To what extent are lemmatisation and annotation relevant for deep learning assignments and textual motifs detection? The case-study of Peter Damian's letters (11 th century)

Valérie Thon, Laurent Vanni, Dominique Longrée

To what extent are lemmatisation and annotation relevant for deep learning assignments and textual motifs detection? The case-study of Peter Damian's letters (11th century)

Valérie Thon[1], Laurent Vanni[2], Dominique Longrée[3]

[1] Université Paris Cité and Université de Liège, France and Belgium – valerie.thon@u-paris.fr

[2] Université de Nice – Sophia Antipolis, France – laurent.vanni@unice.fr

[3]Université de Liège, Belgium – dominique.longree@uliege.be

ABSTRACT

This paper wishes to explore to what extent lemmatisation and morphosyntactic annotation are important for deep learning predictions and textual motif detection. A broader research on the style of Peter Damian's letters (11th century) was the occasion to explore this question. After having trained two deep learning models on a selection of 12 classical authors using the *Hyperdeep* platform, one on lexical forms alone and the other on lemmatised and annotated texts, we introduced to them the medieval letters of Peter Damian in order to not only examine which authors are deemed to be stylistically close to Peter according to both models, but also to compare whether the results are similar and whether the same linguistic structures receive a high activation rate. The results suggest that a dialogue between both methods could be an interesting path to explore in the search for textual motifs, as the first "lexical" model may indicate rough outlines of these motifs, whereas the second model can offer concrete examples and/or variants of the first motifs identified.

KEYWORDS

Deep learning; textual motif; lemmatisation; annotation; Peter Damian.

1. INTRODUCTION AND METHODOLOGY

In this paper, we wish to explore the efficiency of trained deep learning models in the identification of textual motifs (Longrée et al., 2008), analysing how their results might differ when confronted firstly with non-lemmatised texts (having only at their disposal the lexical forms), versus lemmatised and morphosyntactically annotated works. Does the first category already offer sufficiently reliable results, or should we consider a dialogue between the two methods? This question lies more broadly within the framework of V. Thon's thesis project, which is focused on a stylistic study of the epistolary corpus of Peter Damian (1007-1072/73), a reforming hermit of the Central Middle Ages. For the purpose of this stylistic research, 20 letters concerning the "vices" Peter perceived within the Church of his time will be lemmatised and morphosyntactically annotated using the LEI software ("LASLA Encoding Initiative"), developed by the Laboratoire d'Analyse Statistique des Langues Anciennes ("LASLA") of the University of Liège. LEI proposes a semi-automatic procedure which provides for each individual form of a text in treatment one or more possible analyses (lemma and morphosyntactic annotation, where each subordinate sentence also receives a specific code according to its subordinating conjunction); these are then subject to a selection and a systematic verification by a confirmed philologist, who completes and/or corrects the possibilities provided. The results will later be statistically explored with the *Hyperbase Web* platform, developed by Étienne Brunet and Laurent Vanni (UMR Bases, Corpus, Langage, Université Nice Sophia Antipolis).

In order to examine how differently trained deep learning models may assist us in the identification of textual motifs, certainly useful in the context of this broader stylistic study, we will rely here on the tools provided by the same *Hyperbase* platform and confront the epistolary corpus of Peter Damian (1007-1072/73) to a collection of 12 classical authors whose life and works range from the 3rd century BC to the 1st century AD. Why this exact selection of 12 classical authors? Our exploration requires most of all a rich and varied corpus of Latin texts that are also available in a lemmatised and morphosyntactically annotated format; since our 20 letters of Peter Damian have been subjected to a precise lexical and morphosyntactic analysis by means of the semi-automatic LEI procedure, it would also be preferable that the labelling of our comparison corpus has been carried out with the same method. The LASLA, having created a large database of digitised, lemmatised and annotated Latin and Greek texts for the purposes of linguistic and literary studies, offers such a corpus. As can be seen from the selection below, however, the LASLA has mostly focused its attention on classical Antiquity, even though they have been expanding their field of expertise in recent years to late antique and even medieval texts (for example, Peter Damian himself). Adding new texts to this corpus would require time and, although tools for automatic lemmatisation and labelling exist, such as the Collatinus-LASLA project or *MBT (Memory-Based Tagger), TnT*

