Multichannel Convolutional Transformer and Intertextuality: A Latin Case Study

Laurent Vanni, Hadi Mahmoudi, Dominique Longrée, Damon Mayaffre

Multi-channel Convolutional Transformer and intertextuality : a Latin case study

Laurent Vanni, Hadi Mahmoudi, Dominique Longrée, Damon Mayaffre

Abstract The detection of intertextuality is at the heart of many linguistic studies. A lot of efforts are currently underway to provide tools for analyzing relationships between authors. Most of them use standard statistics to compare textual data and find traces of text reuse from one author to another. The main objective of this work is to provide a new approach based on deep learning architectures and Corpus-Driven analysis. Building on previous contributions, we propose a hybrid architecture called Multi-channel Convolutional Transformer (MCT). Using this method, we develop a new tool for intertextuality detection based on authorship attribution. We have empirically demonstrated its efficiency using a Latin corpus. We conclude that our model can highlight complex linguistic patterns as features responsible for the classification decision. We consider these patterns as new categories of intertextuality traces, complementary to the existing ones.

1 Introduction

According to its etymology, the text (*textum/texere*) is perceived as tissue, a woven fabric[1]. This tissue must be described with textual linguistics as a patchwork: as-

Laurent Vanni
CNRS, UCA, UMR7320 BCL, France e-mail: laurent.vanni@univ-cotedazur.fr

Hadi Mahmoudi
UCA, UMR7320 BCL, France e-mail: hadi mahmoudi@etu.univ-cotedazur.fr

Dominique Longrée
ULiège, L.A.S.L.A, Belgique e-mail: dominique.longree@uliege.be

Damon Mayaffre
CNRS, UMR7320 BCL, France e-mail: damon.mayaffre@univ-cotedazur.fr

[1] This work has been supported by the French National Research Agency through the ANR TRACTIVE project ANR-21-CE38-0012-01 ; ANR PREFAB project ANR-22-CE54-0013-02, 2022-2025 and the UCA AAP CSI 2023 through the MCT project

sembly of sentences or syntagms already produced by others, the composition of themes already dealt with elsewhere, text reuse, presuppositions, allusions, traces, resonance, etc. *Every text is an intertext; other texts are present within it to varying degrees and in more or less recognizable forms.* [2]. Computational linguistics is henceforth challenged to implement the complex concept of intertextuality, and has already obtained some noteworthy results. In terms of the *patchwork* text, the *mosaic* text [15], or the *palimpsest* text [5, 12], the detection of literal repetitions in the form of n-grams shared between the target text and the original text, is thereby rightly proposed by, for example, [8] or more recently [4] who use trained embedding vectors. Basically, it involves comparing a target text A' (influenced text) to an original text A (influencing text). Based on similarity distance measures and the extraction of textual overlapping areas, these approaches make it possible to quantify the presence of A in A' and to show, in context, relevant lexical and syntactic intersections as well as conforming semantic relationships. However, the automatic detection of more tenuous linguistic borrowings, as well as the identification of evanescent traces of one text in another, or of unacknowledged stylistic influences between authors, i.e. intertextuality, still partially resist analysis today. Unlike comparative perspectives, aimed at assessing style differentiation between discrete text samples, this contribution, supported by a corpus-driven approach [20] and deep learning models involved in authorship attribution tasks. By hypothesis, a training dataset intuits and materializes the potential intertexts of a target text. Gathering authors's works into a corpus allows for the immersion of a selected one who is suspected of having been influenced by one or the other. The trained model predicts each text of the target author by attributing it to one or more source authors in the reference corpus. Therefore, intertextuality is plural, by definition. To shed light on the linguistic objects or passages of the target text that may have been borrowed, we invoke the automatic feature extraction capability of trained models for automatic text classification. We have chosen to study two complementary architectures. On the one hand, convolutional networks [14], which detect local saliences in texts [21]. On the other hand, transformer-based architecture [1], which works on the basis of self-attention mechanisms and has exhibited impressive performance in recognizing distance relationships between words [22]. Inspired by recent studies that merge the above-mentioned approaches in image analysis [25] and text classification [18, 9], we suggest an approach that uses both convolutional networks and Transformers trained on Multi-channel data which we call Multi-channel Convolutional Transformers (MCT). The feature extraction mechanism implied in MCT provides a new approach to exploring intertextuality in a corpus-driven environment. We empirically compared MCT to other standard statistical methods by applying them to a Latin corpus. Experiments and results show the added value of this architecture for intertextuality studies. Before diving into more details about MCT, in the next section, we present the model and the process of intertextuality detection.

