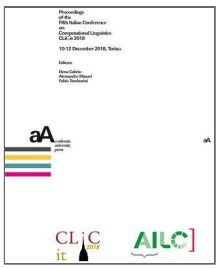


Elena Cabrio, Alessandro Mazzei and Fabio Tamburini (dir.)

Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018
10-12 December 2018, Torino

Accademia University Press

Using and evaluating TRACER for an *Index fontium computatus* of the *Summa contra Gentiles* of Thomas Aquinas

Greta Franzini, Marco Passarotti, Maria Moritz and Marco Büchler

DOI: 10.4000/books.aaccademia.3369
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2018
Published on OpenEdition Books: 8 April 2019
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788831978682

Using and evaluating TRACER for an *Index fontium computatus* of the *Summa contra Gentiles* of Thomas Aquinas

Greta Franzini, Marco Passarotti, Maria Moritz and Marco Büchler

1 Introduction

1 Thomas Aquinas (1225-1274) was a prolific medieval author from Italy : his 118 works, known as the *Corpus Thomisticum*, amount to 8,767,883 words (Portalupi, 1994, p. 583) and discuss a variety of topics, ranging from metaphysical to legal, political and moral theory (Kretzmann and Stump, 1993). The web of references to biblical, ecclesiastical and classical literature that stretches the whole *Corpus Thomisticum* speaks to daunting erudition. In the late 1940s, Humanities Computing pioneer Father Roberto Busa (1913-2011) spearheaded a scholarly effort, known as *the Index Thomisticus*, to manually annotate reuse, both *explicit* (i.e., explicitly introduced by Aquinas as a quote) and *implicit* (i.e., reference to works without quotation), in the texts of Thomas Aquinas (Busa, 1980). Four decades later, Portalupi noted:

> Ancora più difficile sarà [...] il tentativo di confrontare automaticamente tutto Tommaso con tutti i testi di uno o più autori, per rintracciare in modo globale la presenza implicita di una fonte. Per fare questo occorrerebbe che si verificassero due condizioni : in primo luogo, gli autori di cui si studiano le presenze implicite in Tommaso dovrebbero essere informatizzati e interrogabili nella totalità delle loro opere ; in secondo luogo, bisognerebbe disporre di un software molto potente e raffinato. (Portalupi, 1994, p. 583)[1]

2 Today, a once visionary task is conceivable, giving way to studies such as the present, which poses the following research question : to which extent can historical text reuse detection (HTRD) software detect explicit and implicit text reuse in the writings of Thomas Aquinas? To this end, we test the performance of TRACER, a text reuse detection

framework, for the creation of an *Index fontium computatus* (a computed index of text reuse). The *Summa contra Gentiles* (ScG) was chosen as a case study because the critical edition used for the *Index Thomisticus*, the 1961 Marietti *Editio Leonina* (Gauthier et al., 1882), is still in use today and because an ongoing treebanking effort of the text will, in future, provide us with the linguistic data needed to further refine the experiments described here (Passarotti, 2011).

2 Related Work

2.1 The significance of text reuse

3 Text reuse (TR) can be summarily described as the written repetition or borrowing of text and can take different forms. Büchler et al. (2014) separate *syntactic* TR, such as (near-)verbatim quotations or idiomatic expressions, from *semantic* TR, which can manifest itself as a paraphrase, an allusion or other loose reproduction. The study of quotation is key to any philological examination of a text, as it is not only indicative of the intellectual and cultural endowment of an author, but may shed light on the sources used, the relation between works and literary influence. Crucially, quotations may also preserve text that is now lost, thus facilitating efforts of textual reconstruction.[2] Owing to the magnitude of the task, the publication of a work's complete index of references, conventionally known as *Apparatus fontium* or *Index scriptorum*, is rare (Portalupi, 1994, p. 582).

