Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5295–5303
Marseille, 20-25 June 2022

Tracking Textual Similarities in Neo-Latin Drama Networks

Andrea Peverelli,[1,2] Marieke van Erp,[2] and Jan Bloemendal[1]
[1]Huygens ING
[2]KNAW Humanities Cluster, DHLab
Oudezijds Achterburgwal 185
1012 DK Amsterdam, Netherlands
andrea.peverelli@huygens.knaw.nl
marieke.van.erp@dh.huc.knaw.nl
jan.bloemendal@huygens.knaw.nl

Abstract

This paper describes the first experiments towards tracking the complex and international network of text reuse within the Early Modern (XV-XVII centuries) community of Neo-Latin humanists. Our research, conducted within the framework of the TransLatin project, aims at gaining more evidence on the topic of textual similarities and semi-conscious reuse of literary models. It consists of two experiments conveyed through two main research fields (Information Retrieval and Stylometry), as a means to a better understanding of the complex and subtle literary mechanisms underlying the drama production of Modern Age authors and their transnational network of relations. The experiments led to the construction of networks of works and authors that fashion different patterns of similarity and models of evolution and interaction between texts.

Keywords: text reuse, textual similarity, transnationality, Neo-Latin, drama

1. Introduction

One of the defining characteristics of the early Modern Era, burgeoning from the Italian Renaissance period, is the wide international network of exchanges between writers of different nationalities that bears the Latin name of *respublica literaria* (Republic of Letters)[1] Authors feel part of a wider, universal, intellectual community, and their authorial signal can and must be read especially in light of its complex network of interdependent exchanges with other peers.

Tracking instances of textual reuse and similarities between works thus comes as the prime reflection of the complexities of the *respublica literaria*. Relations between authors are primarily expressed through writing and an intense, foregoing discussion upon the reuse of models (that is, the hot topic of *imitatio*).

Besides the vast literary production in the various natural languages of Europe, the early Modern Age is characterised by a wide, if barely known, production in Latin from the first sparks of Humanism in the Italian Peninsula during the XIV century. The production of poetry and prose in Latin increased dramatically, and drama began to be involved in this process in great quantities (at least 10,000 works are known from this period; see Bloemendal, Parente, Smith, 2022, for reference). The Low Countries were at the forefront of this revitalisation, thanks especially to the outstanding work of Erasmus.

Early Modern Latin (or "Neo-Latin") was very different from the one written and spoken in the former medieval centuries. According to (Bloemendal and Norland, 2013), Neo-Latin was characterised by: "a shift [from the Middle Ages] in the use from a pragmatic one (if necessary, new words could be coined, even 'unclassical' ones, and syntactic means could be used as seemed fit), to a principled one, which should aim at writing 'classical' Latin morphologically and syntactically". This, paired with the general methodology typical of Humanism of "going ad fontes" (i.e. to strictly adhere to the original classical texts), makes for a close resemblance of early Modern Latin to the classical standards. It thus comes naturally that comparison with classical authors is, in the topic of textual similarity, particularly meaningful as a means of clustering authors within common ancestries.

Whether conscious text reuse or coincidental resemblance, textual similarity can be viewed in a twofold manner, based on its presence or absence: when present, it is a measure of the closeness between two texts, so that one of them can be read as a means of relations to the other one; when instead absent, it represents their degree of distance (or "dissimilarity"), and it is as important as its counterpart. Moreover, dissimilarity can be a criterion for further inquiries: as a standard measure between two texts, to state their closeness under different literary aspects (style, content, space and time, etc.); or as a marker for a more subtle closeness to be found in a common ancestry back in time or in another unrelated place, in the form of a predecessor, or "pre-text" shared by both texts. The concept of pre-

[1]The concept appears for the first time in an epistolary exchange between the humanists Francesco Barbaro and Poggio Bracciolini at the start of the XV century. See (van Miert, 2018) for a recent reading on the topic

text within evolutionary networks is well explained in *The structure and evolution of story networks*(Karsdorp and Van den Bosch, 2016), by which our research was inspired. According to (Karsdorp and Van den Bosch, 2016): "Story networks consist of stories and links between stories that represent pre-textual relationships. We make the simplifying assumption that stories that are more similar to each other are more likely to stand in a pre-textual relationship than stories that are more distant". While the focus of their paper is on "storied retellings" (well-defined story frames towards which heavy text reuse is ascertained as a starting point), our own verges on a more explorative approach: trying to discover the very existence of a complex network of textual reuses and its internal strategies.
Our research question is formulated within the framework of the TransLatin Project Project, which tries to inquire this very notion and blends perfectly with the aim of our paper: what is the extent of the process of imitation and reception within Neo-Latin drama? Are any authors connected at all? Which ones serve as the strongest pre-texts (in literary terms: "models") for the others? To answer these questions we made our first steps towards a thorough investigation of similarities networks, while being aware of the wide arrange of tools for text reuse detection, through two different methodologies: Cosine Similarity and Bootstrap Consensus Trees.
The contributions of this paper are as follows:

- CURRENS: a new tool for the pre-processing of Latin texts:
- insights into reuse of Neo-Latin Drama;
- new applications of known methodologies, drawn from Information Retrieval and Stylometry, towards the topic of textual similarities.