2 Model

2.1 Multi-Channel CNN and multi-head Attention

It has recently been shown that the combination of Multi-channel CNN architectures and multi-head attention outperforms baseline models in text classification tasks such as Sentiment Analysis [9]. To improve its interpretability and allow the detection of intertextuality traces, the architecture must be adapted to our needs. In linguistics, the order and position of words are crucial in determining the meaning of a sentence. While Recurrent and Convolutional Networks utilize the linear structure of text to detect features in the text, the multi-head Attention layer in the Transformer block considers independently words as single values. Therefore, the words ordering is lost and the attention is constant to the sequence order. The Positional encoding, originally called position embedding in [11] is a technique that compensates for information loss and helps train a model based solely on self-attention. It has recently been shown in image analysis [25] that positional encoding can be replaced by a convolutional embedding (a convolutional feature map) to capture local saliences before processing distant dependencies between words using the multi-head attention embodied in the Transformer block. Our architecture use also the convolutional embedding representing each word, but it is still shifted using the standard positional encoding algorithm. Therefore, the generated embedding vector represents a positional convolutional feature that highlights words based on local salience and position. The main idea of a positional convolutional embedding is to keep the position of each word as information that can be used to analyze complex linguistic patterns. Deep learning models are also known for their ability to handle multidimensional data [7]. Intertextuallity detection tools such as Tesserae platform [8] allow choosing between the complete form of the word, the lemma, the sound, and even a semantic representation (for Latin). The added value of deep learning approaches, compared to standard methods, is the ability to combine these representations and thereafter automatically extract relevant associations. Indeed, a Multi-channel encoding of the text allows the model to detect features from different text representation at the same time. In our study, we focus on three levels of representation (i.e. channels) given by standard annotation tools: the graphic form, the Part-Of-Speech (POS), and the lemma. Multi-channel encoding provides the model with the ability to make a higher level of abstraction of the text in which we expect to retrieve new types of intertextuality traces. However, within standard feed-forward networks the information about the relationships between each channel is shuffled into hidden layers and is therefore lost. To make our Multi-channel architecture interpretable, we train each of the three convolutional layers independently. Then, we use multi-head attention operations to combine the convolutional information into a feature map which represents precisely the relationships between each entity. Technically, the model is trained in two steps and according to the following architecture. First, three standard Convolutional Neural Networks (CNNs) [14] are trained using all three levels of word representation. CNNs consist of an embedding layer, followed by a convolutional layer, and finally Fully Connected layers (FC). The purpose of these sub-models is to manage a pre-training process in each channel and extract local saliences to feed then the Transformer block. The second step is to train the Transformer block based on [1] where the input layer is replaced by concatenating the three convolutional embedding layers. To maintain the integrity of the convolutional markers, the weights of these convolutional embeddings are no longer trainable by the model, which focuses solely on the multi-head attention of the Transformer block. By means of this hybrid network, using convolution and multi-head attention, feature extraction can be refined to a higher level of abstraction, allowing us to develop a new tool for intertextuality extraction based on complex linguistic patterns detection.

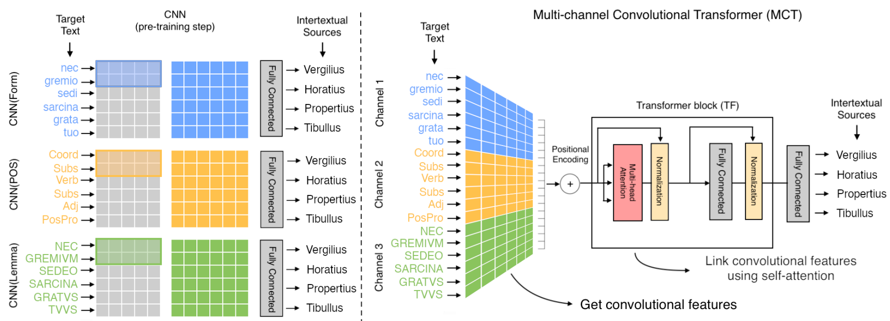

Fig. 1 Multi-channel Convolutional Transformer (MCT). Pre-training step based on CNNs (left part). Transformer with CNNs features as Multi-Channel inputs (right part).