2.2 Text reuse in Thomas Aquinas

4 Like many of his Christian predecessors, Aquinas' body of work teems with references to secular and Christian literature alike. In the ScG (1259-1265) Aquinas cites 170 works both explicitly and implicitly (Gauthier et al., 1882, Vols. IV-XV). Explicit quotations provide information about the source text and the author and/or work, and can either be direct or indirect (Gauthier et al., 1882, vol. XVI, pp. XVI-XXII). Implicit reuses, in the ScG and in general, are more elusive, as they are almost never syntactically nor lexically-faithful to the original text, thus making them hard for both machines and humans to spot (Portalupi, 1994, p. 582).[3] Durantel notes that Aquinas' tendency in TR is to borrow only what is necessary to fit the flow of his narrative without significant semantic or syntactic deviation from the original (Durantel, 1919, p. 63). And yet, Pelster's observation on Aquinas' paraphrastic reuse of Aristotle might suggest greater deviation (Pelster, 1935, p. 331).[4]

5 Roberto Busa's effort in the late 1940s resulted in the creation of the *Index Thomisticus*, a manually-lemmatised version of Thomas Aquinas' *opera omnia* (Jones, 2016). Among the annotations, the *Index Thomisticus* tags tokens forming explicit quotations as QL if literal (*ad litteram*) and QS if a paraphrase (*ad sensum*), and tokens forming *implicit* quotations as QR to indicate a reference or citation alluding to another text. An example quotation in the ScG containing a mixed annotation is :

> [...] ratio(QL) vero (QL) significata(QL) per(QL) nomen(QL) est(QL) definitio(QL) secundum(QR) philosophum(QR) in(QR) IV(QR) Metaph.(QR)[5]

6 The (QL) portion of this example contains the literal quote, while the second (QR) portion provides the reference.

2.3 Historical text reuse detection

7 HTRD is a Natural Language Processing (NLP) task aimed at identifying syntactic and semantic TR in historical sources. The computational analysis of historical languages is particularly challenging as tools at our disposal are often trained on a synchronic rather than diachronic state of a language[6] and on controlled textual corpora. Eger et al. (2015) and Passarotti (2010) tested the performance of seven different taggers, including TreeTagger (Schmid, 1994), for different training sets and tag-sets of medieval (church) Latin texts showing accuracies tightly below 96% and 96.75% for PoS-tagging, and around 90% and 89.90% for morphological analysis, respectively. These results have yet to be generalised to other variants of Latin and can be improved upon with the provision of additional training corpora, treebanked and semantically-tagged, the creation of corpora containing intertexts, or with the expansion of lexical resources, such as the *Latin WordNet* (Minozzi, 2017, p. 130).

8 The extent to which the limitations of these resources and taggers (e.g., correct resolution of homographs) affect HTRD tools, including *Tesserae* (Coffee et al., 2013), *Passim* (Smith et al., 2017)[7] and *TRACER* (Büchler, 2013) is not yet fully understood. Reasons for this are the field's lack of progress caused by "inconsistent standards and the scattering of insights across publications" (Coffee, 2018), the general failure of HTRD studies to publish negative results, and the quasi-absence of gold standards for testing. To our knowledge, the only projects to have published computed results from intertextual studies on historical sources are the *Proteus Project* (English and Latin) (Yalniz et al., 2011), the *Chinese Text Project* (early Chinese) (Sturgeon, 2017), *Commonplace Cultures* (English and Latin) (Gladstone and Cooney, forthcoming), *SHEBANQ* (Hebrew) (Naaijer and Roorda, 2016), *Samtla* (Search and Mining Tools for Language Archives) (language-independent) (Harris et al., 2018), and *Tesserae* (Latin), but of these only the latter discloses tool configurations.

3 Methodology

3.1 Gold Standard

9 To facilitate the classification of automatically-detected reuse, all QL-, QS- and QR-annotated tokens were extracted from the *Index Thomisticus*. Of the total 24,416 sentences constituting the ScG, the 7,396 (30.29%) containing any combination of QL, QS and QR were stored in a tabular file, which we define as the *Index Thomisticus Gold Standard* of TR (hereafter IT-GS). The number of sentences containing only QL tokens (1,139) compared to that of sentences containing only QS tokens (2,270) corroborates expert assertions about Aquinas' paraphrastic style of TR.

3.2 Text acquisition and preparation

10 For the sake of processing efficiency, out of the ScG's 170 source works we began with a set of five readily available texts. These are *Philosophiae Consolationis* and *De Trinitate* of Boethius, *De Deo Socratis* of Apuleius, Cicero's *De Divinatione* and the Moerbeke Latin translation of Aristotle's *Metaphysica*. The texts were acquired from different sources and

cleaned of all paratextual information. The clean texts were then segmentised by sentence, PoS-tagged and lemmatised with the TreeTagger Brandolini parameter file (with an average accuracy of 93.72%), whose tag-set provides the degree of granularity needed in this experiment.[8] Finally, a script was used to format sentences to TRACER requirements.

