By training diachronic embeddings on Croatian online news articles spanning the last 25 years, we successfully detect linguistic shifts pertaining to recent major events, exhibited by existing words acquiring new meanings or completely changing how they are used. These results show that linguistic shifts can also be detected in shorter time spans. We also reveal practical implications of linguistic shifts on sentiment analysis, showing that word meanings from recent periods tend to be more positive, contrasting with research indicating an increase in negativity over the same period.

Limitations

In our experiments, we analyze only five-year periods, revealing some regularities that might be too coarse- or fine-grained for others. We experimented with two-year periods but found them too fine-grained. Future works can vary the duration of periods. We use a lexicon-based context-free lemmatizer (MOLEX), which could be error-prone and introduce noise to the experiments. The distribution of article count per period varies significantly as earlier periods have fewer articles. This fact influences the quality of produced word embeddings and could bias the results. Finally, we only explore a single distributed word embedding method in SGNS, the results of which need not generalize to other methods.

References

Samar Al-Saqqa and Arafat Awajan. 2020. The use of Word2vec model in sentiment analysis: A survey. In *Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control*, AIRC '19, page 39–43, New York, NY, USA. Association for Computing Machinery.

Ana Barić, Laura Majer, David Dukić, Marijana Grbeša-zenzerović, and Jan Snajder. 2023. Target two birds with one SToNe: Entity-level sentiment and tone analysis in Croatian news headlines. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 78–85, Dubrovnik, Croatia. Association for Computational Linguistics.

Michel Bréal. 1904. *Essai de sémantique (science des significations)*. Hachette.

Walter Cullen, Gautam Gulati, and Brendan D Kelly. 2020. Mental health in the COVID-19 pandemic. *QJM: An International Journal of Medicine*, 113(5):311–312.

David Dukić, Marin Petričević, Sven Ćurković, and Jan Šnajder. 2024. TakeLab Retriever: AI-driven search engine for articles from Croatian news outlets. *arXiv preprint arXiv:2411.19718*.

Darja Fišer and Nikola Ljubešić. 2019. Distributional modelling for semantic shift detection. *International Journal of Lexicography*, 32(2):163–183.

Goran Glavaš, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Experiments on hybrid corpus-based sentiment lexicon acquisition. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 1–9, Avignon, France. 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 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Vedrana Janković, Jan Šnajder, and Bojana Dalbelo Bašić. 2011. Random indexing distributional semantic models for Croatian language. In *Text, Speech and Dialogue*, pages 411–418. Springer.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6985–6998. Association for Computational Linguistics.

Nikola Ljubešić and Davor Lauc. 2021. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830.

Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlj, and Senja Pollak. 2020. Zero-shot learning for cross-lingual news sentiment classification. *MDPI*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Diego Rozado, Russell Hughes, and Jamin Halberstadt. 2022. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models. *PLoS One*, 17(10):e0276367.

Vsevolod Rozanov, Tanja Frančišković, Igor Marinić, Maria-Magdalena Macarenco, Marina Letica-Crepulja, Lana Mužinić, Ruwan Jayatunge, Merike Sisask, Jan Vevera, Brenda Wiederhold, and 1 others. 2019. Mental health consequences of war conflicts. *Advances in psychiatry*, pages 281–304.

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, Frank D Zamora-Reina, Felipe Bravo-Marquez, and Nikolay Arefyev. 2024. Sense through time: Diachronic word sense annotations for word sense induction and lexical semantic change detection. *Language Resources and Evaluation*, pages 1–35.

Peter H Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10.

Jan Šnajder, B Dalbelo Bašić, and Marko Tadić. 2008. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing & Management*, 44(5):1720–1731.

Jan Šnajder, Sebastian Padó, and Željko Agić. 2013. Building and evaluating a distributional memory for Croatian. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Sofia, Bulgaria. Association for Computational Linguistics.

Stephen Ullman. 1962. An introduction to the science of meaning. *New York: Barnes & Nobel*.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 8(4):e1253.

Leo Zuanovic, Mladen Karan, and Jan Šnajder. 2014. Experiments with neural word embeddings for croatian. In *Proceedings of the 9th Language Technologies Conference*, pages 69–72.