2.2 Intertextuality detection tool

While the main task of the model is to predict an author, our tool puts the model's prediction into service to exhibit traces of intertextuality by using the feature extraction capabilities imparted by the model. In the case of convolutional layers, information extraction is not straightforward. Usually, the convolution process must be transposed (a.k.a *deconvolution*) to be interpretable. However, this method is inadequate for intertextuality detection because it brings in additional operations and outputs approximations of features [21]. Thus, instead of deconvolution, we chose to use specific convolutional operations that do not reduce the data space. To avoid data compression, we use a *stride*[2] set to one row/word and we apply a *padding*, which allow to keep the *same* number of rows/words to precisely preserve dimensionality. This technique maintains the integrity of the text without adding additional layers, and the feature extraction transforms into a latent vector printout of each word. To

[2] Number of rows traversed per one dimensional convolutional slide

enhance feature interpretation of the convolutional layers we apply a Class Activation Mapping method inspired by computer vision works such as GradCAM [19] and GradCAM++ [6] and recent studies about interpreting predictions of NLP models [24]. The key insight of this method is to use the weights of the last Fully Connected layers (FC) of the neural networks to adjust the convolutional embedding representation. Indeed, the FC layers transforms the data abstraction given by the convolutional feature map into one output vector leading to the class prediction, usually via a softmax transformation. When the FC layers are applied to the whole feature map (the whole flatten vector), the output vector leads to the class prediction for the whole text. Conversely, if the FC layers only act on one word/vector they produce a score corresponding to the contribution of that word for each class. This method provides an impressive aid for the interpretation of results in a multi-class classification environment by adjusting the score of a given convolutional feature according to its contribution to the selected class (which can be then positive or negative). Feature extraction using the Transformer block is more standard. The multi-head attention layer returns both the attention output (the result of all the self-attention computations) and the attention scores (The multi-head attention coefficients over attention axes). While only attention output is typically used to feed the last Fully Connected layers, attention scores can be considered as features responsible for the multi-head attention output. The principal assumption here is that attention scores obtained from multi-channel convolutional features allows to represent both syntagmatic and paradigmatic relations between words [17] where the paradigmatic axis is the axis of position-based selection (or substitution) of words and the syntagmatic axis is the axis of their combinations. Figure 1 shows the global structure of our architecture. The input is a sample of target text encoded in three separate channels and the output contains the class probabilities (source authors classification).The main purpose of the MCT model is perform both authorship attribution and complex linguistic markers mining tasks. The features extraction supplied with this architecture provide new intertextuality detection tool which the next section aim to compare with baseline methods.

3 Compared approaches

Quantitative Intertextuality [10] describes several algorithms for practicing quantitative text reuse detection. Automated procedures used in tools such as *Tracer* [3] and *Tesserae* [8] allow the exploration of the *traces* of a source text contained in a target text. These methods are mainly based on word matching using several dimensions: lexical, morphological, phonological, semantic, etc. To evaluate our model, we chose to compare it to Tesserae, one of the most widely used tools for intertextuality detection. In addition, we also compared our model to the statistical calculation of the z-score[3], one of the standard measures used with corpus-driven ap-

