Estratto dalla bozza del cap.
06- ANALISI DIGITALE DEL TESTO

di Francesco Stella

in stampa (in traduzione inglese)
per Presses Universitaires de Grenoble

2017

Una volta acquisita o prodotta l'edizione del testo, del documento o del gruppo di testi o documenti medievali di cui vogliamo occuparci può cominciare l'attività più importante, che costituisce il fine di gran parte delle ricostruzioni digitali: usare il testo per ricerche sul loro stile e la loro lingua e più spesso per indagini sui loro contenuti storici, memoriali, antropologici, sociali, filosofici, psicologici, ecc. Questa amplissima serie di finalità si realizza con la lettura del testo e la sua schedatura, e da qualche anno si può effettuare anche con strumenti informatici che sono in grado di trattare masse di dati non gestibili dall'osservazione umana e di estrarne le informazioni che via via ci interessano. Questo è il *data mining* (nel nostro caso: *text mining* o *textométrie*), la cui base operativa è sempre un'indagine quantitativa che produce dati statistici. Proviamo a illustrarne le procedure orientandoci a una finalità pratica dellla ricerca e basandoci sulla nostra esperienza, senza alcuna pretesa di fornire un contributo scientifico necessariamente innovativo o esaustivo.

1. La linguistica quantitativa
(omissis)

Storia delle tecniche di analisi
(Omissis)

2. Applicazioni alle fonti letterarie
(omissis)

Processi di elaborazione testuale
(omissis)
In primo luogo le informazioni quantitative estraibili da un testo si riconducono a due dati di base*: trovare una "parola"* in masse testuali molto ampie e *calcolare il numero di occorrenze* di una parola o di un lemma all'interno di un testo. Vedremo che la gran parte delle realizzazioni attive oggi si basa, in ultima analisi, in forma sempre più nascosta e addobbata, sull'indice di frequenze.

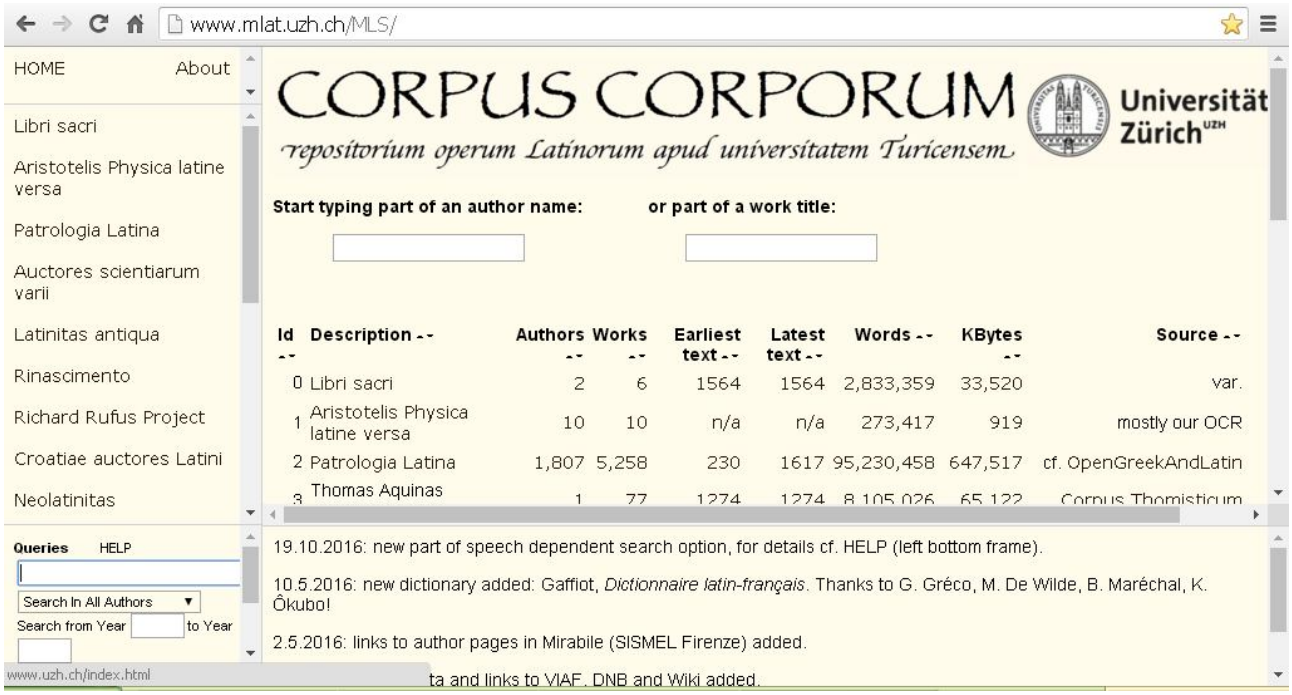
3. La terminologia di base: parola (word), forma (type), occorrenza (token).
Il primo approccio possibile, base di tutti gli altri, è la procedura per "parole", che si può automatizzare e che nelle ricerche più avanzate viene raffinata procedendo per concetti o ambiti semantici, meno facilmente automatizzabili. In questo tipo di studi si intende per "parola" (word) qualsiasi sequenza di lettere separata tramite numeri, spazi o segni non alfabetici da altre sequenze (ad es. *potestati*), per occorrenza (token) le apparizioni di questa parola in un testo o autore (10, 20), per forma (type) le varietà con cui un termine compare (*potestas, potestatis, potestatem* ecc.), per "lemma" il lemma linguistico cui le varie forme sono ricondicibili (*potestas*), significato diverso da quello dell'inglese *lemma* che significa anche "lezione del testo originale o del testo-base
(omissis)
In generale i riferimenti teorici di stilistica computazionale sono quasi tutti orientati al riconoscimento di paternità autoriale di testi "anonimi" e quasi mai all'analisi stilistica vera e propria: da questo punto di vista più che i metodi teorici sono necessari un'ipotesi di ricerca, che può partire solo da una seria conoscenza degli autori studiati, e una pratica strumentale dei software da utilizzare. Un'eccezione sono, per la letteratura mediolatina, Philippart de Foy-D'Angelo 2015 e Stella 2008, Stella 2010, Stella 2012, Stella 2015, Ricci 2015.

Le basi metodologiche più elementari da applicare all'analisi dei *corpora* ma anche di gruppi di testi letterari sono la distribuzione delle frequenze di parole o *clusters* (sequenze di parole) o *n-grams* (sequenze di parole o di caratteri), la misura della varietà lessicale (rapporto type/token o indice di Guiraud), analisi di parole-chiave, concordanze, cooccorrenze. In realtà l'analisi letteraria digitale ha sviluppato ormai numerosi altri sistemi (network analysis, topic modelling, sentiment analysis ecc.) su cui cerchiamo di fornire qualche ragguaglio.

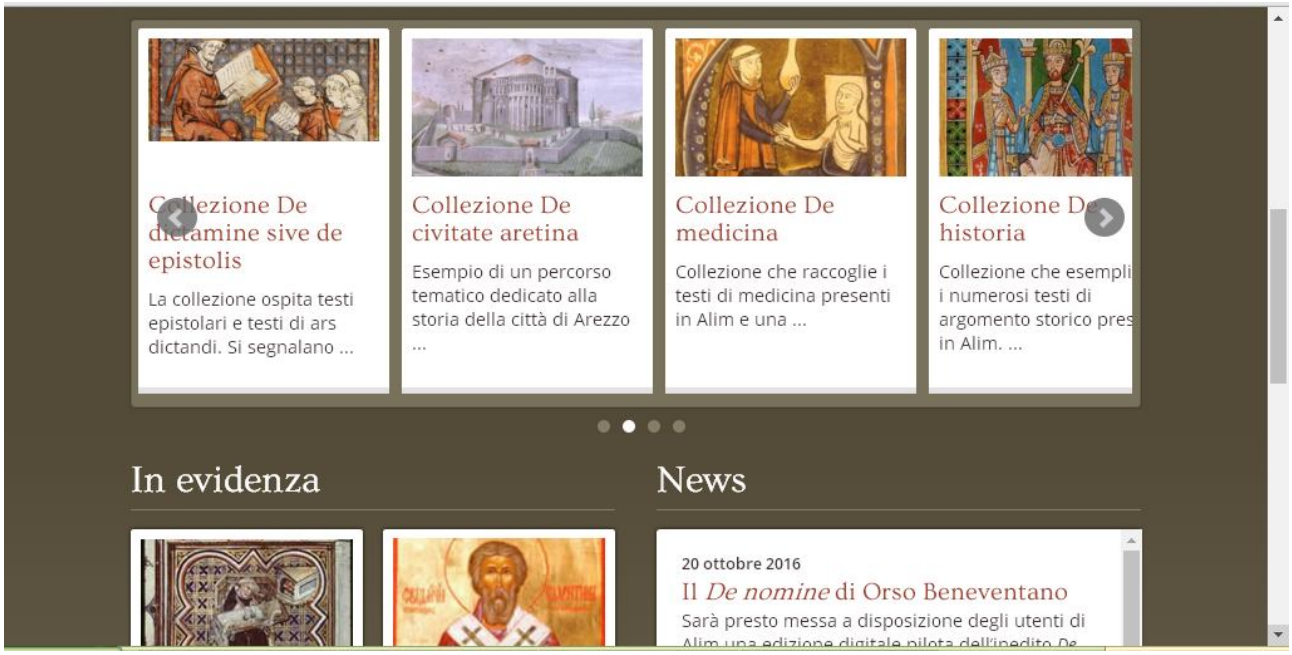
4. Il primo passo: trovare il testo. Le Biblioteche Digitali.
Naturalmente il primo testo su cui eseguire le ricerche è quello dell'edizione che abbiamo curato secondo i criteri esposti nei capitoli precedenti. Se però il testo su cui si vuole effettuare l'analisi non è quello da noi edito o se occorrono altri testi o corpora (cioè gruppi di testi) per consentire comparazioni e ricerche di "big data", allora conviene rivolgersi alle biblioteche digitali del settore, cioè ai siti che raccolgono testi, possibilmente in edizioni filologicamente affidabili, in formato digitale. Di solito il formato è html, che poi dovremo convertire in txt (salvando il file in formato testuale) per sottoporlo ai programmi di analisi testuale.
Nel settore medievistico esistono biblioteche di grande rilievo nelle diverse letterature. Qualcosa si trova in archivi generalisti come *Bibliotheca Augustana*, realizzata ad Augsburg da Ulrich Harsch, che contiene in formato html testi di molte letterature mondiali. https://www.hs-augsburg.de/~harsch/augustana.html

Di estensione analoga la biblioteca *Intratext*, che offre anche la possibilità di effettuare o richiedere statistiche linguistiche e concordanze: www.intratext.com/
Una ricca lista di biblioteche digitali medievistiche si trova nel sito della Central European University di Budapest (http://medstud.ceu.hu/index?id=10&cikk=127). Testi anche in traduzione all'indirizzo di Fordham http://sourcebooks.fordham.edu/Halsall/sbook2.asp#lit2
Per il latino, che raccoglie il 90% delle fonti relative al medioevo, cioè al periodo fra V e XV secolo, le principali sono le raccolte storiche della *Patrologia Latina* (una serie di 220 volumi editi nel XIX secolo da edizioni precedenti, ora disponibile in formato pdf nel sito dei *Documenta Catholica Omnia*, altrimenti consultabile a pagamento nel sito pld.chadwyck.co.uk/), i *Monumenta Germaniae Historica*, consultabili gratuitamente in formato pdf ottico (cioè non ricercabile) o in formato html con ottimo motore di ricerca nel sito www.mgh.de, che raccoglie le edizioni filologiche dal XIX al XXI secolo di qualsiasi testo o documento abbia avuto a che fare con la storia della Germania e del Sacro Romano Impero: entrambe sono consultabili in formato testuale a

