DHQ: Digital Humanities Quarterly

Preview
2024
Volume 18 Number 3

Towards a differentiated digital-hermeneutic analysis tool for the detection of short quotations using the example of the Church Father Jerome

Franziska Schropp <franziska_dot_schropp_at_uni-konstanz_dot_de>, University of Konstanz

Thomas E. Konrad <thomas_dot_eugen_dot_konrad_at_uni-konstanz_dot_de>, University of Konstanz
https://orcid.org/0000-0002-0568-9420

Marie Revellio <marie_dot_revellio_at_uni-konstanz_dot_de>, University of Konstanz

Barbara Feichtinger <barbara_dot_feichtinger_at_uni-konstanz_dot_de>, University of Konstanz

Abstract

Late Latin literature is characterized by numerous references to classical texts and authors. For Jerome of Stridon in particular, manual-hermeneutic research has revealed various intertextuality phenomena usually published in encyclopaedic collections of quotations. In this paper, we present a digital-hermeneutic analysis toolkit primarily designed to detect *short* text-text congruencies that have a high chance of being evaluated as an intentional quotation. We favour a mixed-methods approach, which is based on findings from manual-hermeneutic research. Our aim is to focus on Jerome's citation technique: Based on hermeneutic analysis of confirmed quotations, we formulate differentiated criteria that lead to a deeper understanding of the phenomenon of quoting and thus also have the potential to optimize our toolkit.

A central goal of the DFG project “Quoting as a narrative strategy: A digital-hermeneutic analysis of intertextual phenomena using the example of the letters of Church Father Jerome”[1] is to develop and establish a differentiated classification system for the analysis of quotations and allusions that sheds more light on quoting as a discursive-narrative strategy and enables a deeper understanding of the cultural hybridization processes of Christian Late Antiquity. The epistolary corpus of the biblical scholar Jerome is at the center of this investigation because it is particularly characterized by references to (very often pagan) pre-texts.[2] 1

A methodological focus of the study lies first, on the critical discussion of the consequences of using digital matching processes for citation detection and the understanding of citations in general and second, on the (further) development and optimization of automated detection procedures themselves. Particular attention is paid to the transparency of the applied filtering processes since only this guarantees insightful links to hermeneutic intertextuality analyses. Moreover, procedures are to be developed that can be applied universally to any text (of Latin literature). 2

This paper presents the results of considerations relating to citation theory, which support the development of a highly precise digital-hermeneutic analysis toolkit for *short* citations.[3] Firstly, the digital analysis process is discussed with regard to its links to manual-hermeneutic research as well as relevant contexts within the field of digital humanities. After establishing the necessary terminology, corpus requirements, and principles of optimization, the steps and procedures of the digital-hermeneutic analysis toolkit are introduced. Lastly, we formulate differentiated criteria based on hermeneutic analysis of confirmed *short* quotations that lead to a deeper understanding of the phenomenon of quoting and thus also have the potential to optimize our toolkit. 3

I. Classification of the digital analysis process

Manual-hermeneutic approaches of *Quellenforschung* and intertextuality research have predominantly revealed *longer* text-text congruencies and clearly marked allusions, which — due to the high degree of congruency — can usually unambiguously be evaluated as quotations. Additionally, the digitally supported investigation process, which is also capable of holistically detecting two-word *loci similes* in particular, increasingly brings to light parallels that consist of only a small amount of shared word material and can therefore, for the time being, only be classified as 4

potential quotations.[4]

The first digital-hermeneutic mixed-methods approach to treat these matches analytically was presented by Revellio (2022). Designed for the detection of (previously undetected) *Aeneid* citations and allusions, it also analyses Jerome's citation practice in more detail. 5

Revellio's study thus sheds light *inter alia* on the outer limits of quoting as a phenomenon and discusses the question of to what extent and under what conditions textual coincidences of only two identical words can at all be declared a quotation. In contrast to unambiguous quotations consisting of a few words or half, full, and multiple-verse quotations, these two-word syntagms represent a special category of textual correspondences. Here, rather weak citation signals play a role and are still to be determined in more detail. To evaluate the citation potency of these new *short* congruencies, Revellio created a citation typology [Revellio 2022, 189–290, esp. pp. 189–192], incorporating into her proposed categorization morphological, syntactic, lexical, idiomatic, and phraseological justifications. 6

However, these (hermeneutic) analyses and classification processes are only practicable for a moderate number of computer-generated matches. For a universal detection procedure of citations, which is applicable also to larger corpora, an even more sophisticated (pre-)selection of actual citations from the pool of found matches is mandatory. This goal shall now be achieved through the development of adequate filtering procedures and the implementation of a "calibrated" detection and classification process comprising several phases. 7

II. Principles, preconditions, and definitions

Corpus requirements

One advantage of the detection method presented here is its universal applicability.[5] In principle, any (Latin) text can be used as an input, as long as it is available as an appropriately prepared `.txt` file. However, this also means that accessibility to electronic full texts is essential for the compilation of specific research corpora.[6] That said, at the present time, the acquisition of the required digital texts is not always easy and even structurally problematic. 8

The databases accessible without license agreements usually offer the text of older editions, which does not always reflect current research. The quality of the texts usually required in the discipline is more likely to be ensured in the case of licensed databases since these offer more up-to-date editions. Here, however, time-consuming and (in some cases) cost-intensive coordination and negotiation processes with rights holders associated with various institutes are necessary. In addition — both in the case of license-free and in the case of licensed texts — the accompanying information from critical apparatus is (mostly) missing. 9

For the current project, the required classical texts of current editions (without critical apparatus) were procured on the one hand through an adaptation of an already existing database license, on the other hand through renegotiating a license for the specific use case with regards to methods of text-data-mining.[7] 10

From a science perspective, however, the use of licensed text editions structurally prevents the transparency and reproducibility of the digital analysis processes actually envisaged by the project. After all, in order to keep the citation analysis processes replicable, the texts used should also be allowed to be disclosed.[8] 11

All texts provided by the publishers had to be converted into a uniform format for entry into the algorithmic processing. This was done using custom `Python` code to transform and (re)organize all the required paratextual information from the `XML` and `.txt` formats into the final `.txt` file formats used.[9] The described detection process is based on the `Python` programming language in version 3.8.[10] 12

Terminology

The concept of quoting is based on the lexical understanding of citations as it is applied within the existing research tradition (i.e., the ancient concept of *loci similes*). The entire filtering process is thus fundamentally based on finding *shared words* in Jerome's corpus of letters as the target text on the one hand and in the works of Virgil, Cicero, and other classical authors as source texts on the other. According to the *loci similes* concept, *shared words* are identical word forms in a certain number.[11] 13

The starting point of the calibration of the analytical toolkit is the so-called gold standard of citations collected in 14

(manual-hermeneutic) research.[12] Potential new quotations are thus those found matches that emerge as additional findings (total findings minus the gold standard contained therein) from the computer-assisted detection process. From these, the actual new quotations are ultimately selected or confirmed by means of hermeneutic close reading.

In order to be able to perform the final evaluation (and classification), the digital preselection — by means of the filters we discuss below — has to be optimized in such a way that the number of "false" found matches is kept as low as possible. 15

III. The (digital) detection process

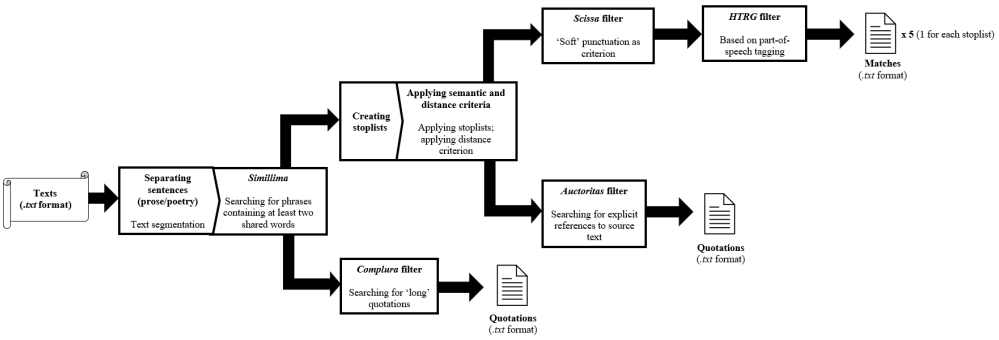

Figure 1. Schematic representation of the digital analysis process

Step 0: Preparing the texts

The texts, in the form of `.txt` files, are subdivided by chapters (for prose) or by verses (for poetry). It is important to emphasize that the analysis process can be carried out with any source text.[13] 16

