How BERT Speaks Shakespearean English? Evaluating Historical Bias in Masked Language Models

Miriam Cuscito[1], Alfio Ferrara[2] and Martin Ruskov[3]

[1]Dipartimento di Lettere e Filosofia, Università degli Studi di Cassino e del Lazio Meridionale, Via Zamosch 43, 03043 Cassino, Italy
[2]Dipartimento di Informatica "Giovanni Degli Antoni", Università degli Studi di Milano, Via Celoria 18, 20133 Milano, Italy
[3]Dipartimento di Lingue, Letterature, Culture e Mediazioni, Università degli Studi di Milano, Piazza Sant'Alessandro 1, 20123 Milano, Italy

Abstract

In this paper, we explore the idea of analysing the historical bias of masked language models based on BERT by measuring their adequacy with respect to Early Modern (EME) and Modern (ME) English. In our preliminary experiments, we perform fill-in-the-blank tests with 60 masked sentences (20 EME-specific, 20 ME-specific and 20 generic) and three different models (i.e., BERTBase, MacBERTh, BL Books). We then rate the model predictions according to a 5-point bipolar scale between the two language varieties and derive a weighted score to measure the adequacy of each model to EME and ME varieties of English.

Keywords

Masked Language Models, Early Modern English, Historical Bias

1. Introduction

Masked language models (MLMs) are deep neural language models which create contextualised word representations, in the sense that the representation for each word depends on the entire context in which it is used. That is to say, word representations are a function of the entire input sentence. Such models are designed to have high predictive capabilities and usually pre-trained on large textual corpora. This makes them closely tied to the domains on which they were trained and dependent on the infrastructure upon which they are based. The presence of various biases in MLMs has been extensively studied, typically with the aim of proposing effective mitigation strategies [1, 2, 3, 4]. However, there are instances where the bias in certain MLMs is not necessarily negative. This is particularly true when the bias manifested in the language reflects its socio-temporal context. This bias could be advantageous for tasks that demand such socio-temporal staging [5, 6].

In this paper, we explore the idea of analysing the bias by focusing on the major syntactic, semantic, and grammatical differences between two varieties of the English language: Early Modern (EME) and Modern (ME). More precisely, we propose a method and a measure of adequacy to test the adherence of MLMs to the natural language variety of interest. In particular, we assess the level of diachronic bias of three MLMs: Bert-Base-Uncased [7] (referred as BERT Base here)[1]; MacBERTh [8][2]; and Bert British Library Books English [9] (BL Books) [3]. In our preliminary experiments, we perform tests with 60 masked questions in which the models have the task to predict the masked word in the sentence.

3rd Workshop on Artificial Intelligence for Cultural Heritage (AI4CH 2024, https://ai4ch.di.unito.it/), co-located with the 23nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024). 26-28 November 2024, Bolzano, Italy
miriam.cuscito@unicas.it (M. Cuscito); Alfio.Ferrara@unimi.it (A. Ferrara); martin.ruskov@unimi.it (M. Ruskov)
https://www.unicas.it/dottorato/elenco-dottorati-di-ricerca-delluniversita-degli-studi-di-cassino-e-del-lazio-meridionale/corso-di-dottorato-in-testi-contesti-e-fonti-dallantichita-alleta-contemporanea/dottorandi/xxxvii-ciclo/miriam-cuscito/ (M. Cuscito); https://islab.di.unimi.it/team/alfio.ferrara@unimi.i (A. Ferrara); https://islab.di.unimi.it/team/martin.ruskov@unimi.it (M. Ruskov)
0009-0003-9585-2803 (M. Cuscito); 0000-0002-4991-4984 (A. Ferrara); 0000-0001-5337-0636 (M. Ruskov)

[1]https://huggingface.co/bert-base-uncased
[2]https://huggingface.co/emanjavacas/MacBERTh
[3]https://huggingface.co/bigscience-historical-texts/bert-base-blbooks-cased

We then rate the proposed responses according to a 5-point bipolar scale between the two language variants and derive a weighted score from the response probabilities and their respective scores on the scale.