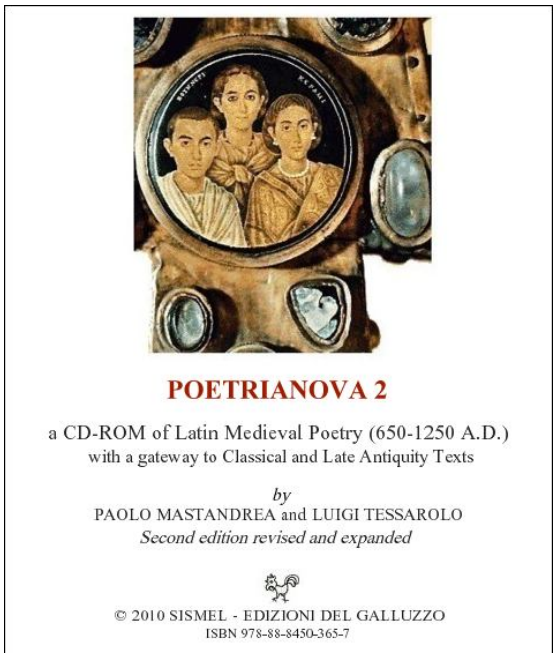
pagamento nel sito *Brepolis* della casa editrice Brepols di Turnhout, insieme al *Corpus Christianorum*, collana di edizioni critiche moderne di testi mediolatini che non è reperibile in formato libero. PL e MGH sono consultabili, insieme ad altre raccolte, nel repository *Corpus Corporum* (mlat.uzh.ch), creato a Zurich da Philipp Roelli, che comprende anche molti altri sub-corpora e rappresenta al momento il più grande repository esistente di testi latini. É anche dotato di lemmatizzatore di latino classco (collegato a quello di Perseus, grande portale di filologia classica dell'niversità di Leipzig) e di analizzatore statistico.



Una delle poche biblioteche digitali latine non ancora "inglobate" nel *Corpus Corporum* è ALIM (alim.unisi.it), Archivio della Latinità Italiana del Medioevo, anch'esso dotato di un analizzatore interno particolarmente adatto alla comparazione testuale (*Lexicon*) e lemmatizzato con lemmatizzatore adattato al lessico mediolatino, che comprende testi della latinità medievale e umanistica provenienti dall'Italia, ricavati da edizioni critiche, ma anche racccolte di documenti (di area toscana, campana, veronese) non reperibili altrimenti e collezioni tematiche predisposte dai team di ricerca. Contiene anche testi in trascrizione inedita non reperibili altrove e indicazioni di novità scientifiche del settore di informatica mediolatina.



Per la poesia il repertorio migliore è *Poetria Nova*, non consultabile online, che si acquisisce a pagamento dall'editore SISMEL e contiene migliaia di testi in formato html copiabili e ricercabili per parole, clausole, caratteristiche metriche e data.

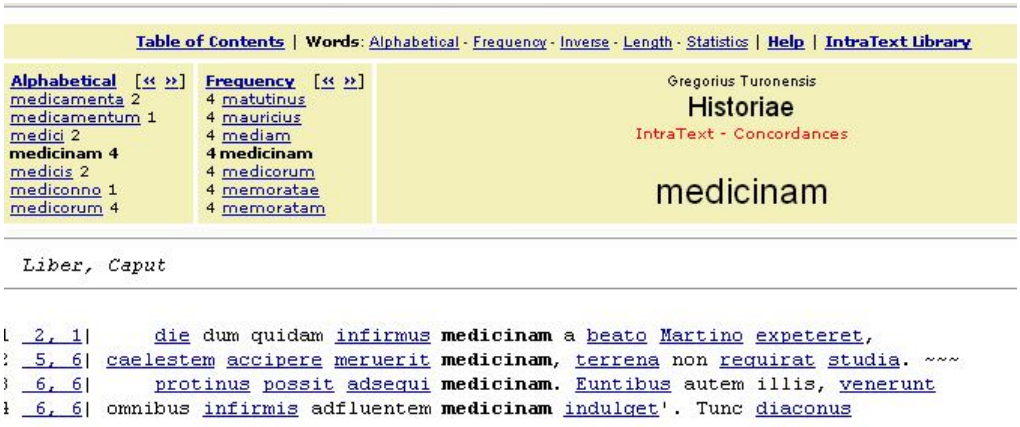


Per le letterature non latine (soprattutto francese e italiana) si trovano molti riferimenti nella pagina di un progetto COST di medievistica digitale realizzata in Polonia: http://scriptores.pl/aktualnosci/workshop/proceedings/presentations.
In particolare, per la **letteratura francese** un corpus di riferimento è la *Base de française médiéval http://bfm.ens-lyon.fr/,* che comprende 153 testi dal IX al XV secolo pronti per essere analizati in TXM (vedi infra) e collegati ad altri archivi di testi e di manoscritti, mentre per quella provenzale va consultato il *Corpus des Troubadours* (http://trobadors.iec.cat/autors_d.asp) e per le opere poetiche in **castigliano, gallego e catalano** rimandiamo a http://www.cantigasdesantamaria.com/, http://icalia.es/troubadours/ca/, http://cantigas.fcsh.unl.pt, http://www.remetca.uned.es. Per la poesia in **neerlandese** medievale si può consultare liederbank.nl, per quella **germanica** antico alto tedesca www.lhm-online.de, per quella **medio-inglese** http://dimev.net e per quella **scandinava** abdn.ac.uk/skaldic, che rappresenta anche il più bel sito di edizione digitale . Tutti questi (e molti altri) archivi di posia medievale stanno per essere collegati in un unico enorme database in costruzione presso la UNED di Madrid in un progetto diretto da Elena González Blancoo (http://linhd.uned.es). Per la letteratura **italiana** i riferimenti sono soprattutto due: la *Biblioteca Italiana* (bibliotecaitaliana.it), codificata in xml, che comprende per i primi secoli protoumanistici e rinascimentali anche testi latini, e il *TLIO* (Tesoro della Lingua Italiana delle origini) curato dall'Opera del Vocabolario Italiano del CNR (ovi.cnr.it/index.php/it/il-corpus-testuale). Ma naturalmente esistono decine di altre biblioteche digitali per la prosa e la documentazione, che spesso vanno consultati uno per uno (si veda il capitolo sulle edizioni digitali di questo stesso manuale).

5. Il secondo passo: indici e concordanze

Lo strumento più basilare e più antico per entrare nell'officina linguistica di un testo è naturalmente l'indice, che esiste da secoli e che fino a pochi decenni fa era compilato a mano, registrando su un foglio a parte il lemma e (nei casi migliori) i passi nei quali occorre all'interno di un testo. Facciamo una prova proprio su un autore medievale. Prendiamo Gregorio di Tours, storiografo del VII secolo nella Gallia Merovingia, le cui *Historiae*, che narrano gli eventi della dinastia merovingia dal IV al VI secolo, troviamo integralmente nella biblioteca digitale *Intratext* (intratext.com/IXT/LAT0783/2/ZF.HTM). Se apriamo il primo capitolo del libro II (il primo è dedicato a una riassunto della storia universale) troviamo la parola *medicina*, riferita alla guarigione che un fedele malato si aspetta da san Martino, e magari ci viene la curiosità di capire quanto sia esteso l'uso di

questa parola in Gregorio, che è anche un modo per ricavare dati sulla storia della malattia nell'alto medioevo: clicchiamo su di essa e vediamo apparire la concordanza delle sue occorrenze, cioè l'indice contestualizzato (= inserito nella frase) delle occorrenze di ogni parola, che può essere per lemmi (come gli indici veri e propri) o per forme, come qui. La concordanza (virtuale) delle *Historiae* relativamente al termine *medicina* dà:



A sinistra leggiamo le indicazioni de passi in cui si trova la parola, a destra i segmenti di frase in cui compare, in modo che se ne possa comprendere meglio il suo significato e il suo uso.
Possiamo trovare concordanze che forniscono un solo rigo di contesto (quelle che si facevano a stampa fino a pochi anni fa di solito avevano questo formato) o porzioni maggiori o minori, ma l'aspetto attuale è di solito questo. Naturalmente il risultato di *medicina* va moltiplicato per il numero di lemmi (o di forme).
A cosa serve una concordanza? Realizzate a stampa fino agli anni '90 del XX secolo, le concordanze – rispetto all'indice semplice - servono a comprendere meglio quale sia l'uso di determinati termini da parte di un autore nel loro contesto sintattico e in confronto ad altre forme nei loro relativi contesti. La concordanza può servire anche a scopo filologico, ad esempio per completare un frammento di autore trovato in un catalogo o in una citazione individuandone il contesto. E naturalmente serve a tutte le finalità condivise con gli indici, in particolar modo l'individuazione di un termine o della sua assenza.
Nell'esempio citato, la concordanza è lemmatizzata, cioè ha cercato tutte le varianti flesse di *medicina* (singolare e plurale: *medicinae, medicinam*, ecc.) e non solo la forma *medicina* come accade in molti software di concordanze automatizzate.
Il risultato è che questo lemma è piuttosto raro nelle *historiae* (solo 4 occorrenze, di cui due nello stesso passo: libro 6 cap. 6) e che dunque o la malattia e la sua guarigione con farmaci ha scarso rilievo per Gregorio e i suoi lettori, anche nell'uso figurato, oppure il suo testo la esprime con altri termini.
In passato per elaborare una concordanza di un testo integrale occorreva schedare il testo parola per parola e creare una scheda per ogni occorrenza ddi ogni parola nel suo contesto. Oggi è sufficiente scaricare un programma (o autonomo e specializzato, come *Concordance*, o autonomo ma integrato in un pacchetto di funzioni, come in lexicon.unisi.it) o integrato in un portale, come in TAPoR (se ne elencano 72, non tutti funzionanti, alla pagina tapor.ca/tools/filterbyattribute?att_val_id=2), copiare e incollare in una finestra il testo, ripulendolo prima da numeri o segni che non si vogliono indicizzare, e cliccare su un pulsante, che in pochi istanti produrrà la sequenza desiderata.
Le concordanze dunque possono essere usate come primo strumento per capire quanto è usata una parola (o un gruppo di parole) e in quali contesti di frase.