[3] z-score refers in this contribution to the specificity score [16]

proaches to detect words that are overused in a given text. The studied corpus brings together a collection of seven classical poets: Catullus, Horace, Juvenal, Lucretius, Propertius, Tibullus and Virgil. Texts are obtained from the L.A.S.L.A.[4] database, which provides a manual tagging of lemmas and part-of-speech (associated with morphological analysis) established by linguists ensuring higher accuracy compared to automatic methods [23]. Composed of more than 60,000 samples[5], this corpus provides an adequate training set to obtain a classifier capable of predicting the label of a new text, possibly related to one of the seven given authors. Before analyzing the text reuse traces detected by our architecture, we tested the accuracy of the MCT models using 10% of the data for validation and 10% for testing. We compared results with FastText [13], three standard CNN models [14], and a transformer-based architecture (TF) [1]. Note that each CNN corresponds to the pre-training step (see section 2.1) of each channel (full Form, POS, Lemma) used by MCT. As expected, MCT achieved the highest performance with 86.95% of accuracy, and the improvements are notable over most of the baseline models we tested[6]. CNNs trained on lemmas and POSs achieved lower but non-random accuracies, extracting information relevant to the authorship attribution task. According to the MCT's accuracy and the fact that the model's architecture is built on top of the other tested CNNs (including Form, POS, Lemma) combined with transformer-based model, we assume that these models are complementary. While each convolutional network captures local saliences on each channel, the multi-head attention embodied in the Transformer block adds a higher level of abstraction that allows the model to reach such a high performance. To question the abstraction level of MCT, we compare the detected linguistic markers with those known by standard methods. With this aim in view, we explore the potential sources of inspiration for a target author in the works of authors present in the training corpus. Among the works of each author, we chose texts available on the Tesserae platform as a test set for our model. We chose Ovid, who was contemporary to Emperor Augustus. Ovid produced abundant and various literary works, either in the form of amorous elegies or subjects of mythological or historical nature. He composed love poems like Propertius who is chosen as the source author for our intertextuality study. In particular, we compared Ovid's *Heroïdes* to Propertius *Elegies*. Using these texts, we evaluated the proportion of markers identified by Tesserae that our model also finds either through one of the convolution layers or the multi-head attention layer. For the convolution layers, the markers correspond to positive scores related to the predicted class (according to our Class Activation Map based algorithm). For the multi-head attention layer, the markers refer to the words that produce or receive high attention scores across any of the channels (The score corresponds to the max values of the multi-head attention scores). Attention scores are normalized to a range between 0 and 100 and only values above 50 are retained. We repeated the same experiment by replacing the Tesserae markers with the z-score calculation allowing us to statistically identify the most overused words

[4] Laboratory for Statistic Analysis of Ancient Languages (L.A.S.L.A) - University of Liège

[5] The corpus is organized into samples of ten words corresponding to the average size of a verse.

[6] When trained on full-forms, FastText reached 85.8% ; CNNs reached 84.06% ; Transformer reached 85.14%

by each author (we set the significance threshold to 5 standard deviations). While the features derived from the z-score are based on word frequency detection, those of Tesserae are drawn from deeper algorithms capable of detecting intertextuality and of text reuse traces. MCT detects the majority of conventional markers obtained both by Tesserae and z-score that contribute to text identification. Indeed, the proportion of classical statistical markers that MCT detect by Convolution or by the Transformer block is always greater than 90%. However, according to the accuracy of the model, the above-mentioned standard statistical markers cannot be the only source of information extraction. Convolutional layers can detect local features composed of multiple words, whereas statistical metrics typically focus on a single word. Furthermore, with MCT, multi-head attention can combine multiple word representations and detect distant dependencies between the convolutional features. To confirm this hypothesis, we reverse the perspective and focuses on the most impactful features detected by our model (i.e., the features with the highest activation score). Using MCT architecture, the convolutional and multi-head attention layers yield activation scores for each word and their related POS and lemma. These scores can be sorted to keep only the highest values for each layer corresponding to the most relevant feature the model uses to predict the class. We observe that the standard statistical markers represent less than 50% of the most relevant features returned by our model. This implies that although MCT places relative importance on well-known markers, it also focuses on other linguistic objects to assign authorship attribution. The next section questions the relevance of the linguistic markers detected by MCT using feature projections as tool for intertextuality detection.