3.3 Text reuse detection with TRACER

11 The HTRD on this corpus was performed (server-side) with TRACER, a language-agnostic framework comprising hundreds of information retrieval (IR) algorithms designed to work with historical and modern languages alike.[9] TRACER is a Java command-line tool driven by an XML configuration file, which users can modify to fit their detection needs. TRACER follows a six-step architecture,[10] which demystifies the detection process by storing the computed output of each step on the disk so that users can more easily follow and locate errors in the processing chain, if any. TRACER is resilient to OCR-noise and capable of detecting both (near-)verbatim quotations and looser forms of TR. The detection of paraphrase requires the use of linguistic resources to help TRACER match a word against its synsets and an inflected form against its base-form. For synonym detection, we extracted synonymous relations from the Latin WordNet. TR identified with TRACER was manually compared against the IT-GS to separate the True (TP) from the False Positives (FP), and to identify False Negatives (FN).

4 Results

4.1 Philosophiae Consolationis

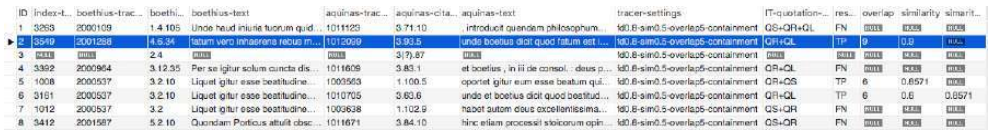
12 To detect both verbatim quotations and paraphrase, TRACER was optimised for recall over precision and configured to work with single words as features, to ignore the top 20% most frequent words,[11] to link text pairs with a minimum overlap of 5 features,[12] to expand the query to synonyms, and to return only those aligned text pairs presenting an overall sentence similarity of at least 50%.[13] Of the eight reuses indicated in the *Editio Leonina*, we were unable to precisely locate one as it alludes to four paragraphs of text ;[14] of the remaining seven, as shown in Figure 1, TRACER identified three (42%). Upon close inspection, two FNs were affected by the 20% threshold of feature removal, for example :

Boethius 1.4.105 Unde haud iniuria tuorum quidam familiarium **quaesivit** : “**Si** quidem **deus**", inquit, “**est**, **unde** **mala**?[15]
Aquinas 3.71.10 , introducit quendam philosophum **quaerentem** : **si** **deus** **est**, **unde** **malum**?[16]

13 Here, the tokens *si*, *est* and *unde* were ignored as they fell within the pool of the 20% most frequent words removed.

14 One reuse was successfully identified on the basis of feature overlap but did not amount to a 50% sentence similarity ; and the fourth reuse could not be identified because of a missing synonymous relation in the Latin WordNet (i.e., *gaudium-beatitudo*)[17] and its insufficient feature overlap. The resulting F1-score is $4{,}6 \bullet 10^{-3}$.

FIGURE. 1: For every TRACER analysis, a MySQL table is created to store and manually-evaluate the results against the IT-GS. The evaluation table for *Philosophiae Consolationis* illustrated here contains a wealth of information, including full citation information for both works, the TRACER settings used for the detection task, the *Index Thomisticus* quotation annotations, the result classification (into True Positive and False Negative), as well as the feature overlap and the overall similarity value of the aligned sentences. The reuse in the highlighted row, for instance, was correctly identified by TRACER on the basis of a 9-word overlap and an overall sentence similarity of 90