(Trigrams'n'Tags) and *TreeTagger*, these tools often necessitate an important manual verification or don't yet reach sufficiently high levels of precision (Verkerk et al., 2020; Longrée and Poudat, 2010). Taking into account not only the available LASLA-texts, but also chronology, genre and style, we arrived at the following selection: Plautus (comedies: *Amphitruo*, *Asinaria*, *Aulularia*, *Bacchides*, *Captivi*, *Casina*, *Curculio*, *Epidicus*), Cornelius Nepos (history: *Vitae*), Cicero (dialogues: *De Officiis I-II-III, Laelius vel de amicitia, Cato maior de senectute*; speeches: *Pro A. Caecina, Pro M. Fonteio, in L. Sergium Catilinam orations I-II-III-IV, Pro lege Manilia, Pro A. Cluentio Habito*), Cesar (history: *Bellum Civile*), Sallust (history: *de Catilinae coniuratione*, *Bellum Jugurthinum*), Livy (history: *Ab Urbe Condita*), Seneca (letters: *Epistulae*; philosophical treatises: *de providentia, de brevitate vitae, de clementia, de otio, de tranquillitate animi, de constantia sapientis, de vita beata*), Quintus Curtius (history: *Historiae Alexandri magni*), Petronius (fictional "novel": *Satyricon*), Tacitus (history: *Historiae*), Pliny (letters: *Epistulae*), Suetonius (history: *Vitae Caesarum*). Plautus is chronologically quite removed from the other authors, but we still wished to include him in our selection for the nature of his Latin, which would have been quite close to spoken Vulgar Latin, both in vocabulary and grammar. Peter's own letters, always intended to be read aloud, might therefore share some subtle morphosyntactic characteristics with Plautus' work. Any poetic text was excluded, mainly because their freer linguistic structure risked altering the results of deep learning. The spelling of the non-lemmatised text files was harmonised, as well as their punctuation and layout (weak punctuation was removed and all strong punctuation was homogenised in the form of full stop punctuation marks).

To carry out the double confrontation between our classical comparison corpus and Peter Damian's letters, we used *Hyperdeep*, a deep learning model integrated in *Hyperbase Web* and trained on a classification task (authorship attribution in our case). The architecture is based on multi-channel convolutional neural networks and allows both prediction of new text and extraction of features responsible for the model's decision (Vanni et al., 2023). Using the *Hyperdeep* platform, we trained two different models on our group of classical authors: the first trained model only has access to the lexical forms of their included texts, whereas the second model can perform more in-depth stylistic analyses and classification tasks, based not only on the lexical forms, but also on the lemma's and the morphosyntactic labelling provided (following LEI procedure). Once both models were trained (with a precision rate of 98.05% and 97.91% respectively), we introduced them to 12 of Peter Damian's letters in order to examine which classical authors were identified as hypothetical literary models and, more importantly, what textual motifs this identification was based on. In the case of the second model, trained on lemmatised and labelled classical texts, the 12 letters were also introduced in lemmatised format (the lemmatisation of the remaining 8 letters will soon be completed). Of course, there are also other stylometric methods that might help us explore our research question, such as *Stylometry with R* ("*Stylo*"). *Stylo* can be used for a number of different purposes, such as the linguistic and stylistic analysis of large textual corpora, the study of an author's personal writing style and authorship study in general, but the *Stylo* package is also capable of classification tasks (such as author attribution). The program also allows users to upload their own annotated text files, but it seems that the LASLA lemmatisation files are not yet completely compatible with this program. For this reason, we have chosen to focus our research solely on *Hyperdeep*'s deep learning models, which are able to exploit the LASLA files and their high precision semi-automatic lemmatisation with morphosyntactic labelling, but it would certainly be useful to explore other stylometric methods in future research.