These results, although preliminary, might suggest a method applicable in the digital humanities when MLMs are employed for the analysis of historical corpora.

2. Related Work

If it is true that language shapes culture while it is shaped by it [10], language models in general – and MLMs in particular – constitute a still partially covered mirror of this dual relationship. Not only can a MLM be tested based on its level of representativeness of the language to determine its reliability, but also it can tell us about linguistic, social, and historical phenomena that concern the culture tied to that specific language. In other words, a MLM could be a valuable tool towards the expansion of the broader social knowledge of a given culture, rightfully becoming part of the basic tools of Cultural Analytics discussed by Manovich [2020]. According to Bruner's [1984] pragmatic-cultural perspective, learning a language also means learning the cultural patterns associated with it. Similarly, analysing the language in its various realisations would mean having the opportunity to visualise the underlying cultural patterns.

Moreover, MLMs can be highly beneficial also for philological [13], pragmatic [14], critical [6], and literary work [15]. However, the effectiveness of these models depends on their ability to adapt to language specificity in its historical dimension. This is typically achieved by training models on historical text corpora. However, the difficulty of accessing large historical documentary collections means that the models available are still few and requires verifying whether they adapt effectively to the historical linguistic context.

BERT is a foundational masked language model (MLM) which to date is the most widely adopted [16]. A number of studies have explored different forms of bias in BERT [1, 2, 3]. Three BERT-based MLMs are of particular interest for our study: *(i)* Bert-Base-Uncased [7], created from a corpus of texts from Wikipedia and BookCorpus and a model of contemporary language, which we use as a control condition in our experiment; *(ii)* MacBERTh [8], pre-trained on texts from 1500 to 1950; and *(iii)* Bert British Library Books English, pre-trained on contemporary texts and fine-tuned on historical texts from the 19th century to the present.

3. Method

To evaluate the adequacy of MLMs on a test set, we define a temporal valence task consisting of a collection of test sentences, each with a masked token (i.e., word). This is a typical fill-in-the-blank task, where the models are required to predict the masked token. Formally, we consider the following three sets: *(i)* we denote with $\mathcal{S}$ the set of all test *sentences*, *(ii)* with $\mathcal{V}$ we denote a set of *vocabulary* words, and *(iii)* with $\mathcal{T} = \{-1, -0.5, 0, 0.5, 1\} \subset \mathbb{R}$, we denote a 5-point bipolar *temporal valence scale*, where -1 represents the farthest historical period and 1 the closest to today.

With the above notation, for each of the masked sentences (denoted as $s \in \mathcal{S}$), we define a function $\rho : \mathcal{S} \to \mathcal{T}$ representing the *sentence temporal valence score*. This function indicates the period from which the masked sentence is typical.

Then, we calculate a *token-in-sentence temporal valence score* $\sigma : \mathcal{V} \times \mathcal{S} \to \mathcal{T}$, indicating the score of a token substituting the sentence mask.

The mentioned *temporal valence scores* are assigned arbitrarily according to the research hypotheses. Taking this study as an example, the criterion used to determine each score was the degree of alignment of certain sentences or tokens with a specific historical period on a philological-linguistic basis. Scholars wishing to delve into language study using this methodological approach can selectively choose the score to assign to their test set based on their specific research needs. The versatility of the proposed methodology is evident in its adaptability to a diverse array of fields of interest. This flexibility enables

BERT Base			MacBERTh			BL books		
token	p	σ	**token**	p	σ	**token**	p	σ
thou	0.712	-1.0	thou	0.987	-1.0	thou	0.573	-1.0
you	0.101	0.0	not	0.008	0.0	you	0.246	0.0
i	0.085	0.0	you	0.004	0.0	he	0.102	0.0
she	0.055	0.0	ye	0.001	-1.0	she	0.040	0.0
he	0.048	0.0	he	0.000	0.0	Thou	0.038	-1.0
β	-0.712		β	-0.988		β	-0.612	
δ	0.856		δ	0.994		δ	0.806	

Table 1
Scores of the models for the historically biased sentence "Why willt [MASK] be offended by that?" (temporal valence $\rho = -1$)

researchers to seamlessly integrate personalized metrics, ensuring a tailored approach to analysis without undermining the inherent consistency of the results.