Step 1: Separation of sentences und text comparison through *Simillima*

These "raw" texts are examined for two-word congruences by means of the script-based tool *Simillima*. These are pairs in which both the source text and the target text contain at least two (distinct) words in exactly the same word form.[14] The actual comparison of texts is based, in essence, on a "matching of certain units" of target and source text: each unit of Jerome's corpus of letters, for example, is compared with each unit of the *Eclogae*. The whole sentence is taken as the matching unit.[15] The output file of the *Simillima* program contains for each pair a match number, the actual text reference,[16] the respective texts (the *shared words* being highlighted), and finally the *shared words* themselves. 17

Step 2: Applying the *Complura* filter

The *Complura* filter returns matches whose shared word material consists of at least four consecutive *shared words*. The pairs found in this way are at the same time already final (albeit partial) results: due to the accumulation of *shared words*, they are unambiguous quotations that do not require further automated analysis.[17] Especially in the case of clustered synsemantic words (grammatical words such as conjunctions, subjunctions, prepositions, pronouns, auxiliary verbs) such citations would also not be detectable in the further digital analysis process.[18] An example of this phenomenon is a quotation of Virgil in Jerome's letter 52, which has been identified by traditional scholars: 18

Verg. *ecl.* 8,57–63	**Hier. *epist.* 52,9,3**
incipe Maenalios mecum, mea tibia, uersus. / omnia uel medium fiat mare. uiuite siluae: / praeceps aërii specula de montis in undas / deferar; extremum hoc munus morientis habeto. / desine Maenalios, iam desine, tibia, uersus. / Haec Damon; uos, quae responderit Alphesiboeus, / dicite, Pierides: **non omnia possumus omnes**.	sed et genus adrogantiae est clementiorem te uideri uelle, quam pontifex Christi est. **non omnia possumus omnes**. alius in ecclesia oculus est, alius lingua, alius manus, alius pes, alius auris, uenter et cetera. lege Pauli ad Corinthios: diuersa membra unum corpus efficiunt. nec rusticus et tantum simplex frater ideo se sanctum putet, si nihil nouerit, nec peritus et eloquens in lingua aestimet sanctitatem. multoque melius est e duobus inperfectis rusticitatem sanctam habere quam eloquentiam peccatricem.
Begin with me, my flute, a song of Maenalus! Nay, let all become mid-ocean! Farewell, ye woods! Headlong from some towering mountain peak I will throw myself into the waves; take this as my last dying gift! Cease, my flute, now cease the song of Maenalus!" Thus Damon. Tell, Pierian maids, the answer of Alphesiboeus; we cannot all do everything. [19]	But it is also a kind of arrogance to want to appear more charitable than Christ's bishop is. Not all of us can do everything. In the Church, one is the eye, another the tongue, a third the hand, still another the foot, another the ears, the stomach, and so on. Read in Paul's letter to the Corinthians: the various members form one body. An uneducated and quite plain brother, if he knows nothing, should not therefore consider himself holy, and neither should an educated and eloquent brother measure his holiness by eloquence. But of the two imperfect kinds, holy modesty is far better than sinful eloquence.[20]

Table 1. Citations with clustered synsemantic words (*non* and *omnis*) would be excluded without the *Complura* Filter. In this example, *non*, *omnia*, *possumus*, and *omnes* are highlighted in both texts.

The *shared words* here mainly consist of stopwords that are already present at the very beginning of a Virgil-Epistulae-specific stoplist, with the exception of *possumus*.[21] Therefore, without passing through the *Complura* filter routine, this well-known quotation would be sorted out as the analysis proceeds. 19

Step 3: Applying filters based on semantic and distance criteria

Since after the first matching process the number of pairs of sentences is clearly too high for the phase of close reading (for example, the matching of the corpus of Jerome's letters with the *Eclogae* yields 408,393 results, that with the *Georgica* 1,217,441 results, and that with the *Aeneid* 4,525,752 results), filtering criteria of semantic quality and distance are applied, as already in Revellio (2022).[22] First, synsemantic words that are usually not sufficient for proving a text–text relationship (i.e. a quotation) are removed with the help of stoplists.[23] If two *shared words* are still contained, their distance is calculated. Then, if a maximum of two words are found between them, the matched pair is kept and written into the output file.[24] 20

After processing this step the algorithm outputs several result lists (one for each stoplist). One goal in the development of the detection process is to determine the one stoplist for which the highest possible number of actual quotations can be found among returned matches, but for which, at the same time, the surplus is as small as possible.[25] The number of potential new quotations remains comparatively high after this step: for the three mentioned works of Virgil alone, the selection of stoplist 6 (at least 100 occurrences) results in 862 pairs for review. Although the number of potential new quotations can thus be decimated compared to the first matching process, the algorithm proves to be deficient at this stage — especially in view of significantly larger text corpora such as those of late-antique authors. 21

Step 4: Applying the *Auctoritas* filter

The *Auctoritas* filter bundles these potential new findings that are clearly recognizable as intentional citations by explicit reference to authors, works, and the like. Thus, for instance, the phrase *sceleratum frigus* (*the villainous cold*) in Jerome's letter 121 is explicitly marked as a citation because the name *Uergilius* is mentioned and because there is actually a citation marker in the form of the predicate *appellet* (*calls*): 22

Verg. *georg.* 2,248–258	**Hier. *epist.* 121,10,5**
pinguis item quae sit tellus, hoc denique pacto / discimus: haud unquam manibus iactata fatiscit, / sed picis in morem ad digitos lentescit habendo. / umida maiores herbas alit, ipsaque iusto / laetior. a, nimium ne sit mihi fertilis illa, / nec se praeualidam primis ostendat aristis! / quae grauis est, ipso tacitam se pondere prodit, / quaeque leuis. promptum est oculis praediscere nigram, / et quis cui color. at **sceleratum** exquirere **frigus** / difficile est: piceae tantum taxique nocentes / interdum aut hederae pandunt uestigia nigrae.	nec hoc miremur in apostolo, si utatur eius linguae consuetudine, in qua natus est et nutritus, cum Uergilius, alter Homerus apud nos, patriae suae sequens consuetudinem **'sceleratum' frigus** appellet. nemo ergo uos superet atque deuincat uolens humilitatem litterae sequi et angelorum religionem atque culturam, ut non seruiatis spiritali intellegentiae, sed exemplaribus futurorum, quae nec ipse uidit, qui uos superare desiderat, siue uidet — utrumque enim habetur in Graeco — ‚praesertim cum tumens ambulet et incedat inflatus mentisque superbiam et gestu corporis praeferat — hoc enim significat *ἐμβατεύων* — ‚frustra autem infletur et tumeat sensu carnis suae carnaliter cuncta intellegens et traditionum Iudaicarum deliramenta perquirens et non tenens caput omnium scripturarum illud, de quo dictum est: caput uiri Christus est, caput autem atque principium totius corporis eorumque, qui credunt, et omnis intellegentiae spiritalis.
Again, richness of soil we learn in this way only: never does it crumble when worked in the hands, but like pitch grows sticky in the fingers when held. A moist soil rears taller grass and is of itself unduly prolific. Ah! not mine be that over-fruitful soil, and may it not show itself too strong when the ears are young! A heavy soil betrays itself silently by its own weight; so does a light one. It is easy for the eye to learn at once a black soil and the hue of any kind. But to detect the villainous cold is hard; only pitch pines or baleful yews and black ivy sometimes reveal its traces.	Let us not be surprised at this in the apostle, that he uses his accustomed language, with which he was born and brought up, since Virgil, our second Homer, following the custom of his homeland, calls the cold 'criminal'. Let no one, then, surpass you and prevail over you who wants to follow the simplicity of Scripture and the worship and adoration of angels, even if you do not serve the spiritual knowledge, but as examples for future times, which the one who wants to despise you has neither seen nor sees himself — for both translations are valid according to the Greek text — especially since he proudly parades around, comes along puffed up, and carries haughtiness before him in spirit and posture — for this is what *ἐμβατεύων* means — for in vain does he puff himself up and boast, when with the sense of his flesh he understands all things carnally, and investigates the silliness of Jewish traditions, and does not hold to the head of all scriptures, that of which it is written: The head of the man is Christ, namely the head and the beginning of the whole body and of those who believe, and the head of all spiritual knowledge.

Table 2. Identification of intentional citations through explicit references with the *Auctoritas* filter, here by the mention of *Uergilius* and the citation marker *appellet* in Jerome. In this example, *sceleratum* and *frigus* are highlighted in both texts. Additionally, *Uergilius*, *alter Homerus apud nos*, and *appellet* are underlined in the text from Jerome.