ID	index-t...	boethius-trac...	boethi...	boethius-text	aquinas-trac...	aquinas-cita...	aquinas-text	tracer-settings	IT-quotation-...	res...	overlap	similarity	similarit...
1	3283	2000109	1.4.105	Unde haud iniuria tuorum quid...	1011129	3.71.10	...introducit quendam philosophum...	Id0.8-sim0.5-overlap5-containment	QS+QR+QL	FN	NULL	NULL	NULL
2	3549	2001288	4.5.34	fatum vero immerens recus m...	1012069	3.93.5	unde boetius dicit quod fatum est ...	Id0.8-sim0.5-overlap5-containment	QR+QL	TP	9	0.9	NULL
3	NULL	NULL	2.4	NULL	NULL	3 (?).87	NULL	Id0.8-sim0.5-overlap5-containment	NULL	NULL	NULL	NULL	NULL
4	3392	2000964	3.12.35	Per se igitur solum cuncta dis...	1011609	3.83.1	et boetius , in iii de consol. : deus p...	Id0.8-sim0.5-overlap5-containment	QR+QL	FN	NULL	NULL	NULL
5	1008	2000537	3.2.10	Liquet igitur esse beatitudine...	1003563	1.100.5	oportet igitur eum esse beatum qui...	Id0.8-sim0.5-overlap5-containment	QR+QS	TP	6	0.6571	NULL
6	3161	2000537	3.2.10	Liquet igitur esse beatitudine...	1010705	3.63.6	unde et boetius dicit quod beatitud...	Id0.8-sim0.5-overlap5-containment	QR+QL	TP	6	0.6	0.6571
7	1012	2000537	3.2	Liquet igitur esse beatitudine...	1003638	1.102.9	habet autem deus excellentissima...	Id0.8-sim0.5-overlap5-containment	QS+QR	FN	NULL	NULL	NULL
8	3412	2001587	5.2.10	Quondam Porticus attulit obsc...	1011671	3.84.10	hinc etiam processit stoicorum opin...	Id0.8-sim0.5-overlap5-containment	QS+QR	FN	NULL	NULL	NULL

4.2 De Trinitate

15 Given the results of the previous analysis, for this second investigation the feature removal and the sentence similarity values were lowered to 10% and 40% respectively, thus optimising for even higher recall (10,349 total sentences aligned). Of the four known reuses, TRACER identified three. The 40% similarity threshold was essential to the identification of one reuse (where the score is 0.4375) ; the FN, which was indeed found on the basis of an eight-word overlap but did not meet the minimum sentence similarity threshold, revealed another missing synonymous relation in the WordNet (i.e., *disciplinatus-eruditus*)[18] and a failed alignment of the variants *temptare* (Boethius) and *tentare* (Aquinas) owing to inconsistent TreeTagger lemmatisation (*tempto* and *tento*, respectively). The F1-score for this analysis was .

4.3 De Deo Socratis

16 This work of Apuleius is quoted twice in the ScG. Of the two reuses, TRACER was able to detect one in full and only parts of the second. The second reuse spans three sentences and is mostly paraphrastic, with only three words annotated in the *Index Thomisticus* as QL (*sunt animo passiva*).[19] To capture the fullest range of reuse diversity, TRACER's feature removal was set to 10%, the overlap to 3 and the overall similarity to 20%. However, as *sunt* (form of the verb *sum* 'to be') is the most frequent word across the texts, TRACER's inbuilt feature removal prevented the detection of the short QL portion of the reuse ; the QR+QS portions, on the other hand, were successfully detected. We counted both results as TPs, resulting in an F1-score of $2{,}6 \bullet 10^{-5}$.

4.4 De Divinatione

17 The only recorded reuse that Aquinas makes of Cicero's text is implicit and alludes to a block of text, making it difficult to manually pinpoint with precision. To detect as loose a similarity as possible, the TRACER search was cast with the same configuration used in the previous analysis. No reuse, however, was found.

4.5 Metaphysica

18 The *Editio Leonina* lists 97 reuses of Aristotle's *Metaphysica*. As previously mentioned, Pelster describes Aquinas' reuse of the Latin translation of the *Metaphysica* as more

paraphrastic than literal. Our manual examination of the texts and the results of TRACER confirmed this observation, in that we could not manually locate seven reuses (due to their strong allusiveness) and a fault-tolerant TRACER configuration (removal of the top 10% most frequent words, overlap of 3 features and an overall sentence similarity of 40%) yielded 19 TPs only (6 out of 15 QL^{20} and 13 out of 75 QR+QS). The F1-score resulting from this analysis is $3{,}8 \bullet 10^{-4}$.

5 Discussion

19 Our results show that the FNs emerging from the computational analyses were largely caused by Aquinas' paraphrastic and allusive TR style, which at times challenged our own ability to spot similarities, even with the help of the critical edition. The allusions that we could identify generally retain the semantics of the alluded-to texts, thus confirming Durantel's insights. While a number of these negative results were also directly tied to *lacunae* in the Latin WordNet and to inconsistent lemmatisation, the flexibility and methodological transparency of TRACER allowed us to locate error sources and accordingly tune configurations to work around these issues (e.g., by increasing the feature overlap and/or lowering the sentence similarity scoring thresholds). Notwithstanding, TRACER's panlingual feature removal parameter affected the retrieval of shorter instances of reuse, particularly those containing forms of the highly frequent verb *sum*.