6. Il terzo passo: l'indice di frequenze

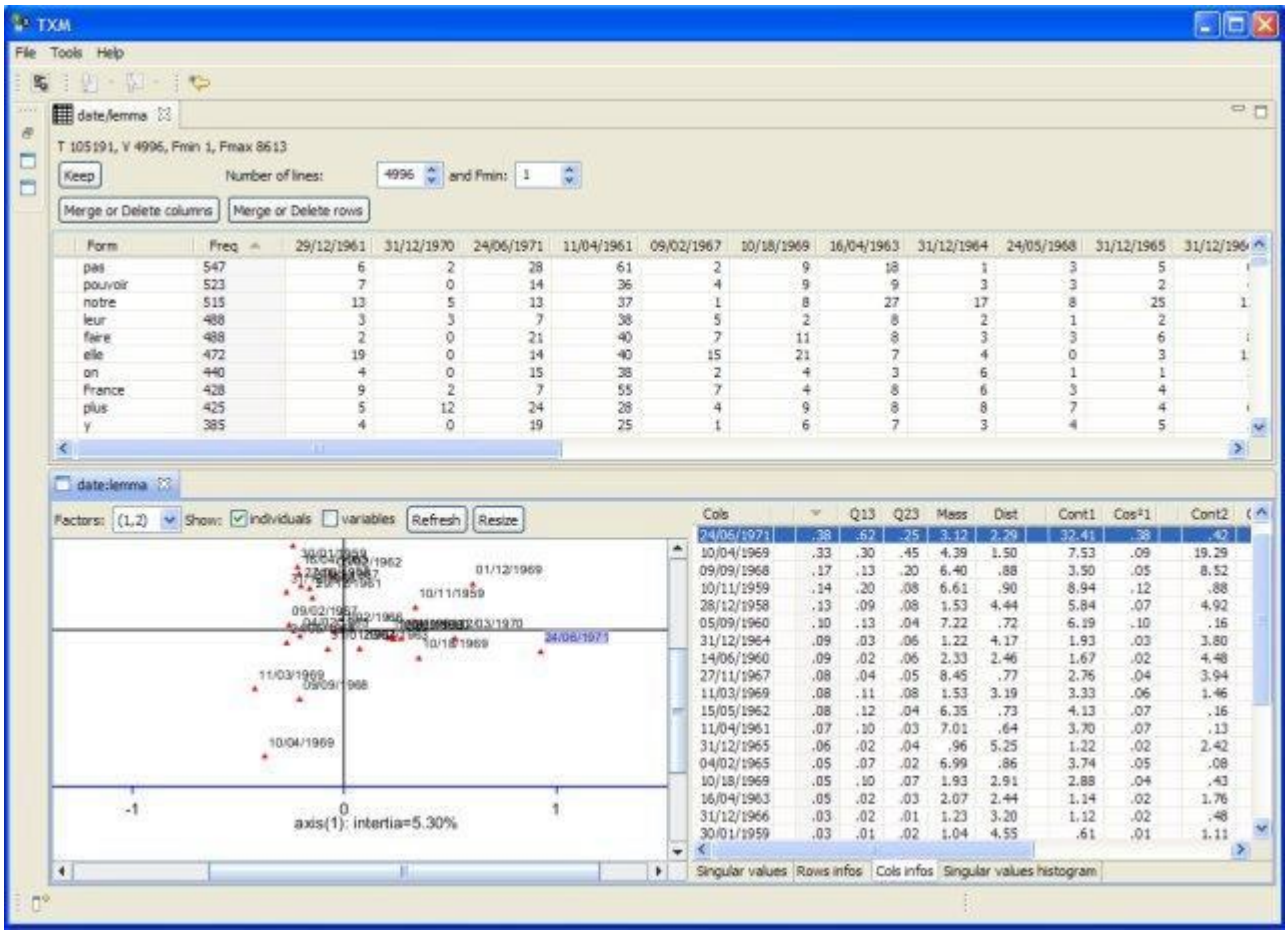
Si comincia ad entrare nella cosiddetta analisi quantitativa quando gli elementi di cui è formato un testo (parole, forme, frasi, parti del discorso, lettere, grafemi) vengono contati e i numeri vengono comparati, perchè solo un termine di confronto ci fa capire quanto valga un dato: non posso sapere se usare "pane" 10 volte in un testo sia poco o tanto finché non so quanto è lungo il testo e quale sia la media di uso di "pane" nei testi di quell'autore o di quel periodo o di quella lingua o letteratura.

L'indice di frequenze è l'operazione-base di ogni ulteriore analisi ed è quella che con maggiore fatica viene svolta a mano (quasi impossibile per testi lunghi) e quella perciò alla quale l'informatica ha dato l'impulso maggiore.

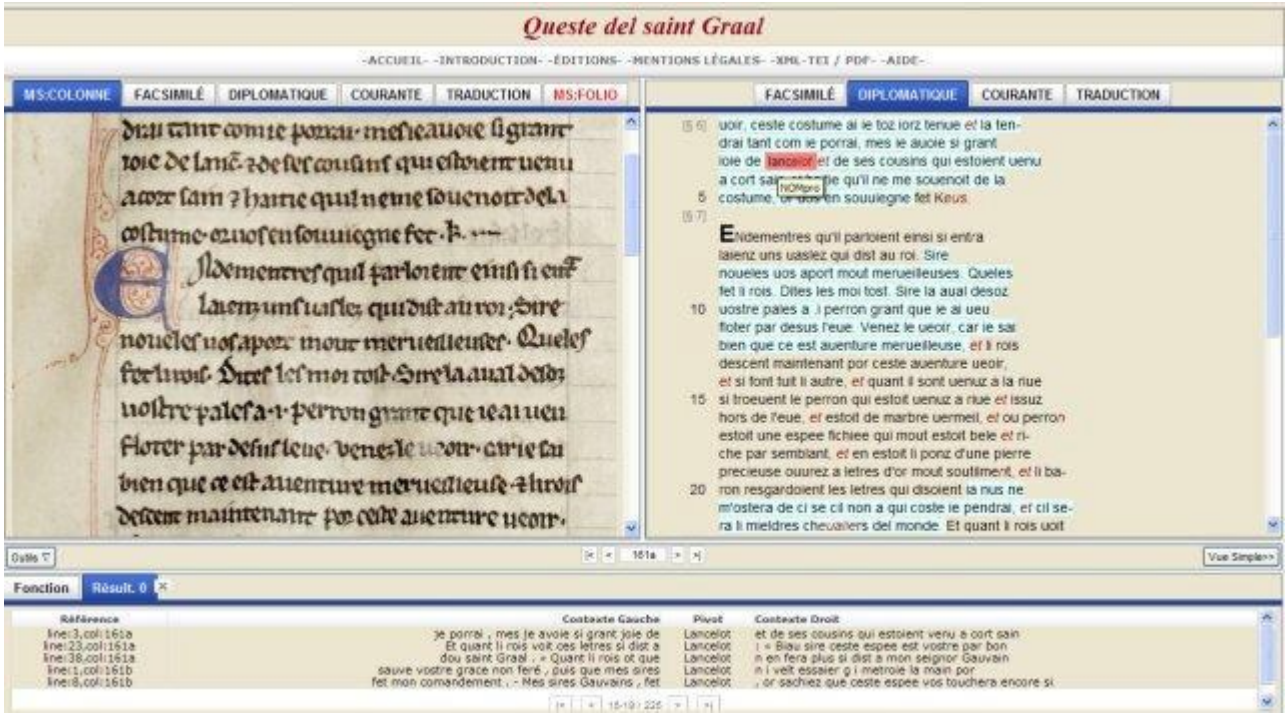
6a. Scelta del software

Per farlo occorre un software (quasi tutte le concordanze di TAPoR lo fanno: (tapor.ca/tools/filterbyattribute?att_val_id=62.), anche se si può fare a mano nel caso che si lavori, ad esempio, solo su due o tre parole in due testi di lunghezza limitata. Ma è un caso che non capita mai.

I tool che possiamo usare sono molti, e nella ricerca sui testi medievali i più frequentemente usati sono *TXM* (che lavora su testi codificati in XML), Stylometry with R e Lexicon. TXM (http://textometrie.ens-lyon.fr/spip.php?rubrique96&lang=fr) è un potente programma open-source, cioè non solo liberamente scaricabile ma con codici liberamente rielaborabili, e lavora su testi sia in formato piano sia (soprattutto) taggati e strutturati in XML. Sviluppato dal CNRS di Lyon (vd. Heiden 2010), produce dei sub-corpora (cioè corpus più piccoli alll'interno di corpus grandi), partizioni di testi da confrontare, concordanze *kwic* (cioè con citazione del contesto di frase), crea un'edizione html di ogni testo, effettua calcoli sulle parole e ne sviluppa diagrammi e cartografie e si collega all'analizzatore di parti del discorso (POS) TreeTagger, esportando poi i dati in formato csv (txt con separatori per i diversi campi) per i testi e SVG per i grafici. Fino a due anni fa si usava solo scaricando un pesante e complesso aggregato di software, da poco è utlizzabile in linea, ma al momento il link al portale è inattivo (http://code.google.com/intl/fr/webtoolkit).



TXM è anche un tool per edizione testo/immagine, come si rileva dal seguente screenshot:
ù

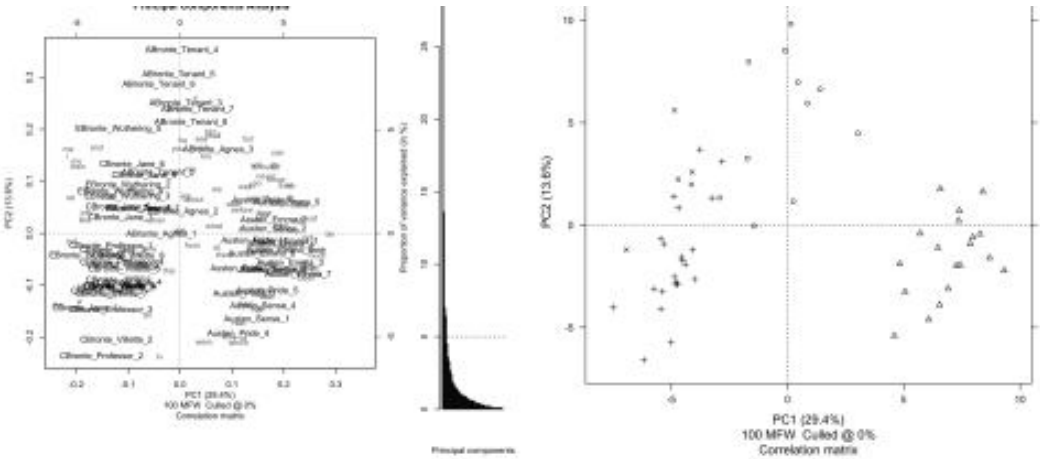


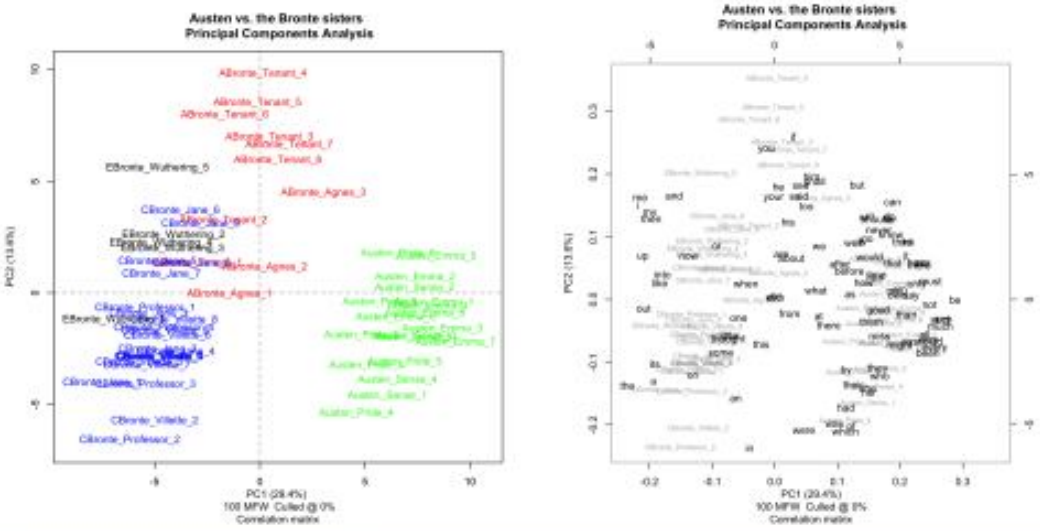
Stylometry with R (descritto in Eder-Rybicki-Kestemont 2016) è il programma forse più usato per analisi vettoriali ed è molto avanzato soprattutto per la capacità di combinare analisi diverse e visualizzarle graficamente, ma come dice il nome richiede che si lavori in

ambiente R (https://sites.google.com/site/computationalstylistics/stylo), dunque scaricando e imparando a usare un software che comporta frequentemente la scrittura di comandi come questi:

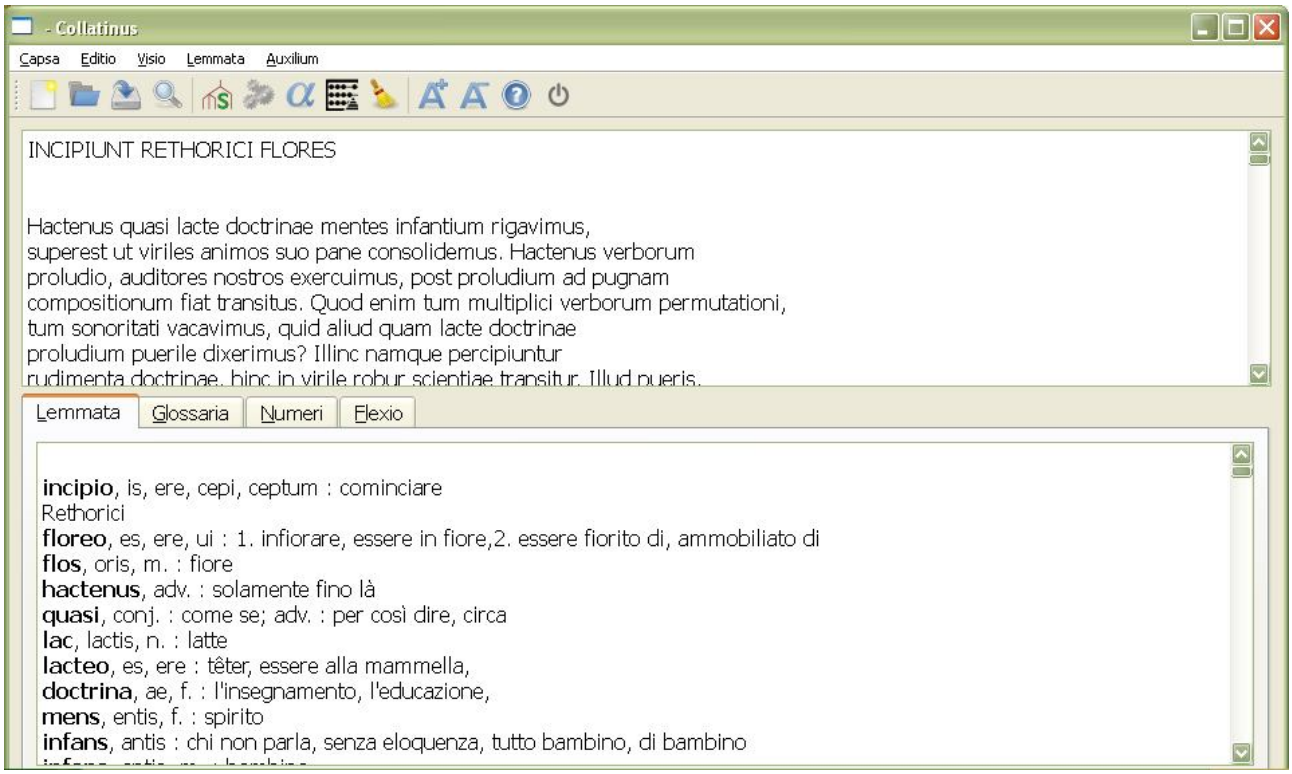
```
stylo(frequencies = culled.freqs, analysis.type = "PCR",
        custom.graph.title = "Austen vs. the Bronte sisters",
        pca.visual.flavour = "technical",
        write.png.file = TRUE, gui = FALSE)

stylo(frequencies = culled.freqs, analysis.type = "PCR",
        custom.graph.title = "Austen vs. the Bronte sisters",
        write.png.file = TRUE, gui = FALSE)

stylo(frequencies = culled.freqs, analysis.type = "PCR",
        custom.graph.title = "Austen vs. the Bronte sisters",
        pca.visual.flavour = "symbols", colors.on.graphs = "black",
        write.png.file = TRUE, gui = FALSE)

stylo(frequencies = culled.freqs, analysis.type = "PCR",
        custom.graph.title = "Austen vs. the Bronte sisters",
        pca.visual.flavour = "loadings",
        write.png.file = TRUE, gui = FALSE)
```

che contribuiscono a produrre visualizzazioni come queste:





Un'altra possibilità è *Collatinus*, pacchetto di software francese elaborato all'interno del mega-progetto *Biblissima,* che al momento funziona solo offline, dopo l'installazione, ma non richiede codifica e comprende un lemmatizzatore che è anche analizzatore morfologico e metrico.

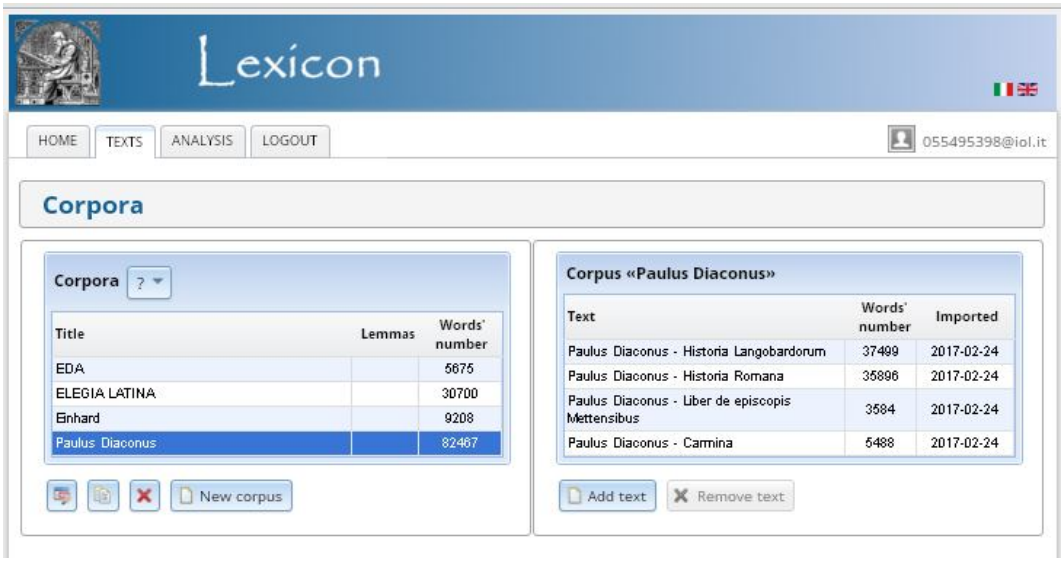


Lexicon (www.lexicon.unisi.it) di Stella-Tessarolo è un software molto più semplice e leggero, che si usa online, e ha il vantaggio di non richiedere installazioni né codifiche di alcun tipo e di essere immediatamente comprensibile, oltre che gratuito. Inseriamo i testi che ci interessano, analizzando le opere di un medesimo autore, Paolo Diacono (VIII secolo), ad esempio la *Historia Langobardorum* e poi la sua *Historia romana*, i *Gesta episcoporum Mettensium* e i *Carmina* (Poesie). Tutti i testi possiamo trovarli in ALIM (Archivio della latinità italiana del medioevo: alim.unisi.it), a dove possiamo scaricarli in 4 formati (txt, html, pdf, xml) e in versione già ripulita e pronta per l'analisi, ma potremmo cercarli anche in MGH (il sito dei *Monumenta Germaniae Historica*), dove li troviamo in pdf ottico e in html ma con tutti gli apparati e le annotazioni, che vanno poi depurati di tutti i segni non pertinenti e convertiti in xtx, oppure in altre biblioteche digitali come *Intratext,* che si estende anche a letterature non latine, o *The Latin Library*: se li prendiamo da uno di questi repository dobbamo ripulire i testi da tutte le imperfezioni (spazi incongrui, parole spezzate, annotazioni o titoli correnti della fonte rimasti nel file, numeri, segni vari ecc.). Un piccolo trucco: conviene salvare la versione stampabile in html (quando si sceglie l'opzione desiderata in "Salva come..."), poi aprire il file html con un editor basico tipo Open office Writer o word, quindi salvarlo come Testo. Se si passa dal pdf si creano invece problemi di spazi e di font che richiedono un lavoro ingente di ripulitura.

6b. Procedure

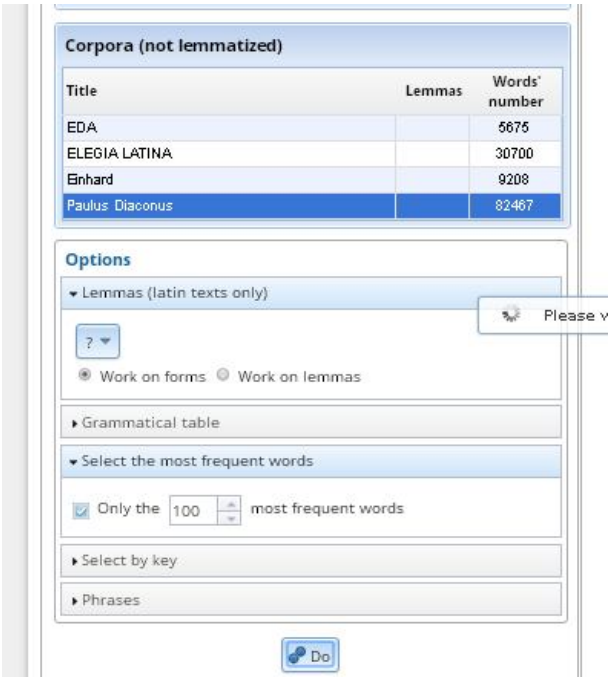
Una volta acquisiti i testi in files txt distinti possiamo incollarli in *Lexicon* (basta andare alla pagina Testi, quindi cliccare su Nuovo, e dove c'è scritto Scegli file andare sul testo desiderato).

Poi andiamo su *Corpora* e creiamo un "nuovo corpus" intitolandolo "Paolo Diacono". Quindi il programma ci abilita ad aggiungere un testo: inseriamo prima uno poi l'altro, provando anche ad aprirli per verificare che non ci siano problemi o segni strani (in tal caso, occorre ripulire il file di origine e ricaricarlo).



Il Corpus "Paulus Diaconus" nel suo complesso conta (ce lo dice il programma senza che nemmeno lo chiediamo) 82487 forme (token). La *Historia Langobardorum* è 37499, la *Historia Romana* 35896, i *Gesta* o *Liber* 3584, i *Carmina* 5488. Se lemmatizziamo i testi (attraverso il Menu "Lemmatizza") potremo avere il numero dei lemmi di ogni testo: ad esempio nel *Liber* sono 1606, dunque 3584 forme e 1696 lemmi, e dunque possiamo calcolare il primo indice stilometrico: il rapporto token/type, che misura la "ricchezza" di forme del testo: in questo caso 2, 23 token/type 0,47). Nella *Historia Romana* i lemmi riconosciuti sono 5973, dunque il rapporto è 6,06 (token/type 0,166), ma naturalmente è un rapporto alterato dalla maggiore dimensione dell'*Historia*: nel caso di rapporti TTR i numeri non possono che essere assoluti, dunque statisticamente non significativi.