3.1 Intertextuality detection

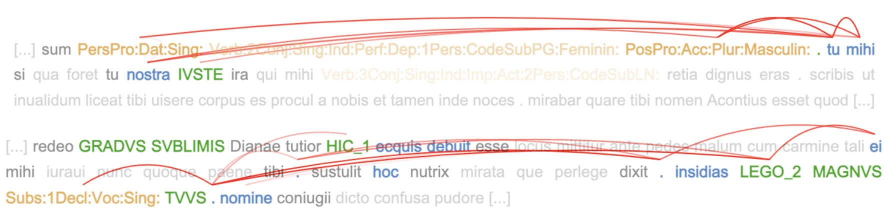

Fig. 2 Examples of features visualisation provided by MCT. Excerpts of Ovid *Heroïdes* : 21, 204-210 and 21, 105-112.

In the following examples, Ovid's *Heroïdes* were divided into samples of fifty words to increase the model's capability to detect long-range dependencies. The model predicts mainly Ovid's *Heroïdes* to Propertius (42% of the samples), then Tibullus (34%), Virgil (19%), and Juvenal, Horace, Catullus, Lucretius (< 2%). Markers highlighted by convolution and distant associations revealed by self-

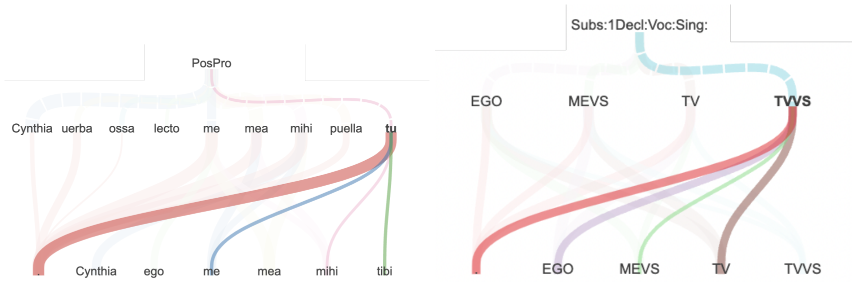

Fig. 3 Cooccurrence analysis on Propertius of the Possessive Adjective-Pronom (on left side) and Substantive of the 1st declension in the Vocative Singular form (on right side). The larger is the line, the higher is the z-score

attention are evaluated using the z-score metric (z in the following). At first sight, and as we shown in section 3, the recognition of the samples could have been the result of the sole fact that it contains standard statistical markers. For example, figure 2 shows the markers related to one of the highest-scoring Ovid samples that the model predicted to be Propertian. In the training corpus, the bigram *"tu mihi"* (highlighted by the convolution which gives the highest feature score) is specific only to Propertius (z=2.52) and Tibullus (z=2.16). Moreover, the sequence *"PosPro ."* (Possessive Adjective-Pronom followed by the end of a sentence) is also specific to the same authors (for Propertius, z=17,91 and for Tibullus, z= 3,74). They have the second and third-highest convolutional feature scores for morphosyntactic labels in the sample, respectively. However, figure 2 shows that the recognition of this segment as Propertian by MCT is also based on more complex, long-distance relationships between word forms, lemmas, and grammatical tags. Primarily, the self-attention markers provided by MCT (red links in the figure) signal the existence of a quadruplet *"PosPro [...] . [...] tu mihi"*. This link is confirmed by a co-occurrence analysis[7] of the tag *PosPro* in Propertius (Figure 3): the word full-form *tu* is a collocation of the tag *PosPro* while the word full-form *mihi* and the *end of sentence* (far left) are specific to passages where the tag *PosPro* and the word full-form *tu* appear together. MCT identifies also a long distance relationship between the word form *nostra* (highly specific to Propertius; z=6,2) and the tag *PosPro* which precedes. A z-score analysis of the sequence *"nostra [...] PosPro"* shows that this sequence achieves the highest z-score in Propertius' work (z=9.35). Another long-distance relationship reported by MCT is the relationship between the tag *PersPro* (personal pronoun) at the beginning of the passage and the local saliences *"PosPro . tu mihi"*. The z-score confirms again this observation: while Propertius obtains the highest z-score in the training corpus for the sequence *"PersPro [...] PosPro"* (z=11), the sequence *"PersProDatSing [...] PosProAccPlurMasc"* is attested only three times in Propertius (references: 1,16,18 ; 1,16,41; 4,11,95). The analysis of this sample clearly shows that MCT is sensitive not only to the specificity of word forms, lem-