A Topical Linguistic Shift

Analyzing Topical Linguistic Shift Embeddings. We create a common vocabulary between periods to measure only the words with sufficient frequency in each period. In total, the five periods share 348,679 words. We curate a list of potential word shifters for each topic (both top shifters by D_c and additional words we expected to shift). We remove words with frequency less than 1,000 over 25 years for each topic list separately. Next, for each word from the curated topic list, we compute D_c and pick 20 candidates with the highest shift score for further analysis. For each candidate, we find its nearest 1,000 noun neighbors by cosine distance. Out of these 1,000, we pick 20 that occur at least 20 times in each period. We analyze the linguistic shift of words using their neighbors for each topic and pick the most interesting words representing the topic shift with its closest and most representative neighbors.

Full Topical Shift Results. We report full results of topical shift on terms pertaining to major events in Table 5.

B Sentiment Shift

Training Setup for Sentiment Classifiers To train the sentiment classifier for Croatian news, we use the STONE and 24sata datasets with the BERTić model (Ljubešić and Lauc, 2021). For the STONE dataset, we utilize only the tone labels, as they capture the overall tone of the headline, aligning with our definition of sentiment. We achieve an F1 score of 0.77 on STONE and 0.73 on the 24sata dataset.

C Hyperparameters and Hardware Details

We train word embeddings and the sentiment regressor on a machine with 2x AMD Epyc 7763 CPUs and 512 GB of RAM. In Table 4, we report the hyperparameters used in word embedding training. Our setup mostly follows that of Hamilton et al. (2016b) with a key difference that we do not restrict our vocabulary but train the embeddings on all the words in the corpus. We discard only punctuation words identified by a part-of-speech tagger and lowercase all the words before training. When training classifiers for sentiment analysis (cf. §3.4), we use the implementation of logistic regression from scikit-learn (Pedregosa et al., 2011) with the default hyperparameters.

We train both BERTić sentiment classifiers on an NVIDIA RTX 3090 GPU with 24GB RAM using CUDA 12.9 and the HuggingFace Trainer[3] library. We employ the default hyperparameters provided by the Trainer and train for 3 epochs with a batch size of 8.

Hyperparameter	Value
vector_size	300
window	4
negative	5
sample	1e-5
alpha	0.02
epochs	5

Table 4: Hyperparameters for word embedding training. The names of hyperparameters in the first column match the argument names when initializing a GENSIM Word2Vec model.