For efficiency reasons, the filter is applied after the separation of sentences and text comparison (*Simillima*).[26] Similar to the output of the *Complura* filter, the result list of the *Auctoritas* filter contains (partially) final results. These are to be classified as quotations due to the explicit reference to the respective source text and therefore do not have to go through any further selection processes. 23

Step 5: Applying the "Scissa" filter

The "Scissa" filter eliminates potential new quotations whose *shared words* — in at least one of the two texts (source or target) — do not belong to one syntactic unit. The indicators for separation of the word material are "soft" punctuation marks (comma, colon, and semicolon).[27] Since colons, commas, and semicolons indicate "weaker" syntactic separations than periods, question marks, and exclamation marks, they are not used in Step 1 to separate sentences, in order to prevent the heavy fragmentation of source and target text. Nevertheless, the caesurae marked in this way can — in this phase of more detailed analysis — be used as indicators for syntactic separation of quoted material. This is illustrated by a match resulting from the matching of Jerome's letters with Virgil's *Georgics* as an example: 24

Verg. *georg.* 1,322–327	Hier. *epist.* 108,28,3
saepe etiam immensum caelo uenit agmen aquarum, / et foedam glomerant tempestatem imbribus atris / collectae ex alto nubes: ruit arduus aether / et **pluuia** ingenti sata **laeta** boumque labores / diluit; implentur fossae et caua flumina crescunt / cum sonitu feruetque fretis spirantibus aequor.	aderant Hierosolymorum et aliarum urbium episcopi et sacerdotum inferioris gradus ac Leuitarum innumerabilis multitudo. omne monasterium uirginum et monachorum chori repleuerant. statimque ut audiuit sponsum uocantem: surge, ueni, proxima mea, speciosa mea, columba mea, quoniam ecce hiemps pertransiuit, **pluuia** abiit sibi, **laeta** respondit: flores uisi sunt in terra, tempus sectionis aduenit et: credo uidere bona domini in terra uiuentium.
Often, too, there appears in the sky a mighty column of waters, and clouds mustered from on high roll up a murky tempest of black showers: down falls the lofty heaven, and with its deluge of rain washes away the gladsome crops and the labours of oxen. The dykes fill, the deep-channelled rivers swell and roar, and the sea steams in its heaving friths.	Present were the bishops of Jerusalem and of other cities, priests of lower rank, and an innumerable multitude of Levites. The whole monastery was filled with crowds of virgins and monks. As soon as she heard the bridegroom speak, ‘Arise, come, my beloved, my fair one, my dove, for behold, winter is over, the rains have ceased,’ she answered joyfully, ‘Flowers have appeared on the earth, the time of cutting is here,’ and, ‘But I believe to behold the kindness of the Lord in the land of the living.’

Table 3. The *Scissa* filter eliminates potential quotations with *shared words* that do not belong to one syntactic unit in either source or target text, here the *shared words* (*pluuia* and *laeta*) are syntactically separated by commas in Jerome’s letter. In this example, *pluuia* and *laeta* are highlighted in both texts. Additionally, *et* is underlined twice on the left and three times on the right.

The *shared words* (*pluuia*; *laeta*)[28] are embedded in an asyndetic structure in Jerome: *pluuia* is the subject of *abiit*, while *laeta* characterizes the subject implicitly contained in *respondit*. The fact that these *shared words* are thus bound to two different predicate verbs (which is indicated accordingly by commas in the underlying print edition) rules out a citation.[29] Having been applied, the “Scissa” filter reduced the number of potential new quotations by 28% for the *Eclogues*, by 22% for the *Georgics* and by 30% for the *Aeneid*.[30] 25

Step 6: Applying the historical text-reuse grammar (*HTRG*) filter[31]

The *HTRG* filter developed by Revellio is placed after the “Scissa” filter: 26

> Potential findings from computer-assisted text comparison, whose matching word material consists of at least two nouns or two verbs as well as a combination of these two word classes, [are] particularly suited to establish a meaning-producing text-to-text relationship.[32] [Revellio 2022, 151]

Accordingly, the *shared words* of each match are analyzed by part-of-speech tagging. Information on whether the combination of *shared words* satisfies the *HTRG* rule (N–N; V–V; N–V; V–N) is added to each match. Thus, this filter does not eliminate matches but increases the “resolution” of the output data.[33] 27

The steps of the digital analysis tools described so far are already performed in an automated way. In what follows, new optimization strategies are presented that are currently in an experimental stage but seem to be target-oriented and promising. By implementing these approaches into the routine, the automation process shall be extended and the amount of work in the close-reading phase shall thus be made manageable by further reducing the number of returned matches. 28

IV. Optimization strategies based on the analysis of published Quotations of Virgil in Jerome

Short quotations make up only a small proportion of the total amount of quotations of Virgil in Jerome’s corpus of letters, a fact that is evident from analyzing the research-based gold standard of published Virgilian quotations from the *Eclogues*, *Georgics,* and the *Aeneid*. At the same time, precisely these *short* quotations are the core result of the digital detection process. Because of this discrepancy, the results have to be considered all the more critically. In order to filter out further non-relevant (i.e., false) matches prior to the phase of hermeneutic close reading, further strategies are presented below, derived from the phenomenological characteristics of known and published (i.e., contained in the gold standard) *short* quotations of Virgil in Jerome. The pairs of quotations discussed below — together with the quotation of Verg. *georg.* 2,256b–257a in Hier. *epist.* 121,10,5 discussed in connection with the *Auctoritas* filter — represent the complete list of Virgilian quotations in Jerome that have two exactly identical word 29

forms and can thus be termed *short*.[34]

30 It should be noted that a match with two *shared words* can in actual fact have other “shared” words of identical form. However, these often belong to the category of synsemantics and are subsequently removed by applying the stoplists. In the case of the *Eclogues*, *Georgica* and *Aeneid*, many of the potential new quotations selected by the algorithm do indeed share other words in addition to two *shared words* [35], which, however, are mostly irrelevant for the evaluation of the match as a quotation. [36] This disregard of synsemantic words among additional *shared words* indeed does mimic a “manual” hermeneutic reading but still justifies the use of the term “short” instead of the term “two-word citation” (where it would always need to be stated whether only *shared words* or all identical word forms are meant).

1. Optimization strategy: Taking into account the direct environment of *shared words*

31 The presented algorithm successfully outputs the quotation of Verg. *ecl.* 4,60–63 in Hier. *epist.* 130,16,3, which is contained in the gold standard, as a match; the *shared words* (*risu*; *matrem*) are retained.[37]

Verg. *ecl.* 4,60–63	**Hier. *epist.* 130,16,3**
Incipe, parue puer, **risu** cognoscere **matrem**: / matri longa decem tulerunt fastidia menses. / incipe, parue puer: qui non risere parentes, / nec deus hunc mensa, dea nec dignata cubili est.	solent enim huiusce modi per angulos musitare et quasi iustitiam dei quaerere: ‘cur illa anima in illa est nata prouincia? quid causae extitit, ut alii de Christianis nascantur parentibus, alii inter feras et saeuissimas nationes, ubi nulla dei notitia est?’ cumque hoc quasi scorpionis ictu simplices quosque percusserint et fistulato uulnere locum sibi fecerint, uenena diffundunt: ‘putasne, frustra infans paruulus et qui uix **matrem risu** et uultus hilaritate cognoscat, qui nec boni aliquid fecit nec mali, daemone corripitur, morbo opprimitur regio et ea sustinet, quae uidemus inpios homines non sustinere et sustinere deo seruientes?
Begin, baby boy, to recognize your mother with a smile: ten months have brought your mother long travail. Begin, baby boy! The child who has not won a smile from his parents, no god ever honoured with his table, no goddess with her bed!	Some of this kind are wont to whisper in silent corners and downright question the justice of God: ‘Why was that soul born in that region? What reason was found that some were conceived by Christian parents, others were born in savage and most wicked peoples who have no knowledge of God?’ As soon as they have shaken some simple people by this, as by the sting of a scorpion, and have made a place for themselves in the tubular wound, they spout their venom: ‘Do you think a very young child, one who barely recognizes its mother by her laughter and her face by her joy, who has done neither good nor bad, is seized by the devil for no reason, or prostrated by jaundice, or endures such things for no reason, which we see godless people do not suffer, but those who serve God do?’

Table 4. Additional material from Virgil (*cognoscat* – *cognoscere*, *infans* – *puer*, *paruulus* – *paruus*) confirms the match as an actual quotation and thereby enhances close reading. In this example, instances of these pairs of words are individually underlined in both texts. Additionally, *risu* and *matrem* are highlighted in both texts.