The remainder of this paper is structured as follows. In Section 3 we explain the criteria that we followed for preparing our corpus. In section 4 we get into details about the experimental setup. In section 5 we discuss the results. Finally, in section 6 we consider future steps and draw conclusions about our whole experiment.

2. Related Work

In the last decade, several tools have been made available for tracking proper text reuse through text alignment or feature extraction for historical languages. The most well-known of these tools (TRACER,[2] Tesserae,[3] and Passim[4]) have also been tested for Latin: one of the last experiments is that of Franzini, Passarotti, Moritz and Büchler (2018), in which a thorough exploration of HTRD (*Historical Text Reuse Detection*) tools can be found. These tools can be quite powerful in detecting precise reuse, both intentional and unintentional, in the forms of quoting and allusion. While more traditional HTRD methods will be employed in the future, we wanted to explore the possibilities of the application of older approaches (Cosine Similarity and Stylometry) to a novel case study, shifting towards a more general textual similarity framework that will serve as a solid base for future inquiries. As for Cosine Similarity, (Manjavacas et al., 2019) approached allusive textual reuse detection on a Latin Biblical corpus from an Information Retrieval standpoint: through an extensive usage of Cosine Similarity scores and Word Embeddings (Manjavacas et al., 2019), they found that custom query algorithms for automatic allusion detection were consistently outperformed by simpler TF-IDF models and that Cosine Similarity can prove a sound basis for inquiring textual reuse. Other studies, such as (Bär et al., 2012) and (Sturgeon, 2018), employed Cosine Similarity and TF-IDF scores, in text reuse and similarity detection with good results, both for contemporary language corpora (the former, which was tested on the METER corpus and the Webis Crowd Paraphrase corpus) and historical language corpora (the latter, which worked on an Early Chinese corpus). As for stylometric approaches, the use of Stylometry for textual similarity and reuse detection is ample. Some experiments have also been conducted upon historical languages, especially Latin (cf. (Eder, 2016)) and Ancient Greek (Gorman and Gorman, 2016).

3. Corpus Preparation

Our corpus was assembled considering three parallel tracks, designed to cover the main aspects of a literary corpus:

- Topical aspect: works pertaining the same subject;
- Authorial aspect: works from the same author;
- Diatopical and diachronical aspects: works from different times and places.

Our aim for this initial set of experiments is to set a stable pipeline and a golden standard to expand upon in the future.
Our corpus is thus built containing 47 works in total, sub-divided as follows.
15 works from early Modern Neo-Latin drama, of which 8 pertain to the topic of "Joseph play" (to satisfy the first aspect), 3 same-author clusters (to satisfy the second aspect), and a range of 4 different nationalities and places of publication (to satisfy the third aspect): 3 authors from Germany, 1 from Poland, 1 from England and 7 from the Netherlands, thus keeping our particular focus on Dutch writing. The diachronic aspect is

[2] `https://www.etrap.eu/research/tracer/` Last visited: 16/1/2021
[3] `tesserae.caset.buffalo.edu` Last visited: 16/1/2021
[4] `https://github.com/dasmiq/Passim` Last visited: 16/1/2021

satisfied by the range of these works: the works were published between 1510 and 1639.
To track the first models of our Modern-era drama corpus and to serve as an additional counter-check proof for the clustering in Subsection 4.2, we added the 6 works from the Latin playwright Terentius, the 20 from Plautus (the known 21st is heavily fragmented and could not serve our purpose) and 6 certain dramas from Seneca, whose corpus authenticity is still highly debated[5]. These texts are gathered from the LASLA corpus[6].
As our Neo-Latin drama texts come through a process of OCR from centuries old prints, they come with errors and imprecisions that can severely impact the processing of a text (van Strien et al., 2020). Furthermore, we needed our texts to be devoid of any unnecessary information (e.g. verse number and character abbreviations), just presenting the bare tokenised script. We thus cleaned the texts in our corpus following a common pipeline of text manipulation for Latin texts:

- Cleaning OCR errors;
- Replacing punctuation;
- Changing everything to lower case;
- Normalizing Latin-related issues with spelling (such as V into U and J into I);
- Replacing para-textual annotation (e.g. characters speaking, line number, verse type).