As an example of *temporal valence score*, given EME (Early Modern English) as the farthest period (i.e., $-1 \in \mathcal{T}$) and ME (Modern English) as closest (i.e., $1 \in \mathcal{T}$), if we consider the sentence $s_1 =$ *"Why wilt [MASK] be offended by that?"* we have $\rho(s_1) = -1$ as s_1 is a representative sentence for EME, and $\sigma(\text{“}thou\text{”}, s_1) = -1$, because in this context "thou" is indicative for EME. On the other hand, $\sigma(\text{“}not\text{”}, s_1) = 0$, because "not" is neutral regarding the two language varieties.

Given a model m, for the masked token in each sentence ($s \in S$), we have the set of $\{w_1, w_2, \ldots, w_n\} \subset \mathcal{V}$ of n words predicted by m for s, that are associated with the vector of corresponding probabilities from this model, shown in Equation 1.

$$\mathbf{p}_m = (p(w_1), p(w_2), \ldots, p(w_n))^T \tag{1}$$

For this set, using the temporal valence score σ, we define a token-in-sentence temporal valence score vector $\mathbf{x}_m$ for m given the sentence s as in Equation 2.

$$\mathbf{x}_m = (\sigma(w_1, s), \sigma(w_2, s), \ldots, \sigma(w_n, s))^T \tag{2}$$

This allows us to define the *bias* of a model regarding the sentence as the dot product of the model-derived probabilities and the token valence scores, providing us with a weighted score as in Equation 3, and effectively getting a single value measurement from the two vectors above.

$$\beta(m, s) = \mathbf{x}_m^T \mathbf{p}_m \tag{3}$$

We can also proceed to define the *domain adequacy* of a model with respect to a sentence s (see Equation 4), based on the difference between the sentence temporal valence score $\rho(s)$ and the model bias $\beta(m, s)$. To do this, we consider the difference between the model bias and the sentence temporal valence (disregarding which one is larger), and project it on the unit interval, making sure that more similar values lead to higher adequacy scores.

BERT Base			MacBERTh			BL books		
token	p	σ	**token**	p	σ	**token**	p	σ
here	0.924	0.0	hither	0.740	-1.0	here	0.556	0.0
back	0.066	0.5	down	0.170	0.0	back	0.224	0.5
there	0.004	-0.5	thus	0.045	-0.5	down	0.091	0.0
forth	0.003	-0.5	in	0.025	0.0	in	0.071	0.0
out	0.003	1	again	0.020	0.0	again	0.056	0.0
β	0.032		β	-0.762		β	0.112	
δ	0.984		δ	0.619		δ	0.944	

Table 2
Scores for the neutral sentence "Have you come [MASK] to torment us before the time?" ($\rho = 0$)

BERT Base			MacBERTh			BL books		
token	p	σ	**token**	p	σ	**token**	p	σ
orientation	0.720	1.0	##ists	0.493	-0.5	##ly	0.582	0.0
misconduct	0.112	1.0	offenders	0.165	1.0	##ists	0.355	0.0
minorities	0.067	1.0	characters	0.130	0.5	##ally	0.024	0.0
partners	0.052	1.0	drunkards	0.117	0.0	men	0.021	0.0
harassment	0.048	1.0	delinquents	0.095	0.5	to	0.018	0.0
β	1.000		β	0.031		β	0.000	
δ	1.000		δ	0.516		δ	0.500	

Table 3
Scores for the sentence "Should men who are known sexual [MASK] be given a platform?", biased towards modernity ($\rho = 1$)

$$\delta(m, s) = 1 - \frac{\mid \rho(s) - \beta(m, s) \mid}{2} \tag{4}$$

Examples of three sentences classified in different periods are provided in Tables 1, 2 and 3, which show the corresponding values for ρ, p, σ, $\beta(m, s)$ and $\delta(m, s)$.