[3] https://huggingface.co/docs/transformers/main_classes/trainer

Topic	Top shift words	Top five noun neighbors				
		2000–2004	**2005–2009**	**2010–2014**	**2015–2019**	**2020–2024**
COVID-19	maska (mask) $D_c = 0.66$	kabanica (raincoat) kombinezon (coverall) značka (badge) lampica (little lamp) šminka (makeup)	lice (face) šilterica (visor cap) kombinezon (coverall) šešir (hat) frak (tailcoat)	rukavica (glove) šminka (makeup) šešir (hat) pancirka (flak jacket) štitnik (protector, shield)	perika (wig) povez (band, patch) šminka (makeup) kaciga (helmet) štitnik (protector, shield)	nošenje (wearing) rukavica (glove) nenošenje (not wearing) distanca (distance) pleksiglas (plexiglass)
	pandemija (pandemic) $D_c = 0.61$	ebola (Ebola) incidencija (incidence) ospice (measles) virus (virus) epidemiolozi (epidemiologists)	SARS (SARS) ebola (Ebola) gripa (flu) ospice (measles) pojavnost (prevalence)	SARS (SARS) ebola (Ebola) ospice (measles) kuga (plague) virolog (virologist)	ebola (Ebola) bolest (disease) ospice (measles) kriza (crisis) epidemiolozi (epidemiologists)	epidemija (epidemic) korona (corona) kriza (crisis) lockdown (lockdown) COVID (COVID)
	varijanta (variant) $D_c = 0.53$	opcija (option) kalkulator (calculator) mogućnost (possibility) solucija (solution) opipavanje (palpation)	opcija (option) solucija (solution) alternativa (alternative) mogućnost (possibility) verzija (version)	opcija (option) solucija (solution) verzija (version) inačica (version) alternativa (alternative)	verzija (version) opcija (option) solucija (solution) alternativa (alternative) vrsta (type, kind)	delta (delta) sojevi (strains) mutacija (mutation) podvrsta (subtype) virus (virus)
European Union	integracija (integration) $D_c = 0.39$	harmonizacija (harmonization) agenda (agenda) aspirant (aspirant) iskorak (step forward) kohezija (cohesion)	unija (union) agenda (agenda) fragmentacija (fragmentation) dobrosusjedstvo (neighborliness) ulazak (entrance)	unija (union) implementacija (implementation) razvoj (development) dobrosusjedstvo (neighborliness) međuovisnost (interdependence)	unija (union) razvoj (development) inkluzija (inclusion) povezivanje (connection) jačanje (strengthening)	implementacija (implementation) uključenost (inclusion) povezivanje (connection) razvoj (development) europeizacija (Europeanization)
	komisija (commission) $D_c = 0.34$	ombudsman (ombudsman) sukladnost (compliance) delegacija (delegation) unija (union) nacrt (draft)	delegacija (delegation) unija (union) mjerodavnost (competence) arbitraža (arbitration) instancija (instance)	unija (union) monitoring (monitoring) ombudsman (ombudsman) povjerenik (commissioner) vijeće (council)	unija (union) prijedlog (proposal) vlada (government) parlament (parliament) vijeće (council)	unija (union) smjernica (guideline) vlada (government) članica (member) parlament (parliament)
	fond (fund) $D_c = 0.32$	portfelj (portfolio) kotacija (quotation) benefit (benefit) refinanciranje (refinancing) socijala (poverty)	alokacija (allocation) benefit (benefit) tranša (tranche) uplaćivanje (payment) stipendiranje (scholarship)	alokacija (allocation) namicanje (raising) kapital (capital) dividenda (dividend) banka (bank)	financiranje (financing) sufinanciranje (co-financing) alokacija (allocation) novac (money) proračun (budget)	alokacija (allocation) ulaganje (investment) sufinanciranje (co-financing) obnova (renewal) samofinanciranje (self-financing)
Technology	vjerodajnica (credential) $D_c = 0.56$	otpravnik (ambassador's deputy) delegacija (delegation) diplomat (diplomat) brzojav (telegram) monsinjor (monsignor)	otpravnik (ambassador's deputy) diplomat (diplomat) useljeništvo (immigration) podtajnik (undersecretary) parafiranje (initialing)	telefaks (fax) adresar (address book) ovjera (certification) fotokopija (photocopy) tiskanica (form)	formular (form) iskaznica (ID card) brzojav (telegram) pošta (mail) veleposlanik (ambassador)	građani (citizens) iskaznica (ID card) lozinka (password) putovnica (passport) dokument (document)
	inteligencija (intelligence) $D_c = 0.51$	kvocijent (quotient) razlikovanje (distinction) instinkt (instinct) jasnoća (clarity) evolucija (evolution)	kvocijent (quotient) empatija (empathy) nadarenost (giftedness) opažanje (perception) habitus (habitus)	kvocijent (quotient) sposobnost (ability) upućenost (familiarity) racionalnost (rationality) erudicija (erudition)	algoritam (algorithm) tehnologija (technology) automatizacija (automation) kvocijent (quotient) robotika (robotics)	tehnologija (technology) algoritam (algorithm) automatizacija (automation) učenje (learning) robotika (robotics)
	privola (consent) $D_c = 0.50$	staratelj (guardian) očevidnik (register) autorizacija (authorization) ovlaštenje (authorization) pozivatelj (caller)	uvjetovanje (conditioning) obvezivanje (commitment) direktiva (directive) konzultiranje (consultation) suodlučivanje (co-decision)	pohrana (storage) odobrenje (approval) ustanoviti (establish) suglasnost (accord) supotpis (co-signature)	pohrana (storage) kolačić (cookie) povjerljivost (confidentiality) suglasnost (accord) stranica (page)	pohrana (storage) suglasnost (accord) kolačić (cookie) dopuštenje (permission) stranica (page)

Table 5: Full topical linguistic shift results with respect to three topics: COVID-19, *European Union*, and *Technology*. We pick three top shift noun words per topic based on the cumulative shift score (second column). For each of the picked words, we show the top five nearest noun neighbors over five periods. Translations are in parentheses.