Martina Paccara, Francesco Stella

Prime analisi computazionali sulla lingua dei testi latini di *SERICA*

1 Introduzione: testometria del latino medievale e moderno e progetti del Centro di Studi Comparati “I Deug-Su”

Francesco Stella

Le applicazioni di linguistica computazionale alle ricerche sullo stile e sul linguaggio letterario hanno conosciuto uno sviluppo impetuoso negli ultimi anni e danno vita ormai a corsi specifici nelle università come quello appena nato a Torino (2023), a riviste specializzate,[1] a un alto numero di post in liste di comunicazione elettronica come *Humanist*. Gli studi sono giunti alla determinazione di procedure statistiche relativamente stabili per le lingue moderne. Le applicazioni di analisi informatica del linguaggio su lingue antiche, benché l’Informatica Umanistica sia nata negli anni ‘40 del secolo scorso proprio con i lavori di padre Roberto Busa sull’*Index Thomisticus*, sono invece rimaste a lungo più rare e irregolari, al di fuori di esperienze isolate o degli esperimenti effettuati a Liegi dal LASLA (*Laboratoire d’Analyse Statistique des Langues Anciennes*), fondato nel 1961: questo ritardo è dovuto probabilmente al tasso di conoscenze tecniche richieste dal tipo di ricerca e all’influenza di un pregiudizio pseudo-umanistico contrario alla stilistica quantitativa. In Italia si segnalano in questo campo soprattutto gli esperimenti di Maurizio Lana[2] e Guido Milanese[3] e le applicazioni più sistematiche di Marco Passarotti prima al CNR di Pisa (con il lemmatizzatore *LemLat*)[4] e poi al CIRCSE di Milano, che ha vinto e gestito negli ultimi tre anni un progetto ERC sulla sintassi latina denominato *Linking Latin*.[5]

1 Fra le tante: *International Journal of Corpus Linguistics, ICAME-Journal, Literary and Linguistic Computing, Natural Language Engineering, Computer and the Humanities, Revue Informatique, Statistique dans les Sciences Humaines, Literary and Linguistic Computing* (poi ridenominata *Digital Scholarship in the Humanities*), *Glottometrics* etc.

2 Elenco completo in https://upobook.uniupo.it/maurizio.lana. Segnaliamo in particolare Lana (2004) e Lana (2011).

3 Dopo alcuni contributi pioneristici si veda ora la sintesi in Milanese (2020).

4 Passarotti (2007).

5 Sito web https://lila-erc.eu/.

https://doi.org/10.1515/9783111718583-016

Al convegno di linguistica latina di Lione del 2009 potevo citare,[6] fra i pochi precedenti in questo campo, solo il lavoro di Xuan Luong e Sylvie Mellet sul corpus degli storici latini[7] che analizzava come l'uso di determinate caratteristiche sintattiche si distribuisce nel *corpus Caesarianum* e negli altri storiografi latini: tempi e modi verbali, parti del discorso e uso dei casi separano nettamente Cesare e i suoi epigoni da Sallustio, Tacito e Curzio Rufo, individuando sottodistinzioni, quali, all'interno del corpo cesariano, quella fra il *Bellum Hispaniense* e le altre opere. E in quegli anni io stesso mi sono cimentato in analisi a fini attribuzionistici sulle *Epistolae duorum amantium*, un epistolario d'amore anonimo del XII o XIII secolo scoperto in un manoscritto di Troyes nel 1975 e attribuito ad Abelardo ed Eloisa, su cui ero stato invitato a pronunciarmi da uno dei team che in varie parti del mondo si impegnava nell'impresa di individuare l'autore di questa raccolta stupefacente.[8] Da lì, con due allievi, ho sviluppato un tentativo di analisi del linguaggio del genere epistolografico latino da Cicerone a Petrarca, individuando costanti e variabili nelle *function words* e in alcuni semantemi che credo aiutino a comprendere l'evoluzione di questa lingua così longeva e condivisa nel tempo e nello spazio.[9]

Altre ricerche linguistiche con finalità critico-letterarie si trovano edite nel volume Brepols del 2020.[10]

Negli ultimi anni il campo ha registrato qualche fruttuoso movimento: singoli studi su testi attribuiti a Gregorio Magno o sugli strati compositivi dell'epistolario di Ildegarde di Bingen hanno coinvolto in via occasionale ricercatori italiani

6 *Latin vulgaire – Latin tardif. IX^e^ colloque international sur le latin vulgaire et tardif* (Lyon, 6-9 septembre 2009). Atti in Biville et al. (2012).

7 Luong/Mellet (2003). L'abstract in inglese recita: "The calculation of intertextual distance is generally performed by studying lexical parameters. We will in the first place examine whether it is possible or not to apply one of the available methods to grammatical parameters, then we explain our own method, based on an ordinal classification table rather than a multiple contingency table. To present this methodology, we use Latin texts extracted from a lemmatized and tagged corpus. The different results will be compared and evaluated". Le basi più solide del metodo restano i lavori seminali di Sinclair (soprattutto Sinclair [1991], Sinclair [2003] e Sinclair/Carter [2004]) e l'esplorazione effettuata da Douglas Biber in Biber (1991), in cui si elencano le frequenze di alcune decine di dati, dalla *Type-Token ratio*, alla presenza di determinati tempi verbali, o congiunzioni, o proposizioni divise per alcune decine di generi più specificamente riconoscibili. Fra gli studi più recenti, McEnery *et al.* (2006) propone una griglia estremamente dettagliata di operazioni statistiche su testi raccolti in *corpora*, ma anche in questa raccolta il metodo adottato per l'analisi diacronica e contrastiva si basa sempre su Biber (1991) e Bybee/Hopper (2001).

8 Stella (2008).

9 Stella (2014)

10 Stella (2020). Sintesi aggiornata alle metodologie più recenti in Tuzzi (2024). Utile anche Nini (2023).

come Edoardo D'Angelo[11] e Luigi Ricci[12], o belgi come Jeroen Deploige, Mike Kestemont e Sara Moens[13] e da poco anche Jeroen De Gussem.[14] Alla Sapienza di Roma un team premiato dal PRIN e diretto da Andrea Cucchiarelli ha lavorato alla definizione dello stile linguistico dell'elegia e a ricerche attribuzionistiche su Sulpicia. Ma i contributi più significativi, destinati a fungere da riferimento, sono giunti da Barbara McGillivray nel 2009 con la sua tesi di dottorato pisana *Computational Approach to Latin Verbs: New Resources and Methods*, poi sviluppata e pubblicata nel 2014 come *Methods in Latin Computational Linguistics*,[15] che all'interno di una raffinata e meticolosa indagine sui preverbi latini dedica 10 pagine al panorama generale della linguistica computazionale latina e alle indagini sul latino con metodi NLP, inevitabilmente considerati da un angolo visuale strettissimo. Due anni fa Philipp Roelli, il matematico mediolatinista di Zurigo cui dobbiamo la più grande biblioteca digitale latina, il *Corpus Corporum*,[16] ha pubblicato *Latin as the Language of Science and Learning*,[17] storia del latino scientifico che dominò l'Europa moderna, che a sua volta sceglie un altro spicchio molto specifico di applicazione, cioè le statistiche comparative di POS, Parts of Speech, che a nostro avviso produce risultati molto limitati in un campo, quello del vocabolario scientifico, in cui è ovviamente il lessico l'elemento distintivo. Questa scelta dimostra come anche in aree così tecniche spesso sono le mode e i modelli metodologici disponibili a imporre scelte che rischiano poi di indirizzare gli studi su strade che non sempre si rivelano produttive per tutti i tipi di ricerca: al CNR di Pisa McGillivray ha appreso soprattutto il metodo di lavoro sulla morfologia, e su quello ha lavorato, concentrandosi su testi iperclassici; Roelli invece ha focalizzato la sua attenzione su tutta la storia del latino utilizzando i metodi sul POS sviluppati nella linguistica internazionale perché sono gli elementi più facilmente analizzabili con le banche dati disponibili come Wordnet.

Più recentemente Timo Korkiakangas dell'Università di Helsinki ha sviluppato un progetto di annotazione linguistica relativo ai documenti latini altomedievali editi da Schiaparelli e destinato alla piattaforma *Perseus*[18] anche se, nono-

11 D'Angelo/Philippart de Foy (2013).
12 Ricci (2013).
13 Deploige *et al.* (2016).
14 De Gussem (2021).
15 McGillivray (2014).
16 Link: https://mlat.uzh.ch/.
17 Roelli (2021).
18 https://researchportal.helsinki.fi/en/publications/edition-and-on-line-publication-codice-diplomatico-longobardo-12-.

stante le dichiarazioni di Open Access dei materiali, l'articolo in merito è ad accesso riservato.

Molto si sta muovendo anche nel campo epigrafico, con il progetto *DiLaDi* (*Digital Latin Dialectology: Tracing Linguistic Variation in the Light of Ancient and Early Medieval Sources*)[19] cui lavora soprattutto Béla Adamik e il più vasto *Computerized Historical Linguistical Database of Latin Inscriptions in the Imperial Age* diretto dallo stesso Adamik all'Università ELTE di Budapest, e altri, una cui rassegna sarà presentata alla seconda edizione del *Digital Latin* di Siena il 4–6 giugno 2025.[20]

Noi al Centro Studi Comparati I Deug-Su di Siena,[21] che con le sue indagini sul latino asiatico risalenti al 2007, è all'origine del progetto ELA[22] e della sua evoluzione SERICA,[23] abbiamo sondato altre possibilità. Nello stesso anno in cui usciva il primo volume degli atti latino-franco-coreani *Hagiographica Coreana*[24] abbiamo pubblicato anche il primo volume e cd-rom del *Corpus Rhythmorum Musicum*,[25] il progetto europeo che ha prodotto la prima edizione critica digitale nel campo mediolatino, ricostruendo, documentando su tutti i manoscritti e analizzando il corpus latino delle più antiche poesie europee musicate. Il *Corpus Rhythmorum Musicum* comprende sia un doppio programma di concordanze delle forme linguistiche, con indice alfabetico e indice di frequenze dirette e inverse non solo dei termini presenti nelle edizioni ma anche di quelli dei singoli manoscritti, sia statistiche dettagliate su decine di fenomeni di differenziazione linguistica del latino altomedievale e spesso popolare di queste "canzoni" rispetto al latino standard e di approssimazione al cosiddetto protoromanzo. Nel 2011 l'edizione è stata trasferita online e dal 2020 il sito è stato rinnovato e viene continuamente implementato con nuove collazioni e nuove edizioni.[26] Questi tool hanno dato origine a ricerche di latinisti italiani ed esteri che hanno contribuito a un timido rinnovamento degli studi.

Negli anni 2012–2017, grazie a ripetuti finanziamenti PRIN, si è lavorato soprattutto ad ALIM, *Archivio della Latinità Italiana del Medioevo*,[27] la più grande biblioteca digitale di testi mediolatini prodotti in Italia, finalizzata inizialmente a fornire

19 https://pric.unive.it/projects/diladi/home

20 https://centroideugsu.unisi.it/2025/02/26/digital-latin-ii-ph-d-international-workshop-4th-6th-june-2025-siena/. Il programma ha coinvolto decine di imprese scientifiche, database, biblioteche digitali e altri tool di analisi linguistica e stilistica: gli Atti sono in preparazione a cura di Paccara/Recupero (forthcoming).

21 Link: https://www.centroideugsu.unisi.it/.

22 Sito web: http://ela.unisi.it.

23 Sito web: https://serica.unipi.it.

24 Stella *et al.* (2007).

25 Stella (2007).

26 Edizione web aggiornata al 2022 al link: https://www.corimu.unisi.it.

27 Link: http://alim.unisi.it/.

materiale per il dizionario di latino medievale italiano (Arnaldi/Smiraglia [2009]), che ha non solo implementato significativamente la propria base di testi e documenti codificati in TEI (inserendola poi in *Corpus Corporum*), ma li ha anche collegati a un analizzatore lessicale, *Lexicon*,[28] che ho avuto l'opportunità di progettare e far realizzare poi a Luigi Tessarolo e Silvia Arrighetti. *Lexicon* è stato il primo programma italiano di analisi testuale *open access* che non richiede la codifica preventiva dei testi e contiene il primo, imperfetto, lemmatizzatore esteso ad alcune migliaia di lemmi del latino medievale. La specificità di *Lexicon,* che comprende caratteristiche diffuse in molti altri programmi come calcolo dei *collocates*, degli *n-grams*, e delle *function words*, è soprattutto la capacità di produrre tavole e diagrammi di *comparazione* (in particolare la comparazione per *overlap* ma anche quella differenziale) fra testi o fra corpora o fra testi e corpora che è strumento indispensabile per analisi diacroniche, come quelle che il Latino consente e anzi impone di realizzare. Grazie a strumenti come questo si è potuta concepire l'idea di un'analisi dell'evoluzione linguistica del latino, unica lingua la cui documentazione copre oltre due millenni e mezzo di storia e una diffusione geografica mondiale. A fianco di questi progetti è nato nel 2018 il già citato *Eurasian Latin Archive*, biblioteca digitale di testi latini relativi al cosiddetto Estremo Oriente, cioè Cina-Corea-Giappone, online dal 2020, con motore di ricerca bilingue, cioè latino e cinese e produzione automatica di statistiche linguistiche multiple che lavora su *Elastic Search* e dunque con metodo diversissimo dalle banche dati precedenti. La sua realizzatrice, Emmanuela Carbé — che nel frattempo ha curato l'organizzazione della XII edizione del convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale, tenutosi a Siena[29] — ha lavorato su un primo nucleo di 70 testi, a partire da un censimento di circa 300 che costituisce la base anche della parte latina di SERICA, un obiettivo che si è rivelato troppo ambizioso perfino per un progetto con un finanziamento così importante. La piattaforma, realizzata da Università di Siena e azienda informatica Questit con il sostegno della Regione Toscana, consente la produzione automatica di dati statistici sulla lingua, grazie agli *ELA Tools* di Carbé e Giannelli,[30] che estraggono informazione dei file XML secondo il *Classical Language Toolkit* e il *Natural Language Toolkit* di Stanford. Il progetto comprendeva anche una parte di analisi comparativa dei dati linguistici all'interno del corpus sino-latino e fra il corpus sino-latino e quello euro-latino, che non è stato possibile attuare nel biennio dell'assegno di ricerca di Carbé, ma di cui nel 2023 Martina Paccara, nel breve corso di una borsa di studio quadrimestrale dell'Univer-

28 Tessarolo *et al.* (2012).

29 Link al sito web del convegno: www.aiucd2023.unisi.it.

30 Carbé/Giannelli (2019).

sità di Torino, ha provato a produrre alcuni *specimina*. Lo scopo, ma le ricerche scientifiche per fortuna producono risultati collaterali che vanno spesso oltre lo scopo iniziale, era di individuare le caratteristiche linguistiche, cioè morfologiche e sintattiche ma soprattutto lessicali e semantiche, che emergono come specifiche del latino "d'Asia" e che arricchiscono il vocabolario latino medievale, già enormemente più ampio di quello classico, accogliendo elementi, aspetti, oggetti e modalità di pensiero di culture così diverse e nello stesso tempo travisandole o integrandole in un mondo linguistico lontano. In particolare, il progetto *Das-Memo*[31] premiato dalla Regione Toscana proponeva fra l'altro fin dal 2018 la determinazione di una sorta di "tasso di innovazione lessicale" che poteva essere misurato solo in va comparativa fra i tesi sino-latini e fra il corpus sino-latino e corpora di latino "europeo".