32 The syntactic relationship of the two nouns, which are in ablative and accusative case respectively, remains unclear at first. Nevertheless, this match can undoubtedly be regarded as an actual quotation. This is confirmed by the surrounding text environment. In it, additional “material” from Virgil’s poem is “recycled” (but not in exactly the same form!): *infans paruulus* can be traced back to the twice appearing *parue puer*, but neither the synonym (*infans* for *puer*) nor the diminutive (*paruulus* < *paruus*) are “visible” for the algorithm. Moreover, the verb *cognoscere* is present in the target text, albeit in inflected form, which, in both target and source text, forms the syntactic connection between the *shared words* (*matrem*: direct object; *risu*: instrumental ablative).

33 This results in the optimization strategy of making “invisible” language material, which can be traced back to the target text, “visible” to the algorithm. As the above *short* quotation shows, this additional material in the target text can be connected to the forms of the source text in three different ways: by lemmatization (cf. *cognoscat* – *cognoscere*), by finding synonyms (cf. *infans* – *puer*), and by detecting etymologically related forms (cf. *paruulus* – *paruus*). Through these approaches more quoted words can be detected. By marking those quotations that consist of only two *shared words* but have further similar meaningful forms in the same sentence, these filters could significantly support the close reading.

2. Optimization strategy: Considering the syntactic complexity of highly frequent phrases

In book 9 of the *Aeneid*, Iris urges the Rutulian king Turnus to attack the Trojan camp with the words *rumpe moras omnis et turbata arripe castra* (*break off delay, and seize the bewildered camp*, 13). The phrase *rumpere moras*, which is frequently employed by poets, is used by Jerome in letter 130, where he writes: *rumpe moras omnes* (*abandon all procrastination*). Hagendahl cautiously counts this passage among the "[l]ess obvious reminiscences of Virgil" and thus points to the difficulty of an exact attribution [Hagendahl 1958].[38] 34

Verg. *Aen.* 9,5–13	**Hier. *epist.* 130,5,3**
ad quem sic roseo Thaumantias ore locuta est: / 'Turne, quod optanti diuum promittere nemo / auderet, uoluenda dies en attulit ultro. / Aeneas urbe et sociis et classe relicta / sceptra Palatini sedemque petit Euandri. / nec satis: extremas Corythi penetrauit ad urbes / Lydorumque manum, collectos armat agrestis. / quid dubitas? nunc tempus equos, nunc poscere currus: / **rumpe moras** omnis et turbata arripe castra.'	urbs tua, quondam orbis caput, Romani populi sepulchrum est, et tu in Libyco litore exulem uirum ipsa exul accipies? quam habitura pronubam? quo deducenda comitatu? stridor linguae Punicae procacia tibi fescennina cantabit. **rumpe moras** omnes. perfecta dilectio foras mittit timorem. adsume scutum fidei, loricam iustitiae, galeam salutis, procede ad proelium. habet et seruata pudicitia martyrium suum. quid metuis auiam? quid formidas parentem? forsitan et ipsae uelint, quod te uelle non credunt.
To him, with roseate lips, thus spoke the child of Thaumas: 'Turnus, what no god dared to promise to your prayers, see—the circling hour has brought unasked! Aeneas, leaving town, comrades and fleet, seeks the Palatine realm and Evander's dwelling. Nor does that suffice; he has won his way to Corythus' furthest cities, and is mustering the Lydian country folk in armed bands. Why hesitate? Now, now is the hour to call for steed and chariot; break off delay, and seize the bewildered camp!	Your city, once the head of the world, is the tomb of the Roman people, and will you, even in exile, accept a man, also exiled, on the coast of Libya? What matchmaker shalt thou have? By what companion shalt thou be led home? The hiss of the Punic language will sing cheeky Fescennine wedding songs for thee. Abandon all procrastination. Perfect love sends fear at the door. Seize the shield of faith, the armor of righteousness, the helmet of salvation, march out to battle. Even the preservation of chastity holds its own martyrdom. What do you fear your grandmother? What do you fear your mother? Maybe they even want for you what they don't think you want.

Table 5. The syntactic similarities validate the match as an actual quotation and improve the close reading of the text. In this example, *rumpe* and *moras* are highlighted in both texts. Additionally, *omnis* and *omnes* are underlined.

However, here again the "additional" linguistic material is striking: Jerome does not only write *rumpe moras*, but strengthens his request — like Virgil in Verg. *Aen.* 9,13 — with *omnes* (= *omnis* [39]). Both instances thus have the same syntactic complexity (predicate + object + attribute adjective). Other instances of the phrase *rumpere moras* therefore seem to be rather unlikely as pre-texts, such as Verg. *Aen.* 4,569 (*heia age, rumpe moras – Up then, break off delay*) and Verg. *georg.* 3,43 (*rumpe moras; uocat ingenti clamore Cithaeron – break with slow delay! With mighty clamour Cithaeron calls*), the earliest instance of the phrase. 35

The following returned match, which cannot be considered a quotation based on these considerations, illustrates *ex negativo* that recycling syntactic complexity increases the plausibility of a quotation: 36

Verg. *georg.* 1,401–403	**Hier. *epist.* 108,17,1**
at nebulae magis ima petunt campoque recumbunt, / solis et occasum seruans de **culmine summo** / nequiquam seros exercet noctua cantus.	Uerum haec possunt communia esse cum <non> paucis et scit diabolus non in **summo** uirtutum **culmine** posita.
But the mists are prone to seek the valleys, and rest on the plain, and the owl, as she watches the sunset from some high peak, vainly plies her evening song.	But these things can be common to a few, and the devil knows that they are not on the highest pinnacle of virtue.

Table 6. Greater syntactic complexity in Jerome's text activates the figurative meaning of *culmen* (*summit*), unlike Virgil's literal use. In this example, *culmine* and *summo* are highlighted in both texts. Additionally, *de*, *in*, and *uirtutum* are underlined.

It is obvious that despite the presence of two *shared words* (*summo*; *culmine*) there is no reference to Virgil's line. This is due to the different nature of the prepositional phrases in which the *shared words* are found. Besides the semantic difference in the prepositions themselves (*de* vs. *in*), it is then, above all, the greater syntactic complexity of Jerome's phrasing; the identical *culmine* is complemented by an attribute genitive (*uirtutum*). In particular, the addition of this attribute genitive "activates" the figurative meaning of *culmen* (*summit*), whereas in Virgil *culmen* refers to a roof ridge. [40] 37

The analysis of syntactic complexity thus highlights a difference of source and target text,[41] which seems to make semantic shifts detectable for the algorithm. Based on these considerations, a filter (to be implemented after part-of-speech tagging) is being developed that analyzes the syntactic structure underlying the *shared words* and checks whether the syntactic complexity is similar in both source and target text. Increased syntactic complexity in one of the two texts considerably reduces the plausibility of a citation. Additionally, the absence of complexity (i.e., especially the absence of one or more attributes) seems to be an indication that there is no citation, especially in simple adverbial ablative phrases like *sole orto* and *toto orbe*. Frequent adverbials of place and time could be eliminated in this way. 38

3. Optimization strategy: Taking into account the manuscript tradition

A fundamental problem that largely eludes the digital analysis tools is the consideration of the manuscript tradition. The algorithm presented here can only take into account a single wording. For example, for the citation pairs of Hier. *epist.* 126,2,2/129,4,3 = Verg. *Aen.* 4,42–43 shown below, the digital analysis produces only two *shared words* (*lateque*; *Barcaei*). In fact, the manuscripts of the *Aeneid* also give the reading *uagantes* (*roaming*) in addition to *furentes* (*raging*), which Jerome obviously refers to[42] — and not just in this passage.[43] If *uagantes* was chosen in the underlying digital text of the *Aeneid*, the digital analysis process could have detected (depending on the respective stoplist) up to three *shared words*. 39