20 The manual evaluation of TRACER results against the IT-GS for the creation of an *Index fontium computatus* was time-consuming, not least because of a number of reference inaccuracies in the critical edition itself (in one case, the reference is off by ten lines). Nevertheless, the creation of the index is proving essential to the assessment of TRACER's fitness for purpose on Latin texts.

21 As far as the usability of the tool is concerned, TRACER's detection power is offset by its cumbersome setup, which is unfriendly to those who are not familiar with the command line, NLP basics and/or Java (stack traces). This issue is being addressed with the development of a user manual (Franzini et al., 2018).

6 Conclusion

22 This article describes a computational text reuse study on Latin texts designed to evaluate the performance of TRACER, a language-agnostic IR text reuse detection engine. The results obtained were manually evaluated against a gold standard and are contributing to the creation of an *Index fontium computatus* to both assess TRACER's efficacy and to provide a test-bed against which analogous IR systems can be measured and thus compared to TRACER. Our study shows that despite the known limitations of existing linguistic resources for Latin, the diverse spectrum of paraphrastic reuse encountered and its own language-agnosticism, TRACER is equipped to detect a wide range of explicit text reuse in the ScG, be that short or long, verbatim or paraphrastic, and implicit reuse only if coupled with explicit. To increase the detection accuracy, we are implementing a black/white list to give users the power to control words or multi-word expressions to be ignored or retained in the detection ; furthermore, we plan on re-running these analyses with the disambiguated linguistic annotation currently being

added to the text of the ScG (Passarotti, 2015) to measure its impact on this particular IR task.

23 The data used and generated in the current study is available from : https://github.com/CIRCSE/text-reuse-aquinas.

Acknowledgments

24 The authors would like to thank Eleonora Litta for proofreading this article and the anonymous reviewers for their valuable comments. This research was funded by the German Federal Ministry of Education and Research (No. 01UG1409).

BIBLIOGRAPHY

David Bamman and Gregory Crane. 2008. The Logic and Discovery of Textual Allusion. In *Proceedings of the ACL Workshop LaTeCH - Language Technology for Cultural Heritage Data*. ACL. http://hdl.handle.net/10427/42685.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, SEQUENCES '97, pages 21–29, Washington, DC, USA. IEEE Computer Society. http://dl.acm.org/citation.cfm?id=829502.830043.

Marco Büchler, Philip R. Burns, Martin Müller, Emily Franzini, and Greta Franzini. 2014. Towards a Historical Text Re-use Detection. In Chris Biemann and Alexander Mehler, editors, *Text Mining*, pages 221–238. Springer International Publishing, Cham. http://link.springer.com/10.1007/978-3-319-12655-5_11.

Marco Büchler. 2013. Informationstechnische Aspekte des Historical Text Re-use. PhD Thesis. http://www.qucosa.de/fileadmin/data/qucosa/documents/10851/Dissertation.pdf.

Roberto Busa. 1980. The annals of humanities computing : The Index Thomisticus. *Computers and the Humanities*, 14(2) :83–90, October. http://www.jstor.org/stable/30207304.

Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde, and Sarah L. Jacobson. 2013. The Tesserae Project : intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, 28 :221–228. https://doi.org/10.1093/llc/fqs033.

Neil Coffee. 2018. An Agenda for the Study of Intertextuality. *Transactions of the American Philological Association*, 148 :205–223. https://muse.jhu.edu/article/693654.

Gregory Crane. 1991. Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, page 243–245. https://doi.org/10.1093/llc/6.4.243.

Jean Durantel. 1919. *Saint Thomas et le Pseudo-Denis*. Librairie Félix Alcan, Paris. http://archive.org/details/cuasaintthomaset00dura.

Steffen Eger, Tim vor der Brück, and Alexander Mehler. 2015. Lexicon-assisted tagging and lemmatization in Latin : A comparison of six taggers and two lemmatization methods. In *In*

Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 105–113. http://www.aclweb.org/anthology/W15-3716.

Don Fowler. 1997. On the Shoulders of Giants : Intertextuality and Classical Studies. *Materiali e discussioni per l'analisi dei testi classici*, 39 :13–34. http://www.jstor.org/stable/40236104.