E questa comparazione, tecnicamente del tutto inedita,[32] mi pare abbia trovato grazie a Martina Paccara un suo *bunch* di metodi e qualche primo risultato.

2 Applicazioni di metodi computazionali alla ricerca sui testi latini di SERICA

Martina Paccara

Obiettivo di questo contributo è comprendere quali metodi computazionali possano essere applicati ai testi dei database ELA e SERICA, grazie all'uso di quali software e con quali risultati.[33] Uno studio di portata generale sull'intero corpus di testi, esaminati da differenti punti di vista (lessicale, morfologico, sintattico etc.), avrebbe richiesto tempi di realizzazione molto più lunghi: presenterò dunque in questa sede solo una prima ricognizione su metodi e tecniche utilizzabili,

31 Il progetto *Data-Mining e analisi statistica su fonti testuali storiche del periodo medievale e moderno* (*Das-Memo*) è sintetizzato al link http://dasmemo.unisi.it/.

32 Il concetto di *Lexical innovation rate* non è tematizzato nella letteratura delle *Digital Humanities*, anche perché è calcolabile solo su corpora linguistici molto estesi nel tempo. L'espressione è usata però nel contributo di sociolinguistica computazionale Danescu-Niculescu-Mizil (2013). In generale è utilizzabile anche Armstrong (2016). Sul piano storico il parallelo migliore è con l'inglese post-coloniale come studiato da Patrizia Anesa in Anesa (2018). Nessuna applicazione nota alla storia del latino, anche se proprio nell'autunno 2025-26 è stato annunciato un workshop su questo argomento fra le attività del dipartimento di Digital Humanities del King's College di Londra.

33 I metodi quantitativi sono impiegati in ambito umanistico dalla linguistica computazionale, in particolare dalla linguistica dei *corpora* (cf. Freddi [2014]) e dalla stilistica, che assume in questo contesto il nome di stilometria: cf. Herrmann *et al.* (2021). Sono inoltre comunemente noti e utilizzati come strumenti di *authorship attribution*: cf. Juola (2008).

accompagnata da qualche esempio di applicazione, affinché possa fungere da stimolo per ulteriori e più approfondite ricerche.

2.1 Lavoro Preliminare

In primo luogo, ogni indagine di tipo stilometrico o computazionale richiede una fase di lavoro preliminare sui testi, che vanno ripuliti di componenti accessori quali immagini, simboli, numeri di pagina, indicazione dei capitoli e, più in generale, degli elementi che pertengono esclusivamente al formato editoriale. A seconda del tipo di analisi da condurre può anche risultare utile eliminare la punteggiatura: se si agisce a livello di *character n-grams* o di *cluster* è necessario rimuovere i segni diacritici; se si agisce sul piano sintattico è invece fondamentale mantenere la distinzione del testo in periodi. Quando presente, è infine consigliabile eliminare la marcatura in TEI, perché la maggior parte dei software lavora sul *plain text* con codifica UTF-8.

Per quanto concerne nello specifico i documenti di ELA e SERICA, al momento delle mie ricerche era ancora in corso la fase di ripulitura, cosa di cui resta traccia nei frequenti errori di battitura, nella presenza di caratteri cinesi, di abbreviazioni non sciolte e di numerose lettere accentate nel corpo del testo. Conseguentemente, il campo di indagine è stato limitato ai testi non necessitassero ancora di una revisione completa o che fossero già stati sufficientemente rivisti. A tal scopo ho attinto in parte dal database di ELA e in parte ho utilizzato le prime bozze dei testi di SERICA.

Ho operato dunque una selezione del materiale che tenesse conto anche del criterio quantitativo, scegliendo di lavorare sui testi più significativi a livello statistico: ho escluso dalle indagini le opere che constavano di un numero di *tokens* inferiore a 7000. Per questo motivo sono rimasti esterni al presente studio i cataloghi e, in larga parte, le lettere isolate. Anche in presenza di raccolte di lettere di dimensioni consistenti, ho preferito attenermi a un principio di esclusione per evitare che sintagmi ripetuti e formule fisse tipici dell'epistolografia avessero un peso eccessivo sui risultati delle indagini. Ho poi scartato i testi che contenevano troppi termini traslitterati (come l'*Ars gramaticae Japonicae linguae* di Diego Collado) o abbreviati (come il *De re literaria Sinensius commentarius* di Theophil Gottlieb Spitzel, costellato di riferimento a autori, opere, volumi, capitoli, versi di opere letterarie latine e non). Ho infine tralasciato i testi di argomento scientifico (algebra, computo e astronomia) che nella trasformazione in *plain text* subivano una perdita di informazione sensibile a causa della presenza di tabelle, schemi e operazioni, rimanendo inoltre di difficile confronto per la cospicua presenza di numeri (in alcuni casi romani, in altri arabi). Il corpus di testi individuato è presentato nella Tabella 1. Si tratta di 20 testi per un totale di 734 217 occorrenze e 78 515 forme di parola uniche.

Tabella 1: Corpus di testi esaminati.[34]

Titolo	Titolo abbreviato[35]	Data	Autore	Origine	Genere	Argomento	Parole
Historia Mongalorum quos nos Tartaros appellamus	*Historia Mongalorum*	1247	Giovanni di Pian del Carpine	It	Resoconto di viaggio Trattato	Descrizione etnografica	20172
Itinerarium	*Itinerarium*	1255	Guglielmo di Rubruck	Be	Resoconto di viaggio	Descrizione etnografica	36795
Relatio	*Relatio*	1338	Giovanni de' Marignolli	It	Resoconto di viaggio	Descrizione etnografica	7627
Relatio de mirabilibus orientalium Tatarorum	*Relatio de mirabilibus*	1330	Odorico di Pordenone	It	Resoconto di viaggio	Descrizione etnografica	15199
Marci Pauli Veneti, Historici fidelissimi juxta ac præstantissimi, De Regionibus Orientalibus Libri III (redazione Z)	*Milione*	XIV	Marco Polo	It	Traduzione latina resoconto di viaggio	Descrizione etnografica	56568
Magni Tamerlanis vita	*Magni Tamerlanis vita*	1553	Pietro Perondino	It	Biografia		7333

34 L'indicazione dell'edizione di riferimento è fornita per il singolo testo tra i metadati inseriti nei siti http://ela.unisi.it/ e https://serica.unipi.it/.
35 D'ora in avanti ci si riferirà alle opere elencate con l'abbreviazione del titolo fornita nella seconda colonna.

Catechismus cristianae fidei A P. Michele Rogerio Collecta	*Catechismus Collecta*	1586 1592	Alessandro Valignano Michele Ruggieri	It It	Trattato Traduzione trattato (metà prefazione Daxue)	Religione Filosofia	13546 24355
De christiana expeditione apud Sinas suscepta ab Societate Jesu. Ex P. Matthaei Riccii eiusdem Societatis commentariis Libri V.	*De christiana expeditione*	1615	Matteo Ricci, Nicolas Trigault	It-Fr	Trattato Relazione	Etnografia	164820
Regni Chinensis descriptio	*Regni Chinensis descriptio*	1639	Nicolas Trigault et al.	Fr	Trattato	Descrizione etnografica	40499
De bello tartarico historia	*De bello tartarico historia*	1654	Martino Martini	It	Storiografia	Guerra	20355
Flora sinensis	*Flora sinensis*	1655	Michał Boym	Po	Trattato	Botanica	10245
Historica narratio de initio et progressu missionis societatis Jesu apud Chinenses ac praesertim in regia Pequinensi	*Historica narratio*	1655	Johann Adam Schall	De	Relazione	Descrizione etnografica	58058
Historia Tartaro Sinica Nova	*Historia Tartaro Sinica Nova*	1673	François de Rougemont	Be	Storiografia	Descrizione etnografica	50578
Confucius, Sinarium Philosophus, sive Scientia Sinensis latina	*CPS*	1687	Prospero Intorcetta, Christian Herdtrich, François de Rougemont, Philippe Couplet	It	Traduzione trattato	Filosofia	139427

(continua)

Tabella 1 (continua)

Titolo	Titolo abbreviato	Data	Autore	Origine	Genere	Argomento	Parole
De magno Sinarum Imperio	*De magno Sinarum Imperio*	1697	Erico Reland	Se	Trattato	Descrizione etnografica	8446
Icon Regia Monarchae Sinarum nunc regnantis ex gallico versa	*Icon Regia*	1699	Joachim Bouvet	Fr	Biografia		16369
Ad virum nobilem de cultu Confucii philosophi et progenitorum apud Sinas	*De cultu Confuci*	1700	Robertus Pazmany[36]	It	Relazione epistolare	Filosofia, religione, etnografia	7763
Brevis Relatio eorum quae spectant ad declarationem Sinarum imperatoris Kamhi circa Coeli, Cumfucii et Avorum cultu	*Relatio decl. Kahmi*	1701	Thomas Pereira et al.	vv.	Relazione epistolare	Filosofia, religione	13359
Historica notitia rituum ac ceremoniarum Sinicarum	*Historica notitia*	1711	François Noël	Be	Trattato	Religione, etnografia	26232

36 Il testo è stato trascritto a partire da un'edizione a stampa del 1700 disponibile in open access al link https://books.google.it/books?id=ofNaAAAAQAAJ&hl. La lettera non è firmata e il nome del suo autore non compare nel frontespizio dell'edizione, ma i metadati presenti in Google Books riportano una non motivata attribuzione a Jean Déz. L'errore è facilmente dimostrabile sulla base di un'ulteriore versione a stampa del medesimo testo disponibile anch'essa in open access al link https://www.google.it/books/edition/Ad_Virum_Nobilem_De_Cultu_Confucii_Philo/M4xXAAAAcAAJ?hl.

Auspicabilmente le ricerche successive estenderanno i testi inclusi nelle operazioni di analisi e creeranno dei *subcorpora* per genere, argomento o data.

2.1.1 Lessico

Il lessico è il primo elemento a essere stato analizzato in ambito computazionale in quanto "not only are words (and n-grams) easily identifiable and countable, compared to figures of speech, themes, or syntactic patterns, they are also much more frequent than most other textual characteristics and are obviously, though not unproblematically, meaningful".[37] Per questo motivo, le indagini sul piano lessicale dispongono di un numero consistente di strumenti di facile accesso. Esistono infatti numerosi software a interfaccia grafica progettati per questo scopo, tra cui i più utilizzati per il latino sono:

- *Collatinus*;[38]
- *Lexicon*;[39]
- *Voyant Tools*;[40]
- *Antconc*;[41]
- *Hyperbase Web*.[42]

Tra i software menzionati, *Collatinus* e *Lexicon* includono una funzionalità auspicabile nell'analisi della lingua latina, non disponibile in *Voyant Tools* o *Antconc*: la lemmatizzazione. Anche il pacchetto degli *ELA tools* consente di operare su testi lemmatizzati, tuttavia, a differenza degli strumenti precedentemente menzionati, limita la ricerca ai testi già presenti nel sito, senza consentire all'utente di inserirne di nuovi. Va inoltre notato che la lemmatizzazione suddetta risulta più imprecisa, dal momento che, in caso di ambiguità tra più lemmi per una forma, gli *ELA tools* associano automaticamente la forma al lemma più frequente. Il software non fornisce infine un conteggio totale delle forme non lemmatizzate ed è dunque difficile stimare l'efficacia della lemmatizzazione. Nonostante i suoi vantaggi, *Lexicon* presenta dei limiti nelle funzioni di visualizzazione, risultando più rudimentale rispetto a *Voyant Tools*. D'altro canto, *Antconc* si configura come un valido sostituto in caso di *corpora* o testi di dimensioni consistenti che *Lexicon*

37 Rybicki *et al.* (2016) 125.
38 Ouvrard/Verkerk (2012). Per alcuni esempi di utilizzo cf. Ouvrard/Verkerk (2014).
39 Tessarolo *et al.* (2012). Per alcuni esempi di utilizzo cf. Stella (2020).
40 Sinclair/Rockwel (2016).
41 Anthony (2022).
42 Vanni (2015).

non riesce ad analizzare completamente. Per quanto riguarda *Hyperbase web*, per i testi latini non appartenenti al Database LASLA, dopo una fase beta, è stata rilasciata la prima versione ufficiale.[43]

a) Varietà e innovazione lessicale

Uno degli aspetti, in cui l'analisi dei testi presenti nei database ELA e SERICA si rivela potenzialmente più interessante, consiste nell'indagine sulla varietà e innovazione lessicale.