Verg. *Aen.* 4,31–44	**Hier. *epist.* 126,2,2**
Anna refert: ‘o luce magis dilecta sorori, / solane perpetua maerens carpere iuuenta, / nec dulcis natos Veneris nec praemia noris? / id cinerem aut manis credis curare sepultos? / esto: aegram nulli quondam flexere mariti, / non Libyae, non ante Tyro; despectus Iarbas / ductoresque alii, quos Africa terra triumphis / diues alit: placitone etiam pugnabis amori? / nec uenit in mentem quorum consederis aruis? / hinc Gaetulae urbes, genus insuperabile bello, / et Numidae infreni cingunt et inhospita Syrtis; / hinc deserta siti regio **lateque** furentes / **Barcaei**. quid bella Tyro surgentia dicam / germanique minas?	hoc autem anno, cum tres explicassem libros, subitus impetus barbarorum, de quibus tuus dicit Uergilius: **lateque** uagantes **Barcaei** et sancta scriptura de Ismahel: contra faciem omnium fratrum suorum habitabit, sic Aegypti limitem, Palaestinae, Phoenices, Syriae percucurrit ad instar torrentis cuncta secum trahens, ut uix manus eorum misericordia Christi potuerimus euadere. quodsi iuxta inclitum oratorem silent inter arma leges, quanto magis studia scripturarum, quae et librorum multitudine et silentio ac librariorum sedulitate, quodque uel proprium est, securitate et otio dictantium indigent!
Anna replies: ‘O you who are dearer to your sister than the light, are you, lonely and sad, going to pine away all your youth long, and know not sweet children or love’s rewards? Do you think that dust or buried shades give heed to that? Grant that until now no wooers moved your sorrow, not in Libya, nor before then in Tyre; that Iarbas was slighted, and other lords whom the African land, rich in triumphs, rears; will you wrestle also with a love that pleases? And does it not come to your mind whose lands you have settled in? On this side Gaetulian cities, a race invincible in war, unbridled Numidians, and the unfriendly Syrtis hem you in; on that side lies a tract barren with drought, and Barcaeans, raging far and wide. Why speak of the wars rising from Tyre, and your brother’s threats ...?	In this year, however, when I had already explained three books, a sudden onslaught of the barbarians, about whom your Virgil says: ‘the far-roaming Barkaeans’ and the holy scripture about Ishmael: ‘opposed to the face of all his brothers he shall live’, overtook the borders of Egypt, Palestine, Phoenicia, Syria and swept away everything like a torrent, so that we could escape their clutches just by the mercy of Christ. But if, according to the famous orator, laws are silent in war, how much more the occupation with the Scriptures, which requires a multitude of books, rest as well as busyness of the scribes and finally, as a special characteristic, the security and leisure for the person dictating!

Table 8. Digital analysis tool finds only two *shared words* (*lateque*, *Barcaei*) but manuscripts of the *Aeneid* offer alternative readings (e.g., *uagantes* instead of *furentes*), which Jerome references. In this example, *latque* and *Barcaei* are highlighted in both texts.

Verg. *Aen.* 4,31–44	**Hier. *epist.* 129,4,3**
Anna refert: ‘o luce magis dilecta sorori, / solane perpetua maerens carpere iuuenta, / nec dulcis natos Veneris nec praemia noris? / id cinerem aut manis credis curare sepultos? / esto: aegram nulli quondam flexere mariti, / non Libyae, non ante Tyro; despectus Iarbas / ductoresque alii, quos Africa terra triumphis / diues alit: placitone etiam pugnabis amori? / nec uenit in mentem quorum consederis aruis? / hinc Gaetulae urbes, genus insuperabile bello, / et Numidae infreni cingunt et inhospita Syrtis; / hinc deserta siti regio **lateque** furentes / **Barcaei**. quid bella Tyro surgentia dicam / germanique minas?	ab Ioppe usque ad uiculum nostrum Bethleem quadraginta sex milia sunt, cui succedit uastissima solitudo plena ferocium barbarorum, de quibus dicitur: contra faciem omnium fratrum tuorum habitabis et quorum facit poeta eloquentissimus mentionem: **lateque** uagantes **Barcaei**, a Barca oppido, quod in solitudine situm est, quos nunc corrupto sermone Afri Baricianos uocant. hi sunt, qui pro locorum qualitatibus diuersis nominibus appellantur et a Mauritania per Africam et Aegyptum Palaestinamque et Phoenicem, Coelen Syriam et Osrohenen, Mesopotamiam atque Persidem tendunt ad Indiam.
Anna replies: ‘O you who are dearer to your sister than the light, are you, lonely and sad, going to pine away all your youth long, and know not sweet children or love’s rewards? Do you think that dust or buried shades give heed to that? Grant that until now no wooers moved your sorrow, not in Libya, nor before then in Tyre; that Iarbas was slighted, and other lords whom the African land, rich in triumphs, rears; will you wrestle also with a love that pleases? And does it not come to your mind whose lands you have settled in? On this side Gaetulian cities, a race invincible in war, unbridled Numidians, and the unfriendly Syrtis hem you in; on that side lies a tract barren with drought, and Barcaeans, raging far and wide. Why speak of the wars rising from Tyre, and your brother’s threats ...?	From Jaffa to our little village of Bethlehem is a distance of 46 miles, adjoining which is a vast wasteland, full of savage barbarians, about whom it is said: ‘opposed to the face of all thy brethren shalt thou live,’ and the mention of which our most eloquent poet thus forms: ‘the far-roaming Barkaeans,’ namely, from the city of Barka, which is situated in the wasteland, whom the Africans now call, in a neglected manner of speech, Barikians. These are the ones whom they call by different names according to the nature of the places, stretching from Mauritania through Africa and Egypt, Palestine and Phoenicia, Syria Coele and Osrhoene, Mesopotamia and Persia to India.

Table 9. Another example for this challenge. In this example, *latque* and *Barcaei* are highlighted in both texts. Additionally, the phrase *facit poeta eloquentissimus mentionem* is underlined in the text from Jerome.

Both pairs of quotations show that especially in the phase of manual-hermeneutical evaluation of potential new quotations special attention must also be paid to the possibility of different readings within the manuscript tradition. Thus, consultation of the critical apparatus is essential.[44] The case of the variants *furentes*/*uagantes* shows that divergences within the manuscript tradition can go far beyond similar lemmata and synonyms. Digital texts that preserve the breadth of the manuscript tradition as much as possible and present it in an algorithm-compatible form can thus be seen as an urgent desideratum in the field of digital text analysis. Since the current digital *Teubner* editions do not (yet) meet these specific needs, we test the definition of minimal deviations in letter sequences as an immediately implementable replacement strategy. It is also possible that this approach will prove to be more efficient from a procedural point of view.[45] 40

4. Optimization strategy: Repeated quotations as evidence for source-text familiarity

The phrase *experto credite* (*trust the one who has experienced it*) from book 11 of the *Aeneid* is used by Jerome in Hier. *epist.* 84,3,5 as *credite experto*. 41

Verg. *Aen.* 11,278–287	**Hier. *epist.* 84,3,5**
ne uero, ne me ad talis impellite pugnas: / nec mihi cum Teucris ullum post eruta bellum / Pergama nec ueterum memini laetorue malorum. / munera quae patriis ad me portatis ab oris / uertite ad Aenean. stetimus tela aspera contra / contulimusque manus: **experto credite** quantus / in clipeum adsurgat, quo turbine torqueat hastam. / si duo praeterea talis Idaea tulisset / terra uiros, ultro Inachias uenisset ad urbes / Dardanus et uersis lugeret Graecia fatis.	quod autem opponunt congregasse me libros illius super cunctos homines, utinam omnium tractatorum haberem uolumina, ut tarditatem ingenii lectionis diligentia conpensarem! congregaui libros eius, fateor; et ideo errores non sequor, quia scio uniuersa, quae scripsit. **credite experto**, quasi Christianus Christianis loquor: uenenata sunt illius dogmata, aliena a scripturis sanctis, uim scripturis facientia. legi, inquam, legi Origenem et, si in legendo crimen est, fateor — et nostrum marsuppium Alexandrinae chartae euacuarunt — : si mihi creditis, Origeniastes numquam fui; si non creditis, nunc esse cessaui.
Do not, do not urge me to such battles! I have no war with Teucer's race since Troy's towers fell, and I have no joyful memory of those ancient ills. The gifts that you bring me from your country, take them rather to Aeneas. We have faced his fierce weapons, and fought him hand to hand: trust one who has experienced it, how huge he looms above his shield, with what whirlwind he hurls his spear! Had Ida's land borne two others like him, the Trojans would even have stormed the towns of Inachus, and Greece would be mourning, with fate reversed.	But this they oppose, that I have gathered together the books of that one about all men; oh, if I had the volumes of all commentators, in order to compensate the slowness of the mind by the diligence of the reading! I have collected his books, I confess; and therefore, I do not follow errors, because I know all that he has written. Trust the one who has experienced it, as if I speak to Christians as a Christian; poisoned are his doctrines, strange in relation to the sacred writings, because they do violence to them. Yes, I have read, I say, I have read Origen, and, if in reading there is an offense, I confess it — and my purse emptied Alexandrian leaves — : If you believe me, I was never a follower of Origen; if you do not believe me, I have now ceased to be one.

Table 10. Despite the shortness, this is a quotation because Jerome previously quoted the lines in full. In this example, *experto* and *credite* are highlighted in both texts.