4. Evaluation

We test our metrics with three BERT-based linguistic models we consider relevant for the varieties of the English language of interest: *(i)* Bert-Base-Uncased, *(ii)* MacBERTh, and *(iii)* BL Books. In accordance with the objectives of this study, the choice of models reflects a specific interest in language; therefore, they can be replaced to best fit any other specific interest in diachronic language analysis. For the test we used 60 word-masked sentences, specifically created for this study. To create the test set, we relied on different types of written language: contemporary standard, journalistic language, social media non-standard, and Early Modern language.

The elements to be masked were selected based on their belonging to specific word classes known to have suffered more exposure to the diachronic variation of the English language: pronouns, verbs, adverbs, adjectives, and nouns. Of the 60 sentences[4], 20 are selected to be suggestive for the EME variety of English, further 20 – as suggestive for ME, and final 20 are generic. Once the test set was complete, a *temporal valence score* was assigned to each sentence (see ρ in Section 3) based on their level of chronological markedness.

The test set was administered to the three MLMs, and the suggested words with their probability were collected. The resultant vocabulary was marked independently from the models that provided it by setting the *token-in-sentence temporal valence score* (i.e. σ) to each word, based on an estimation of proximity of the token's meaning to a certain linguistic variety in the context in which it appeared. Notably, during this phase, our decision was to work on a sentence level (contextually) rather than on a set level (globally). The method proved highly effective in avoiding the risk of semantic flattening, given that almost every word has shown some level of contextual semantic specificity if taken contextually rather than globally. An example is the pronoun *you* in "*fare you well, sir*", which is globally neutral and yet acquires a strong diachronic value if evaluated in its context, in which it appears to be utmost archaic.

Once β and δ were calculated, we proceeded with the analysis of the data and the collection of results. The distribution of the bias score β and the domain adequacy score δ for the sentences in the three groups (i.e., EME, Neutral, and ME) is shown in Figures 1 and 2, respectively.

[4]For transparency and reproducibility purposes, the following anonymous link contains the complete test set with the corresponding values produced during evaluation:
https://tinyurl.com/bert-shakespearean

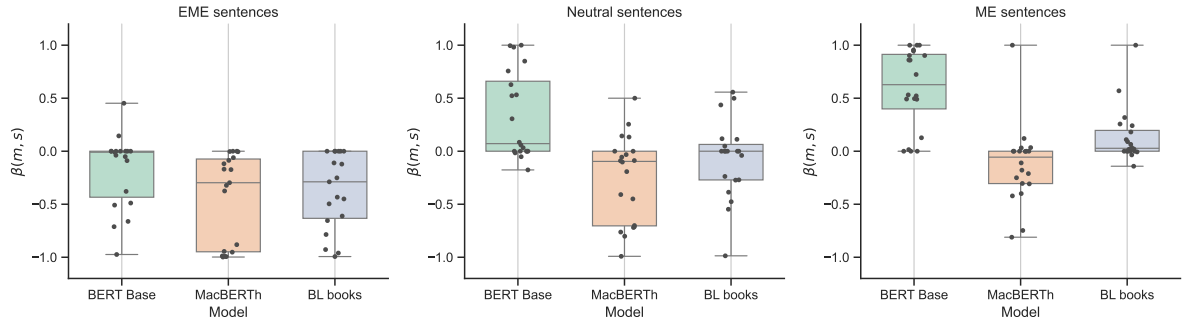

Figure 1: Distribution of the *bias* $\beta(m, s)$ of the three models with respect to the three test sets.