Type-Token ratios

Lo studio della diversità morfologica e lessicale si basa sul calcolo della *Type-Token ratio* (TTR), un indicatore della varietà linguistica di un testo. Vi sono varie tipologie di TTR, tra cui si segnalano in particolare la *morphological diversity* ("diversità morfologica" o "TTR delle forme") e la *lexical diversity* ("diversità lessicale" o "TTR dei lemmi"). La diversità morfologica è data dal rapporto tra *type* e *token*[44] e offre una misura di quante forme diverse vengono utilizzate in un testo rispetto al numero complessivo di occorrenze. La diversità lessicale si calcola dividendo il numero di lemmi per il totale dei *token* e fornisce un indice della diversità di lemmi utilizzati in un testo. Un problema comune di queste misure è che i campioni di testo contenenti un gran numero di *token* danno valori inferiori per TTR poiché è spesso necessario per lo scrittore o il parlante riutilizzare le stesse *function words*,[45] che sono in numero molto più limitato dei lemmi. Una conseguenza di ciò è che la diversità lessicale è meglio utilizzata per confrontare testi di uguale lunghezza. Se si osservano infatti le misure di TTR delle forme e dei lemmi di due testi di diversa lunghezza quali la *Regni Chinensis descriptio* (40499 *token*) e la *Magni Tamerlanis vita* (7333 *token*) si nota che la TTR delle forme e dei lemmi è nettamente inferiore per il primo testo (rispettivamente 29% e 12%) rispetto al secondo (53% e 34%) (cf. Tabella 2). Un'alternativa è quindi il calcolo della *lexical density* ("densità lessicale" o TTR delle PoS), che consiste nel numero di *token* con POS lessicale (ovvero nomi, verbi, aggettivi e spesso anche avverbi) su numero totale di *token*. Questa misura stima dunque la presenza delle parole lessicali in un testo in rapporto alla sua estensione. Nonostante dia conto della

43 La versione 1.0 non era ancora disponibile al tempo delle mie ricerche, che sono state dunque condotte utilizzando la versione beta.

44 Con *type* si fa riferimento alle differenti forme delle parole presenti in un testo, mentre per *token* al numero complessivo delle occorrenze.

45 Spesso chiamate anche *stop words* e, in italiano, parole grammaticali, parole vuote o parole funzione.

complessità linguistica di una composizione scritta, ha lo svantaggio di non fornire una stima della varietà lessicale.

Tabella 2: TTR delle forme e dei lemmi della *Regni Chinensis descriptio* e della *Magni Tamerlanis vita.*

	Nicholas Trigault et al. ***Regni Chinensis descriptio***	**Pietro Perondino** ***Magni Tamerlanis vita***
Occorrenze	40 499	7333
Lemmi	4719	2380
Forme	11748	3890
TTR delle forme	0,29	0,53
TTR dei lemmi	0,12	0,34
TTR delle PoS	0,67	0,71

Lexical Innovation Rate/Tasso di innovazione lessicale

Il dato più interessante da calcolare in presenza di un corpus di questo tipo consiste tuttavia nell'apporto lessicale innovativo dei suoi testi, che ci si attende cospicuo in narrazioni relative a luoghi, culture e popolazioni molto distanti nello spazio. A questo scopo ho utilizzato una misura che Francesco Stella[46] propone di denominare *Lexical innovation rate* "tasso di innovazione lessicale" e che consiste nel rapporto tra il numero di forme non lemmatizzate dal software e il totale di forme di parola uniche nel corpus (cf. Figura I).

Figura I: Tasso di innovazione lessicale.

Questa misura ci sembra particolarmente adatta allo studio dei testi dei database ELA e SERICA poiché mira a fornire una stima della loro originalità lessicale a partire da un dato empirico: la difficoltà di lemmatizzazione riscontrata dai software.

Per il conteggio delle forme non lemmatizzate ho fatto ricorso a *Lexicon* e *Collatinus*. Dal confronto tra i due software è emerso che *Collatinus* non solo riesce a lemmatizzare una percentuale più alta di parole sia nei testi classici che nei testi medievali (cf. Tabella 3), ma fornisce anche il numero totale degli elementi non lemmatizzati senza contare due volte le forme ripetute. Inoltre l'elenco delle

46 Cf. *supra*, n. 27.

forme non lemmatizzate è disponibile in *Collatinus* per tutti i testi, mentre in *Lexicon* solo per quelli più brevi.

Tabella 3: Efficacia della lemmatizzazione di *Lexicon* e *Collatinus* nei testi classici e medievali.

	Forme non lemmatizzate		Forme totali	Percentuale di forme non lemmatizzate	
	Collatinus	***Lexicon***		***Collatinus***	***Lexicon***
Gaio Cornelio Tacito *Germania*	14	71	2694	0,51%	2,6%
Valafrido Strabone *Vita Mammae*	27	64	2910	0,93%	2,2%

Ho dunque analizzato il corpus di testi elencati nella Tabella 1 con *Collatinus* e sommato il numero di forme non riconosciute per ogni testo. Il totale è di 10668 forme non note, per un tasso di innovazione lessicale del 13,58% (cf. Tabella 4).

Tabella 4: Innovazione lessicale nel corpus.

Tot. occorrenze	734217
Tot. forme di parola unica	78515
Tot. occorrenze non lemmatizzate	13148
Tot. tipi non lemmatizzati	10668
Tasso di innovazione lessicale	13,58%

Nonostante il dato in apparenza elevato, sono necessarie alcune precisazioni. Un'analisi manuale dei risultati ottenuti per alcuni testi campione (cf. Tabella 5a; 5b; 5c), rivela che la percentuale di forme non lemmatizzate risulta "sporca". Vi è infatti un gruppo di parole costituito da errori di battitura (che auspicabilmente saranno stati corretti nelle versioni dei testi caricate ora sul sito); una percentuale di falsi positivi dovuta a forme presenti nei dizionari/lessici regionali medievali non come lessemi di origine orientale ma semplici forme tarde o locali; e una percentuale di errori del software. Queste tre categorie costituiscono in media il 50% del totale: si può di conseguenza considerare che l'innovazione lessicale si aggiri intorno alla metà del dato iniziale, al 6,8%. Più nello specifico, questa innovazione è costituita quasi totalmente (più del 90%) da nomi propri, toponimi o aggettivi di nazionalità.

Tabella 5a: Innovazione lessicale in Martino Martini, *De bello Tartarico historia.*

Forme non lemmatizzate: 435/6749				
Nomi di persona o aggettivi di nazionalità	**Innovative**	**Parola/grafia medievale**	**Errori battitura**	**Errori Software**
269 (61,8% – **94%**)	17 (3,9% – **5%**)	58 (13%)	72 (16%)	19 (6%)
286 (65%)		148 (35%)		

Tabella 5b: Innovazione lessicale in Giovanni de' Marignolli, *Relatio.*

Forme non lemmatizzate: 342/3248				
Nomi di persona o aggettivi di nazionalità	**Innovative**	**Parola/grafia medievale**	**Errori battitura**	**Errori Software**
178 (52% – **92,2%**)	15 (0,3% – **7,8%**)	124 (36%)	25 (7,3%)	0
194 (56,7%)		148 (43,3%)		

Tabella 5c: Innovazione lessicale in Prospero Intorcetta *et al.*, *CPS.*

Forme non lemmatizzate: 916/24205				
Nomi di persona o aggettivi di nazionalità	**Innovative**	**Parola/grafia medievale**	**Errori battitura**	**Errori Software**
359 (39% – **88,6%**)	46 (5% – **11,4%**)	258 (28,16%)	181 (19,7%)	72 (7,8%)
405 (44,2%)		511 (55,78%)		

Al di là del dato percentuale, l'interesse di questo tipo di analisi risiede nella possibilità di visualizzare in modo rapido e immediato la totalità delle forme potenzialmente innovative, che possono costituire un'importante base di studio. Si consideri a mo' di esempio la *Flora sinensis* di Michał Boym, uno dei primi trattati di storia naturale cinese. *Collatinus* consente l'individuazione di 270 parole non lemmatizzate, di cui ben 185 risultano, al netto dei falsi positivi, realmente innovative (in questo caso addirittura il 68,5% del totale) (cf. Tabella 5d).

Di queste, una significativa percentuale è costituita da nomi di frutti e animali provenienti dall'Oriente: il libro include, infatti, descrizioni della flora e della fauna asiatica, di cui viene fornito sia il nome cinese che quello latino (se disponibile) o portoghese (quando non esiste un equivalente latino o è necessario chiarire l'oggetto descritto). In portoghese sono anche alcune parole connesse alla descrizione di elementi della flora. Ulteriori elementi di interesse risiedono in alcuni aggettivi e

Tabella 5d: Innovazione lessicale in Michał Boym, *Flora Sinensis.*

Forme non lemmatizzate: 270/4612						
Nomi di persona o aggettivi di nazionalità	Innovative Frutti-animali	Altro	Tratti medievali Parola	Grafia	Errori battitura	Errori Software
106 (39,3%)	59 (21,8%)	20 (7,6%)	29 (10,7%)	20 (7,4%)	28 (10,3%)	8 (2,9%)
185 (68,5%)			49 (18,1%)		36 (13,3%)	

sostantivi quali *alboflavus*, *durefactus*, *marcimorus*, *ramiferus*, *squinatia* che non si trovano in alcun vocabolario di latino (anche medievale) e dunque rappresentano esempi di quelle acquisizioni lessicali la cui individuazione è uno degli scopi della ricerca (cf. Tabella 6). Anche in questo caso, il software semplifica notevolmente il lavoro, sopperendo alla fallibilità dell'intuizione dello studioso, che potrebbe non prestare la giusta attenzione a termini che suonano a lui comprensibili e familiari per somiglianza a successive evoluzioni della lingua. Inoltre, è interessante notare che tra i termini di uso esclusivamente medievale si trovano alcuni vocaboli che appaiono nei dizionari regionali di Boemia e Ungheria, come *caravana*, *corroborativus*, *mappalis*, *mollificativus* o *saccharatus* (cf. Tabella 7).[47]

Tabella 6: Forme innovative non lemmatizzate nella *Flora Sinensis.*

Frutti e piante	Animali
Ananas – FamPoLoMie; Areca/ae/am – Pimlam (Betel); PaCyao – Banana; Bellota (ghianda); Brinhoes (melanzana); Carambole; Carciochis/orum (scorza della palma simile a cardo); Cievko – Goyava; Duliam; Garyaphilla (chiodi di garofano); Giambolane (prugna); Giangame (ciliegia indiana); Giaca – Giaka – PoLoMie; Fanyaycu – Papaya; GiamBo; HuCyao (pepe); KaGiu – KiaGiu (anacardio); KueyPi (zafferano); LumYen (litchi); Manko – Manga – Mangas; PeFoLim (radice sinica o "pane indiano") Rhabarbari/o/um – Socuir; SemKiam (zenzero); SuPim (cachi); Tamerae/as/is; YaTa (Annona)	Cabelo (tipo di serpente); Camello Toki (tipo di gallina); Ciamviki (tipo di gallina); Cobra/ as (tipo di serpente); Gento (tipo di serpente); HaijMá (cavallo marino); Hiam (cervo muschiato); HivenPao (pantera); FumHoam (ape); LoMeoQuey (tartaruga); MasFum (femmina ape); Rhinocerontis Sumxu (tipo di topo); Yeki (gallina)

47 La famiglia di Michał Boym era originaria dell'Ungheria.

Tabella 6 (continua)

Parole portoghesi	Altro
Almiscar (carne di rene); Cayro (polpa del cocco); Cocobarca (polpa dello Giaca frutto); Giagra (zucchero di cocco); raiz; mangiarBianco	alboflavum, durefacti, marcimorum, propullulat, ramifera/am/os, scudis/o, squinatiam

Tabella 7: Tratti medievali nella *Flora Sinensis.*[48]

Boemia e Ungheria	Diffusione varia
Alembici, caravanis, corroborativum, mappale, mollificativis, prolifica, saccharatus, ullibi	aggratulantur, artericis, camphoram, comproducit, cordialem, defossionem, grossitie, Iulepe, muscatam, porcellanas, porosaque, porosissimus/a, sclopetum, stomachale, unctuosum/i, ustibilis

b) Lessico e sintagmi più ricorrenti

Le analisi finora condotte hanno avuto come focus gli elementi stranianti del lessico. Tuttavia è altrettanto possibile esaminare le occorrenze, le forme, i sintagmi e le locuzioni più ricorrenti nei testi. Nella sezione successiva, a titolo di esempio si sottoporrà a indagine il *Confucius philosophus Sinarum*, la traduzione di tre dei quattro libri del confucianesimo (*Daxue* "Il grande studio", *Zhongyong* "Il giusto mezzo" e *Lunyu* "Dialoghi") ad opera di Prospero Intorcetta, Christian Herdtricht, François de Rougement e Philippe Couplet.[49] Il software di riferimento sarà *Lexicon*, i cui apporti saranno integrati tramite l'utilizzo di altri programmi qualora consentano una riuscita migliore delle indagini lessicali o presentino funzionalità più specifiche.

48 Mi riservo di condurre in altra sede un'analisi lessicografica più approfondita delle occorrenze riportata nella Tabella 7.

49 Nel frontespizio, l'opera è presentata come lavoro collettivo dei quattro padri. Il volume si apre con la dedica al sovrano Luigi XIV e una mappa dell'impero cinese ad opera di Couplet; seguono una prefazione redatta da Intorcetta e Couplet sulla cultura cinese (*Proemialis Declaratio*) e una biografia di Confucio (*Confucii Vita*) scritta da Intorcetta. Le traduzioni latine dei tre libri non sono firmate dagli autori. Chiudono l'opera una della storia cinese (*Tabula Chronologica*) ed una sinossi sui caratteri principali dell'impero cinese (*Imperii Sinarum et Rerum in eo Notabilium Synopsis*) compilati da Couplet. L'analisi linguistica ha incluso la prefazione, la biografia di Confucio e le tre traduzioni.

Distribuzione delle forme e dei lemmi nel documento

Una prima analisi del *CPS* tramite *Voyant Tools* restituisce uno sguardo d'insieme sul testo facilmente intellegibile dal punto di vista grafico. L'osservazione del *word-cloud* delle parole più frequenti del testo — privato preventivamente delle *function words* — consente di individuare la presenza di alcuni di temi chiave anche in assenza di lemmatizzazione: spiccano infatti le forme declinate dei sostantivi *familia, discipulus, animus, imperium, regnum, homo, virtus, populus, coelum* (cf. Figura IIa).