A final layer of cleaning involved the process of manipulating the actual content of the texts:

- Stop words filtering, based on the Perseus Project list[7] and then heavily modified and expanded;
- Non-semantic words filtering (conjunctions, subjunctions, pronouns, auxiliaries, some very common adverbs);
- Lemmatisation. These two final steps were oNeo-Latiny implemented in the Cosine Similarity part of our analysis (Subsection 4.2).

This whole process was done automatically using our custom-built program *CURRENS* that builds upon the tokeniser and enclitics exception list from the CLTK pipeline[8], and the LemLat lemmatiser amended with in-house developed modules and expanded stopwords from the Perseus project[9]. CURRENS is available on Github.[10] The results of the pre-processing can be seen in 1.

4. Experimental Setup and Analysis

In this section, we present our experiment setup and analysis. The experimental setup sketches our general approach to analysing Neo-Latin texts. We then explain how we use Cosine Similarity 4.2 and Stylometry 4.3, by constructing a Bootstrap Consensus Tree, to compare different texts and what these different analysis methods bring.

4.1. Experimental Setup

Our first analysis is inspired by (Karsdorp and Van den Bosch, 2016), where we calculate the cosine similarity for every text in our corpus and produce a sparse correlation matrix, in order to express the closeness between texts and authors in terms of vector representation in an n-dimensional space model. This served as a basis to build a network representation that revealed a pattern of evolution shaped by the "PA-TA"(preference-based and temporal-based) attractiveness, basically a heavy-tailed, mostly chronological distribution of similarities that resemble real life evolutionary growth networks (and thus confirming the findings of (Karsdorp and Van den Bosch, 2016)).
Our second analysis involves a stylometric approach (Eder, 2017). We computed a Delta-distance Bootstrap Consensus Tree and produced the Principal Components Analysis (PCA) for our corpus of 47 works. Combining the results from these analyses, we drew another network that revealed a new and unexpected clustering, unveiling similarities unknown before. The computer was also able to correctly identify classical models and age- or generation-defined clusters, as found in previous literary inquiries, thus confirming the general structure and the evolution of Early-Modern Neo-Latin drama.

4.2. Cosine Similarity

As a first step in our twofold experiment, we calculated the cosine similarity scores for each pair of texts in our corpus, which needed a final layer of manipulation: we thus lemmatised the texts, since calculating the cosine similarity between tokenised corpora, for a highly inflected language such as Latin, would bear too many false negatives (for example, the tokens "Deus" and "Deorum" would be held separated and would not contribute towards the final cosine similarity score, when they are clearly the concept pointing to the same word - "lemma" level - realised in two different ways - "token" level -); on the other hand, to prevent the inflation of the final score due to false positives, we eliminated most of the lemmas that do not possess a high semantically-informative content and that usually occur in the form

[5]We followed the selection in (**?**; **?**). For an overview on Seneca's corpus authenticity, see (**?**)
[6]http://web.philo.ulg.ac.be/lasla/
[7]http://www.perseus.tufts.edu/hopper/stopwords
[8]http://cltk.org/
[9]http://www.perseus.tufts.edu/hopper/
[10]https://github.com/AndrewPeverells/CURRENS

of textual invariants (stopwords and function words, together with some very frequent Latin words, such as a few adverbs - *ut, iam, saepe...* - nouns - *res* - and verbs - mostly auxiliaries and derivates: *habeo, sum, fio, possum...* -).[11] As interjections are an important part of theatrical writing, these were kept.

Firstly, we transformed the lemmatised corpus into a Word Vector Space model (an n-dimensional space for representing documents and/or words as vectors, needed in order to compute the TF-IDF - *term frequency/inverse document frequency* - scores as the logarithmically scaled product of vectors); secondly, we turned it into a matrix of TF-IDF features; finally we computed the actual cosine similarity scores for our texts. We then generated the co-occurrence tables for every text, for a better in-depth explanation of the word likeness between works, divided into spreadsheets with a precise ratio: one for the general co-occurrences between the two sub-corpora (Neo-Latin Modern plays and Classical plays), and the other group for the highest cosine similarity scoring texts. Every spreadsheet is accompanied by a second sub-sheet bearing some general statistics for the particular pair analysed: type-token ratio (TTR), medium word length and lemma dispersion.

4.3. Stylometry

For the second part of the experiment, we opted for a stylometric analysis through the R package *stylo* (Eder et al., 2016). We went back a step in the corpus preparation procedure to maintain the stopwords/function words in the texts, as they are the vital part of every stylometric analysis, and we kept our corpus tokenised. We then produced a Bootstrap Consensus Tree, spanning through different parameter tests.