Figure 3 shows that for all three test sets, MacBERTh is most aligned with EME, whereas BERT Base is always most aligned as ME. BL Books shows a tendency towards a more neutral language than the other two models in marked sentences, whilst surprisingly it aligns to ME in neutral sentences.

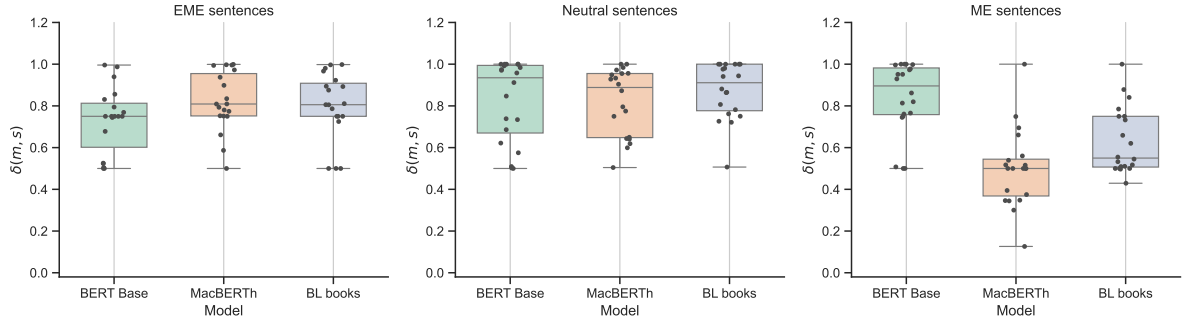

Figure 2: Distribution of the *domain adequacy* $\delta(m, s)$ of the three models with respect to the three test sets.

Figure 2 shows that MacBERTh has best domain adequacy for EME, and BERT Base has best domain adequacy for ME. In the case of the neutral test set, domain adequacy is no less informative. Although the sentences do not inherently carry their expectations regarding language, models appear consistently well-suited to a neutral context, and none of them pushes for strong specificity of their corresponding trained domain. In effect this leaves the sentences close to their original neutrality

This preliminary study provides an illustration of the nature and functioning of the MLMs predictive behaviour. The presence or absence of markedness in the sentences enables all three MLMs to select the type of element which best fits the co-text. So, while for diachronically marked sentences, models without training in that domain attempted to suggest probable solutions, sometimes resulting in a form of linguistically inconsistent mimicry, in unmarked sentences, the models perform exceptionally well, and linguistic inaccuracies are rare.

5. Conclusion

Both notions of *bias* (β) and *domain adequacy* (δ) provide important insights of the nature of the models. The first, β, indicates a tendency in terms of temporal valency. In other words, the interpretation of its value should be considered within the context of the specific dichotomy of language varieties. On the other hand, δ reflects the adequacy for an individual language variety. It successfully captures model tendencies when completing historically predetermined sentences. However, due to its inclusion of ρ, δ is less informative when there is no bias originating from the sentence, i.e. when completing temporally-neutral sentences.

Notably, our measures demonstrate that MacBERTh is better at representing the EME historical context than BL Books. This could possibly be explained by the nature of the models in question. First, MacBERTh is a model created from scratch and trained on texts spanning a time range that takes into account the evolution of the English language from EME to ME. BL Books, on the other hand, was only fine-tuned on texts from the modern period, so it has no direct exposure to EME. It does perform better on ME than MacBERTh and worse than BERT Base. Thus, MacBERTh demonstrates a strong linguistic consistency, given the wide range of language varieties it is trained on, but in tasks related to ME yields worse results than other more specialised models. Simply put, having a specific, narrower domain poses fewer problems when working within it but reveals clear gaps when moving outside that domain.

The notion that LMs can serve as a window into the history of a population is not new, but there is a growing interest in exploring the relationships between these models and the socio-linguistic and socio-cultural contexts [17, 18, 19, 20]. It is equally imperative to establish a procedural framework to address the lack of evaluative methods for these models, as previously hinted at in this text. This is particularly useful when no direct links could be drawn between the corpus used to train the model and the social context of the test set.