Figura IIa: Word cloud delle forme più frequenti nel *CPS*.

Delle singole forme o lemmi si può visualizzare anche la distribuzione nel corso dell'opera (cf. Figura IIb). Ad esempio il termine *Confucius* diventa nettamente preponderante nella parte finale dell'opera, in cui effettivamente è tradotta la sezione dei *Dialoghi* dei *Quattro libri*. In questa sezione diminuisce la presenza di *virtus* e aumenta quello della *familia*. Questo metodo consente dunque di esplorare i temi specifici di ciascun libro, identificando quali lemmi sono più presenti nelle diverse sezioni della filosofia confuciana.

Liste di frequenza

L'indagine sui lemmi — anche in questo caso ad esclusione delle *function words* — può essere affinata con *Lexicon*, che consente di visualizzarne le frequenze percentuali (cf. Tabella 8). In cima alla lista si trovano, come è naturale, il verbo essere, i pronomi e gli aggettivi dimostrativi. A seguire, degna di nota è la frequenza del lemma *virtus* (0,5%) che sembra superare anche la frequenza del nome di Confucio. Tuttavia è bene notare che il vocabolario del software non con-

Tabella 8: Lista di frequenza degli 80 lemmi più ricorrenti nel *CPS*.

Lemma	Ricorrenza	Frequenza %	Lemma	Ricorrenza	Frequenza %
edŏ sŭm	2075	1.488	ăgŏ	295	0.212
quī quīs	2041	1.464	impĕrātor	281	0.202
sŭm	1969	1.412	sŭpĕrŭs	273	0.196
ĭs	979	0.702	ănĭmŭs	263	0.189
quī quĕō	929	0.666	tōtŭs tŏtŭs	261	0.187
omnĭs	858	0.615	tempus	260	0.186
sŭŭs sī	846	0.607	ĭtŏ ĭtem	254	0.182
virtūs	709	0.509	quīsquĕ quŏque	251	0.18
confucius	688	0.493	dĕindĕ	247	0.177
rēs	666	0.478	prīmŭs	246	0.176
sŭŭs	625	0.448	sē sūs sŭŭs sŭŏ	242	0.174
quī quīs quăm	580	0.416	priscŭs	240	0.172
possum	563	0.404	rex	238	0.171
āio	562	0.403	sē sibŭs	238	0.171
hŏmo	555	0.398	ūnŭs	236	0.169
caelŭm	497	0.356	rēspondĕŏ	224	0.161
dīcŏ	458	0.328	măgister	220	0.158
sŭŭs sŭŏ	438	0.314	făcĭŏ	207	0.148
ălĭŭs allĭŭm	433	0.311	mŏdŭs mŏdŏ	206	0.148
hăbĕŏ	430	0.308	ălĭŭs	203	0.146
ĭs ĕō	424	0.304	nĭhil	202	0.145
vĕro vērŏ vērŭs	419	0.301	fīlĭŭs	195	0.14
sē sŭŭs	400	0.287	ūtŏr/o ŭt	195	0.14
magnŭs	394	0.283	vīrŭs vĭr	193	0.138
rătĭo	375	0.269	rex rĕgŏ	192	0.138
princeps	370	0.265	tandem	192	0.138
quīdăm	360	0.258	verbŭm	186	0.133
ădĕō	358	0.257	interprĕs	183	0.131
impĕrĭŭm	334	0.24	quīspĭam	182	0.131
fămĭlĭă	333	0.239	quīvīs quamvīs	182	0.131
ĭtŏ ĕō ĭta	325	0.233	sūs sŭŭs sŭŏ	181	0.13
annŭs annŏ	317	0.227	terră	180	0.129
inquăm	317	0.227	tantŭs	179	0.128
discĭpŭlŭs	315	0.226	audĭŏ	177	0.127
scīlĭcet	312	0.224	nĕŏ nē	172	0.123
çu	307	0.22	rītŭs	172	0.123
pōpŭlŭs pŏpŭlŭs	307	0.22	ĭgĭtur	170	0.122
regnŭm	304	0.218	sūs sŭŭs	169	0.121
vĭdĕŏ	304	0.218	nostĕr	168	0.12

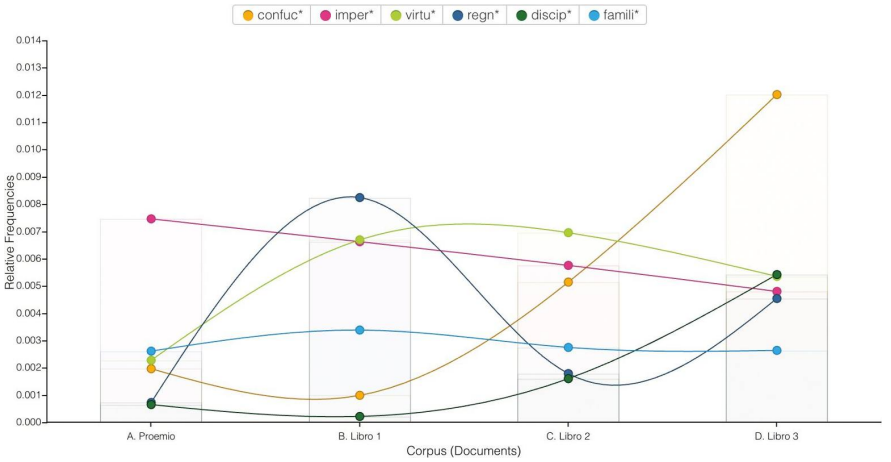

Figura IIb: Distribuzione di alcuni lemmi selezionati nel *CPS*.

tiene il lemma *Confucius, i* e, di conseguenza, il programma non riesce a ricondurre le forme declinate del lemma (*Confuci, Confucio, Confucium* etc.) al lemma base: risulterà molto più presente sommando i valori dell'indice di frequenza delle varie forme flesse. Oltre a una presenza abbondante del lessico della parola e del dire (*aio, vero, dico*), tema cardine è però quello dell'uomo e del saggio perfetto, come denota l'abbondanza di termini quali *homo, rex, magnus, primus, princeps, sapiens* e *vir*. Sono poi particolarmente presenti una serie di lemmi quali *imperium, regnum, populus, familia, rex*, che evidenziano la dimensione sociopolitica del messaggio di Confucio.

Word n-grams

I *word n-grams* ("n-grammi parola") — chiamati "locuzioni"' da *Lexicon* per mantenere la terminologia del programma *Analisi lessicale* (1998) di cui *Lexicon* è un'evoluzione — sono sequenze di parole ripetute. Tramite *Lexicon* è possibile ricercare *n-grams* di qualsiasi estensione, anche in forma di lemmi. Avvalendomi di *Antconc*[50] ho effettuato la ricerca dei *2-grams* e *3-grams* più ricorrenti nel *CPS*. Tra i *2-grams*, degni di nota sono *Confucius ait* e *Confucius respondit*. Per il resto i

50 *Lexicon* non supporta la ricerca su un testo di così vaste dimensioni.

sintagmi che contano il maggior numero di occorrenze sono costituiti da traslitterazioni dal cinese e espressioni esplicative e di snodo logico del discorso come *id est*, *et tamen, et in, quod si, in hoc* (cf. Tabella 9a). I *3-grams* presentano qualche spia più indicativa dell'orientamento moralistico del discorso di Confucio, con *virtutis ac sapientiae* in prima posizione. Seguono altre locuzioni costituite da nomi cinesi, frequentemente composti di due parti, e nessi quali *et tamen non* e *in hunc modum* (cf. Tabella 9b).

Tabella 9a: *2-grams* più ricorrenti nel *CPS.*

Type	Frequenza
confucius ait	211
id est	146
et tamen	90
et in	88
sic ait	84
çem çu	72
quod si	63
çu lu	63
in hoc	62
est est	60
confucius respondit	59
hoc est	59

Tabella 9b: *3-grams* più ricorrenti nel *CPS.*

Type	Freq
virtutis ac sapientiae	22
et tamen non	15
yao et xun	15
anno ante christum	14
in hunc modum	14
ven vam et	14
confucius ait vir	13
in regno lu	13
libro xu kim	13
et vu vam	12
discipulus çu lu	11
est confucius ait	11
inter se mutuo	11

Tabella 9b (continua)

Type	Freq
secus ac si	11
sic tamen ut	11
çem çu ait	11
confucius ait si	10
discipulus çu cham	10
et haec quidem	10
haud secus ac	10
çu hia ait	10

L'analisi degli *n-grams* è stata successivamente estesa all'intero corpus per individuare, se presenti, eventuali pattern nell'utilizzo della lingua da parte degli autori gesuiti. I *2-grams* si sono rivelati poco significativi perché formati per lo più da *function words* giustapposte in espressioni proprie dell'uso comune e non caratteristiche. Fa eccezione il sintagma *Confucius ait* che tuttavia, pur contando numerose occorrenze, è utilizzato solamente in 2 dei testi del corpus (come si evince dal *range*); più interessante è invece l'*apud sinas*, che si trova in ben 12 dei 20 testi (cf. Tabella 9c). Escludendo le *function words* e selezionando un *range* minimo di testi (7) in cui le locuzioni devono apparire per poter essere conteggiate si ottengono risultati più interessanti, quali il *flexis genibus* e i vari sintagmi che esprimono l'idea di totalità: *rerum omnium, toto regno, totius imperii, utraque partes, omnes homines, orbis terrarum, rerum omnium* (cf. Tabella 9d). Anche nel caso dei *3-grams*, i sintagmi più frequenti afferiscono per lo più a un solo testo (cf. Tabella 9e). Se si ordinano i risultati per *range* si ottengono espressioni più diffuse, ma comunque caratteristiche di meno della metà dei testi del corpus (cf. Tabella 9f). Degne di interesse sono locuzioni quali *in hunc modum* e *non secus ac* che in questi testi sono molto più utilizzate della media, pur non potendosi ritenere tipiche esclusivamente del linguaggio dei gesuiti.

Collocates

Di ogni parola si possono ricercare anche i *collocates* ("co-occorrenze") intesi come associazioni abituali e privilegiate di due o più parole all'interno di una frase. *Lexicon* consente questo tipo di analisi, ma permette di ordinare solamente o in base alla ricorrenza — e questo comporta che in cima alla lista compaiano sempre i lemmi più frequenti in generale, come il verbo essere (cf. Tabella 10°) — o in base alla distanza — e in questo caso termini che appaiono anche solo due volte nell'immediata prossimità della parola chiave si trovano per primi (cf. Tabella 10b). A questo problema ovvia *Hyperbase* il quale calcola i *collocates* sulla

Tabella 9c: *2-grams* più ricorrenti nel corpus con *function words.*

Type	Frequenza	Range
et in	709	20
et cum	404	19
confucius ait	401	2
id est	368	14
in hoc	327	18
in quo	277	18
in hac	275	18
et ad	263	19
apud sinas	260	12
usque ad	256	16
est et	252	20
est in	250	18
et de	242	19
in ea	238	17
non est	224	19
in eo	222	16
in qua	217	18
p matthaeus	213	5
et non	210	20
est ut	204	19
quod in	203	16
ab eo	190	17

Tabella 9d: 2*-grams* più ricorrenti nel corpus senza *function words.*

Type	Freq	Range
rerum omnium	69	10
toto regno	68	7
procul dubio	60	7
dici potest	44	9
prae caeteris	43	8
eodem tempore	41	9
flexis genibus	37	9
maiorum suorum	33	7
toto imperio	28	7
multis aliis	27	8
totius imperii	27	10
utraque parte	27	8
fieri potest	26	10

Tabella 9d (continua)

Type	Freq	Range
plus minus	26	8
imperii sui	25	8
matthaeus riccius	25	7
omnes homines	25	8
orbis terrarum	25	8
domum suam	23	8
ob causam	23	9
rebus omnibus	23	7
alio nomine	22	7

Tabella 9e: *3-grams* più ricorrenti nel corpus in ordine di frequenza.

Type	Freq	Range
rituum tomus capitulum	109	1
liber rituum tomus	102	1

Tabella 9f: *3-grams* più ricorrenti nel corpus in ordine di range.

Type	Frequenza	Range
in hunc modum	71	9
non secus ac	31	9
factum est ut	28	8
de quo supra	26	8
et in ea	22	8
qui sunt in	18	8
ab eo qui	13	8
est et non	12	8
in quo est	12	8
non modo non	11	8

base dello *z-score*, una misura della probabilità di un dato fenomeno (cf. Tabella 10c).[51] Quando si è interessati a sapere quali parole tendono a comparire nella

51 Lo *z-score* o *standard-score* è definito come il numero di deviazioni standard rispetto alla media di un punto informativo. Esso si ottiene sottraendo alla variabile aleatoria la sua media e dividendo il tutto per la deviazione standard (σ): Z = X – M / σ.

stessa frase di un termine "x" (*e.g. vir*), il programma divide il corpus in due parti: le frasi che contengono *vir*, da un lato, e quelle che non contengono *vir*, dall'altro. Per ogni parola che compare nel primo gruppo di frasi, il software calcolerà se ha un numero di occorrenze superiore alla media in questo sottoinsieme del corpus. Le parole con uno *z-score* più elevato saranno classificate come *collocates* di *vir*. Come si evince dal grafico che il software permette di visualizzare, i veri *collocates* di *vir* sono *perfectus*, *probus, sapiens, eximius* nonché *solidae virtutis* (cf. Figura III). Anche *Antconc* fornisce risultati concordanti (cf. Tabella 10d).

Tabella 10a: *Collocates* di *vir* nel *CPS* calcolati tramite *Lexicon* e ordinati in base alla ricorrenza.

Parole	Ricorrenza	Distanza media
ĕdŏ – sŭm	104	25.654
sŭm	81	40.16
quī quīs	80	38.275
āio	68	25.853
confucius	66	18.015
ĭs	55	36.727
sī	53	37.528
quī -quĕō	52	39.731
omnĭs	49	44.878

Tabella 10b: *Collocates* di *vir* nel *CPS* calcolati tramite *Lexicon* e ordinati in base alla distanza media.