Per avere dati statisticamente rilevanti dobbiamo lavorare sulle frequenze relative, cioè sul rapporto fra il numero di forme presenti e il numero di occorrenze totali: andiamo al Menu analisi e clicchiamo su "Frequenze". Selezioniamo il Corpus "Paulus Diaconus", indicando se vogliamo le 100 parole più frequenti, o le 150 o le 50.



C'è grande discussione su quale sia la soglia di affidabilità del numero se si hanno finalità attribuzionistiche (vd. Moens-Kestemont 2015), ma per comodità scegliamo 50 e clicchiamo su "Esegui": avremo una tabella che possiamo ordinare in senso e comincia così (ordinandola dal più alto al più basso): basta cliccare sul titolo "Frequenze":

Paulus Diaconus (*corpus*)

Word	Occurrence	Frequency %
a	550	0.667
ab	359	0.435
ac	162	0.196
ad	924	1.12
annis	106	0.129
anno	164	0.199
apud	221	0.268
aput	83	0.101
atque	184	0.223
autem	133	0.161
bello	82	0.099
bellum	173	0.21
contra	191	0.232
cui	88	0.107
cuius	123	0.149
cum	1053	1.277
de	595	0.722
dum	183	0.222
ei	185	0.224
eius	377	0.457
eo	198	0.24
eorum	93	0.113
eos	116	0.141
erat	124	0.15
esse	111	0.135
esset	103	0.125

Ricorrenza (Occurrence) indica il numero di volte che la parola ricorre nel corpus, detto anche frequenza assoluta (o tokenizzazione), un dato che però è poco significante perché varia secondo la dimensione del testo, mentre Frequenza indica la percentuale delle ricorrenze relativamente alla dimensione del testo (cioè al numero di ricorrenze totali di tutte le parole), detta anche Frequenza relativa, ed è il dato effettivamente significativo, perché la rilevanza di *medicina* detto 7 volte in un testo di 100 pagine o 7 volte in testo di 3 pagine cambia molto: la frequenza assoluta è la stessa, ma nel testo più breve la frequenza relativa è molto più alta e quindi significativa.
La prima visualizzazione delle frequenze è in ordine alfabetico, ma cliccando su *Frequency %* si ottiene l'ordine discendente che ci serve per le nostre analisi.

Paulus Diaconus (*corpus*)

Word	Occurrence	Frequency %
et	2119	2.57
in	1636	1.984
est	1230	1.492
cum	1053	1.277
ad	924	1.12
qui	821	0.996
ut	600	0.728
de	595	0.722
a	550	0.667
per	417	0.506
eius	377	0.457
non	360	0.437
ab	359	0.435
quae	348	0.422
se	338	0.41
quod	332	0.403
sunt	332	0.403
ex	316	0.383
post	273	0.331
quam	269	0.326
eum	254	0.308
etiam	252	0.306
sed	248	0.301
quoque	230	0.279
vero	229	0.278
hoc	223	0.27
apud	221	0.268

In Paolo Diacono le parole più frequenti sono le stesse in quasi ogni testo di una lingua: in questo caso *et, in, est, cum, ad, qui* ecc. Sono di solito parole-funzione (**function-words**), preposizioni, articoli e congiunzioni, molto più usate di parole con un significato proprio. Il loro ordinamento relativo però cambia di testo in testo e può essere rivelatore di qualche caratteristica. Le frequenze in *Historia Langobardorum* e *Carmina* sono infatti:

Paulus Diaconus - Historia Langobardorum

Word	Display	Occurrence	Frequency %
et	20	1063	2.835
in	65	828	2.208
cum	38	530	1.413
est	26	500	1.333
ad	139	491	1.309
qui	57	453	1.208
de	5	448	1.195
ut	305	291	0.776
eius	170	222	0.592
a	315	202	0.539
per	498	191	0.509
quae	158	184	0.491
quod	7	182	0.485
non	411	168	0.448
vero	1094	166	0.443
se	653	164	0.437
rex	190	161	0.429
ab	115	157	0.419
eum	2301	150	0.4
sunt	58	150	0.4
hoc	25	144	0.384
quoque	211	132	0.352
sed	487	124	0.331
ex	12	121	0.323
quam	594	120	0.32
haec	579	117	0.312
quomodo	101	116	0.309

La differenza fra la media di Paolo Diacono e *l'Historia Langobardorum è* minima, ma anche questo è un dato: potrebbe indicare una costante dello stile di Paolo Diacono. Muta la posizione reciprica di *est* e *cum*. Cambia la frequenza delle poesie:

Paulus Diaconus - Carmina

Word	Display	Occurrence	Frequency %
et	53	113	2.059
est	28	87	1.585
in	45	60	1.093
ut	1310	49	0.893
non	192	33	0.601
tibi	390	31	0.565
quae	1682	30	0.547
qui	232	29	0.528
per	127	27	0.492
ad	11	25	0.456
de	3	23	0.419
haec	163	22	0.401
quod	255	22	0.401
si	1637	22	0.401
iam	1480	21	0.383
sed	188	21	0.383

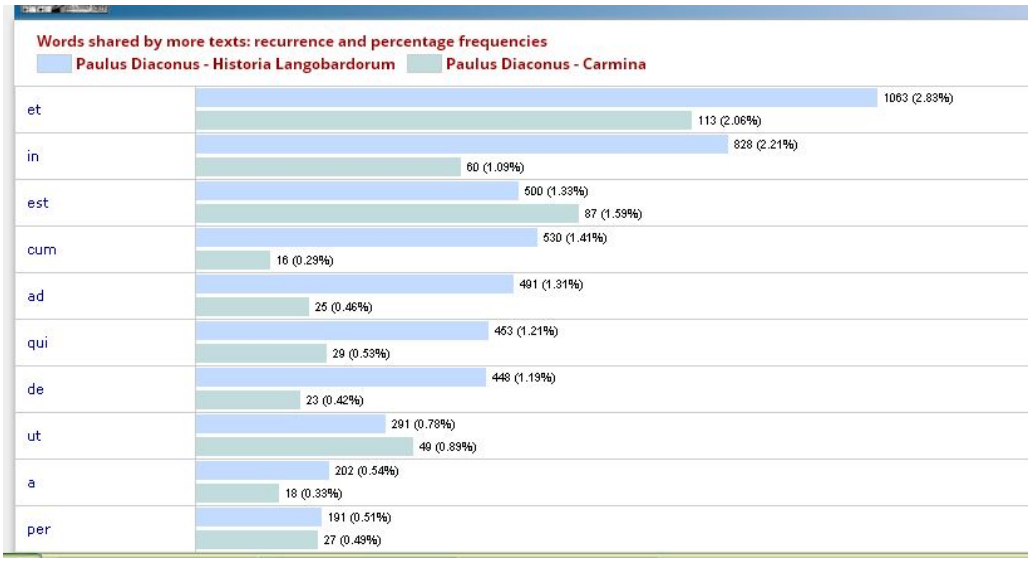
L'ordine è *et, est, in, ut, non, tibi, quae, qui*, quindi le differenze sono sensibili, in particolare è più frequente la forma est, risale molto *ut,* che indica la presenza di finali e completive, e la presenza di *tibi* indica ad esempio che si tratta di poesie relazionali, indirizzate a un destinatario specifico, a differenza naturalmente di quanto può avvenire nella *Historia.* Questo può essere evidenziato tramite il tasto comparison, dove possiamo scegliere tre configurazioni, tipiche di questo software: A-B, cioè le forme di A che NON sono presenti in B, B-A (l'opposto) e AxB, cioè le forme comuni.

Questo è il grafico di A-B:

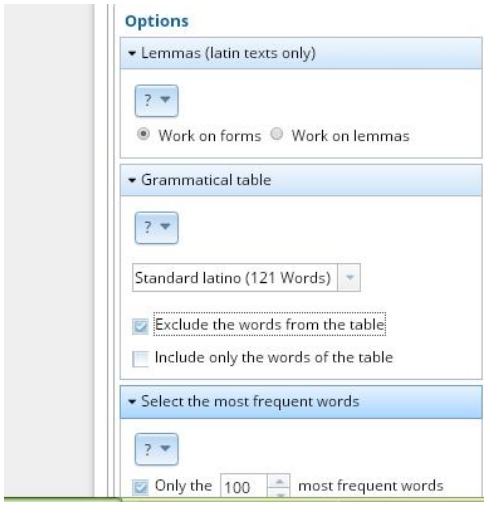
Words of A absent from B: 63 (tra le 100 di maggiore frequenza) ?

Words	Occurrence	Frequency %
eius	222	0.592
vero	166	0.443
se	164	0.437
rex	161	0.429
eum	150	0.4
quomodo	116	0.309
ei	106	0.283
tunc	105	0.28
quia	101	0.269
regis	93	0.248
eo	91	0.243
etiam	88	0.235
langobardorum	88	0.235
autem	85	0.227
aput	83	0.221
ita	77	0.205
usque	77	0.205
erat	75	0.2
langobardis	74	0.197
his	73	0.195
francorum	72	0.192
sibi	72	0.192
contra	71	0.189

Significa che, fra le 100 parole più frequenti dell'*Historia Langobardorum, eius, vero, se, rex, eum* ecc. non sono presenti nella poesie: a volta può essere per motivi metrici (*autem*), a volte perché il tema (ad es. *rex* e *regis,*oppure *langobardorum*) non compare in poesia, a volta perché la congiunzione *quia* introduce un tipo di frase che in poesia è raro o almeno lo è nella poesia di Paolo Diacono. É anche interessante che *Langobardi* non compaia spesso al nominativo ma sempre come specificazione o destinazione, al pari di *Francorum*. Cioè il popolo non è soggetto della storia ma specificazione delle sue personalità, almeno nella forma linguistica. Anche l'altissima frequenza di *contra* apre una finestra sulle scene di aggressività bellica dell'HL, mentre la poesia di Paolo è più vicina al genere lirico.
La forma grafica della comparazione è questa:



Qui le differenze, anche se minime, sono più evidenti: ad es. l'uso di *cum* (sia congiunnzione che preposizione) è molto più alto nell'HL, il che conferma la scarsa presenza di sintassi subordinativa nelle poesie, ma è anche sull'uso della preposizione *de,* che si giustifica probabilmente con la distanza fra linguaggio della prosa e della poesia. Naturalmente le osservazioni cambiano molto se invece delle 100 parole più frequenti incluse le function words si lavora sulle parole semantiche: questo si può fare in Lexicon spuntando la casella relativa



Ecco il risultato relativo alle Poesie:

Paulus Diaconus - Carmina

Word	Display	Occurrence	Frequency %
pater	231	18	0.328
semper	888	18	0.328
versus	0	11	0.2
cuncta	740	10	0.182
omnia	706	10	0.182
prima	137	10	0.182
simul	146	10	0.182
amor	942	9	0.164
habet	835	9	0.164
pectora	585	9	0.164
pectore	363	9	0.164
potens	1484	9	0.164
spes	882	9	0.164
corda	505	8	0.146
epitaphium	2273	8	0.146
nunc	151	8	0.146
regna	959	8	0.146
vestra	2769	8	0.146
vitae	260	8	0.146
amore	402	7	0.128

Le parole (non lemmatizzate) semanticamente rilevanti con maggiore frequenza sono *pater*, semper, versus, cuncta, omnia, prima, simul, amor, pectus, ognuna delle quali rivela un elemento della scrittura di Paolo: il riferimento "paterno" al proprio interlocutore, la propensione metapoetica (versus), l'assolutizzazione (cuncta, omnia), l'enfasi affettiva (amor, pectus) e tante altre cose che potremmo ricavare da questa lista.
Nell'*Historia Langobardorum* i termini più importanti sono invece *bellum, vir, enim, dicitur*: tema guerresco, visuale maschile, inclinazione alla spiegazione, atteggiamento dello storiografo che riporta le notizie (*dicitur*) anziché narrativo, preminenza del ruolo regale (*rex, regnum*), mentre in comune con l'HL è la visuale paterna ma dall'altro lato (*filius*). Le scoperte che si possono fare con l'analisi delle frequenze lessicali sono moltissime.