Within this evaluation, we created a dedicated test set for each model under scrutiny, drawing upon approaches used for evaluation of bias in MLMs. In creating our test sets, we built our sentences both on logical-semantic and logical-syntactic tasks. Future work could try to create a test set for model interrogation that is culture-oriented, delving into socio-culturally significant elements such as customs, historical events, and attitudes towards social groups – elements recognised as belonging to social knowledge. Alternatively tests could be derived from word in context datasets, such as TempoWiC and HistWiC [21, 22, 23].

Alternatively, the temporal valence of word tokens could be derived not simply from the sentence where they emerge, but from the wider historical context, e.g. from a large corpus, representative for the period [24, 25]. This would allow automating not only the calculation of the token temporal valence σ, but also the identification of sentences that are representative for each historical period. As a consequence, the dependence on manual expert evaluation would be strongly reduced, which would result in both higher reproducibility and wider generalisability of the approach.

This study aims not only to propose a methodology for assessing language models but also to put forth hypotheses for expanding the available tools to humanities scholars interested in studying complex socio-cultural phenomena with an approach which begins by interpreting textual clues and inferring their connections to reality. As such it is also applicable to contexts beyond diachronic, but also across dialects or professional jargons.

Acknowledgments

The research leading to these results has received funding from MUR, PRIN2022 project "MetaLing Corpus: Creating a corpus of English linguistics metalanguage from the 16th to the 18th century", ref.: 202233C93X, funded by the European Union under the programme NextGenerationEU.

References

[1] D. de Vassimon Manela, D. Errington, T. Fisher, B. van Breugel, P. Minervini, Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 2232–2242. doi:10.18653/v1/2021.eacl-main.190.

[2] J. Ahn, A. Oh, Mitigating language-dependent ethnic bias in BERT, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 533–549. doi:10.18653/v1/2021.emnlp-main.42.

[3] M. Mozafari, R. Farahbakhsh, N. Crespi, Hate speech detection and racial bias mitigation in social media based on BERT model, PLOS ONE 15 (2020) e0237861. doi:10.1371/journal.pone.0237861.

[4] D. Nozza, F. Bianchi, D. Hovy, HONEST: Measuring Hurtful Sentence Completion in Language Models, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2398–2406. doi:10.18653/v1/2021.naacl-main.191.

[5] P. J. Corfield, Fleeting gestures and changing styles of greeting: researching daily life in British towns in the long eighteenth century, Urban History 49 (2022) 555–567. doi:10.1017/S0963926821000274.

[6] A. Morollon Diaz-Faes, C. Murteira, M. Ruskov, Explicit references to social values in fairy tales: A comparison between three European cultures, in: M. Hämäläinen, E. Öhman, F. Pirinen, K. Alnajjar, S. Miyagawa, Y. Bizzoni, N. Partanen, J. Rueter (Eds.), Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages, Association for Computational Linguistics, Tokyo, Japan, 2023, pp. 62–75. URL: https://aclanthology.org/2023.nlp4dh-1.8.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[8] E. Manjavacas, L. Fonteyn, Adapting vs. Pre-training Language Models for Historical Languages, Journal of Data Mining & Digital Humanities NLP4DH (2022). doi:10.46298/jdmdh.9152.

[9] S. Schweter, Pretrained Language Models on British Library Corpus, 2024. doi:10.5281/zenodo.10715629.

[10] L. Boroditsky, How Language Shapes Thought, Scientific American 304 (2011) 62–65. doi:10.1038/scientificamerican0211-62.

[11] L. Manovich, Cultural analytics, The MIT Press, Cambridge, Massachusetts, 2020. URL: https://mitpress.mit.edu/9780262037105/cultural-analytics/.

[12] J. Bruner, Pragmatics of Language and Language of Pragmatics, Social Research 51 (1984) 969–984. URL: https://www.jstor.org/stable/40970973.