2. A CONFRONTATION BETWEEN TWO DEEP LEARNING MODELS

2.1. MODEL 1 (LEXICAL FORMS)

First, 12 letters of Peter Damian were introduced one by one to the model trained on the non-lemmatised texts of the classical corpus, and it was giving the task of assigning these letters to one or more of the authors of this group. In order to check the stability of the results on a larger population, we also proposed to this model almost all of Peter's letters in a single file (*Patrologia Latina*; 107 letters were known at the time of this edition); we should note that we were unable to do this stability-check for the second model since not all of Peter's letters are lemmatised. According to deep learning predictions, the individual 12 letters and the large corpus of the combined letters are all stylistically close to Pliny, very often identified first, and to Seneca; the recognition rates for the other authors in our "classical" corpus almost never rise above 10%. By way of illustration, we show below the general profile generated by *Hyperdeep* for the almost-entirety of Peter's epistolary corpus in one file, compared to the classical authors (Figure 1). Pliny and Seneca are recognized at 39% and 34% respectively; the next author in the list is Cicero, but his recognition rate is only situated at 7.1%:

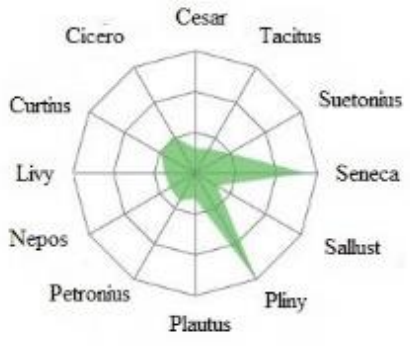

Figure 1: recognition rates of Peter's epistolary corpus

Within the classical corpus, Peter's work is clearly associated with that of the two other letter-writers: Pliny and Seneca. Their hypothetical identification seems to rest primarily on a shared vocabulary. For Pliny, for example, many connectors used at the beginning of a sentence are highlighted by the first *Hyperdeep* model, such as *unde* ("wherefore"), *nam* ("because"), *ut* ("so that"), *sicut* ("like"), *ergo* ("therefore"), *et* ("and/also"); some pronouns, especially of the first person, also tend to light up (*me*, *mi*, *mihi*), just like the vocative *domine* ("lord"), frequent with Pliny but used in a religious way with Peter, and different forms of the verb *dicere* ("to say"). Let us note that all these recognized forms are also characteristic of the epistolary genre in general. Connectors seem to have less weight in the recognition of Seneca; only *ergo* ("therefore") and *quia* ("because") occur regularly in the passages highlighted. There is, on the other hand, a whole lexicon of a rather 'philosophical' nature which lights up for him in Peter's epistolary corpus (who nevertheless uses it in a religious sense): *mundus* ("world"), *persona* ("person"), *deus* ("god"), *gradus* ("degree"), *homo* ("man"), *sapiens* ("(the) wise"). We have also found several times among the words recognized for Seneca the verbal form *ait* ("he/she says") as well as a whole series of demonstrative pronouns (*iste*, *ille*, *hic*), whose presence can be linked again to the dialogical character of the epistolary genre itself. When the latter occur, they are often preceded by a coordinating conjunction, such as *et* or, in one case, *autem*: *et iste quomodo vivet*, *et illud psalmistae*, *et illa signanter*, *et illud Ysaie*, *et illum blasphemantes*, *autem iste illa*. In some cases, the pronoun is surrounded by two verbs: *possunt ista congruere*, *facientibus ista consentiunt*, *facit errare ista perire* – as an example, we show below (Figure 2) the activation rate of the formula *possunt ista congruere* in *Letter 162*, written by Peter Damian around 1069-1072 and addressed to the archpriest and papal chancellor Peter, who is asked to support our hermit in his battle against clerical marriage and domestic partnership.