Greta Franzini, Emily Franzini, Kirill Bulert, Marco Büchler, and Maria Moritz. 2018. TRACER : A User Manual. https://tracer.gitbook.io/-manual/.

R. A. Gauthier, L. J. Bataillon, A. Oliva, T. de Vio Cajetan, Commissio Leonina, and Dominicans. 1882. *Sancti Thomae Aquinatis Doctoris Angelici Opera Omnia iussu edita Leonis XIII P.M.* Ex Typographia Polyglotta S.C. de Propaganda Fide, Rome.

Clovis Gladstone and Charles Cooney. forthcoming. Opening New Paths for Scholarship : Algorithms to Track Text Reuse in ECCO. *Digitizing Enlightenment*.

Martyn Harris, Mark Levene, Dell Zhang, and Dan Levene. 2018. Finding Parallel Passages in Cultural Heritage Archives. *Journal on Computing and Cultural Heritage*, 11(3) :15 :1–15 :24. http://doi.acm.org/10.1145/3195727.

Francis John Haverfield. 1916. Tacitus during the Late Roman Period and the Middle Ages. *The Journal of Roman Studies*, 6 :196–201. https://doi.org/10.2307/296272.

Richard D. Janda and Brian D. Joseph. 2005. On Language, Change, and Language Change – Or, Of History, Linguistics, and Historical Linguistics. In Brian D. Joseph and Richard D. Janda, editors, *The Handbook of Historical Linguistics*, pages 3–181. Wiley-Blackwell, Oxford.

Steven E. Jones. 2016. *Roberto Busa, S. J., and the Emergence of Humanities Computing : The Priest and the Punched Cards*. Routledge, March.

Christopher P. Jones, editor. 2017. *Apuleius. Apologia. Florida. De Deo Socratis*, volume 534 of *Loeb Classical Library*. Harvard University Press, Loeb Classical Library.

Norman Kretzmann and Eleonore Stump, editors. 1993. *The Cambridge Companion to Aquinas*. Cambridge University Press, Cambridge ; New York, May.

Stefano Minozzi. 2017. Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'Information Retrieval. In Paolo Mastandrea, editor, *Strumenti digitali e collaborativi per le Scienze dell'Antichità*, number 14 in Antichistica, pages 123–134. http://doi.org/10.14277/6969-182-9/ANT-14-10.

Martijn Naaijer and Dirk Roorda. 2016. Parallel Texts in the Hebrew Bible, New Methods and Visualizations. *CoRR*, abs/1603.01541. http://arxiv.org/abs/1603.01541.

Marco Passarotti. 2010. Leaving behind the less-resourced status. The case of Latin through the experience of the Index Thomisticus Treebank. In *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages LREC 2010, Valletta, Malta, 23 May 2010*.

Marco Passarotti. 2011. Language Resources. The State of the Art of Latin and the Index Thomisticus Treebank Project. In Marie-Sol Ortola, editor, *Corpus anciens et Bases de données, frenchALIENTO. Échanges sapientiels en Méditerranée*, volume 2, pages 301–320. Presses universitaires de Nancy, Nancy.

Marco Passarotti. 2015. What you can do with linguistically annotated data. From the Index Thomisticus to the Index Thomisticus Treebank. In Vijgen Roszak Piotr, editor, *Reading Sacred Scripture with Thomas Aquinas. Hermeneutical Tools, Theological Questions and New Perspectives*, pages 3–44. Brepols.

F. Pelster. 1935. Die Uebersetzungen der aristotelischen Metaphysik in den Werken des hl. Thomas von Aquin : Ein Beitrag. *Gregorianum*, 16(3) :325–348. http://www.jstor.org/stable/23567607.

Enzo Portalupi. 1994. L'uso dell'"Index Thomisticus" nello studio delle fonti di Tommaso d'Aquino : Considerazioni generali e questioni di metodo. *Rivista di Filosofia Neo-Scolastica*, 86(3) :573–585. http://www.jstor.org/stable/43062344.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

David A. Smith, Ryan Cordell, and Abby Mullen. 2015. Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History*, 27(3) :E1–E15. http://dx.doi.org/10.1093/alh/ajv029.

Donald Sturgeon. 2017. Unsupervised identification of text reuse in early Chinese literature. *Digital Scholarship in the Humanities*. https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqx024/4583485.

James Turner. 2014. *Philology : The Forgotten Origins of the Modern Humanities*. Princeton University Press, Princeton and Oxford.