Paulus Diaconus - Historia Langobardorum

Word	Display	Occurrence	Frequency %
bellum	214	39	0.104
vir	6905	39	0.104
enim	859	40	0.107
dicitur	100	41	0.109
fecit	6060	41	0.109
ibi	962	41	0.109
rege	2358	41	0.109
regno	249	41	0.109
suam	3966	41	0.109
mox	813	42	0.112
vel	6348	42	0.112
filius	206	43	0.115
omnes	731	43	0.115
esset	3897	46	0.123
sicut	380	47	0.125
alboin	283	48	0.128
beati	2843	49	0.131
quibus	163	51	0.136
dux	5341	53	0.141
eis	75	54	0.144
eorum	230	54	0.144
loco	856	54	0.144

Facciamo una controprova con la *Vita Karoli* di Eginardo il grande biografo di Carlo Magno (intorno all'830). Ecco le sue parole più frequenti, dal testo scaricabile da *The Latin Library*:

Word	Display	Occurrence	Frequency %
adeo	1309	11	0.139
annos	1117	11	0.139
apud	999	16	0.202
bello	1206	14	0.177
bellum	1614	15	0.19
contra	1208	15	0.19
deinde	1439	9	0.114
die	5647	9	0.114
ea	469	11	0.139
ei	877	16	0.202
eis	2070	13	0.164
eius	262	40	0.506
enim	48	9	0.114
eo	82	25	0.316
eos	1013	13	0.164
erant	766	10	0.127
eum	270	17	0.215
fecit	3151	9	0.114
francorum	221	16	0.202
illius	650	11	0.139
karoli	2	10	0.127
nihil	372	10	0.127
nomine	328	9	0.114
ob	78	9	0.114
omnes	18	9	0.114
omnia	424	8	0.101

:

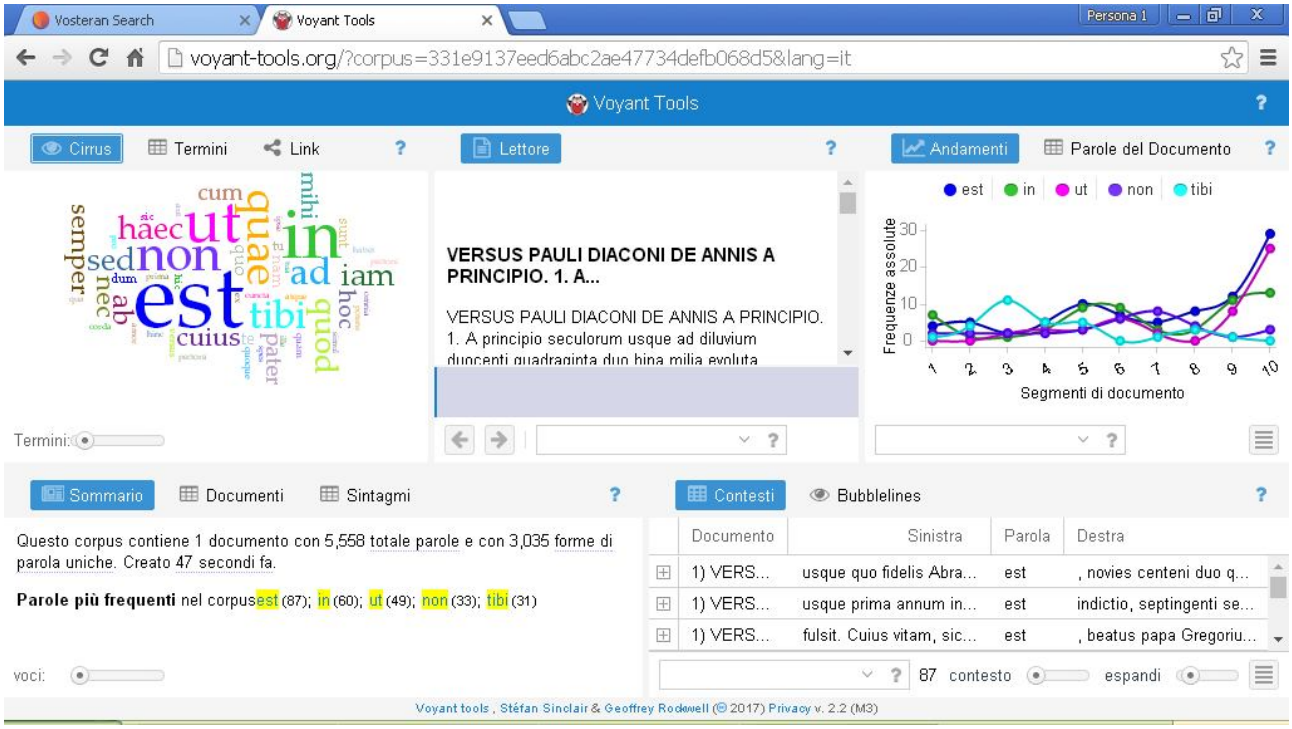
Anche qui la guerra è il tema dominante, nonostante di guerre Eginardo parli solo nella prima parte del suo testo: anche qui *contra* è una delle preposizioni più frequenti, anche qui il popolo è designato quasi sempre al genitivo, dunque come specificazione di una personaggio, mentre *nomen* ('nome' ma anche 'titolo') ci trasmette l'attenzione di Eginardo ai ruoli politici e alle personalità.

La ricerca si può effettuare anche con selezioni più raffinate, per chiave (key) quindi ad es. *reg-* per evidenziare il lessico relativo al re e al regno, o per locuzione (phrase), che analizza i clusters, le sequenze di parole, e individua le espressioni più ricorrenti e i sintagmi, per uno studio più linguistico o stilistico.

6c. Altri tipi di tools.

Lexicon, pensato da un filolofo e un informatico umanista, prevede la possibilità di selezionare il tipo di lessico, la comparazione per inclusività, esclusività e sovrapponibilità che interessano gli specialisti. Ma altri software presentano visualizzazioni molto più spettacolari ed evidenti, che vengono spesso usate nel giornalismo: ad esempio *Voyant Cirrus* (http://voyant-tools.org/view=cirrus), dove basta incollare il testo dell'HL in txt e cliccare sul pulsante per ottenere in pochi secondi la celebre "nuvola" di parole a colori

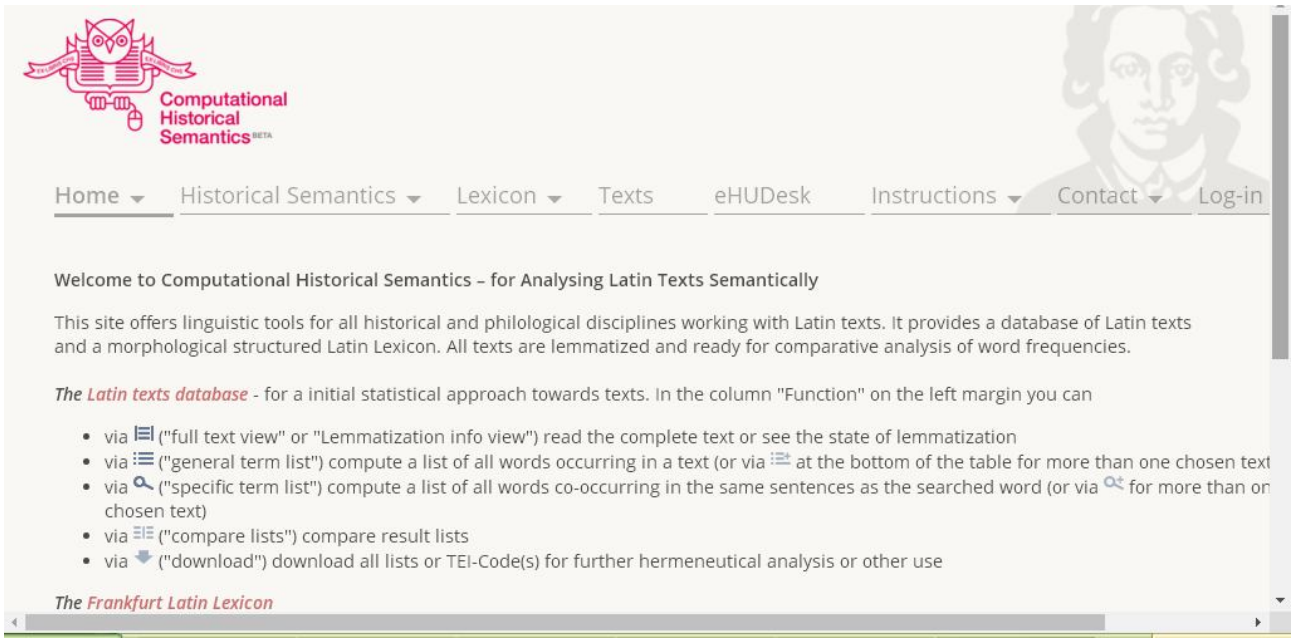
mentre una versione più articolata del software (http://voyant-tools.org/lang=it) produce una visualizzione multipla: la palla di parole in alto a sinistra, mentre al centro si può leggere il testo e a destra abbiamo il grafo che presenta l'andamento delle frequenze distribuito lungo il percorso del testo e in basso a sinistra l'indice delle frequenze in forma testuale e discorsiva.