[13] L. van Lit, Among Digitized Manuscripts. Philology, Codicology, Paleography in a Digital World, Brill, Leiden, The Netherlands, 2019. doi:10.1163/9789004400351.

[14] M. Ruskov, Who and How: Using Sentence-Level NLP to Evaluate Idea Completeness, in: N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, O. C. Santos (Eds.), Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, Communications in Computer and Information Science, Springer Nature Switzerland, Cham, 2023, pp. 284–289. doi:10.1007/978-3-031-36336-8_44.

[15] A. Piper, H. Xu, E. D. Kolaczyk, Modeling Narrative Revelation, in: A. Šeļa, F. Jannidis, I. Romanowska (Eds.), Proceedings of the Computational Humanities Research Conference 2023, volume 3558 of *CEUR Workshop Proceedings*, CEUR, Paris, France, 2023, pp. 500–511.

[16] F. Periti, S. Montanelli, Lexical Semantic Change through Large Language Models: a Survey, ACM Comput. Surv. 56 (2024) 282:1–282:38. URL: https://dl.acm.org/doi/10.1145/3672393. doi:10.1145/3672393.

[17] R. M. M. Hicke, D. Mimno, T5 meets Tybalt: Author Attribution in Early Modern English Drama Using Large Language Models, in: A. Šeļa, F. Jannidis, I. Romanowska (Eds.), Proceedings of the Computational Humanities Research Conference 2023, volume 3558 of *CEUR Workshop Proceedings*, CEUR, Paris, France, 2023, pp. 274–302. URL: https://ceur-ws.org/Vol-3558/#paper2757.

[18] H. Usui, K. Komiya, Translation from historical to contemporary Japanese using Japanese t5,

in: M. Hämäläinen, E. Öhman, F. Pirinen, K. Alnajjar, S. Miyagawa, Y. Bizzoni, N. Partanen, J. Rueter (Eds.), Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages, Association for Computational Linguistics, Tokyo, Japan, 2023, pp. 27–35. URL: https://aclanthology.org/2023.nlp4dh-1.4.

[19] N. Pedrazzini, B. McGillivray, Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers, in: M. Hämäläinen, K. Alnajjar, N. Partanen, J. Rueter (Eds.), Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities, Association for Computational Linguistics, Taipei, Taiwan, 2022, pp. 85–95. URL: https://aclanthology.org/2022.nlp4dh-1.12.

[20] A. Palmero Aprosio, S. Menini, S. Tonelli, BERToldo, the historical BERT for Italian, in: R. Sprugnoli, M. Passarotti (Eds.), Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, European Language Resources Association, Marseille, France, 2022, pp. 68–72. URL: https://aclanthology.org/2022.lt4hala-1.10.

[21] 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, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1267–1273. doi:10.18653/v1/N19-1128.

[22] D. Loureiro, A. D'Souza, A. N. Muhajab, I. A. White, G. Wong, L. Espinosa-Anke, L. Neves, F. Barbieri, J. Camacho-Collados, TempoWiC: An Evaluation Benchmark for Detecting Meaning Shift in Social Media, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3353–3359. URL: https://aclanthology.org/2022.coling-1.296.

[23] F. Periti, H. Dubossarsky, N. Tahmasebi, (Chat)GPT v BERT Dawn of Justice for Semantic Change Detection, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 420–436. URL: https://aclanthology.org/2024.findings-eacl.29.

[24] G. D. Gasperis, P. Pavone, S. Bolasco, A Strategy to Identify the Peculiarity of a Lexicon in the Analysis of a Corpus, in: G. Giordano, M. Misuraca (Eds.), New Frontiers in Textual Data Analysis, Springer Nature Switzerland, Cham, 2024, pp. 105–118. doi:10.1007/978-3-031-55917-4_9.

[25] S. Bolasco, T. De Mauro, L'analisi automatica dei testi: fare ricerca con il text mining, number 922 in Studi superiori Statistica, 1a edizione ed., Carocci, Roma, 2013.