Parole	Ricorrenza	Distanza media
nŭm	2	0
lin	2	0.5
constŏ	2	1.5
ordĭnārĭŭs	2	1.5
tristŏr	2	1.5
aspernŏr	2	2
hoei	2	2

Tabella 10c: *Collocates* di *vir* calcolati tramite *Hyperbase* in base allo *z –score.*

Mot	Z-Score	Probability
perfectus	10.55	4.24E-24
probus	7.11	1.94E-12
eximius	6.93	6.62E-12
sapiens	6.79	1.6E-11
censendus	5.55	2.61E-8
fortitudini	4.89	8.71E-7
solidae	4.68	2.35E-6
defert	4.67	2.58E-6

Confronto tra testi

Altra possibilità offerta da *Lexicon* è quella del confronto tra testi. Vista la compatibilità di genere letterario ho scelto di adottare come termine di paragone gli scritti morali di Seneca.[52] I trattati sono stati selezionati in numero tale da costituire un corpus equiparabile al *CPS* relativamente all'ampiezza: *De beneficiis*, *De brevitate vitae*, *De consolatione ad Helvium, ad Marciam*, *Ad Polybium*, *De constantia*, *De clementia*, *De ira*, *De otio*, *De tranquillitate animi*, *De vita beata* per un totale di 124699 rispetto alle 139931 del *CPS* (cf. Tabella 11a).

Ho dunque ricercato i lemmi di A assenti in B escludendo le *function words*. In cima all'elenco compaiono numerosi nomi cinesi e traslitterazioni cinesi (non riportati per motivi di spazio) che seguono immediatamente gli ovvi *Confucius* (e relative forme flesse) e *Christus* (cf. Tabella 11b). D'altro canto il *CPS* non contiene riferimenti ad alcuni personaggi dell'antichità latina quali *Caesar, Hercules, Cato*, presenti nei testi del filosofo antico (cf. Tabella 11c). Nel *CPS* spiccano i lemmi *perfectio* e *oboedentia*, mentre il lessico di Seneca presenta più sostantivi astratti (*nequitia, sevitia, libertas, patientia*).

Tra il lessico comune ai due campioni analizzati, i termini *discipulus* e *doctrina* sono più usati nel *CPS*, insieme a un vocabolo relativamente inusuale quale *erga* (cf. Tabella 11d). Anche l'idea di perfezione ricorre con maggiore insistenza. Si parla di *riti*, di *sacrificium*, di *superstitio*, mentre è irrisorio l'utilizzo di *ira, fortuna, veneficium, contumelia*: il discorso di Confucio è meno improntata agli elementi negativi tipici della trattazione senecana (cf. Tabella 11e).

52 I testi sono stati scaricati dalla sezione *Latinitas Antiqua* di *Corpus Corporum*.

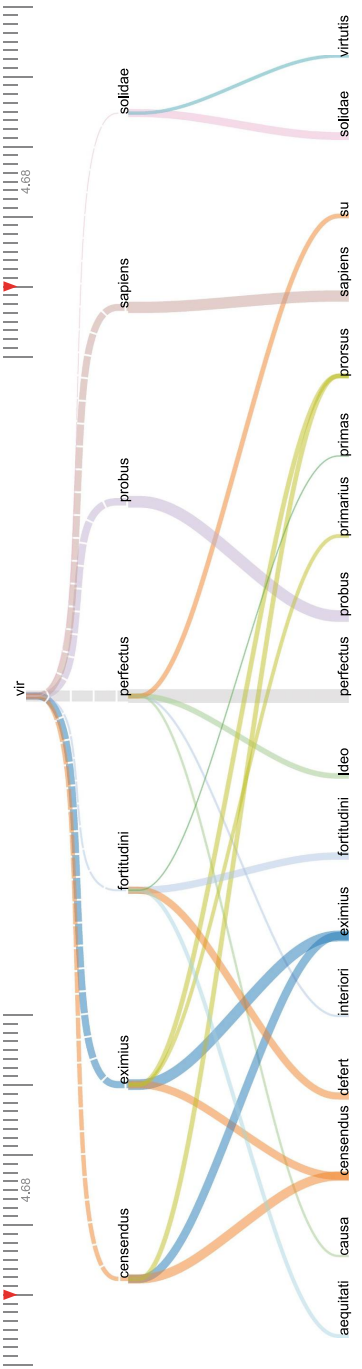

Figura III: Visualizzazione della distribuzione dei *collocates* di *vir* nel CPS tramite *Hyperbase*.

Tabella 10d: *Collocates* di *vir* nel *CPS* calcolati tramite *Antconc.*

Collocate	Freq(Scaled)	FreqLR	FreqL	FreqR	Likelihood	Effect
perfectus	570	36	16	20	220.149	5.767
probus	410	24	3	21	142.898	5.657
sapiens	400	20	1	19	112.685	5.430
eximius	70	8	4	4	58.580	6.623
ait	5010	28	26	2	44.085	2.269
confucius	6890	30	22	8	35.725	1.908
ideo	490	8	7	1	27.596	3.815
gravis	250	6	0	6	25.089	4.371
praestans	60	4	1	3	24.824	5.845
integer	60	4	0	4	24.824	5.845
solidae	80	4	2	2	22.499	5.430

Tabella 11a: Testi messi a confronto tramite *Lexicon.*

	A	B
Testi	De beneficiis, De brevitate vitae, De consolatione ad Helvium, De consolatione ad Marciam, De consolatione ad Polybium, De constantia, De clementia, De ira, De otio, De tranquillitate animi, De vita beata	Confucius Philosophus Sinarum
Autori	Lucio Anneo Seneca	Prospero Intorcetta, Christian Herdtrich, François de Rougemont, Philippe Couplet
Fonte	CC	SERICA
Occorrenze	**124 699**	**139 427**
Lemmi	7755	7631
Forme	24695	24205
Forme non riconosciute	412	1049
TTR of forms	0,172	0,19
TTR of lemmas	0,05	0,06

Tabella 11b: Lemmi di A – lemmi di B.

Lemmi	Ricorrenza	Frequenza %	Lemmi	Ricorrenza	Frequenza %
Confucius	688	0.493	gŭbernŏ	51	0.037
priscŭs	240	0.172	sīmĭlĭter	50	0.036
quīspĭam	182	0.131	paragraphŭs	48	0.034
prŏbŭs	95	0.068	praecĭpŭŭs	47	0.034
Confucii	86	0.062	necdum	45	0.032
Christŭs	80	0.057	mĕmŏrŏ	44	0.032
rēvēră	80	0.057	necnon	44	0.032
sinarum	77	0.055	sinae	42	0.03
ōdă – ōdē	68	0.049	vĭcissim	39	0.028
nīmīrum	66	0.047	ĭdentĭdem	38	0.027
existŏ	65	0.047	īlĭco	38	0.027
Confucio	64	0.046	ĕquĭdem	37	0.027
constanter	57	0.041	[…]		
perfectĭo	57	0.041	observantĭă observŏ	30	0.022
Confucium	56	0.04	sŭpĕrŭs	30	0.022
Christĭānŭs	54	0.039	Eurōpă	29	0.021
mūsĭcă/ŭs	54	0.039	ipsemet	29	0.021
sinicae	52	0.037	ŏboedĭentĭă	28	0.02

Tabella 11c: Lemmi di B – Lemmi di A.

Parole	Ricorrenza	Frequenza %	Parole	Ricorrenza	Frequenza %
numquīs – numquam	152	0.122	Cătо	18	0.014
Caesar	101	0.081	servŭs servŭs servŏ	18	0.014
ădĭcĭŏ	68	0.055	tormentŭm	18	0.014
servŭs	57	0.046	dēbĭtor	17	0.014
oblĭgŏ	43	0.034	hostīlĭs	17	0.014
Dĕŭs – dīves	39	0.031	pătĭentĭă	17	0.014
condĭcĭo	35	0.028	prōĭcĭŏ	17	0.014
sŭpervăcŭŭs	31	0.025	rĕsistŏ	17	0.014
sānŏ	30	0.024	cădŏ – caedŏ	16	0.013
concŭpiscŏ	28	0.022	consŭlātŭs	16	0.013
ēmittŏ	27	0.022	dēlĭcātŭs dēlĭquŏ	16	0.013
flĕŏ	24	0.019	Ĕpĭcūrŭs	16	0.013
Hercŭles – Hercŭles	24	0.019	fĕrĭtas	16	0.013
con	23	0.018	lĭbĕt – lĭbŏ	16	0.013
īrācundŭs	23	0.018	summŏvĕŏ	16	0.013
ĕtĭamnunc	22	0.018	ădŭlescens	15	0.012
nēquĭtĭă	22	0.018	audācĭă	15	0.012
saevĭtĭă	21	0.017	custōdĭŏ	15	0.012
occŭpātĭo	20	0.016	Pompēiŭs	15	0.012

Tabella 11c (continua)

Parole	Ricorrenza	Frequenza %	Parole	Ricorrenza	Frequenza %
dīdūcŏ	19	0.015	Ălexandĕr	14	0.011
lībertas – lībertŭs	19	0.015	Chrȳsippŭs	14	0.011
Marcĭă	19	0.015			

Tabella 11d: Parole comuni ad A e B più frequenti in A.

Parole	A%	B%	A%|B%	Parole	A%	B%	A%|B%
discĭpŭlŭs	0.226	0.002	140.863	signĭfĭcŏ	0.002	0.061	0.026
doctrīnă	0.105	0.001	130.578	quattŭor	0.002	0.06	0.02
ergā	0.07	0.001	87.648	mānŭs	0.001	0.028	0.029
quippe	0.068	0.001	84.965	rătĭōnālĭs	0.001	0.027	0.03
planŭs- plānē	0.062	0.001	77.81	măgister	0.005	0.158	0.03
Ĭtŏ-ĭtem	0.182	0.002	75.723	săcrĭfĭcĭŭm sacrificiŭm	0.001	0.026	0.031
hūiusmŏdi	0.058	0.001	72.444	consentānĕŭs	0.001	0.024	0.033
perfectŭs-perfĭcĭŏ	0.058	0.001	72.444	dēgŏ	0.001	0.024	0.033
dēclārŏ	0.051	0.001	63.5	instar	0.002	0.071	0.034
praefectŭs-praefĭcĭŏ	0.099	0.002	61.711	intĕgrĭtas	0.001	0.024	0.034
ĭŏn	0.044	0.001	55.451	ŏboedĭŏ	0.001	0.024	0.034
rĭtŭs	0.123	0.002	51.277	subdŏ	0.004	0.116	0.035
quōdammŏdo	0.001	0.038	0.021	stătŭŏ	0.001	0.023	0.035
dĭcĭo	0.001	0.037	0.022	conclūdŏ	0.001	0.022	0.037
percontŏr/o	0.001	0.035	0.023	sŭperstĭtĭo	0.001	0.022	0.037
instĭtūtĭo	0.001	0.034	0.024	rătĭōnālĭs	0.001	0.027	0.03

Tabella 11e: Parole comuni ad A e B più frequenti in B.

Parole	A%	B%	A%|B%	Parole	A%	B%	A%|B%
ingrātŭs	0.001	0.134	0.011	abdūcŏ	0.019	0.001	26.835
contŭmēlĭă	0.001	0.044	0.016	incurrŏ	0.018	0.001	24.599
bĕnĕfĭcĭŭm vĕnēfĭcĭŭm	0.012	0.565	0.022	sōlācĭŭm	0.035	0.001	24.599
ignoscŏ	0.001	0.032	0.022	advŏcŏ	0.017	0.001	23.48
īră	0.004	0.149	0.024	excŭtĭŏ	0.017	0.001	23.48
īrascŏr	0.003	0.111	0.026	tangŏ	0.017	0.001	23.48
vulnus	0.001	0.026	0.027	pŭtŭspŭtŏ	0.05	0.002	23.108
nŏcĕŏ	0.002	0.071	0.03	effŭgĭŏ	0.016	0.001	22.362
pātrĭmōnĭŭm	0.001	0.022	0.033	glădĭŭsglădĭŭm	0.016	0.001	22.362
rĕmĕdĭŭm – rĕmĕdĭŏ	0.001	0.02	0.036	irrītŏirrĭtŏ	0.016	0.001	22.362

c) Distanza tra testi

Un'altra possibilità di analisi lessicale, spesso impiegata nelle ricerche di *authorship*[53] *attribution* è l'analisi della distanza tra testi sulla base di tecniche di statistica multivariata[54] che fanno uso delle *most frequent words*.

Tra i programmi che permettono di lavorare tramite interfaccia grafica i migliori a mio avviso sono *Lexos*,[55] e il pacchetto *Stylo*[56] di *R*.[57] *Lexos* ha il vantaggio di offrire un flusso di lavoro integrato tramite strumenti di pre-elaborazione (selezione dei testi, segmentazione, scrub), analisi e visualizzazione. D'altra parte *Stylo with R* fornisce, oltre all'output grafico, la lista delle configurazioni immesse, l'elenco degli elementi presi in esame nell'analisi e la descrizione delle relazioni tra nodi e archi nel grafico.[58]

Cluster analysis

Tra gli strumenti di analisi disponibili in *Lexos* e *Stylo with R* mi concentrerò sul *tool* per la *Cluster analysis*, un'analisi che mira a raggruppare unità tra loro eterogenee (i testi in questo caso) in sottoinsiemi tendenzialmente omogenei. Le unità statistiche vengono suddivise in gruppi (i *cluster*)[59] a seconda del loro livello di

53 Per un contributo recente di ampio respiro sulle tecniche di *authorship attribution* cf. Nini (2023).

54 La statistica multivariata (*multivariate statistics*) è un ramo della statistica che si occupa dell'analisi simultanea di più variabili, al fine di comprendere le relazioni complesse che possono sussistere tra di esse. A differenza della statistica univariata (*univariate statistics*), che si concentra su una singola variabile alla volta, la statistica multivariata esplora le interconnessioni tra molteplici variabili, Rientrano tra le tecniche di statistica multivariata: analisi della correlazione canonica (*Canonical-Correlation Analysis*) e analisi delle componenti principali (*Principal Component Analysis*); analisi fattoriale (*Factor analysis*); analisi delle corrispondenze (*Correspondence Analysis*); analisi dei cluster (*Cluster analysis*); analisi discriminante (*Linear Discriminant Analysis*); analisi di regressione multidimensionale (*Regression analysis*). Per un'introduzione alla statistica multivariata applicata all'analisi dei dati linguistici si veda Baayen (2008), capp. 5–6.