Due to the shortness of the phrase — and also since it is not the only possible source[46] — this match does not look like an obvious quotation at first glance. Nevertheless, the instance in Jerome is accepted as a quotation of Virgil by scholars. The key indicator for this verdict is the reuse of the quotation, albeit in shorter form: Jerome has already quoted these lines of the *Aeneid* in full in Hier. *epist.* 50,4,2. For the evaluation of potential new quotations, this observation results in the strategy of comparing several instances (if any) of *short* quotations. If a short quotation occurs elsewhere as part of a longer quotation, familiarity with the pre-text seems plausible and thus increases the match's potential to be an actual quotation. 42

V. Concluding remarks and outlook

From the analysis of published *short* quotations of Virgil in Jerome consequences arise in two respects. On the one hand, concrete optimization strategies for the algorithm can be deduced from phenomenological properties of the text passages. These are: 43

(a) the lemmatization of the close environment of quotations,
(b) the analysis of syntactic complexity, especially of compact frequent phrases,
(c) the consideration of the manuscript tradition, and
(d) the testing for iteration of quotations or quotation fragments.

On the other hand, independent of the aptitude and effectiveness of the filters developed on the basis of these considerations, the analysis of the *short* quotations of Virgil in Jerome also expands the horizon of the manual-hermeneutic reading process: strikingly, none of the quotations result only from the sum of the *shared words*. Indeed, an intentional quotation is only recognizable when:

(a) either further quoted words (inflected or replaced by synonyms) or an explicit reference (e.g. naming the source author) are found in the environment,
(b) high-frequency phrases are quoted in their exact syntactic structure (and thus at the same time are *longer* quotations),
(c) editorial changes obscure further *shared words* (precisely a problem of computer-aided access to digital texts), or
(d) finally the reuse of a quotation fragment can be proven.

These findings, which also shed light on Jerome's citation technique, are of central importance for any further systematization of citation recognition processes, as they define initial "minimum standards" for citations and thus make the structured parameters of citation theory visible and transparent — and easier to operationalize. 44

Primary Sources 45

Conte 2019 Conte, G. B. (ed.) (2019) *P. Vergilius Maro. Aeneis. Editio Altera*. Berlin/Boston: Walter de Gruyter.

Fairclough 1999(a) Fairclough, H. R. (transl.) (1999) *Virgil. Eclogues. Georgics. Aeneid: Books 1-6. Revised by G. P. Goold*. Loeb Classical Library, 63. Cambridge, Mass.: Harvard University Press.

Fairclough 1999(b) Fairclough, H. R. (transl.) (1999) *Virgil. Aeneid: Books 7-12. Appendix Vergiliana. Revised by G. P. Goold*. Loeb Classical Library, 64. Cambridge, Mass.: Harvard University Press.

Hilberg 1996(a) Hilberg, I. (ed.) (1996) *S. Eusebii Hieronymi Epistulae. Pars I: I–LXX*. 2 edn. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Hilberg 1996(b) Hilberg, I. (ed.) (1996) *S. Eusebii Hieronymi Epistulae. Pars II: LXXI–CXX*. 2 edn. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Hilberg 1996(c) Hilberg, I. (ed.) (1996) *S. Eusebii Hieronymi Epistulae. Pars III: CXXI–CLIV*. 2 edn. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Ottaviano and Conte 2013 Ottaviano, S. and Conte, G. B. (eds.) (2013) *P. Vergilius Maro. Bucolica et Georgica*. Berlin/Boston: Walter de Gruyter.

Notes

[1] Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project no. 382880410.

[2] Cf., for example: [Luebeck 1872], [Hagendahl 1958, 91–328], [Godel 1964], [Voß 1969], [Burzacchini 1975], [Burzacchini 1978], [Jakobi 2006], [Cain 2013], [Adkin 2011], and [Feichtinger 2021]. A synopsis of Jerome's own statements on the Christian use of classical literature are collected by Mohr (2007).

[3] On the term *short* citation, which is used throughout the paper, see Section IV.

[4] Analytical approaches to the digital detection of intertextuality phenomena have been presented by Bamman and Crane (2008), Hohl Trillini and Quassdorf (2010), and Schubert and Heyer (2010) with the "Citationgraph" of the *eAqua* project, Büchler et al. (2014) with the "Tracer" of the *eTrap* project, and Manca et al. (2011) within the project *Musisque Deoque*. The approach of Coffee et al. (2013), which focuses on two-word *loci similes*, pursues, within the *Tesserae* project, an operationalization approach specifically pertaining to classical philology [Burns 2017], [Diddams and Gawley 2017], [Coffee et al 2020]. However, problems arise in the concrete application of the filter settings (lack of transferability of once calibrated settings, lack of explicability and transparency of the selection and insufficient adjustment possibilities), so that this project environment is not used for the development of a differentiated classification system of citations.

[5] The present project, unlike *Tesserae*, does not rely on pre-existing corpora. At the same time this circumstance ensures that the criteria arising from the demands for transparency in the filtering process are satisfied (see Section I, n. 4).

[6] Although, in theory, the Copyright Act attempts to enable precisely this with regulations on text and data mining for the purposes of scientific research, in practice — in our experience — negotiations remain necessary.

[7] In the first case, the texts of the relevant classical authors were made available directly by the publisher without further negotiations within the framework of an existing license. In the second case, the situation was more difficult because the question of which institution holds the rights to the electronic (!) texts remained unclear until the end, even for the publishers involved. In addition, the cost of making the texts available within the existing license offer was estimated to be very high, which made negotiations about the cost framework necessary.

[8] The copyright exceptions for text-mining methods (especially for research purposes) are limited by regulations surrounding the publication of the texts within the context of research results.

[9] For this step in particular, it would be important for publishers to provide texts in a systematic and consistent manner in order to avoid time-consuming adaptation processes for singular and divergent text formats.

[10] The individual analysis steps, so far, are stored in individual scripts. The aim is to combine the scripts into a *pipeline*, not least for the sake of user-friendliness. It is our intention to publish the code base on the GitHub development platform.

[11] Thus, the citation pair Verg. *ecl.* 4,60–61 (*Incipe, parue puer, **risu** cognoscere **matrem**: / matri longa decem tulerunt fastidia menses – Begin, baby boy, to recognize your mother with a smile: ten months have brought your mother long travail*; all translations for Virgil's works are taken from Fairclough's two-volume *Loeb Classical Library* edition revised by Goold, 1999–2000) and Hier. *epist.* 130,16,3 (*putasne, frustra infans paruulus et qui uix **matrem risu** et uultus hilaritate cognoscat, qui nec boni aliquid fecit nec mali, daemone corripitur, morbo opprimitur*

regio et ea sustinet, quae uidemus inpios homines non sustinere et sustinere deo seruientes? – Do you think a very young child, one who barely recognizes its mother by her laughter and her face by her joy, who has done neither good nor bad, is seized by the devil for no reason, or prostrated by jaundice, or endures such things for no reason, which we see godless people do not suffer, but those who serve God do?; all translations for Jerome's letters are by the authors) contains exactly two *shared words* (*risu – risu*; *matrem – matrem*). Other similar word forms such as declined, conjugated, and derived forms of the same lemma (*matri – matrem*; *cognoscere – cognoscat*; *parue – paruulus*) as well as synonyms (*puer – infans*) although particularly relevant for the hermeneutical reading process, initially are undetected in the digital detection process.

[12] Cf. on the establishing of optimization rules: [Revellio 2022, 128–135].

[13] On the universal applicability of the detection processes, see Corpus requirements, n. 5.

[14] In this routine, the matching process with *Simillima* replaces *Tesserae* in its basic function of text comparison, [Revellio 2022, 123]. Nevertheless, the process is structurally clearly comparable to *Tesserae*.

[15] During the algorithmic procedure the end of a sentence is indicated by a period, question mark, and exclamation mark. Of course, the modern punctuation is not an unproblematic criterion, whose employment and effects on the results of the study must always be reflected; [Revellio 2022, 116, n. 384]. Separation of sentences is also performed when a colon and quotation marks are combined, i.e., at the beginning of quoted speech. If, on the other hand, a period, question mark, or exclamation mark occur within a parenthesis (marked, for example, by brackets), separation of sentences is suppressed in order to keep the parenthesis and the surrounding sentence together as one unit. However, insertions marked by dashes can only be addressed programmatically in a few cases, because dashes can also occur individually (whereas brackets always occur in pairs). Therefore, the separation of sentences is not suppressed for an insertion marked by dashes with a period, question mark, or exclamation mark. This can be considered unproblematic, since insertions with such punctuation marks are usually long and it is therefore unlikely that quote-constituting split word material can be found around this insertion (distance criterion). Sentences of direct speech which are interrupted by speech-introducing words such as *inquit*, *ait*, *dixit*, or *dicit* are also held together. The fact that only the most common speech introductions can be considered here is negligible, because it is rather unlikely that a quotation within a literal speech is interrupted (if at all) by an insertion other than these standardized variants. Additionally, in order to improve the algorithm, periods as part of abbreviations should be treated as exceptions during the separation of sentences. These exceptions have so far not been implemented, as *Georgics*, *Eclogues*, *Aeneid* and Jerome's letters do not contain any abbreviations. Naturally, with a view to other corpora (e.g., Cicero), this is necessary and proposed. Moreover, the sentence-tokenization feature of the *Classical Language Toolkit* (CLTK) is currently being tested as an alternative.