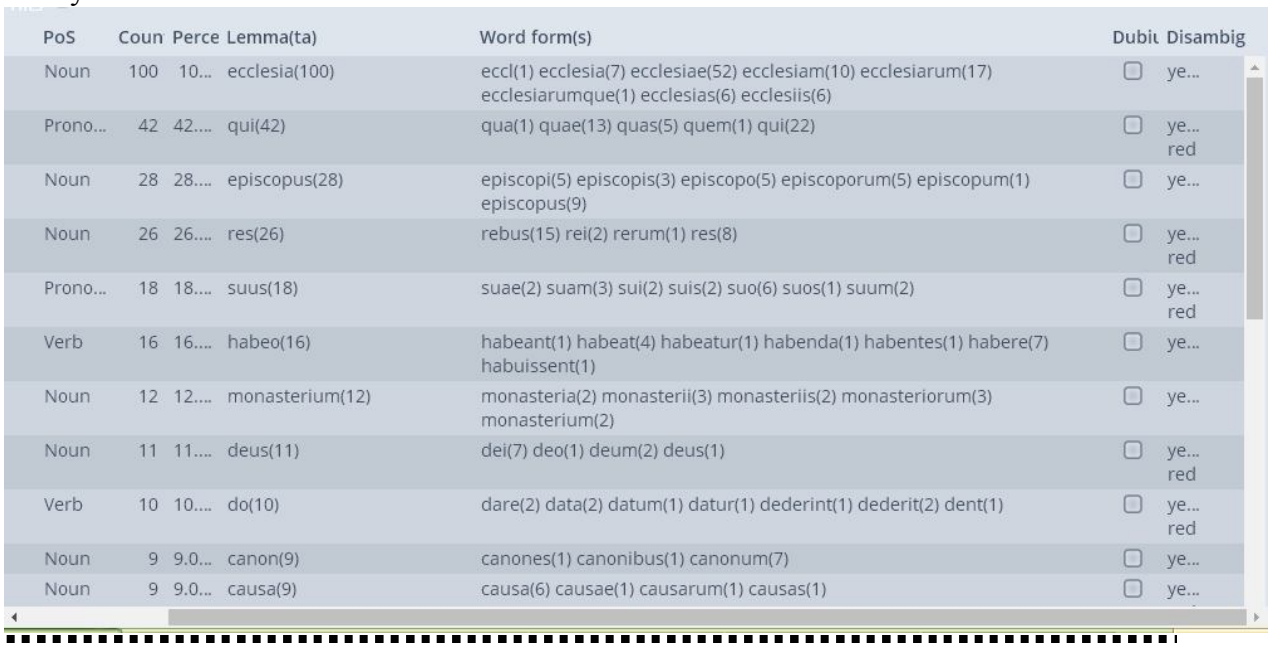
In generale fra i set di tools indicizzati nei portali come TAPoR quelli della categoria voyant tools sono fra i più efficaci per eseguire ricerche separate e rappresentarle graficamente: http://docs.voyant-tools.org/tools/.

7. Strumenti di analisi semantica delle fonti medievali: PALM e Computational Historical Semantics

Casi esemplari di archivi testuali di storia medievali con possibilità di analisi semantica, anche se ad accesso riservato a iscritti tramite registrazione, sono il progetto *Computational Historical Semantic* e *PALM*. Il primo (comphistsem.org), diretto da Dessi Schmid all'Università di Francoforte, consente di analizzare un corpus di testi storiografici in formato full text o di estrarne statistiche di frequenza non solo per parola ma anche per campi semantici predeterminati dal team: ad es. *ecclesia* include un ampio set di parole riferibili alla realtà storica della chiesa medievale.



Il funzionamento è macchinoso e il caricamento assai lento, ma la base di dati è molto ampia, pur essendo specializzata, e include spesso documenti non facilmente reperibili in altri archivi. Soprattutto, *Comphistsem* è l'unico archivio che permetta la semantizzazione delle ricerche attraverso la selezione di *superlemma* invece che *lemma* o *word*. Questi sono i risultati della ricerca di cooccorrenze del lemma *ecclesia* nei *Canones* di Abbone di Fleury:



PoS	Coun	Perce	Lemma(ta)	Word form(s)	Dubi	Disambig
Noun	100	10...	ecclesia(100)	eccl(1) ecclesia(7) ecclesiae(52) ecclesiam(10) ecclesiarum(17) ecclesiarumque(1) ecclesias(6) ecclesiis(6)		ye...
Prono...	42	42....	qui(42)	qua(1) quae(13) quas(5) quem(1) qui(22)		ye... red
Noun	28	28....	episcopus(28)	episcopi(5) episcopis(3) episcopo(5) episcoporum(5) episcopum(1) episcopus(9)		ye...
Noun	26	26....	res(26)	rebus(15) rei(2) rerum(1) res(8)		ye... red
Prono...	18	18....	suus(18)	suae(2) suam(3) sui(2) suis(2) suo(6) suos(1) suum(2)		ye... red
Verb	16	16....	habeo(16)	habeant(1) habeat(4) habeatur(1) habenda(1) habentes(1) habere(7) habuissent(1)		ye...
Noun	12	12....	monasterium(12)	monasteria(2) monasterii(3) monasteriis(2) monasteriorum(3) monasterium(2)		ye...
Noun	11	11....	deus(11)	dei(7) deo(1) deum(2) deus(1)		ye... red
Verb	10	10....	do(10)	dare(2) data(2) datum(1) datur(1) dederint(1) dederit(2) dent(1)		ye... red
Noun	9	9.0...	canon(9)	canones(1) canonibus(1) canonum(7)		ye...
Noun	9	9.0...	causa(9)	causa(6) causae(1) causarum(1) causas(1)		ye...

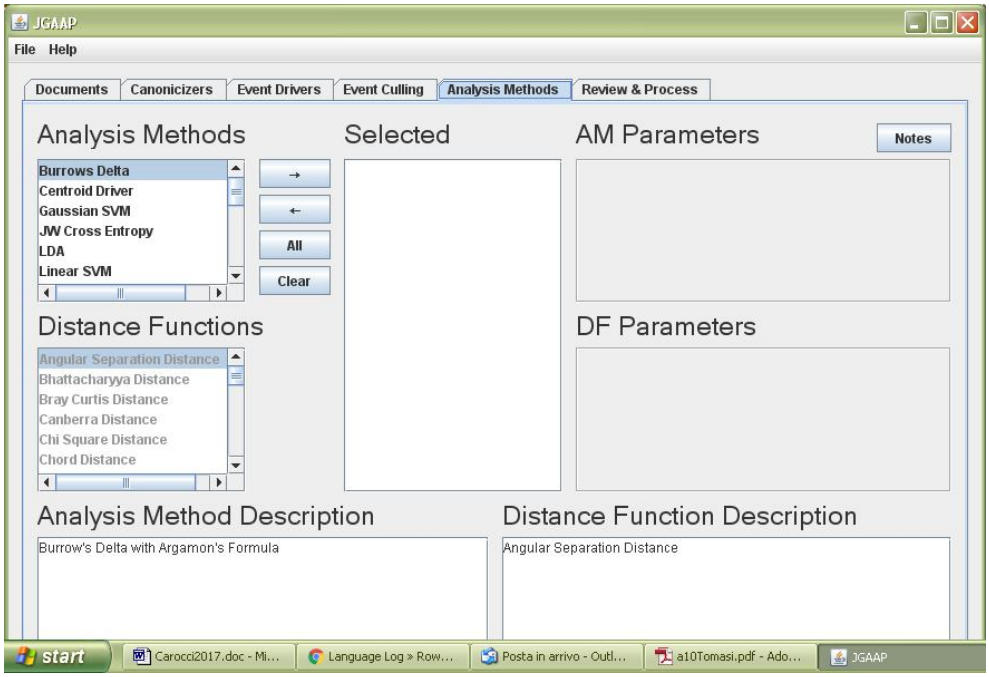
A sinistra troviamo l'indicazione del PoS (part of Speech), che può essere più o meno utile per ricerche linguistiche, stilistiche e attribuzionistiche; poi la colonna col numero di occorrenze assoluto, poi il numero relativo della percentuale, poi la "famiglia lessicale" individuata con specificazione dei numeri di occorrenze di ogni forma, infine la spunta per l'eventuale disambiguazione. I risultati si possono poi comparare con quelli di altri testi. Dunque possibilità analoghe a quelle che offre *Lexicon* in ALIM, con la differenza di un funzionamento meno perspicuo e meno fluido.

PALM (Plateforme d'Analyse Linguistique Médiévale: http://palm.huma-num.fr/PALM/), diretta fino al 2015 dal medievista Genet per il laboratorio LAMOP (CNRS-Sorbonne) e ora dismesso, è stato un progetto di raccolta di testi tardomedievali (XIII e XIV secolo) relativi alla teoria politica in latino, inglese e francese dotato di un analizzatore TXM interno, con l'intenzione di rendere possibili ricerche semantiche di storia e filosofia politica medievale e un disambiguatore di forme allografe, molto numerose nei vernacoli medievali.

8. Ricerche attribuzionistiche. Shakespeare, Ildegarde, Abelardo.
Uno dei campi che più ha dato spazio alle analisi statistiche è la ricerca attribuzionistica, che usa e produce comunque dati stilometrici per identificare autori di opere di paternità discussa o sconosciuta. Le letterature antiche e moderne sono ricche di testi anonimi o pseudonimi di cui gli studiosi hanno cercato di inviduare l'autore fondandosi su argomenti storici e filologici (attestazioni di altre fonti, datazione e localizzazione dei testimoni in caso di manoscritti, presenza di citazioni ecc.), ma l'analisi stilistica è stato da sempre uno di questi argomenti, se non il principale. Molte ipotesi attributive sono state proposte o confutate sulla base di considerazioni linguistiche e formali, e certamente gli strumenti informatici consentono oggi di avere una base oggettiva per supportare le ipotesi di paternità letteraria.

(omissis)

In generale per questo tipo di indagini si usano software di comparazione che individuino quella che Labbé ha chiamato "distanza intertestuale" e che può essere misurata da diversi punti di vista: la percentuale di forme lessicali condivise fra due testi o autori, la percentuale di alcune parti del discorso (ad es. quante preposizioni usano), la percentuale di hapax o il rapporto di Guiraud (= ampiezza lessicale, cioè numero di types/quadrato della dimensione del corpus, cioè dei tokens). Tutte operazioni che di solito sono state concettualizzate già nella filologia tedesca dell'Ottocento e che in teoria si potrebbero fare a mano (e si possono fare su testi brevi), ma che l'informatica ora ci consente di realizzare su masse testuali imponenti e non controllabili a mano. Tutti questi metodi sono riuniti in un solo programma chiamato JGAAP («Java Graphical Authorship Program), sviluppato da Patrick Juola, che è diventato celebre per aver scoperto, proprio con questo software, che l'autrice del romanzo "The cuckoo's calling" scritto con lo pseudonimo di Robert Galbraith era J.K. Rowling, la creatrice di Harry Potter. Il programma è liberamente scaricabile con una breve guida e assomma una serie di metodi di analisi tutti matematici che si possono usare contemporaneamente o uno per uno per mettere a confronto due o più testi fra loro, ma finora il suo uso in Italia (ma anche all'estero) è stato scarsissimo se non inesistente, per evidenti difficoltà di funzionamento.