- As for the computed Distance, we decided to choose *Eder's Delta* (Evert et al., 2017), which is particularly suited for highly-inflected languages and not too long word vectors (the texts in our corpus very rarely exceed 15,000 tokens);
- As for the most frequent word (MFWs) analysed, we run through several trials, and found that the clustering begun to fall off at around 500 MFWs, gradually reuniting every work together in a single branch. We thus chose a comfortable plateau of 200 MFWs, that could properly show a meaningful branching of the clusters;
- As for other important parameters, we kept the Consensus of our tree at 0.5, left the pronouns out, and employed no culling of the MFWs.

5. Results and Discussion

(Karsdorp and Van den Bosch, 2016) propose that the evolution of textual networks is to be based on two dimensions: *temporal attractiveness* (TA), the principle for which authors tend to prefer more recent models, and *model-based attractiveness* (MA), that involves elements from the context (such topic or the importance of an author). Our networks follow these two principles.

From our cosine similarity experiment (see fig.3), we can see that texts tend to follow a TA evolutionary fashion, exhibiting works that are closer in time as their highest-scoring models. Moreover, another key element incurs in the earlier stages of the network: a first cluster is clearly visible, composed of authors (Macropedius-Crocus-Gnapheus) of Dutch origin and active in the Netherlands. This shows the relative importance of the spatial aspect, which is closely related to the temporal one, thus transforming the TA into a T-SA (*temporal-spatial attractiveness*). However, this model of T-SA is especially true for the initial elements of the corpus, while the probability of works straying off their closest ones as models gets increasingly higher with the evolution of the network. This is due to the growing importance of context (MA): as time passes, authors are given more choice for their inspiration.

Another crucial aspect is that of hubs, or, in our case, key turning points. We drew a graph from our cosine data (see fig.1), introducing a minimum threshold of 0.3 to filter out the weakest scores. The resulting (*out-degree*) network, displaying the outgoing edges, clearly shows that some texts serve as central hubs of reuse, or "models": works from a first, earlier period (1510-1556) appear to be strong models for later authors, and a clear-cut clustering also stands out, with one very tight group (Crocus-Macropedius-Diether) and another cluster (Macropedius-Gnapheus-Foxe) that looks loosely connected to the first one. Moreover, each cluster has its key central hub that serves as a cornerstone: in the first one, Diether is well connected to both works from an earlier period and to later texts, while in the other one Foxe serves this purpose. In general, Crocus, Macropedius, Diether and Foxe were the highest scoring authors, both in cosine similarity and out-degree values, so we can (relatively safely) assume their importance and renown in the greater *respublica literaria*.

From the data gathered in our second part of the experiment, we can draw some new and complementary considerations. We drew a Bootstrap Consensus Tree with our full corpus (also comprising Plautus, Terence and Seneca) through a built-in algorithm from the R package *stylo* (fig.2). Three main aspects stand out. Firstly, the authorial signature is very strong: all three groups of same-author works were correctly clustered together. This came in spite of the first aspect that we wanted to inquire: topic seems to be completely irrelevant to this kind of analysis, as Joseph plays are mixed with non-Joseph plays without any discerning ratio. Moreover, it is particularly interesting since this

[11] The importance of lemmatisation in cosine similarity score measuring for textual similarity is also confirmed by (Manjavacas et al., 2019), as "lemmatization boosts the performance of nearly all models"