Ismet Zeki Yalniz, Ethem F. Can, and R. Manmatha. 2011. Partial Duplicate Detection for Large Book Collections. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 469–474. http://doi.acm.org/10.1145/2063576.2063647.

NOTES

1. Our English translation reads : 'It will be even harder to automatically compare all of Thomas against all of the texts of one or multiple authors to check for the presence of implicit sources. Such a task would only be possible under two conditions : firstly, the texts of the authors quoted by Thomas would have to be digitised and searchable in their entirety ; secondly, one would need very powerful and sophisticated software'.

2. One notable example is the fragmentary survival of Alexandrian scholarship at the hands of Roman philologists (who wrote commentaries known as *scholia*) and grammarians (Turner, 2014, p. 16).

3. For problems with implicit quotations, see (Haverfield, 1916, p. 197) and (Fowler, 1997, p. 15). For automatic allusion detection, see (Bamman and Crane, 2008).

4. "Da Thomas die Schriften des Aristoteles [...] gewöhnlich nur dem Gedanken nach, nicht wörtlich anführt." In English : 'Since Thomas usually quotes paraphrastically, not literally.'

5. Book 1, chap. 12, n. 4. Our English translation reads : '[...] according to the philosopher in Metaph. IV, the meaning of a name is its definition'.

6. See joseph2005 for the dichotomy.

7. https://github.com/dasmiq/passim

8. The Brandolini tag-set was manually mapped against that of Morpheus (Crane, 1991), which TRACER uses as a reference. Ambiguously-lemmatised word forms were not disambiguated.

9. https://doi.org/21.11101/0000-0007-C9CA-3

10. The six steps are : *Preprocessing*, *Featuring*, *Selection*, *Linking*, *Scoring* and *Postprocessing*.

11. The parameter, known as *feature density*, is a language-independent measure used to decontaminate the texts and to contain the number of results based on chance repetition; an 80%

feature density means that TRACER ignores or removes the most frequent types that cover 20% of the tokens.

12. For a 24k sentence corpus such as this, an overlap of 5 is statistically significant (Büchler, 2013, p. 134).

13. The value was chosen on the basis of previous experiments as a good trade-off between precision and recall. The similarity measure used is Broder's *containment*, which is particularly suited to documents or sentences of uneven length (Broder, 1997).

14. This reuse would have doubtless been overlooked by TRACER too owing to the absence of features to compare.

15. Our English translation reads : 'It is not wrong that a certain acquaintance of yours has questioned : 'If in fact God exists,' he asks, 'where is evil from?"

16. Our English translation reads : '(Boethius) introduces a certain philosopher who asks : 'If God exists, where is evil from?'.'

17. Incidentally, this relation is also not mapped in BabelNet (bn :00042905n) nor in ConceptNet (http://conceptnet.io/c/la/gaudium) (as of 8 June 2018).

18. Also not present in neither BabelNet nor ConceptNet.

19. *[daemones] [...] sunt animo passiva* or 'demons are emotional in mind' (Jones, 2017, pp. 372-373).

20. The QL quotations in the ScG seem to refer to a different Latin translation than that available to us, which would explain why some instances of QL went undetected.

ABSTRACTS

This article describes a computational text reuse study on Latin texts designed to evaluate the performance of TRACER, a language-agnostic text reuse detection engine. As a case study, we use the *Index Thomisticus* as a gold standard to measure the performance of the tool in identifying text reuse between Thomas Aquinas' *Summa contra Gentiles* and his sources.

Questo articolo descrive un'analisi computazionale effettuata su testi latini volta a valutare le prestazioni di TRACER, uno strumento "language-agnostic" per l'identificazione automatica del riuso testuale. Il caso studio scelto a tale scopo si avvale dell'*Index Thomisticus* quale gold standard per verificare l'efficacia di TRACER nel recupero di citazioni delle fonti della *Summa contra Gentiles* di Tommaso d'Aquino.

AUTHORS

GRETA FRANZINI

Università Cattolica del Sacro Cuore – greta.franzini[at]unicatt.it

MARCO PASSAROTTI

Università Cattolica del Sacro Cuore – marco.passarotti[at]unicatt.it

MARIA MORITZ

Georg-August-Universität Göttingen – mmoritz[at]etrap.eu

MARCO BÜCHLER

Georg-August-Universität Göttingen – mbuechle[at]etrap.eu