[16] In the case of poetry, the line in which a sentence "begins" is taken as the text reference.

[17] On this filter approach, [Revellio 2022, 158]. By focusing on the number of consecutive *shared words* and disregarding their actual order, which may be different in the target and source text, there are indeed false (i.e., irrelevant) matches, but only in small numbers. For the *Georgica*, *Eclogae* and *Aeneid*, for example, the *Complura* filter yields 76 correctly identified matches and only five "false" matches, which can be rejected at first glance with little manual effort.

[18] This is due to the removal of stopwords, see Step 3.

[19] All translations for Virgil's works are taken from Fairclough's two-volume Loeb Classical Library edition revised by Goold, 1999–2000).

[20] All translations for Jerome's letters are by the authors.

[21] On stoplists, see Step 3.

[22] On the deduction of these criteria and discussion thereof, [Revellio 2022, 136–145].

[23] On the application of such stoplists, [Revellio 2022, 156–159]. In addition to the standard list of the *Perseus* project, seven corpus-based stoplists consisting of the most frequent words are used. Despite the combination of Virgil's three works with Jerome's *Epistulae*, the stoplists show a strong dominance of Jeromean word material with respect to the autosemantics contained by the most-frequent-words approach. For example, in the first stoplist containing the words occurring at least 250 times, the word *Christi* (445 times) appears more frequently than many synsemantics such as *quoque* (*too*, 441 times) or *tu* (*you*, 433 times). For comparison, the word *Aeneas* only appears in the list of words that occur at least 150 times. Approaches of a differentiated weighting of the two texts or of taking into account the frequency of a word within the text (cf. Zou et al., 2006) by way of which such disproportions could be corrected have been considered here but could not yet be implemented.

[24] The distance of 2 was preferred to the distance of 3 based on previous observations [Revellio 2022, 144]. On the implementation, see [Revellio 2022, 156–159].

[25] So far, for Virgil's *Eclogae*, *Georgica*, and *Aeneid*, stoplist 6, which contains words with more than 100 occurrences in the *Eclogae*, *Georgica*, *Aeneid*, and Jerome's *Epistulae*, has turned out to be optimal in this regard.

[26] The filter searches the sentences of the target text for authorial information (e.g., *Uergilius*, *poeta*, *Aeneis*, etc.) and returns all matches for the respective sentence if a match is found. However, if the filter is applied immediately after matching (*Simillima*), *Uergilius* alone for Hier.

epist. 121,10,5 already produces 382 matches with various passages of Virgil, of which, except for Verg. *georg.* 2,256, all are the result of only two synsemantics such as *et* or *in* and cannot be considered as pretexts. Therefore, the application after removal of the stoplist of the *Perseus* project is advisable in order to exclude as many as possible, but no author-specific synsemantics. If corpus-based stoplists were applied, just quotations, whose words fall under the most frequent words of a certain author, would no longer be detectable — even if this author, his work, or the like is mentioned explicitly in the environment.

[27] After applying the filter, pairs are kept where the split word material is separated by the same number of "soft" punctuation marks. In this case, there could be a longer quotation containing, for instance, a subordinate clause.

[28] With *et* (underlined), the pair contains, strictly speaking, another word shared by both texts (two instances in Virgil, three in Jerome). As a frequent word (and synsemantic), however, the conjunction does not factor into the evaluation of the quotation (see Step 6).

[29] The fact that the *shared words* do not belong together syntactically in Virgil either (*pluuia ingenti* is ablative; *sata laeta* is the direct object of *diluit*) is irrelevant with regard to the separation of the word material in the target text.

[30] The numbers refer to returned matches after stoplist 6 is applied. The filter is currently being tested at an early stage. It also can only be applied to potential new quotations that have exactly two *shared words* after application of the stoplists. Beyond an extension of the filter's general capabilities, there is potential for further optimization in two respects. Currently, only an unequal number of punctuation marks between the *shared words* is used as a decisive criterion (i.e., a match is rejected as soon as one of the two texts has more commas than the other). In addition, it would also be possible to set a maximum limit for the number of allowed "soft" punctuation marks between the *shared words* since the plausibility of the finding being an actual quotation decreases even with the same high number of intervening "soft" punctuation marks. Second, there is a need to optimize the filter also with respect to cases where one of the two *shared words* is contained more than once in a phrase. For example, when applying the filter to Verg. *ecl.* 10,52–54 (*certum est in siluis inter spelaea ferarum / malle pati tenerisque* ***meos*** *incidere* ***amores*** */ arboribus: crescent illae, crescetis,* ***amores*** *– Well I know that in the woods, amid wild beasts' dens, it is better to suffer and carve my love on the young trees. They will grow, and you, my love, will grow with them*; *shared words*: *meos*; *amores* [2x]) two text segments are examined for their "soft" punctuation marks: *incidere* between *meos* and the first instance of *amores* as well as *incidere amores / arboribus: crescent illae, crescetis,* between *meos* and the second instance of *amores*. This results in 0 or 2 commas and 0 or 1 colon. At the moment, this finding is retained as soon as the text of Jerome also contains either 0 or 2 commas and either 0 or 1 colon. Thus the find is only eliminated if both combinations do not match the text material of Jerome.

[31] On the development of the historical text-reuse grammar, [Revellio 2022, 146–152] as well as [Revellio 2022, 159–160] on the deduction and implementation of the *HTRG* filter.

[32] "Potentielle Funde des computergestützten Textvergleichs, deren übereinstimmendes Wortmaterial der Wortartenstruktur nach aus mindestens zwei Nomina oder zwei Verben sowie aus der Kombination dieser beiden Wortarten besteht, [sind] besonders prädestiniert dafür, eine sinnproduzierende Text-Text-Beziehung zu etablieren." On the discussion of this criterion, [Revellio 2022, 146–152].

[33] On the implementation of the filter, [Revellio 2022, 159–160].

[34] The complete list of *all* Virgilian quotations in Jerome will be the subject of a further publication.

[35] These include words such as *a*, *ab*, *ac*, *ad*, *ante*, *atque*, *autem*, *caelo*, *centum*, *contra*, *corpore*, *cui*, *cum*, *de*, *dies*, *dum*, *ea*, *enim*, *est*, *et*, *etiam*, *ex*, *frater*, *haec*, *hanc*, *hic*, *his*, *hoc*, *hominum*, *iam*, *illa*, *ille*, *in*, *inter*, *ipse*, *ita*, *manu*, *me*, *mihi*, *mundi*, *ne*, *nec*, *neque*, *nihil*, *non*, *nos*, *nunc*, *oculos*, *omnes*, *omnia*, *per*, *possumus*, *post*, *procul*, *quae*, *quam*, *quem*, *qui*, *quibus*, *quid*, *quidquid*, *quis*, *quo*, *quorum*, *sanguine*, *se*, *sed*, *si*, *sint*, *sit*, *siue*, *sua*, *sub*, *sunt*, *super*, *te*, *terra*, *uerbo*, *uix*, *unde*, *ut*.

[36] For the *Eclogues*, the following numbers result: Of 89 potential new quotations (stoplist 6 applied), 87 have exactly two *shared words*; of these, 46 share additional words (53%). Similarly in the case of *Georgica*: Of 167 potential citation pairs (stoplist 6 applied), 153 have exactly two *shared words*; of these, 122 share further material (80%). Finally, the following numbers are obtained for the *Aeneid*: Of 606 potential new quotations (stoplist 6 applied), 583 have exactly two *shared words*; of these, 414 share further words (71%).

[37] In relation to the list of potential new quotations after application of stoplist 6.