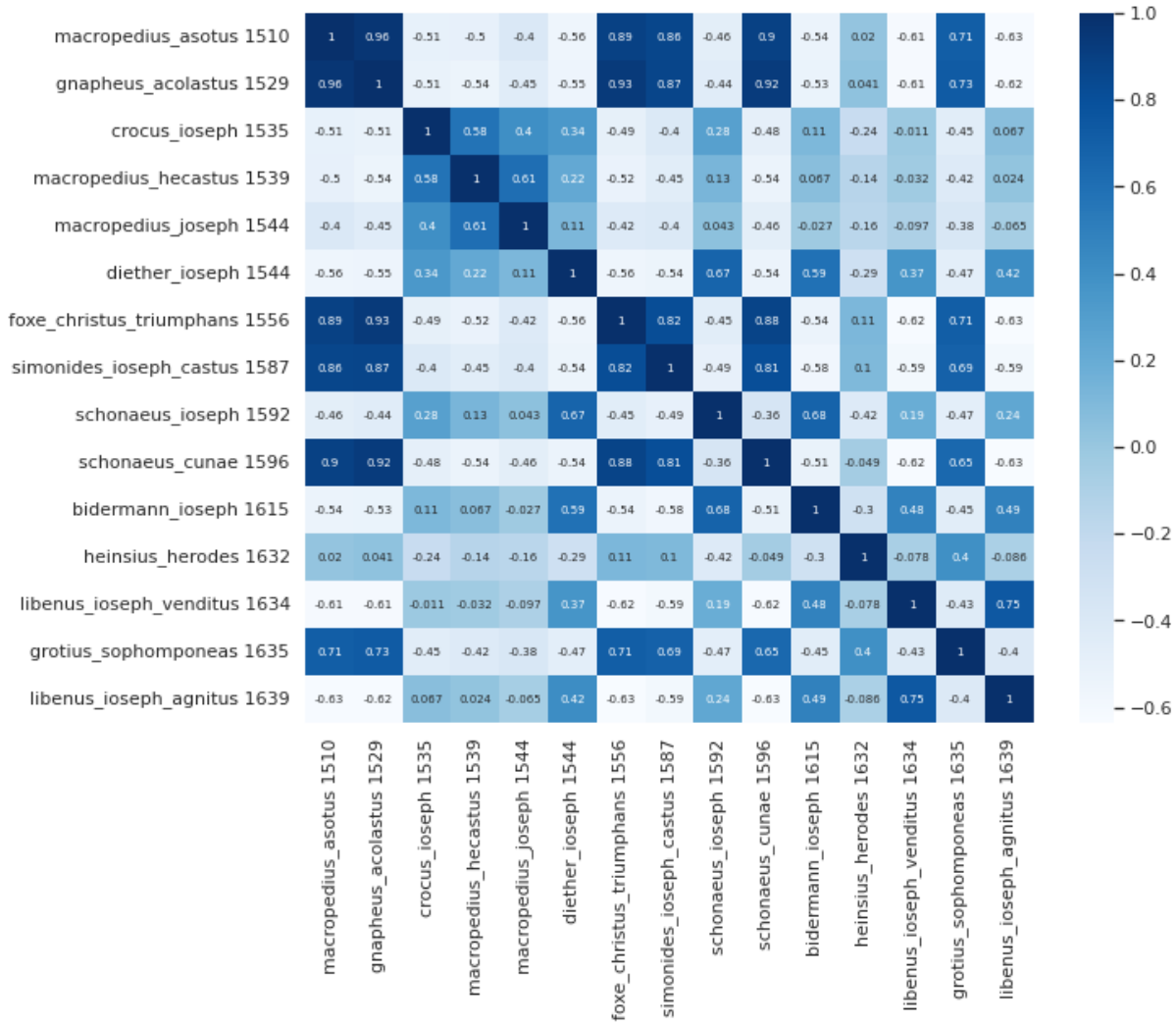

Figure 1: Cosine Similarity scores heat map.

preeminence of the authorial aspect over topic was not really clear in the first part of the experiment: from cosine similarity scores, sometimes same-topic texts over-scored same-author clusters (as in the case of the Joseph by Schonaeus, which scored really low when compared to the Cunae, another of his works), while sometimes works from the same author had a higher score than same-topic works from other authors (as in the case of the Joseph and the Hecastus, both from Macropedius). Secondly, the T-SA dimension is maintained, but with new and interesting additions: the algorithm automatically drew two very distinct clusters, separating the XVI century works from the XVII century ones (although Bidermann seems to be an exception). This goes in pair with a third consideration, regarding classical authors: Plautus was put aside from everything else, while Terence seem to have a higher influence on the XVI century cluster and Seneca on the XVII century one. This clear-cut subdivision is confirmed by literary studies on the matter. (Bloemendal and Norland, 2013) identifies a three-staged evolution of Dutch Neo-Latin drama: a first one (roughly 1500-1550) that serves as a proving ground for new authors that revealed to be very influential in later periods; a second one (1550-1600), characterised by the use of Terence as the primary model; and, finally, a third one, more akin to Baroque literature, that shifted heavily towards a more Senecan style.

There are still two notable exceptions to our Consensus Tree: Bidermann (1615) seems to fit better in the XVI century cluster, and the Adelphoe resulted as the oNeo-Latiny separated terentian work in all of our tests, more akin to authors in the XVII century cluster. The first anomaly is maybe due to Bidermann's Jesuit background: within the XVII century cluster, oNeo-Latiny one (Libenus) out of three authors pertain to the Jesuit Order, which is much more concentrated in the XVI century cluster. The second anomaly, the Adelphoe by Terence, still needs more investigation.

6. Conclusion and future work

This paper describes an exploration towards building a functioning pipeline for assessing textual similarities in