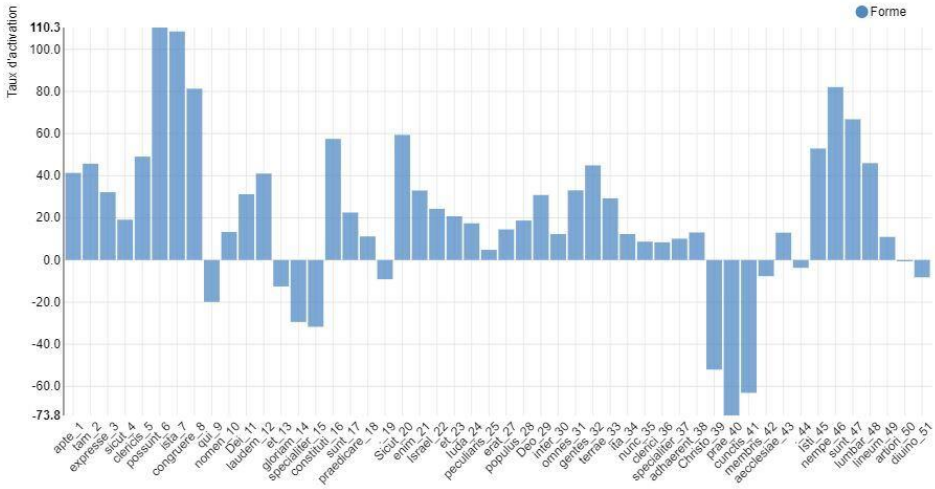

Figure 2: recognition rate for "possunt ista congruere" (Letter 162)

A search in *Hyperbase*'s "LASLA" database, containing almost all of the Latin texts morphosyntactically labelled by this same laboratory, tells us that some of these verbs are already slightly overused by Seneca in his work (*posse*, *facere*, *consentire*, *errare*, *perire*), but that the structure "verb – *ista* – verb" in particular is also characteristic of his style. The "LASLADamianus" database, created by us as a means to compare the already lemmatised letters of Peter with the ancient authors, confirms this first observation and also reveals that this same syntactic combination is not statistically specific of Peter's own style. Since previous work has highlighted the possible sensitivity of deep learning to morphosyntax (Vanni

et al., 2018a and 2018b; Thon et al., 2022), it is possible that *Hyperdeep* identified in Peter's letters a sequence "verb – demonstrative pronoun – verb" which echoes the style of Seneca. It is difficult to say whether this is a conscious borrowing on the part of our hermit. The editor of his letters, Kurt Reindel, does not cite the works of Seneca at any point among the sources used by Peter; similarly, the 12th century catalogue of the library of Fonte Avellana, the hermitage in northern Italy where Peter was not only monk but also prior, does not mention manuscripts containing texts belonging to Seneca; secondary literature also suggests that the dissemination of his work in general would have been very limited until the last decade of the 11th century and the beginning of the 12th century (Mayer, 2015, p. 277-278). The apocryphal correspondence between Seneca and St. Paul, on the other hand, was quite popular during the Middle Ages, appearing in multiple manuscripts from the 9th to the 12th century and onwards (Fürst, 2014, p. 213) – but again, no direct link between Peter's work and this collection can be found, and these 'false' letters of Seneca were not included in our comparison corpus. In any case, whether Peter Damian consciously borrowed the "verb – *ista* – verb" sequence from Seneca or not, deep learning was able to reveal its presence.

2.2. MODEL 2 (LEMMATISED AND ANNOTATED)

In order to further explore these first results based solely on lexical forms, which are already quite promising, we have recreated our primary corpus of 12 ancient authors, but this time including only their works lemmatised and labelled with a complete morphosyntactic analysis by the LASLA by means of their LEI interface. After having trained a new deep learning model on this second, labelled, corpus representing Antiquity, we reintroduced the same 12 letters of Peter Damian – also lemmatised – to this collective of authors in order to verify 1) whether *Hyperdeep*'s predictions remain stable when passing from predictions based on lexical forms alone to labelled and annotated texts, and 2) whether the enriched training corpus (lexical forms, part of speech and lemma) would allow us to refine our first results.