[7] Co-occurrences are the above-mean frequency (standard deviation) of occurrence of two terms. The visualization (Figure 3) is given by the Hyperbase Web platform: http://hyperbase.unice.fr

mas, morphosyntactic tags, or specificity of contiguous multidimensional elements (local saliences) but also to long-distance relationships. This allows us to cast a new glance at the phenomenon of intertextuality. The sample corresponds to an excerpt of a response from Cidyppe, a young woman, to a letter sent by Acontius, a young man who loves her (Ovid, *Heroïdes*, 21, 204-210). Most of the relations highlighted by MCT are here related to the two lovers, corresponding to the writer and the addressee: *tibi* (to you), *meos* (my), *tu* (you), *mihi* (for me), *nostra* (ours). The segment *"tu mihi"* is quite characteristic of colloquial texts where speaker and addressee have a close relationship: in the LASLA corpus, we found the highest z-scores in the comedies of Plautus (z=6,8). This kind of colloquial language is also characteristic of elegy: the attribution of this letter passage to Propertius by MCT is therefore fully coherent with what we know about Propertius' and Ovid's works. Another sample shows that MCT is also sensitive to sentence rhythm phenomena (Figure 2). In this case, intertextuality is undeniable and would be easily retrieved with standard methods: the sample includes what appears to be a direct quotation from Propertius by Ovid. The segment *"magne poeta"* (tagged as *"MAGNVS Subs:1Decl:Voc:Sing"* in the excerpt) is attested only once in the training corpus, in the work of Propertius (reference: 1,7,24), where it appears in the same place in the verse. Besides, the verification in a larger corpus than those of the L.A.S.L.A. (the corpus of the Packard Humanities Institute[8]) shows that this segment is not attested elsewhere in the classical Latin literary corpus. The recognition of a potential influence of Propertius on Ovid's *Heroïdes* also relies on the specificity of certain word forms, lemmas, and morphosyntactic tags, as *LEGO* (highest z-score of the training corpus = 4,2), *TVVS* (highest z-score = 16,5) or *Subs:1Decl:Voc:Sing*, a substantive of the 1st declension in the vocative singular form (*poeta* in the passage; highest z-score = 9,5). However, as illustrated in figure 2, under self-attention (red links), MCT is sensitive to the pattern: *"LEGO [...] TVVS"* (a co-occurrence confirmed by the z-score=4,8 and exclusive to Propertius with 3 occurrences in: 1,20,7; 2,24b, 21; 3,3,20) ; *"LEGO [...] TVVS ."* (z=2,6); *"Subs:1Decl:Voc:Sing TVVS"* (z=3,4); *"Subs:1Decl:Voc:Sing [...] ."* (z=5,4). The bond between these elements is confirmed by a collocation analysis: *TVVS* is a specific collocate of the tag *Subs:1Decl:Voc:Sing:* and the dot form (end of sentence) of the pair (Figure 3). As in the first passage, MCT points out features which highlight the close relationship between speaker and addressee, and are therefore features characteristics of the elegiac genre and of the Propertius style. We noticed that self-attention establishes links between three endings of sentences, isolating a first sentence consisting of 6 words and a second of 5. In this passage, these two sentences are overlapped by two verses which form an elegiac distich, the first verse counting 6 feet, the second 5. Such a structure is quite representative, even if not exclusive to Propertius' poetry (z=3,8). A phenomenon that, as far as we know, has never been detected. Therefore, MCT's feature extraction can be considered a powerful heuristic tool for intertextuality detection, covering both micro and macro-structural perspectives, not only at the level of words, lemmas, or morphosyntactic features but also at the level of word order, rhythm, and metrics. The same method

[8] https://Latin.packhum.org

can be successfully applied to the other authors of the training corpus, even with a far lower prediction, for instance Tibullus or Virgil who could also have influenced Ovid.