[38] The same problem arises in Pliny the Younger; in *ep.* 5,10 to Suetonius he uses the words *rumpe iam moras* with the goal of persuading his addressee to publish a literary work. Cf. Schwerdtner: "Da Plinius' *rumpe iam moras* nicht mit letzter Sicherheit einer einzigen Quelle zugewiesen werden kann, ist es problematisch, die Stelle als Vergilzitat bezeichnen zu wollen, auch wenn Plinius sonst bevorzugt aus Vergil zitiert und bereits in *ep.* 5,8 auf das Georgicaproömium zurückgreift" [Schwerdtner 2015, 250–255, esp. p. 254].

[39] Thus, the total number of identical word forms would amount to three; the quotation would then not be listed among those quotations of Virgil in Jerome with exactly two identical words. On this problem of the textual tradition, see 3. Optimization strategy.

[40] The same phenomenon is even more noticeable in another returned match consisting of Verg. *georg.* 1,401–403 and Hier. *epist.* 116,5,2. Again, there are the same two *shared words culmine* and *summo*. However, in addition to the semantically different prepositions (*de* vs. *in*) as well as the attribute genitive (*auctoritatis*) activating the figurative meaning, the expression is even more complex because of a second attribute (*caelesti*).

Verg. *georg.* 1,401–403	**Hier. *epist.* 116,5,2**
at nebulae magis ima petunt campoque recumbunt, / solis et occasum seruans de **culmine summo** / nequiquam seros exercet noctua cantus	immo uero sanctam scripturam in **summo** et caelesti auctoritatis **culmine** conlocatam de ueritate eius certus ac securus legam
But the mists are prone to seek the valleys, and rest on the plain, and the owl, as she watches the sunset from some high peak, vainly plies her evening song	but of course I should read the holy scriptures, set on the highest, heavenly summit of validity, firmly convinced of their truth

Table 7. In this example, *culmine* and *summo* are highlighted in both texts. Additionally, *de* is underlined in the text from Virgil and *in*, along with the phrase *et caelesti auctoritatis*, is also underlined in the text from Jerome.

[41] Virgil: preposition + propositional object + attribute adjective; Jerome: preposition + attribute adjective I [+ conjunction] + attribute adjective II + attribute genitive + propositional object.

[42] The status as an intentional quotation of the *Aeneid* cannot be questioned because of the references to *Uergilius* (Hier. *epist.* 126,2,2) and the *poeta eloquentissimus* (129,4,3).

[43] Another occurrence of the phrase within the works of Jerome can be found in *In Isaeam* 5,21,13–17. Hagendahl (1958, p. 230 n. 4) already points to the fact that *uagantes* is present in a number of late manuscripts of the *Aeneid*. Indeed, the critical apparatus of Conte's *Teubner* edition lists *uagantes* as another reading instead of *furentes*, which can, besides Jerome, also be traced back to two *Codices Bernenses* from the 9th and 10th century.

[44] Cf. the correspondence of *omnes* and *omnis* above in Verg. *Aen.* 9,13 = Hier. *epist.* 130,5,3 also see 2. Optimization strategy, n. 39.

[45] Variants are considered comprehensively, but more extensive editorial changes such as omissions, conjectures, etc. cannot be addressed.

[46] Cf., for example, Ov. *ars* 3,511 and Sen. *Thy*. 81.

Works Cited

Adkin 2011 Adkin, N. (2011) “Catullus in Jerome? Notes on the Cohortatoria de paenitentia ad Sabinianum (Epist. 147)”, *VChr*, 65 (4), pp. 108–424.

Bamman and Crane 2008 Bamman, D. and Crane, G. (2008) “The logic and discovery of textual allusion”, *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*. Available at: http://hdl.handle.net/10427/42685 (Accessed: 13 October 2022).

Büchler et al 2014 Büchler, M., Burns, P. R., Müller, M., Franzini, E. and Franzini, G. (2014). “Towards a Historical Text Re-use detection” in Biemann, C. and Mehler, A. (eds.) *Text mining. From ontology learning to automated text processing applications: Festschrift in honor of Gerhard Heyer*. Cham: Springer, pp. 221–238.

Burns 2017 Burns, P. J. (2017) “Measuring and mapping intergeneric allusion in Latin poetry using Tesserae”, *Journal of Data Mining and Digital Humanities*, pp. 1–15. Available at: https://jdmdh.episciences.org/3821 (Accessed: 13 October 2022).

Burzacchini 1975 Burzacchini, G. (1975) “Note sulla presenza di Persio in Girolamo”, *GIF*, 27, pp. 50–72.

Burzacchini 1978 Burzacchini, G. (1978) “Marginalia hieronymiana”, *BstudLat,* 8, pp. 270–272.

Cain 2013 Cain, A. (2013) “Two allusions to Terence, Eunuchus 579 in Jerome”, *CQ*, 63 (1), pp. 407–412.

Coffee et al 2013 Coffee, N., Koenig, J.-P., Poornima, S., Forstall, C. W., Ossewaarde, R. and Jacobson, S. (2013). “The Tesserae Project. Intertextual analysis of Latin poetry”, *Literary and Linguistic Computing*, 28 (2), pp. 221–228.

Coffee et al 2020 Coffee, N., Forstall, C., Galli Milić, L. and Nelis, D. (2020) *Intertextuality in Flavian Epic Poetry*. Berlin/Boston: Walter de Gruyter.

Diddams and Gawley 2017 Diddams, C. and Gawley, J. (2017) “Measuring the presence of Roman rhetoric. An intertextual analysis of Augustine's De Doctrina Christiana IV”, *Mouseion*, 14 (3), pp. 391–408.

Feichtinger 2021 Feichtinger, B. (2021) “*Quid facit cum psalterio Horatius?* (Hier. *ep.* 22,29,7). Untersuchung zu Hieronymus' Umgang mit klassischen und biblischen Referenzen am Beispiel von Epistula 3 ad Rufinum”, *VChr*, 75, pp. 389–454.

Godel 1964 Godel, R. (1964) “Réminiscences de poètes profanes dans les lettres de St-Jérôme”, *MH*, 21 (1), pp. 65–70.

Hagendahl 1958 Hagendahl, H. (1958) *Latin fathers and the classics. A study on the apologists, Jerome and other Christian writers*. Göteborg: Almquist & Wiksell.

Hohl Trillini and Quassdorf 2010 Hohl Trillini, R. and Quassdorf, S. (2010) “A 'key to all quotations'? A corpus-based

parameter model of intertextuality", *Literary and Linguistic Computing*, 25 (3), pp. 269–286.

Jakobi 2006 Jakobi, R. (2006) "Argumentieren mit Terenz. Die Praefatio der 'Hebraicae Quaestiones in Genesim'", *Hermes*, 134 (2), pp. 250–255.

Luebeck 1872 Luebeck, E. (1872) *Hieronymus quos nouerit scriptores et ex quibus hauserit*. Leipzig: Teubner.

Manca et al 2011 Manca, M., Spinazzè, L., Mastandrea, P., Tessarolo, L. and Boschetti, F. (2011) "Musisque Deoque: Text Retrieval on Critical Editions", *Journal for Language Technology and Computational Linguistic*, 26, pp. 129–140.

Mohr 2007 Mohr, A. (2007) "Jerome, Virgil, and the captive maiden. The attitude of Jerome to classical literature" in Scourfield, J. H. D. (ed.) *Texts and culture in late antiquity. Inheritance, authority, and change*. Swansea: Classical Press of Wales, pp. 299–322.

Revellio 2022 Revellio, M. (2022): *Zitate der* Aeneis *in den Briefen des Hieronymus. Eine digitale Intertextualitätsanalyse zur Untersuchung kultureller Transformationsprozesse*. Berlin/Boston: Walter de Gruyter. Available at: https://doi.org/10.1515/9783110760828 (Accessed: 13 October 2022).

Schubert and Heyer 2010 Schubert, C. and Heyer, G. (2010). "Neue Methoden der geisteswissenschaftlichen Forschung – Eine Einführung in das Portal eAQUA" in Schubert, C. und Heyer, G. (eds.) *Das Portal eAQUA – Neue Methoden in der geisteswissenschaftlichen Forschung I*, Leipzig: Universität Leipzig, pp. 4–9.

Schwerdtner 2015 Schwerdtner, K. (2015) *Plinius und seine Klassiker. Studien zur literarischen Zitation in den Pliniusbriefen*. Berlin/Boston: Walter de Gruyter.f

Voß 1969 Voß, B.R. (1969) "Vernachlässigte Zeugnisse klassischer Literatur bei Augustin und Hieronymus", *RhM*, 112 (2), pp. 154–166.

Zou et al 2006 Zou, F., Wang, F. L., Deng, X., Han, S. and Wang, L. S. (2006) "Automatic construction of Chinese stop word list", *Proceedings of the 5th WSEAS International Conference on Applied Computer Science*, pp. 1010–1015.