In general, the recognition scores of Peter's letters correspond for the most part to what *Hyperdeep* had predicted based on lexical forms alone: in almost all cases, and with considerably high percentages, Pliny and Seneca remain the first authors to be identified as hypothetical models. Let us go back to the example of *Letter 162*, used in the previous section. When examined by the first model trained, it was considered stylistically close to Pliny (with a recognition rate of 38%), followed closely by Seneca (35%) and, in third place, to Quintus Curtius at only 7.1%. When enriched with morphosyntactic information, *Hyperdeep* still associates the same letter with Seneca's style (at a remarkable 53% recognition rate) and, to a lesser extent, with Pliny (14%), followed closely by Quintus Curtius (11%). Even though the exact percentages differ, our original trio has remained the same, suggesting that, even when working on forms alone, *Hyperdeep* already has a significant precision in its predictions.

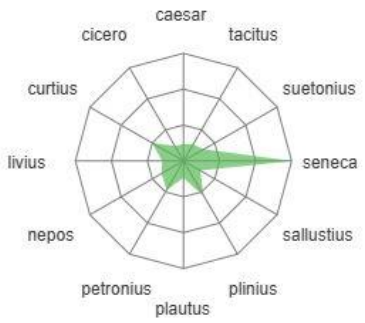

Figure 3: recognition rate (Letter 162)

Even more interesting is the fact that, whereas deep learning based on lexical forms seemed to identify the sequence "verb – *ista* – verb" as characteristic of Seneca's writing, the predictions dependent on the lemmatised LASLA-files allow us to push further this result by highlighting other more complex structures appearing to be variants of the same sequence. Among the key passages identified for Seneca in *Letter 162*, for example, we find the two following sentences: *ait dominus ad eum: surge, vade* and *ait omnis homo qui audit verba mea*. By means of illustration, we show here the activation rates for the first example:

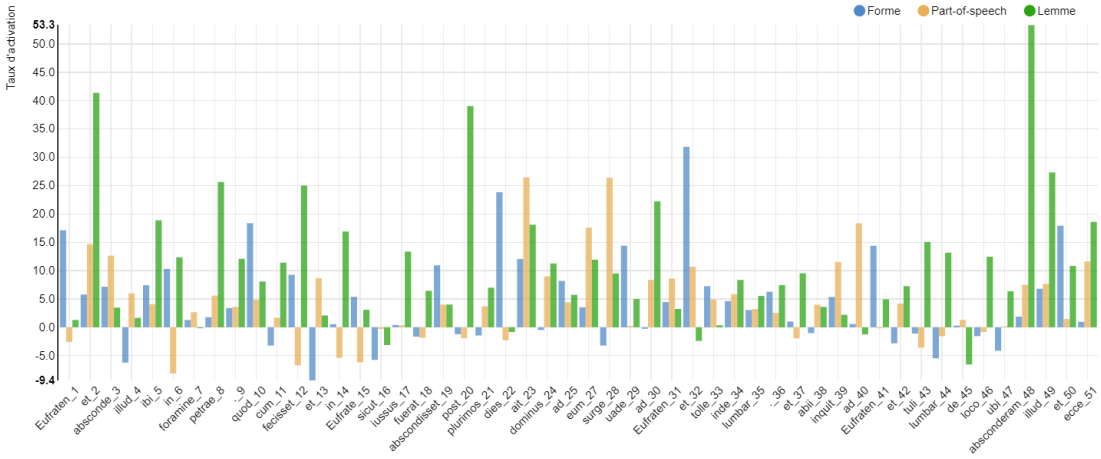

Figure 4: activation rate for the sequence "ait dominus: ad eum vade, surge"