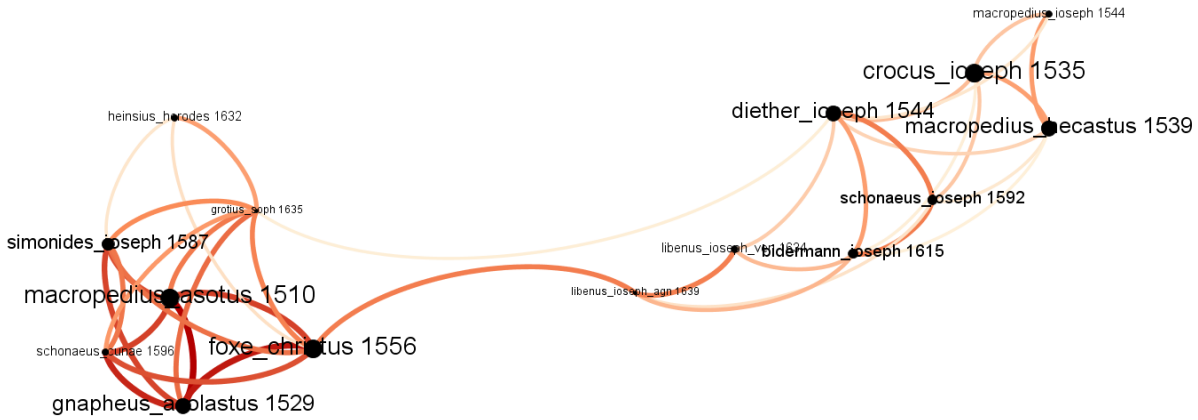

Figure 2: Out-Degree Network of the Neo-Latin works.

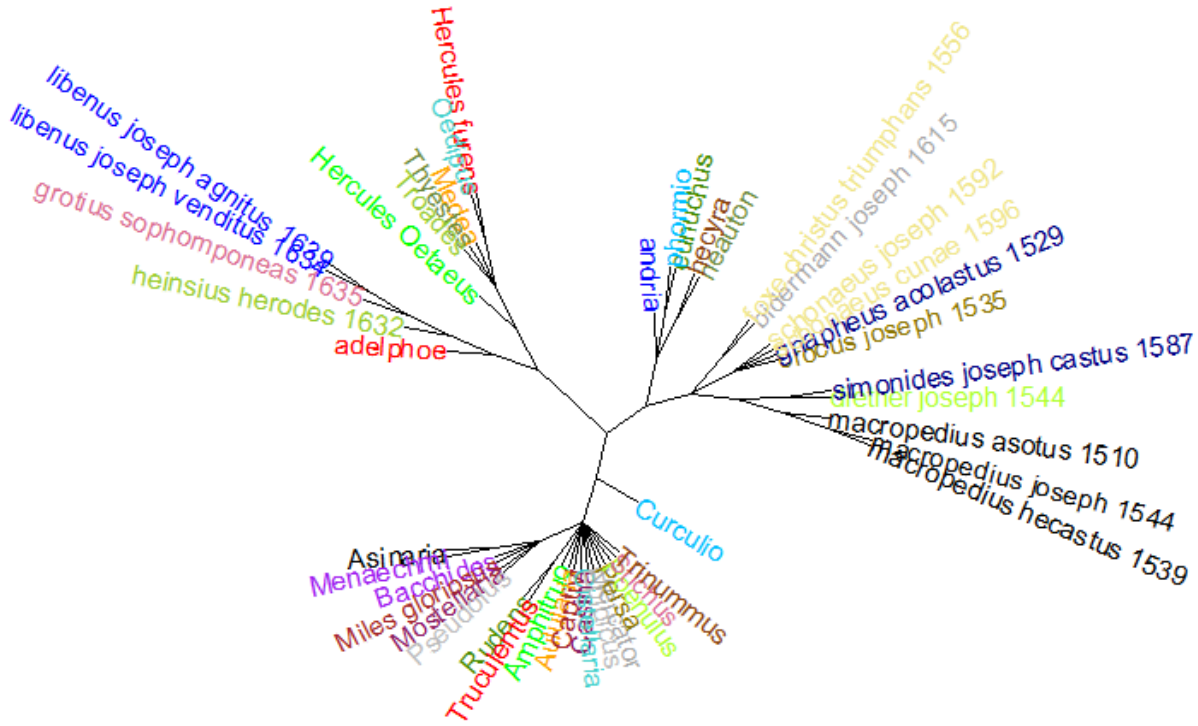

Figure 3: Bootstrap Consensus Tree of 49 Latin and Neo-Latin Dramas using 2-202 Most Frequent Words, Eder's Delta distance, Consensus 0.5.

Neo-Latin texts. Through this experiment we wanted to test textual similarity and reuse through the three main aspects of literary works: the authorial aspect (works from the same authors, thus tracking internal reuse); the topical aspect (works pertaining the same topic, thus tracking similarities throughout a same-topic subcorpus); and the diatopic-diachronic aspect (thus tracking the reuse of other authors' texts through time and space and the individuation of "models"). We hence built our test corpus in such a way that it covered every one of the three aspects we wanted to inquire, also inserting the works of classical drama authors (Plautus, Terence and Seneca) as a counter-check for the second part of our experiment.