4 Conclusion

Using a corpus-driven approach in a multi-class/author classification context, we introduced and tested a new approach for detecting intertextuality traces using deep learning models. In particular, we have shown that feature projection of Multi-channel Convolutional Transformers is a powerful tool to detect complex linguistics markers made from local saliences (convolutional features) and their long-term dependencies (self-attention). The architecture we propose can detect not only the relevant statistical markers, but also more complex structures composed of word/entity from three word representation levels: full Forms, POS (partial or full morphosyntactic analysis) and Lemma. The complex linguistic patterns we have detected with MCT and exemplified with the Propertius/Ovid case, strongly encourage us to further develop deep learning models to explore new categories of intertextuality.

References

[1] N. Apoorv. Text classification with transformer. In *GitHub repository* https://github.com/keras-team/keras-io, 2020.
[2] R. Barthes. *Text (theory of)*. Encyclopaedia universalis, Paris, 1973.
[3] M. Buchler. *Informationstechnische Aspekte des Historical Text Re-use*. PhD thesis, Universitat Leipzig, 2013.
[4] P. J. Burns, J. A. Brofos, K. Li, P. Chaudhuri, and J. P. Dexter. Profiling of intertextuality in Latin literature using word embeddings. In *Proceedings of NAACL-HLT2021*, pages 4900–4907, Online, June 2021. ACL.
[5] R. Capozzi. Palimpsests and laughter: The dialogical pleasure of unlimited intertextuality in the name of the rose. *Italica*, 66, 12 1990.
[6] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE WACV*, pages 839–847, 2018.
[7] Y. Chen, K. Wang, W. Yang, Y. Qing, and et al. A multi-channel deep neural network for relation extraction. *IEEE Access*, 8:13195–13203, 2020.
[8] N. Coffee, J.-P. Koenig, P. Shakti, R. Ossewaarde, and et al. The tesserae project: Intertextual analysis of latin poetry. *Literary and Linguistic Computing*, 28: 221–228, 06 2013.
[9] Y. Feng and Y. Cheng. Short text sentiment analysis based on multi-channel cnn with multi-head attention mechanism. *IEEE Access*, 9:19854–19863, 2021.

[10] C. W. Forstall and W. J. Scheirer. *Quantitative intertextuality : analyzing the markers of information reuse / Christopher W. Forstall, Walter J. Scheier.* Springer, Cham, Switzerland, 2019. ISBN 9783030234133.

[11] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, 2017.

[12] G. Genette. *Palimpsestes : la littérature au second degré*. Collection Poétique. Éditions du Seuil, Paris, 1982. ISBN 2-02-006116-3.

[13] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[14] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP 2014*, pages 1746–1751, Doha, Qatar, Oct. 2014. ACL.

[15] J. Kristeva. *Sèméiotikè*. Seuil, Paris, 1969.

[16] P. Lafon. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1:127–165, 1980.

[17] G. Lapesa, S. Evert, and S. Schulte im Walde. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of SEM 2014*, pages 160–170, Dublin, Ireland, Aug. 2014. ACL.

[18] P. Li, P. Zhong, K. Mao, D. Wang, X. Yang, Y. Liu, J. Yin, and S. See. Act: an attentive convolutional transformer for efficient text classification. *Proceedings of the AAAI 2021*, 35(15):13261–13269, 05 2021.

[19] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE ICCV 2017*, pages 618–626, 2017.

[20] E. Tognini-Bonelli. Corpus linguistics at work. *Computational Linguistics*, 28: 583–583, 2002.

[21] L. Vanni, M. Ducoffe, D. Mayaffre, Precioso, and et al. Text deconvolution saliency (tds): a deep tool box for linguistic analysis. In *Proceedings of ACL 2018*, pages 548–557, Melbourne, 2018.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, and et. al. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[23] P. Verkerk, Y. Ouvrard, M. Fantoli, and D. Longrée. L.A.S.L.A. and Collatinus: a convergence in lexica. *Studi e saggi linguistici 2020*, pages 1–26, 2019.

[24] E. Wallace, M. Gardner, and S. Singh. Interpreting predictions of NLP models. In *Proceedings of the EMNLP 2020*, pages 20–23, Online, Nov. 2020. ACL.

[25] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. Cvt: Introducing convolutions to vision transformers. In *IEEE/CVF ICCV 2021*, pages 22–31, 2021.