Even though they appear to be different from each other at first glance, they are in essence quite similar to the sequence "verb – demonstrative or relative pronoun – verb" suggested by our previous deep learning model. In fact, both examples start out with a verb (in this case: *ait* ("he/she says"), already specific to Seneca as a lexical form alone), followed closely thereafter by a demonstrative or relative pronoun (*eum*/*qui*), immediately followed again by another verb (*vade*/*audit*), who both primarily light up on the grammatical plane. Other letters of Peter's test-corpus present different realisations of this same textual motif as well. *Letter 61*, for instance, written around 1059 and asking pope Nicolas II to take action against bishops who are married or living in concubinage, is also estimated to be close to Seneca (59%), to Quintus Curtius (12%) and to Pliny (8.7%) according to our second model. Among the words and morphosyntactic combinations identified for Seneca in particular, the underlined words of the following sentence were activated: *apostolo testante qui ait qui adheret meretrici unum corpus efficitur*, which gives us again the "verb – relative pronoun – verb" sequence with *ait qui adheret*, where *ait* introduces once more the rest of the motif (we might therefore consider the pattern "*ait* – pronoun – verb" to be a specific variant of the more general sequence identified). Another example offered by *Letter 61* is not introduced by *ait* specifically: *Facti siquidem culpam habet qui quod potest neglegit emendare*. The sequence, however, is still present; the two verbs *habet* ("he has") and *neglegit* ("he neglects") are activated for Seneca, not only on a grammatical level, but also because of their lemma's. They encircle the relative pronoun *quod* ("which"), which is mainly recognised for its grammatical quality as a pronoun. Using the proven methods of the more traditional TDA (Textual Data Analysis), we were also able to verify that the textual motif "verb – demonstrative or relative pronoun – verb" is still highly specific to Seneca's oeuvre even if other elements are inserted between the verbs and the pronoun. It should however be noted that three of our four examples are of a biblical nature (see most notably Jer 13, 6-10; Mt 7, 24; 1 Cor 6, 16). Peter Damian was of course an important ecclesiastical figure who was intimately familiar with Holy Scripture, and his personal writing style was most probably influenced by this as well. A possible explanation might be that the grammatical pattern "verb – pronoun – verb" often results in quite a colloquial phrasing ("he who hears", "he who adheres", "he who can do this"), such as we have seen above: *homo qui audit*; *qui adheret*; *qui quod potest*, and that such language is of course similar to the writing style of the Bible, and also leaves its traces in Peter's own Latin. It might therefore certainly be interesting to explore, in future research, the possible similarities between Seneca's epistolary corpus and the biblical (mostly New Testament) writing style.

3. CONCLUSION

These results, however provisional they may be, indicate first of all the remarkable precision and sensitivities of deep learning predictions when working with lexical forms alone; according to our first trained model on ancient texts, indeed, Pliny and Seneca were the authors deemed most similar to Peter Damian's letters. The hypothetical identification in itself makes sense, especially since all three of them have practised and are represented by the epistolary genre. The passages brought forward by the activation rates suggested that the identification relied mostly on the individual or combined specificities of lexical forms, but that *Hyperdeep* in this case may be able to sense some morphosyntactic information as well (such as verbs, for example). The second model allowed us to refine these results. Trained on the same corpus, but this time lemmatised and annotated, it not only produced a similar prediction (Seneca and Pliny), but offered far more specific information on the activation rates. More precisely, it allowed us to identify more complex variants of the previously suggested motif "verb – demonstrative or relative pronoun – verb", three of them starting with *ait* ("he/she

says"), already characteristic in itself as a lexical form for Seneca. The confrontation with the second trained model therefore shows the interest of going past forms alone to focus on the lemma's and on finer morphosyntactic annotations as well, and suggests that a dialogue between the two methods may be an interesting path to explore in the search for textual motifs: a model trained on forms alone in an effort to find rough outlines of these motifs, and a second model trained on three linguistic layers that would offer more concrete examples and/or variants of the motifs first identified.

4. BIBLIOGRAPHY

[1] Brunet, Étienne and Vanni, Laurent. «*Deep learning* et authentification des textes». Texto ! Textes et cultures 24.1 (2019): 1-34.

[2] Eder, Maciej, Rybicki, Jan and Kestemont, Mike. «Stylometry with R: A Package for Computational Text Analysis». The R Journal 8.1 (2016): 107-122.