55 Kleinman *et al.* (2019).

56 Eder *et al.* (2015). Sul pacchetto *Stylo* cf. anche Eder *et al.* (2016).

57 Per l'utilizzo di R come ambiente software per il calcolo statistico si veda Baayen (2008), cap. 1.

58 Nella descrizione di un elemento grafico sono nidificate le dichiarazioni di nodi e archi. Un nodo viene dichiarato con l'elemento *node* e un arco con l'elemento *edge*. Ogni *edge* deve definire i suoi due *endpoint* con gli attributi XML *source* e *target*. Il valore di *source* e *target* deve essere l'identificatore di un nodo nello stesso documento. L'attributo XML *type* è facoltativo e dichiara se l'*edge* è diretto, non diretto o reciproco (diretto dall'origine alla destinazione e dalla destinazione all'origine). Una variabile di peso fornisce un valore per ogni osservazione in un set di dati. Il peso di un arco in un multigrafo è la somma dei gradi dei suoi vertici estremi. Per un'introduzione agli strumenti di visualizzazione dei dati si veda Baayen (2008), cap. 2.

59 La nozione di *cluster* non può essere definita con precisione, il che è uno dei motivi per cui esistono così tanti algoritmi di *clustering*. Il comune denominatore è quello di un gruppo di *data*

"somiglianza", calcolato a partire dai valori che una serie di variabili prescelte (in questo caso i token, i *word n-grams* o i *characters n-grams*) assumono in ciascuna unità. Dal momento che la punteggiatura influenza la tokenizzazione (che si basa sugli spazi bianchi, assenti tra parola e segno diacritico), è consigliabile utilizzare preventivamente lo strumento di scrub dei testi che permette di eliminare punteggiatura, maiuscole, numeri e altri elementi selezionabili di volta in volta. È necessario scegliere un *linkage-method* ("metodo di raggruppamento"), ovvero il metodo con cui i cluster sono uniti due a due,[60] e una *metric* ("metrica"), ovvero un metodo di calcolo di distanza tra i due vettori rappresentanti i testi. Viene perciò costruita una gerarchia di cluster di numero (de)crescente, visualizzabile mediante una rappresentazione grafica detta dendrogramma.

Per illustrare il procedimento tramite un esempio concreto, ho selezionato dal *corpus* i testi che rientrano nel genere dei resoconti di viaggio: l'*Itinerarium* di Guglielmo di Rubruck, la versione latina del manoscritto Z del *Milione* di Marco Polo, la *Relatio* di Giovanni de' Marignolli, la *Relatio de mirabilibus* di Odorico di Pordenone. A titolo di confronto ho incluso nell'analisi anche dei testi esterni al corpus: due resoconti di viaggio in Terra Santa di XII-XIII secolo (il *De Hierosolymitana peregrinatione* di Pietro di Blois e la *De Hierosolymitano itinere historia* di Pietro Tudebode), due itinerari di età precedente (le anonime *Navigatio Brendani* e *Peregrinatio Egeriae*) e un testo di genere letterario differente (il *De amicitia* di Marco Tullio Cicerone).[61]

Per effettuare la *Cluster analysis* ho scelto la distanza del coseno[62] che viene consigliata da *Lexos* per l'analisi di corpora e il metodo di raggruppamento "pesato", migliore degli altri per comparare testi di lunghezza diversa. Sia *Stylo with*

objects. L'analisi dei cluster in sé quindi non è costituita da un algoritmo specifico, ma è un compito che può essere svolto tramite vari algoritmi, che differiscono in modo significativo nella loro comprensione di ciò che costituisce un cluster e di come trovarlo in modo efficiente. Per una introduzione all'analisi dei cluster applicata alla linguistica si veda Moisl (2015).

60 Si distingue tra *single-linkage* (si prende il cluster che contiene il punto — ad esempio la frequenza di un termine — più vicina al cluster che si sta analizzando e lo si unisce ad esso), *complete-linkage* (si prendono i due punti più lontani — uno per ogni cluster — tra il cluster considerato e ciascuno degli altri cluster e il cluster con la distanza minore al cluster considerato è unito a quello), *average-linkage* (una via di mezzo tra i due: prende la distanza media tra tutti i punti in ogni cluster e usa la più corta distanza media per decidere con quale cluster deve essere unito quello preso in esame) e *weighted-linkage* (si esegue il calcolo tramite il metodo *average*, ma si ponderano le distanze sulla base del numero di termini nel *cluster*). Per una trattazione approfondita si veda Moisl (2015), cap. 4.

61 I testi elencati sono stati scaricati da *Corpus Corporum*.

62 La *cosine similarity* ("similarità del coseno") è una tecnica per la misurazione della similitudine tra due vettori effettuata calcolando il coseno tra di loro. Il coseno di uno dei due angoli interni adiacenti all'ipotenusa è definito come il rapporto tra le lunghezze del cateto adiacente all'angolo e dell'ipotenusa. Per una trattazione approfondita si veda Moisl (2015), cap. 4.

R che *Lexos* restituiscono risultati coerenti sulla base delle 500 *most frequent words* (MWF): dal punto di vista lessicale il *De amicitia* — come da previsione — si isola dagli altri testi, il *Milione* e l'*Itinerarium* presentano forti analogie, che si riscontrano in maniera più lieve con la *Relatio* e i due testi relativi a Gerusalemme. La *Navigatio Brendani* e la *Peregrinatio Egeriae* fanno gruppo a sé insieme alla *Relatio de mirabilibus*, che in maniera inattesa sembra differire nel lessico dagli altri testi della stessa epoca (cf. Figura IVa e IVb).

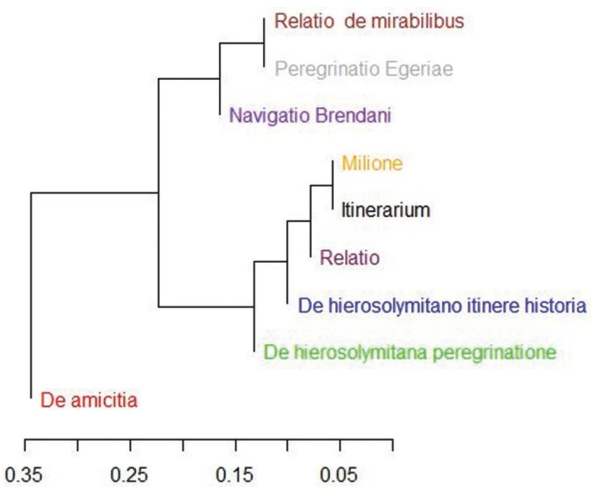

Figura IVa: Risultato *Cluster analysis* tramite *Stylo with R* (*cosine metric*, 500 mfw).

Tuttavia, se si apportano lievi modifiche ai parametri, si osservano delle variazioni nel dendrogramma. Ad esempio, impostando la metrica come euclidea[63] i due resoconti di viaggio in Terra Santa fanno cluster a sé (cf. Figura IVc), mentre scegliendo il metodo di raggruppamento "completo" la *Relatio de mirabilibus* si colloca tra gli altri itinerari di XII–XIII secolo (cf. Figura IVd).

Il problema preliminare all'applicazione della *Cluster analysis* consiste dunque nel comprendere e stabilire quali siano i parametri più adatti per la tipologia di testi da sottoporre a indagine. Secondariamente, risulta complesso comprendere su quali basi e calcoli si fondano i risultati ottenuti che, tuttavia, possono fornire interessanti spunti di indagine in merito al perché alcuni testi siano più simili tra loro rispetto ad altri.

63 Nel calcolo della distanza euclidea, i due vettori sono visualizzati come lati in un triangolo e l'ipotenusa tra queste due linee è misura della distanza tra i due documenti. Per una trattazione approfondita si veda Moisl (2015), cap. 4.

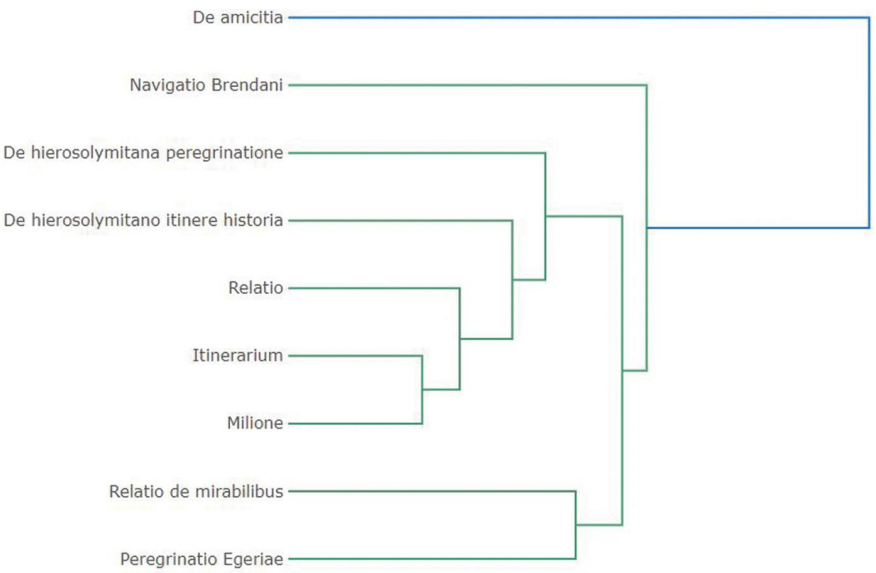

Figura IVb: Risultato *Cluster analysis* tramite *Lexos* (*cosine metric*, 500 mfw, *weighted linkage*).

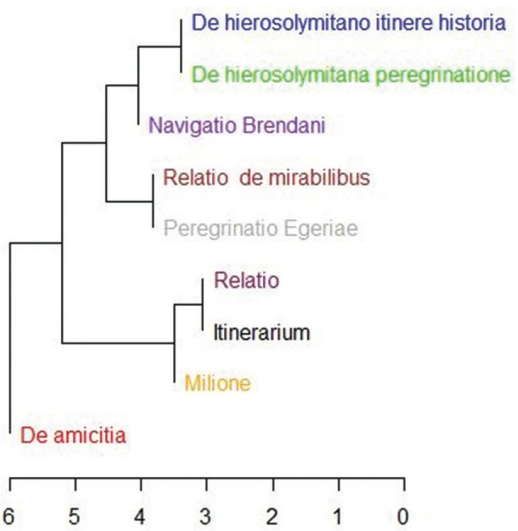

Figura IVc: Risultato *cluster analysis* tramite *Stylo with R* (*euclidean metric*, 500 mfw).

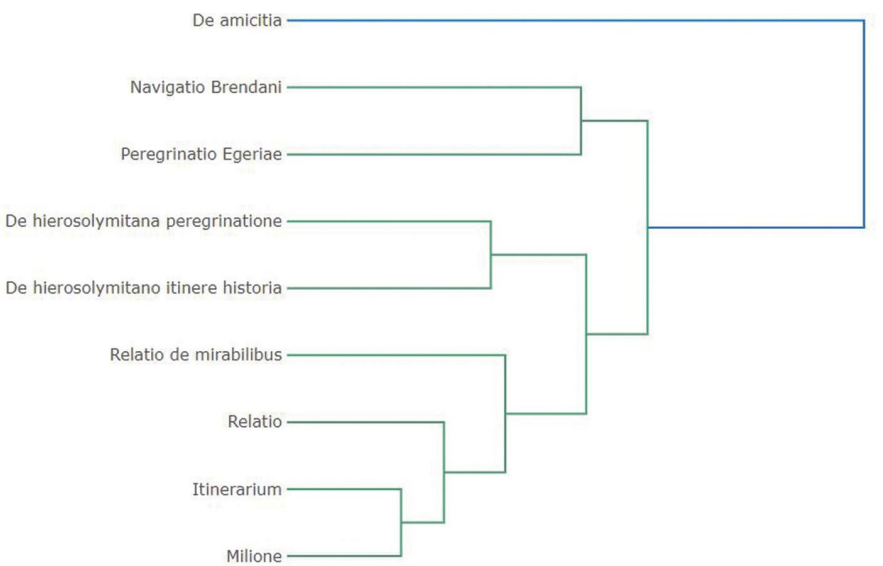

Figura IVd: Risultato *cluster analysis* tramite *Lexos* (*cosine metric*, 500 mfw, complete linkage).

2.1.2 Morfologia

È possibile analizzare i testi con metodi quantitativi anche dal punto di vista morfologico. Questo tipo di indagine, mettendo in risalto le differenze nell'utilizzo della morfologia tra le diverse opere, consente di meglio caratterizzare lo stile di un testo o le predilezioni e idiosincrasie del suo autore.

Preliminari a questo tipo di analisi sono la lemmatizzazione e il *PoS-tagging* dei testi. Il software che consente risultati migliori tra quelli attualmente disponibili è *UDpipe*,[64] che a differenza del più classico Treetagger produce un output con etichette di *PoS* conformi allo *Universal PoS tagset*, cioè lo standard oramai accettato dalla comunità scientifica. L'output del *PoS-tagging* consiste in un'analisi frase per frase e parola per parola degli elementi morfo-sintattici del periodo, per il cui conteggio e elaborazione ho utilizzato Excel.

a) Frequenza delle voci morfologiche in uno o più testi

Per illustrare le tecniche e i sistemi tramite cui è possibile studiare la morfologia con metodi computazionali ho fatto nuovamente ricorso ai resoconti di viaggio,

64 Wijffels *et al.* (2017).

così da indagare se le somiglianze che sono state individuate sul piano lessicale nascondano anche una somiglianza dal punto di vista morfologico. Dei quattro resoconti di viaggio presenti nel corpus, ho calcolato la frequenza percentuale di *PoS* e casi (cf. Tabella 12a e b).

Tabella 12a: Frequenza percentuale utilizzo PoS negli Itineraria del corpus.