Although our focus is on drama pieces and on Neo-Latin, this pipeline can be applied to any kind of Latin text, as its parameters are the same. This is thanks, especially, to our tool, CURRENS, which can be used to pre-process a Latin work in a customised fashion, depending on which module is needed in one's analysis. We employed it in its entirety to generate clean, tokenised texts to work upon, and tweaked its modules

in order to get a two-fold type of data from our initial corpus: one raw, tokenised, as presented in the original texts (the full 49, comprising classical authors); and one lemmatised and deprived of stop words and function words (just the 15 Neo-Latin texts of the XVI-XVII centuries period that we gathered as an initial exploration). The latter was used in the first part of our experiment, involving cosine similarity, for which we employed a TF-IDF vector space model to calculate its score for every text in our corpus. From these results, we built a network showing the out-degree values for the processed texts, and a heat map showing the correlation distribution between the 15 samples. From this, two noteworthy results stand out:

- The evolution of similarity patterns in our corpus tends to follow a S-TA/MA model (spatial-temporal attractiveness / modal attractiveness): in the early (chronological) stages of the network, authors tend to prefer texts closer in time and space as their models of reuse (S-TA); conversely, as time passes by, the dispersion of this preferential attachment increases dramatically, with authors preferring other texts based on more aleatory reasons such as topic, style or vicinity (MA).

- The emergence of text clusters is modeled around central hubs (our "models"), represented by particularly fortunate authors: our corpus, in particular, split into two clear-cut clusters, with Andreas Diether in one and John Foxe in the other serving as central hubs of reuse, well connected with both authors from the early age and authors from later stages, around which the other texts seem to gravitate. This shows the relative importance of these authors and the end of the early period of our network (1544-1556) as a testing ground for later literary imitation.

The other type of processed data (raw, tokenised text) that resulted through the use of CURRENS was instead used for the second part of our experiment. We generated a BCT (Bootstrap Consensus Tree) of the whole 49 works that make up our corpus, combining together the Neo-Latin works and the texts from classical authors as a counter-check for the clustering method that we employed (Eder's Delta, 0.5 consensus strength, 200 MFWs). From this, we could draw further considerations:

- The authorial signal is stronger than the topical aspect. Internal style within same-author clusters takes over features of same-topic style. The cosine similarity experiment gave mixed results in this regard.

- S-TA is maintained, but with new clusters that define an age-dependent evolution of style: the algorithm automatically recognised two very distinct groups, one in the XVI century and one in the XVII century, with classical authors arranged as clear models (Terence for the first group and Seneca for the second; Plautus was set apart as too distant). This is confirmed by literary critique studies, that report a similar generation-like evolution of Neo-Latin drama and model selection.

- Two main exceptions stand out: the alien presence of the Adelphoe by Terence in the XVII century group and that of Bidermann (1615) in the XVI century cluster. These need more evidence.

We thus answered to the original questions: we demonstrated that the process of imitation and reception within Neo-Latin drama is extensive, and it happened on many layers (spatiality, temporality, modality); we tracked connections between authors and checked the reuse of models, both contemporary and ancient: finally, we demonstrated the existence of hubs of reuse, thus gaining more insight on the importance of some authors in the Early Modern Era and the reflection of classical drama writers on this very age.

As a further step to improve our model of textual similarity for Neo-Latin texts, we plan to improve on the basis we have set, as well as employ new methodologies for our next experiment. First of all, an expansion of our corpus, with new Neo-Latin texts from the XVI-XVII centuries, will be a constant background operation, as the TransLatin Project moves forward and enables more texts to be digitised and analysed. Secondly, a word embeddings analysis for our corpus will be conducted, to improve upon the foundations of the cosine similarity experiment that we already conducted. Finally, for a more different approach, we would like to implement a topic modelling analysis to better inquire the topical aspect of our pipeline and have a deeper understanding of how textual reuse works in conjunction with topic variation.

The code and the data for this paper is available on GitHub at `https://github.com/AndrewPeverells/Translatin`

Acknowledgements

This research is conducted within the framework of the TransLatin project, funded by the Dutch Research Council (NWO). The authors would also like to thank Dirk van Miert, which provided invaluable assistance during the research.

7. Bibliographical References

Bär, D., Zesch, T., and Gurevych, I. (2012). Text reuse detection using a composition of text similarity measures. In *Proceedings of COLING 2012*, pages 167–184.

Bloemendal, J. and Norland, H. (2013). *Neo-Latin Drama in Early Modern Europe*. Brill.

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with r: a package for computational text analysis. *The R Journal*, 8(1).

Eder, M. (2016). A bird's-eye view of early modern latin: Distant reading, network analysis, and style variation. *Early Modern Studies After the Digital Turn*, page 63.

Eder, M. (2017). Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, 32(1):50–64.

Evert, S., Proisl, Jannidis, F., Reger, I., Pielström, S., Schöch, C., and Vitt, T. (2017). Understanding and explaining delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32.