[3] Fürst, Alfons. «*Epistulae Senecae ad Paulum et Pauli ad Senecam*». In *Brill's Companion to Seneca. Philosopher and Dramatist*, edited by Andreas Heil and Gregor Damschen, 213-214. Leiden-Boston: Brill, 2014.

[4] Longrée, Dominique and Poudat, Céline. «New Ways of Lemmatizing and Tagging Classical and post-Classical Latin: the LATLEM project of the LASLA». In *Proceedings of the 15th International Colloquium on Latin Linguistics*, edited by Peter Anreiter and Manfred Kienpointner, 683-694. Innsbruck, 2010.

[5] Longrée, Dominique, Xuan, Luong and Mellet, Sylvie. «Les motifs : un outil pour la caractérisation topologique des textes». In *Actes des JADT 2008, 9èmes Journées internationales d'Analyse statistique des Données Textuelles*, edited by Serge Heiden and Bénédicte Pincemin, 733-744. Lyon: Presses Universitaires de Lyon, 2008.

[6] Martini, Paola Supino. «L'inventario del secolo XII della biblioteca di Santa Croce di Fonte Avellana». In *Studi sulle società e le culture del Medioevo per Girolamo Arnaldi*, edited by Ludovico Gatto and Paola Supino Martini, 629-642. Firenze: All'Insegna del Giglio, 2002.

[7] Mayer, Ronald. «Seneca *Redivivus*: Seneca in the Medieval and Renaissance World». In *The Cambridge Companion to Seneca*, edited by Shadi Bartsch and Alessandro Schiesaro, 277-288. Cambridge: Cambridge University Press, 2015.

[8] Reindel, Kurt (éd). *Die Briefe des Petrus Damiani. Teil 1 – 4*. Die Briefe der deutschen Kaiserzeit (MGH). München: Monumenta Germaniae Historica, 1983-1993.

[9] Thon, Valérie, Vanni, Laurent and Longrée, Dominique. «Le deep learning auxiliaire de l'ADT dans le choix de textes à étiqueter en vue d'un corpus de comparaison : à propos de l'étude stylistique des lettres de Pierre Damien». In *Proceedings of the 16th International Conference on Statistical Analysis of Textual Data (2 vols.)*, edited by Michelangelo Misuraca, Germana Scepi and Maria Spano, 834-841. Napels-Cosenza: VADISTAT Press/Edizioni Erranti, 2022.

[10] Vanni, Laurent, Corneli, Marco, Longrée, Dominique, Mayaffre, Damon and Precioso, Frédéric. «Hyperdeep : deep learning descriptive pour l'analyse de données textuelles». Lexicometrica (2020): 1-12.

[11] Vanni, Laurent, Corneli, Marco, Mayaffre, Damon and Precioso, Frédéric. «From text saliency to linguistic objects: learning linguistic interpretable markers with a multi-channels convolutional architecture». Corpus 24 (2023). https://doi.org/10.4000/corpus.7667.

[12] Vanni, Laurent, Ducoffre, MMélanie, Mayaffre, Damon, Precioso, Frédéric, Longrée, Dominique, Elango, Veeresh, Buitrago, Nazly Santos, Gonzalez, Juan, Galdo, Luis and Aguilar, Carlos. «Text Deconvolution Saliency (TDS): a deep tool box for linguistic analysis». In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 548-557. Melbourne, 2018.

[13] Vanni, Laurent, Mayaffre, Damon and Longrée, Dominique. «ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables». In *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*, edited by Domenica Fioredistella Iezzi, Livia Celardo and Michelangelo Misuraca, 459-466. Rome: UniversItalia, 2018.

[14] Verkerk, Philippe, Ouvrard, Yves, Fantoli, Margherita and Longrée, Dominique. «L.A.S.L.A. and Collatinus: a convergence in lexica». In *Studi e saggi linguistici*, edited by Laura Tesconi, 1-26. Pisa: Edizioni ETS, 2020.