	Relatio	Relatio de mirabilibus	Itinerarium	Milione	Dev. Standard
ADJ	8,45	6,47	5,67	6,67	1,17
ADP	11,81	9,68	10,67	9,82	0,98
ADV	6,96	8,76	7,42	8,21	0,81
AUX	2,89	4,12	3,23	3,88	0,57
CCONJ	7,89	7,14	7,60	9,64	1,10
DET	4,83	10,90	7,50	8,27	2,50
NOUN	30,50	20,50	22,04	23,46	4,42
PRON	4,04	6,24	7,73	4,95	1,60
PROPN	1,82	0,50	0,78	0,41	0,65
SCONJ	3,02	4,13	4,57	4,01	0,66
VERB	15,94	18,41	20,40	18,34	1,82

Tabella 12b: Frequenza percentuale utilizzo dei modi negli Itineraria del corpus.

	Relatio	Relatio de mirabilibus	Itinerarium	Milione	Dev. Standard
IND	36,23	48,89	50,86	55,59	8,27
SUB	5,15	9,08	10,94	9,81	2,52
INF	15,64	15,69	14,71	10,28	2,57
PTC	38,95	23,58	20,69	19,96	8,91

Questi dati forniscono una panoramica oggettiva delle caratteristiche morfologiche dei testi. Ad esempio, dal confronto delle quattro opere tra di loro si nota che la *Relatio* di Giovanni de' Marignolli è il testo più straniante: c'è un largo uso del participio a scapito delle subordinate esplicite con congiunzione subordinante. Inoltre l'indicativo è poco utilizzato così come il congiuntivo. L'autore usa molti aggettivi, preposizioni, nomi. La versione latina del manoscritto Z del *Milione* invece utilizza preferenzialmente l'indicativo e la congiunzione coordinante mentre ricorre di rado a infiniti e participi.

b) Distanza tra testi sulla base della morfologia

Allargando la prospettiva a tutti i resoconti di viaggio già considerati nella *Cluster analysis*,[65] ho calcolato la presenza di alcune caratteristiche morfologiche in tutti i testi (cf. Tabella 12c).

Ho utilizzato poi *Clustvis*,[66] un *tool* online che, tramite una tecnica chiamata *Principal Component Analysis* (PCA),[67] mi ha permesso di visualizzare i testi su un grafico bidimensionale in base alle loro caratteristiche morfologiche. L'idea fondamentale della PCA è di proiettare dei dati multidimensionali su due dimensioni nel modo "migliore" possibile, cioè preservando più possibile la varianza. Nel nostro caso, a partire da un dataset costituito di 9 unità (i testi) e 16 variabili (ind, sub, inf etc.), la PCA identifica delle componenti astratte la cui correlazione con le singole variabili è presentata nella tabella denominata *component loadings* (cf. Tabella 12d). Delle componenti principali si selezionano poi le due (PC1, PC2) che consentono una rappresentazione con la minore perdita di informazione possibile. Si tratta quindi delle due componenti con varianza più elevata. Nei nostri campioni, queste due componenti "migliori" rappresentano insieme il 56% dell'intera variazione (25,5% e 30,3%). In altre parole solo più della metà dell'intera variazione.

Tramite questa analisi si constata che sulla base delle caratteristiche morfologiche i testi si distribuiscono e agglomerano in maniera decisamente diversa rispetto quanto era stato osservato per il lessico (cf. Figura V): la *Relatio* di Giovanni de' Marignolli si stacca dalle altre relazioni di viaggio, i due itinerari in Terra Santa mostrano differenze profonde tra di loro e con gli altri testi, il *De amicitia* fa comunque gruppo a sé mentre la *Peregrinatio Egeriae* si avvicina al Milione e all'*Itinerarium*. In conclusione la PCA delle componenti morfologiche non dà risultati omogenei a quelli della *cluster analysis* delle componenti lessicali, mostrando come morfologia e lessico non vadano necessariamente di pari passo.

2.1.3 Sintassi

Le ricerche quantitative possono essere applicate anche a elementi della sintassi – campo nel quale le ricerche di linguistica latina sono state più vivaci e frequen-

65 Cf. *supra*, cap. 1C, cluster analysis.

66 Metsalu/Vilo (2015).

67 Secondo il metodo di Roelli cf. Roelli (2021).

Tabella 12c: Caratteristiche morfologiche del gruppo selezionato di *Itineraria*.

	De amicitia	Relatio	Relatio de mirabilibus	Itinerarium	Milione	Navigatio Brendani	Peregrinatio Egeriae	De Hyeroso limitana peregrinatione	De Hyeroso limitano itinere historia
IND	36.28	36.23	48.89	50.86	55.59	53.63	54.54	56.15	42.52
SUB	17.52	5.15	9.08	10.94	9.81	8.76	11.48	12.64	10.17
INF	23.75	15.64	15.69	14.71	10.28	11.46	10.10	9.17	14.54
PTC	19.78	38.95	23.58	20.69	19.96	21.77	22.82	19.24	28.92
NOM	32.45	23.39	33.38	26.01	31.01	29.03	38.67	31.26	2.53
GEN	11.37	15.44	9.74	9.82	9.65	14.11	9.43	21.27	1.37
ACC	27.51	26.68	29.62	36.39	30.61	29.32	24.87	24.84	3.42
ABL/DAT	28.13	34.31	27.03	27.64	28.48	27.46	26.85	22.59	2.68
ADJ	6.86	8.45	6.47	5.67	6.67	6.81	6.92	4.01	7.95
ADP/PREP	6.75	11.81	9.68	10.67	9.82	10.49	10.30	8.02	8.61
ADV	9.57	6.96	8.76	7.42	8.21	7.36	11.09	4.21	8.43
CONJ	11.84	10.91	11.27	12.17	13.65	10.05	11.73	11.22	11.38
NOUNS	21.62	32.32	21.00	22.82	23.87	28.03	24.00	34.27	24.40
PRON	9.35	4.04	6.24	7.73	4.95	5.11	6.45	5.01	6.47
VERB	18.90	15.94	18.41	20.40	18.34	19.66	16.18	17.43	19.55
AUX	4.92	2.89	4.12	3.23	3.88	2.61	5.03	3.21	2.97

Tabella 12d: *Component loadings.*

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
IND	0,09	-0,32	-0,22	-0,45	-0,07	-0,2	0,02	-0,03	-0,07
SUB	-0,36	-0,17	-0,17	0,24	-0,16	0,01	-0,02	-0,48	0,31
INF	-0,27	0,15	0,18	0,43	0,24	0,1	0,07	0,25	-0,14
PTC	0,23	0,35	0,26	0,13	-0,06	0,12	-0,3	0,13	0,29
NOM	-0,04	-0,43	0,27	0,02	-0,09	-0,21	0,15	0,14	0,63
GEN	0,24	-0,32	0	0,38	-0,04	0	0,09	-0,08	-0,19
ACC	0,02	-0,37	0,25	0,01	0,43	0,08	-0,06	0	-0,03
ABL/DAT	0,1	-0,3	0,42	0,13	0,22	0,08	0,17	-0,14	-0,18
ADJ	0	0,39	0,34	-0,06	-0,01	-0,03	0,51	-0,37	0,1
ADP/PREP	0,29	0,05	0,3	-0,31	0,27	-0,09	-0,45	-0,26	-0,02
ADV	-0,29	0,11	0,34	-0,25	-0,16	-0,31	0,08	-0,26	-0,19
CONJ	-0,19	-0,12	0,05	-0,32	-0,07	0,87	0,02	-0,11	0,06
NOUNS	0,38	-0,05	-0,13	0,28	-0,21	0,08	-0,03	-0,49	-0,21
PRON	-0,42	-0,01	-0,01	0,16	0,15	-0,1	-0,55	-0,28	0,06
VERB	-0,19	0,06	-0,32	-0,1	0,6	-0,07	0,24	-0,1	-0,09
AUX	-0,32	-0,16	0,26	-0,03	-0,37	-0,04	-0,11	0,19	-0,47

tate negli ultimi anni.[68] Obiettivo della prossima sezione sarà mostrare qualche strumento che consenta questo tipo di ricerche a livello di interfaccia e il suo funzionamento.

a) Strumenti

Anche per il *parsing* sintattico è possibile fare ricorso a *UDpipe*. In primo luogo è necessario scegliere un modello per il latino tra i cinque disponibili, corrispondenti ad altrettante *treebanks* latine di *training*. Per il latino medievale il modello

68 Penso in particolare allo sviluppo e ampliamento delle *treebanks*. Una *treebank* è un *corpus* di testi annotati con informazioni sintattiche. In una *treebank* ogni parola è annotata con informazioni sulla sua struttura grammaticale e sulle relazioni sintattiche con le altre parole della frase. Questa annotazione è spesso rappresentata sotto forma di albero sintattico, dove ogni nodo dell'albero rappresenta una parola e i collegamenti tra i nodi rappresentano le relazioni sintattiche.

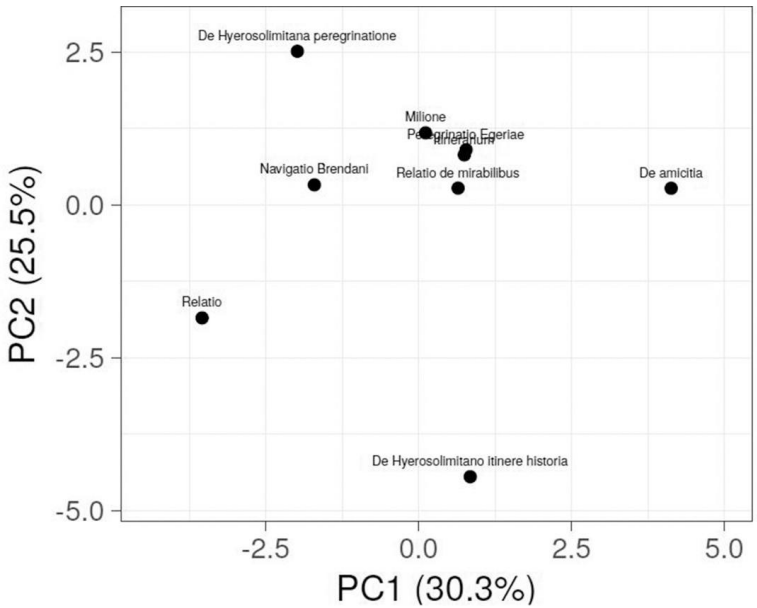

Figura V: PCA degli Itineraria.

preferibile è quello dell'*Index Thomisticus Treebank* (IT-TB).[69] Le etichette attribuite a ciascuna parola attraverso il processo di *parsing* riflettono i legami sintattici che intercorrono tra i membri della frase. Tuttavia, in ambito sintattico non conta il singolo nodo, bensì le relazioni tra nodi: il conteggio del numero di occorrenze di una singola etichetta quale *root* o *amod* non risulta significativo. Alcuni strumenti online, come *Tündra*,[70] permettono l'esecuzione di *query* più sofisticate, come la ricerca di uno specifico costrutto sintattico.

b) Esempi di ricerca di singoli costrutti sintattici

Ad esempio, possiamo analizzare i due casi del *cum* narrativo e dell'ablativo assoluto[71].

Supponendo di voler individuare quante volte in un testo viene utilizzato il *cum* narrativo, dovremmo cercare i *token* con POS verbale e modo congiuntivo che dipendono da un *token* con lemma *cum* e PoS di congiunzione subordinante (cf. Figura VIa).

69 Le altre *treebanks* sono la *Latin Dependency Treebank* (LDT), la *PROIEL Project treebank*, la *Late Latin Charter Treebank* (LLCT) e la *UDante treebank*.

70 Martens (2013).

71 Per un'applicazione stilometrica a fini attribuzionistici della ricerca di questi due costrutti in un *corpus* di autori classici si veda Field (2016).

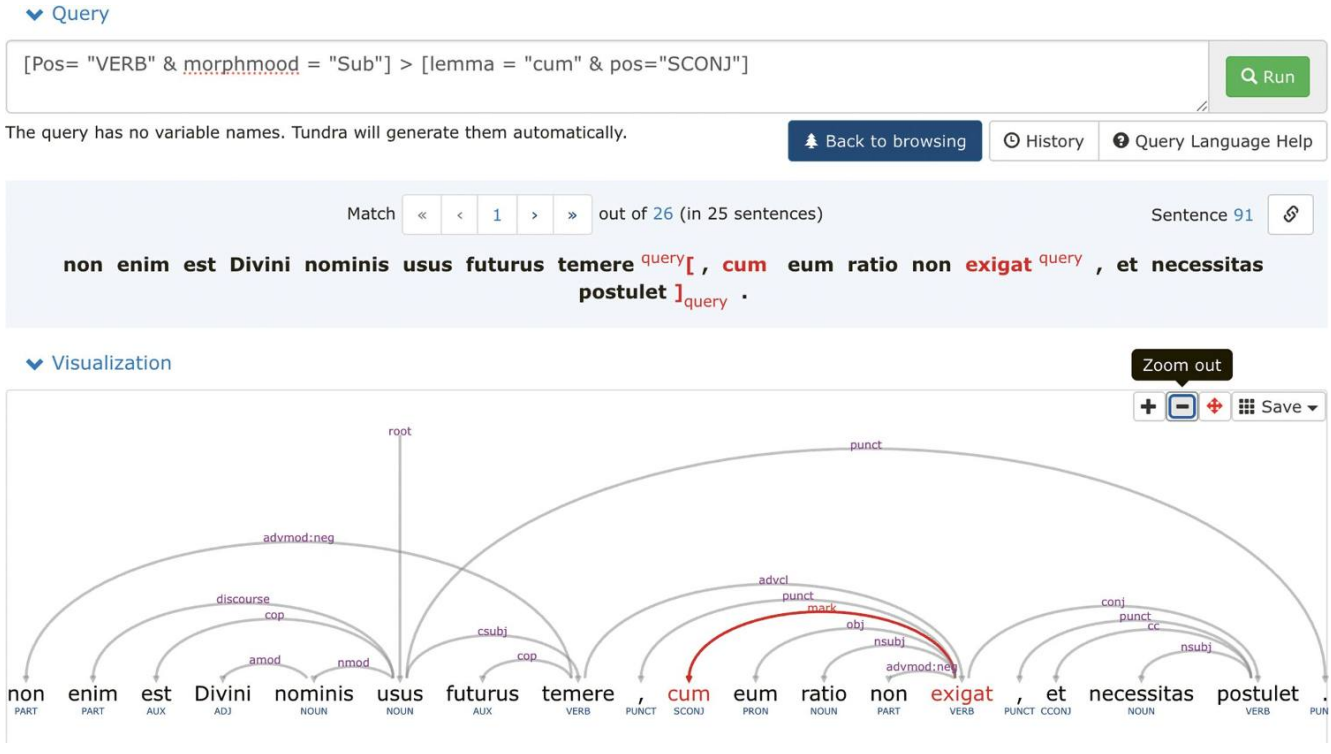

Figura VIa: Ricerca nel *Catechismus* dei *cum* narrativi con *Tündra.*

Ancora più complesso il caso dell'ablativo assoluto, la cui ricerca richiederebbe di individuare un termine con PoS verbale, forma verbale di participio, caso ablativo, numero singolare (o plurale) e relazione sintattica "advcl" (*adverbial clause modifier*) da cui dipende un termine con PoS di determinativo — nome o pronome — caso ablativo e numero singolare (o plurale) con relazione sintattica "nsubj" ovvero *nominal subject* (cf. Figura VIb)[72].