Gorman, V. B. and Gorman, R. J. (2016). Approaching questions of text reuse in ancient greek using computational syntactic stylometry. *Open Linguistics*, 2(1).

Karsdorp, F. and Van den Bosch, A. (2016). The structure and evolution of story networks. *Royal Society open science*, 3(6):160071.

Manjavacas, E., Long, B., and Kestemont, M. (2019). On the feasibility of automated detection of allusive text reuse. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 104–114.

Sturgeon, D. (2018). Digital approaches to text reuse in the early chinese corpus. *Journal of Chinese Literature and Culture*, 5(2):186–213.

van Miert, D. (2018). Towards a conceptual history of the republic of letters in the modern period. Cultural History Seminar.

van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the impact of ocr quality on downstream nlp tasks. In *Proceedings of ICAART 2020*.

A. Appendices

Authors/Titles	**Year/Age**	**Tokens in text**	**Types (distinct words)**	**Lemmas (after cleaning)**	**Type/token ratio (TTR)**	**Mean word length (chars)**
Macropedius *Asotus*	1510	11,450	7,936	2,257	41.26	5.39
Gnapheus *Acolastus*	1529	8,823	5,964	3,646	41.32	5.17
Crocus *Joseph*	1535	6,964	4,702	2,756	39.57	5.09
Macropedius *Hecastus*	1539	12,577	8,958	2,094	36.38	5.33
Macropedius *Joseph*	1544	12,013	7,116	4,479	37.28	5.45
Diether *Joseph*	1544	16,475	9,609	6,430	39.03	5.49
Foxe *Christus Triumphans*	1556	10,082	6,840	4,251	42.16	5.33
Simonides *Joseph Castus*	1587	9,628	6,915	4,418	45.89	5.35
Schonaeus *Joseph*	1592	12,420	8,282	3,745	30.15	5.13
Schonaeus *Cunae*	1596	5,504	3,423	2,237	40.64	5.35
Bidermann *Joseph*	1615	13,224	11,170	6,001	33.10	5.20
Heinsius *Herodes*	1632	9,280	7,430	1,647	47.19	5.55
Libenus *Joseph Venditus*	1634	5,484	4,582	1,491	50.64	5.41
Grotius *Sophomponeas*	1635	6,907	5,261	1,683	53.48	5.39
Libenus *Joseph Agnitus*	1639	6,284	4,807	1,287	50.37	5.44
Terence	II century B.C.					
Adelphoe		8,711	4,310	2,745	31.52	4.72
Andria		8,413	4,362	2,709	32.20	4.80
Eunuchus		9,010	5,608	2,932	32.54	4.82
Heauton		8,832	4,529	2,812	31.84	4.77
Hecyra		7,301	4,264	2,390	32.74	4.81
Phormio		8,971	4,394	2,900	32.33	4.73
Plautus	III century B.C.					
Amphitruo		8,425		2,749	32.63	4.90
Asinaria		14,747	9,223	4,508	30.57	4.91
Aulularia		3,929	2,315	1,627	41.41	4.96
Bacchides		10,030	8,862	3,307	32.97	4.96
Captivi		8,350	4,044	2,883	34.53	4.91
Casina		7,271	4,173	2,520	34.66	4.74
Cistellaria		5,397	3,124	2,057	38.11	4.85
Curculio		2,300	4,818	1,124	48.87	4.87
Epidicus		6,546	5,005	2,327	35.55	4.87
Menaechmi		9,133	6,159	2,879	31.52	4.85

Mercator		8,766	6,002	2,915	33.25	4.86
Miles Gloriosus		12,811	7,226	3,886	30.33	4.98
Mostellaria		9,777	5,600	3,081	31.51	4.84
Poenulus		10,858	8,579	3,486	32.11	4.93
Pseudolus		11,369	7,281	3,579	31.48	4.85
Rudens		11,450	8,642	3,543	30.94	4.93
Stichus		6,394	3,998	2,477	38.74	5.02
Trinummus		9,834	6,554	3,262	33.17	4.96
Truculentus		8,226	5,028	2,942	35.76	4.88
Persa		7,954	4,556	2,699	33.93	4.73
Seneca	I century C.E.					
Hercules Furens		3,592	2,879	2,335	65.01	5.57
Hercules Oetaeus		10,292	7,818	4,290	41.68	5.42
Medea		6,349	4,957	3,034	47.79	5.53
Oedipus		5,792	4,709	3,439	59.38	5.56
Thyestes		6,220	4,360	3,410	54.82	5.46
Troades		6,698	5,235	3,520	52.55	5.50
Overall		419,861	102,995	53,812	12.82	5.10

Table 1: Dataset statistics. The titles of the plays are in italics.