In generale, non esiste tuttavia un software che analizzi globalmente i testi dal punto di vista sintattico mettendone in evidenza i pattern nascosti.

2.1.4 Prospettive sulla sintassi: il topic modelling

L'ultimo livello praticabile di analisi del testo è quello semantico, che è anche il più complesso da studiare tramite analisi automatica. La sezione che segue presenterà il caso del *topic modelling*, modello statistico utilizzato per l'individuazione di strutture semantiche nascoste in un testo o in una raccolta di testi. Il modello lavora sul riconoscimento di cluster di parole o espressioni che si associano a costanti di contenuto. Due sono le tecniche di *topic modelling* principali:

– *Latent Semantic Analysis* (LSA);[73]
– *Latent Dirichlet Allocation* (LDA).[74]

72 Per una casistica completa dei criteri di ricerca automatizzata dell'ablativo assoluto in testi annotati morfologicamente e testi non annotati cf. Field (2016) 61–62.

73 La tecnica LSA si basa sull'ipotesi distributiva, che afferma che la semantica delle parole può essere colta osservando i contesti in cui le parole appaiono. L'assunto fondamentale è infatti che la semantica di due parole sarà simile se tendono ad apparire in contesti simili. Partendo dal presupposto che documenti simili conterranno approssimativamente la stessa distribuzione di parole, la LSA calcola la frequenza con cui le parole ricorrono nei documenti e nell'intero *corpus*. In questa tecnica, le informazioni sintattiche (*e.g.* l'ordine delle parole) e semantiche (*e.g.* la molteplicità di significati di una data parola) vengono ignorate e ogni documento viene trattato come un insieme di parole.

74 La tecnica LDA si basa anch'essa sull'ipotesi distributiva (ovvero argomenti simili fanno uso di parole simili) e l'ipotesi statistica mista (ovvero i documenti parlano di più argomenti per i quali è possibile determinare una distribuzione statistica). Lo scopo di LDA è mappare ogni documento del *corpus* in relazione a una serie di argomenti che coprano una buona parte delle parole nel documento. Anche la tecnica LDA ignora le informazioni sintattiche e tratta i documenti come insiemi di parole. Presuppone inoltre che a tutte le parole nel documento possa essere assegnata una probabilità di appartenenza a un argomento. La principale differenza è che LSA non presuppone alcuna distribuzione e quindi porta a rappresentazioni vettoriali più opache di argomenti e documenti.

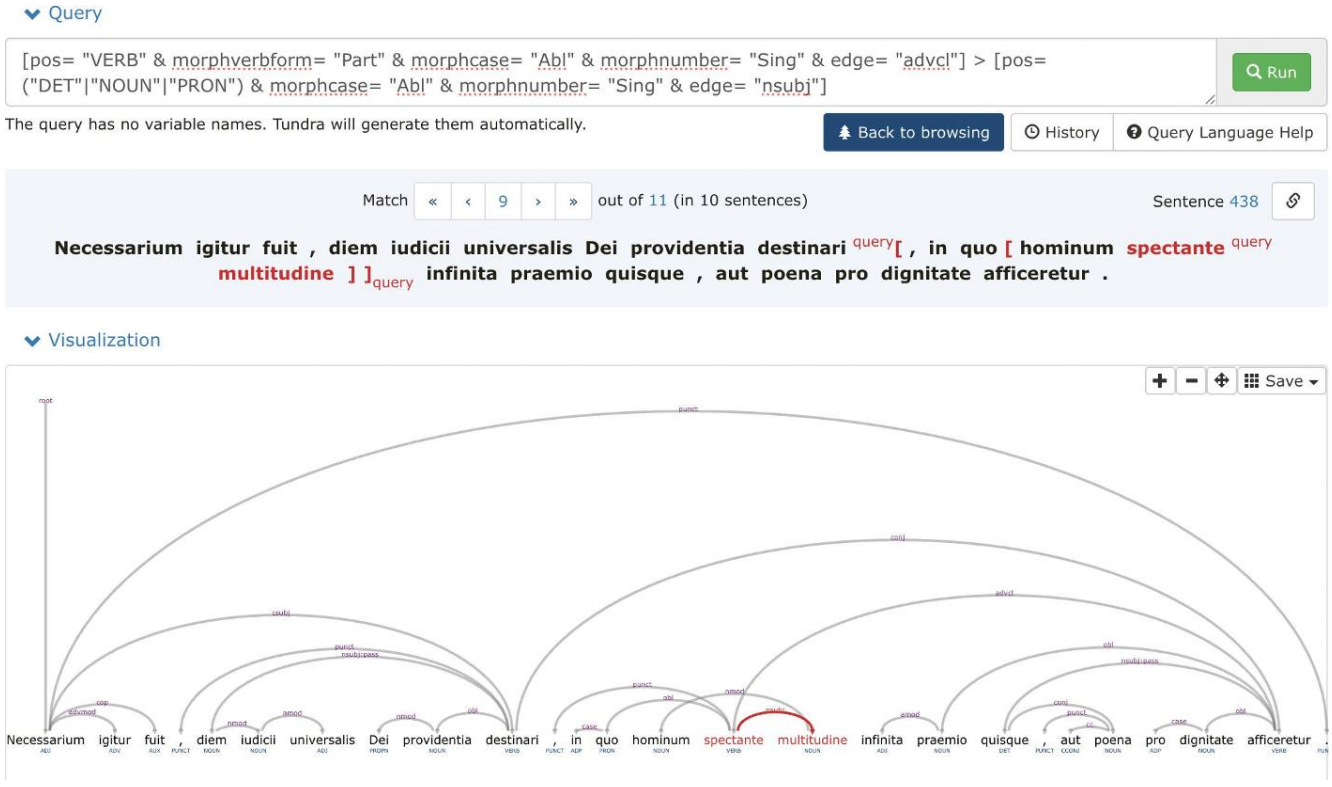

Figura VIb: Ricerca nel *Catechismus* dell'ablativo assoluto con *Tündra*.

a) Strumenti

Sulla tecnica LDA si basa *ToPān*,[75] software pensato per il *topic modelling* delle lingue classiche che non richiede conoscenze di programmazione. *ToPān* utilizza file strutturati come CVS e presenta la possibilità di eseguire *data import*, lemmatizzazione e *parsing*, ripulitura del testo dalle *function words* e visualizzazione dei risultati in un unico ambiente.

Una volta effettuate queste operazioni preliminari è necessario scegliere il numero di *topic* da individuare nel testo e regolare i due parametri che controllano la somiglianza tra documento e argomento:

- α: un valore basso di α assegnerà meno argomenti a ciascun documento, mentre un valore alto di α avrà l'effetto opposto;
- β: un valore basso di β utilizzerà meno parole per modellare un argomento mentre un valore alto utilizzerà più parole, rendendo così gli argomenti più simili tra loro.

b) Esempi

A titolo illustrativo ho processato l'*Icon regia* di Joachim Bouvet scegliendo di individuarvi 20 *topic*. I risultati sono fruibili tramite visualizzatore o in tabelle. Lo strumento di visualizzazione affianca a una rappresentazione della distanza e sovrapposizione tra *topic* (cf. Figura VIIa), una vista riassuntiva del singolo *topic* limitata ai suoi 25 termini più caratteristici. Se si osserva ad esempio il *topic* 14, si nota che è connotato da termini quali *mathematicus, calculus, geometria, elementum*. I *topic*, nonostante quel che ci si aspetterebbe, non hanno un nome: è l'insieme di termini che deve suggerire a chi conduce l'indagine quale sia l'argomento del *topic* (cf. Figura VIIb).

3 Conclusioni

In sintesi, il lavoro computazionale sui corpora di ELA e SERICA presenta diverse sfide, dovute in parte alla natura dei testi oggetto di esame, in parte agli strumenti attualmente a disposizione:

- l'impossibilità di una lemmatizzazione automatica completa a causa della presenza nei testi di grafie medievali difformi, ma soprattutto di numerose traslitterazioni dal cinese per lo più di nomi propri e aggettivi relativi alle popolazioni orientali;

75 Köntges (2016).

Figura VIIa: Rappresentazione della distanza tra topic tramite scaling multidimensionale.

- errori conseguenti nell'analisi morfologica e sintattica automatizzata da parte del software;
- grande variazione diacronica nel corpus;
- grande varietà di temi, generi letterari, ambiti di afferenza dei testi;
- difficoltà nell'individuazione di termini di paragoni adeguati nella letteratura occidentale in lingua latina per alcune tipologie di testi (*e.g.* le relazioni sugli usi e costumi orientali).

Nonostante queste difficoltà, l'approccio quantitativo al corpus presenta numerosi vantaggi e interessanti possibilità. Innanzitutto, il grande numero di testi del database che non hanno ancora ricevuto attenzioni critiche induce ad auspicare

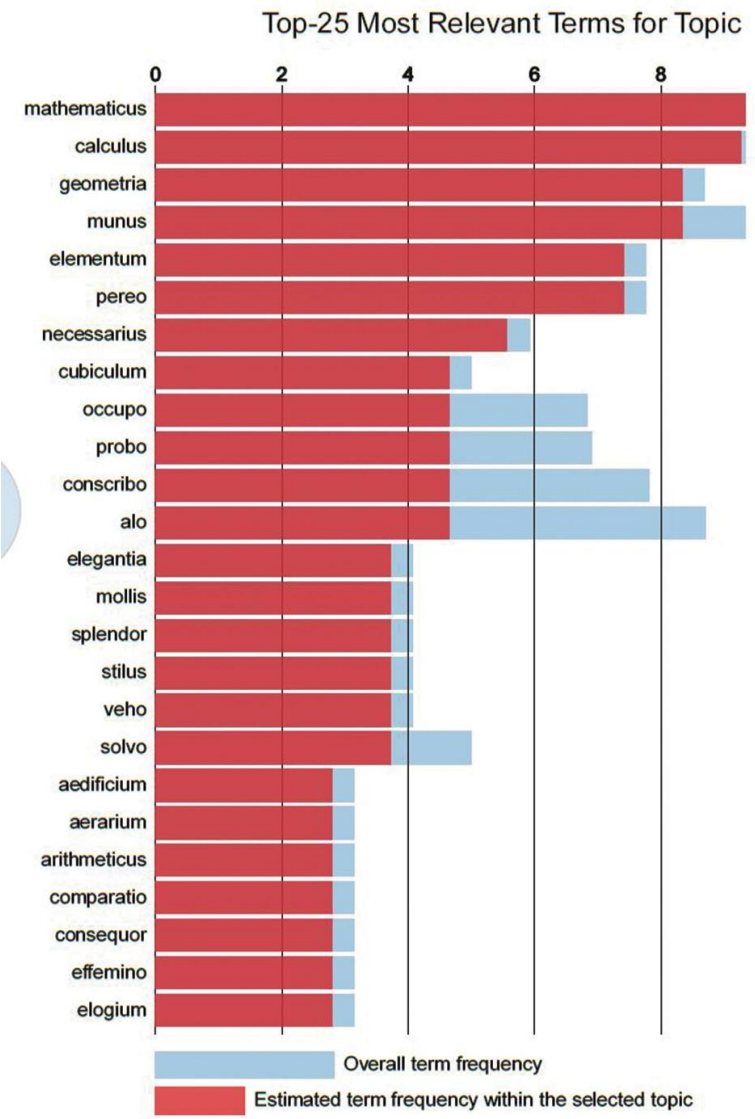

Figura VIIb: Topic 14.

quanto meno un'indagine complessiva che ne metta in luce le caratteristiche comuni. In un'ottica più ampia, il confronto di questo corpus di argomento orientale con altri *corpora* latini medievali e moderni potrebbe fornire importanti spunti sull'influsso esercitato dal contatto con l'Oriente sulla lingua, sulla morfologia o

sulla sintassi degli autori. Inoltre, l'analisi dei dati quantitativi potrebbe rivelarsi altrettanto vantaggiosa nello studio specifico di singoli testi.

La tipologia di analisi che mi sembra più promettente è quella lessicale, in quanto è necessario un contatto molto intenso e prolungato con una lingua diversa perché si verifichino cambiamenti nella morfologia e nella sintassi. D'altro canto, alcuni dei missionari gesuiti hanno trascorso in Oriente lunghi periodi, apprendendo e studiando le lingue autoctone: quanto di questo processo di acculturazione si è ripercosso sul loro latino? D'altro canto, sicuramente degna di nota è la forma di adeguamento del lessico latino alla descrizione di nuove realtà attraverso risemantizzazione o neoplasia.

Infine, un'ultima possibilità di indagine che non mi è stato possibile condurre, ma che potrebbe offrire interessanti riscontri è quella della comparazione intertestuale. Manca tuttavia un software che permetta un confronto tra un testo target e i testi classici e mediolatini del *Corpus Corporum* alla ricerca di affinità di sintagmi, locuzioni o intere espressioni. Da questo punto di vista sarebbe auspicabile una collaborazione con il progetto *Tesserae* dell'Università di Buffalo.