Una buona descrizione dell'interfaccia è in Canettieri 2013, che l'ha testato con pieno successo su opere medievali in lingue romanze, oltre che su Montale prima di Tomasi-Condello, e impiegato per contribuire alla definizione dell'autorialità del *Fiore,* che si conferma non dantesco. Ne riportiamo qui la sua limpida descrizione: «JGAAP utilizza un'architettura modulare, i cui livelli base sono l'uniformazione/regolarizzazione grafica del testo (*Canocization*), l'elemento stilometrico che si vuole processare (*Event Set Generation*), la modalità di selezione dell'elemento (*Event Culling*) e l'analisi statistica dei dati acquisiti (*Analysis*). Ognuno di questi livelli viene gestito da un'unica generica classe

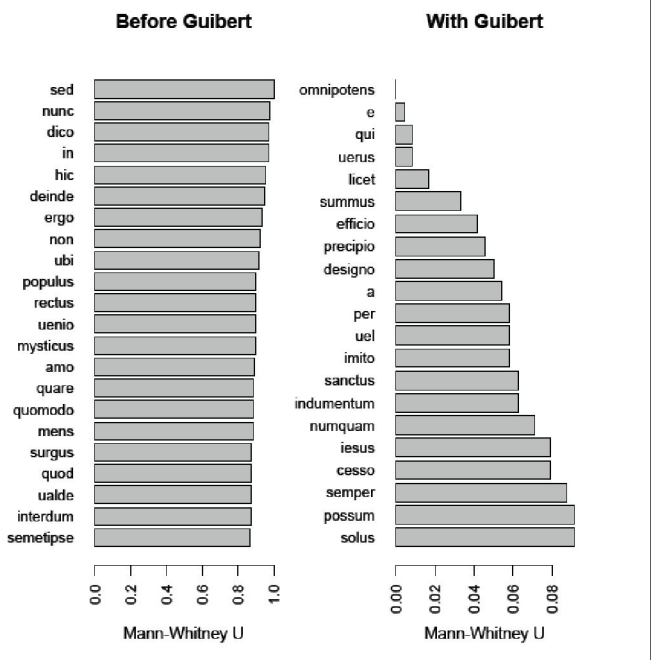
Java: così il modulo *Canocization* viene gestito dalla classe *Canonicizer,* il modulo *Event Set Generation* dalla classe *Event Drivers*, il modulo *Event Culling* dalla classe *Event Cullers* e il modulo *Analysis* dalla classe *Analysis Methods*. Fra gli *Event Drivers* si possono selezionare singoli caratteri o *n* caratteri contigui raccolti da una finestra scorrevole (*Characters* e *Characters Grams*), così come singole parole o *n* parole contigue (*Words* e *Word Grams*), la prima parola di ogni frase (*First Word in Sentence*), le parole con un numero variabile di lettere o vocali (*M- etter Words* e *Vowel M- Letter Words,* dove *m* e *n* sono parametri variabili), le parole vuote o parole-funzione utilizzate nello studio di Mosteller e Wallace sui *Federalist Papers* (*MW Function Words*), le parole rare, come quelle usate una o due volte in ogni documento (*Rare Words*), la lunghezza della frase misurata in parole (*Sentence Lenght*), i suffissi, intesi come le ultime 3 lettere di ogni parola (*Suffices*), le sillabe per parola, con un sistema molto semplificato in cui ogni vocale o gruppo di vocali viene computato come una sillaba (*Syllables Per Word*), ecc. Per selezionare la modalità di analisi si attiva la funzione *Event Culling,* che permette di scandagliare in tutti i doccumenti gli *n* fenomeni più rari o gli *n* fenomeni più frequenti o solo i fenomeni presenti in tutti i campioni (*Least Commons Events, Most Common Events, Xtreme Culler*), ecc. Infine è possibile scegliere fra gli innumerevoli tipi di analisi statistica, fra i quali ricorderemo *Burrow's Delta, Support Vector Machine* (*SVM*), nella versione gaussiana (*Gaussian SVM*) e in quella lineare (*Linear SVM), Linear Discriminant Analysis* (LDA), *Markov Chain Analysis, Native Nayes Classifier, PCA, SPCA, WEKA.* Alcuni metodi richiedono la selezione delle funzioni di distanza (*Distance Functions*), come la *Cross Entropy,* la *Lempel-Ziv-Welch nearest Neighbor Classifier* e molte altre.» (Canettieri)

9. *Due avvertenze. Valore statistico dei campioni, prova del Chi quadrato, interrogazioni.*

A livello di leggi statistiche naturalmente lavorare su campioni che non possono sempre essere scelti, come invece si fa nelle indagini demoscopiche o in quelle epidemiologiche o farmacologiche, pone delle limitazioni. In statistica linguistica infatti si dice (ma è anche questa una convenzione: vd. Sinclair 2004) che la soglia di credibilità statistica richiede un campione di almeno un milione di bit, ossia di caratteri; ma la dimensione dei dati letterari è imposti dalla storia i poemi di Virgilio sono quelli e non posso aumentarli o modificarli per renderli più adatti all'indagine. Quando la quantità dei campioni è ridotta e in generale per qualsiasi indagine statistica è tuttavia possibile sottoporre i campioni a testi di "significatività statistica", che in sostanza eliminano l'ipotesi che i risultati della ricerca siano dovuti al caso. Il più usato di questi test è quello detto del χ (chi, lettera greca) quadrato di cui esistono peraltro molte versioni, che in sostanza quantifica la differenza fra numeri osservati e numeri attesi (secondo la formula: quadrato della differenza fra numero osservato meno numero atteso diviso numero atteso). Se il dato ottenuto è superiore alla soglia del 5% si considera attendibile. Per questo conteggio esistono in rete numerosi calcolatori o fogli excel (ad esempio www2.unipr.it/~bottarel/epi/assoc/chi_qua.xls oppure socscistatistics.com/tests/chisquare/Default2.aspx o quantpsy.org/chisq/chisq.htm.). La controprova del Chi quadro è stata già applicata a fenomeni letterari, come le statistiche sul cursus (sistema di disposizione ritmica delle parole) dei testi medievali da Tore Janson (*Prose rhytm* 1975 e Orlandi 1978): ma siamo convinti però che quando si applica questo tipo di contabilità a dati non meramente linguistici bensì stilistici occorra tenere conto del fatto che non si tratta di dati "neutrali" ma di dati intenzionali, cioè di prodotti artistici che sono stati creati come sono in maniera quasi completamente volontaria, dunque la statistica a nostro avviso può essere considerata significativa anche su campioni piccoli e anche indipendentemente dal test del Chi quadro.

Ultima raccomandazione: pur essendo procedure semplici (più semplici a farsi che a descriverle) ha senso cimentarvisi solo in due casi: per imparare la tecnica oppure, meglio, per testare un'ipotesi critica. Non ha alcuna utilità imparare il manuale di Windows tutto intero e senza uno scopo concreto, meglio apprendere le operazioni che di volta in volta ci sono necessarie: allo stesso modo per la statistica letteraria è indispensabile e interrogare i programmi con una domanda precisa, con uno obiettivo chiaro e definito, ma soprattutto partire da una conoscenza dei testi su cui si lavora e dei problemi critici e filologici che li accompagnano.

Altro programma di scarso uso, ma più diffuso di JGAAP è il citato *Stylometry with R*, sempre nel pacchetto R, che realizza e rappresenta graficamente i confronti abituali.
É stato usato recentemente per analizzare la corrispondenza della scienziata, mistica e monaca del XII secolo Ilderde di Bingen, parte delle cui lettere sembra influenzata o redatta dal segretario Guiberto di Gembloux. Mike Kestemont e Sara Moens, assistiti dallo storico Jeroen Deploige, hanno applicato la lemmatizzazione e l'analisi delle parole vuote (function words) all'epistolario di Ildegarde, calcolando in particolare il cosiddetto Delta di Burrows, cioè la distanza fra un testo-campione casuale e una serie di testi oggetto di comparazione, secondo un protocollo ormai invalso dopo l'articolo di Argamon (Argamon 2008). Riproduciamo solo un diagramma che analizza la frequenza delle parole prima dell'intervento di Guiberto e dopo secondo il test chiamato U-test o test di Wilcoxon-Mann-Withney che si può usare direttamente online (http://www.socscistatistics.com/tests/mannwhitney/).



Uno dei risultati del lavoro di Kestemont-Moens è stato la dimostrazione, peraltro ancora fragile, della preminenza del fattore-autore rispetto al fattore-tema nel determinare il lessico di un testo. Questo sarebbe un risultato di straordinaria importanza se venisse provato da un'indagine a tappeto su un campione molto esteso, impresa che aspetta i critici letterari e i linguisti di ogni settore.
Per il latino medievale un caso di cui ci siamo occupati direttamente è quello delle *Epistolae duorum amantium,* una raccolta di 113 lettere (o 116, secondo i conteggi) d'amore scambiate fra un uomo e una donna, che sono state scoperte intorno al 1974 e pubblicate per la prima volta nel 1975 come possibili lettere giovanili di Abelardo ed Eloisa, un famoso filosofo francese del XII secolo e una sua coltissima allieva che poi ci hanno lasciato letteea autentiche di alto livello intellettuale ed emotivo. Molti hanno

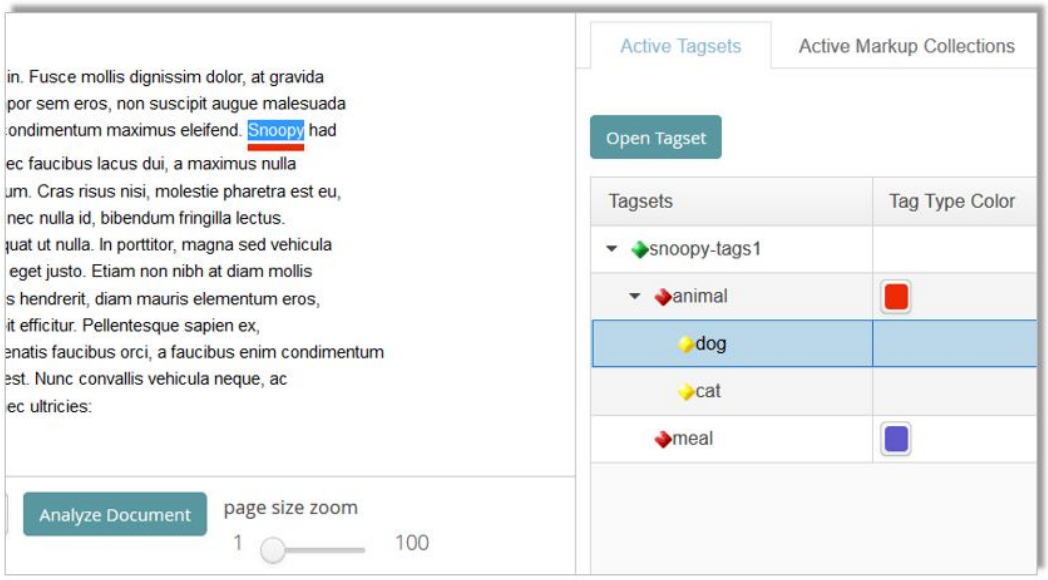
provato a sostenere o confutare l'attribuzione di queste lettere alla celebre coppia, e a noi è stata chiesta un'analisi lessicale che abbiamo realizzato con il tool successivamente messo in linea come *Lexicon* e illustrato sopra (Stella 2008), mettendo a confronto le frequenze delle function words dell'amante-uomo con Abelardo e quelle dell'amante-donna con Eloisa e inserendo testi-controprova per essere più sicuri che il risultato fosse significativo o comunque non fosse casuale o comune a tutti gli autori, e infine calcolando i tassi di sovrapposizione o esclusione delle forme lessicali (cioè le percentuali di uso non uso delle stesse parole da parte dei due testi). Il risultato, semplificando un'indagine abbastanza complessa, fu che gli autori erano due (contro l'ipotesi di un falso redatto da un autore solo), con evidente differenziazione lessicale e semantica (affettiva) del linguaggio usato, ma che non c'erano prove sufficienti per un'attribuzione ad Abelardo delle epistole dell'uomo. In precedenza un autore statunitense aveva lavorato con 5 function words basandosi però su cifre assolute (cioè ad esempio il numero totale di *enim* in Abelardo e nell'amante-uomo), ma cifre assolute ovviamente possono dipendere da quanto è lungo il testo-campione. Gli unici dati statistici utili sono quelli relativi, ossia le percentuali. I numeri assoluti ovviamente sono utili se si prendono campioni della stessa lunghezza, ma in letteratura questo equivale quasi sempre a privarsi di dati indispensabili e non sostituibili.

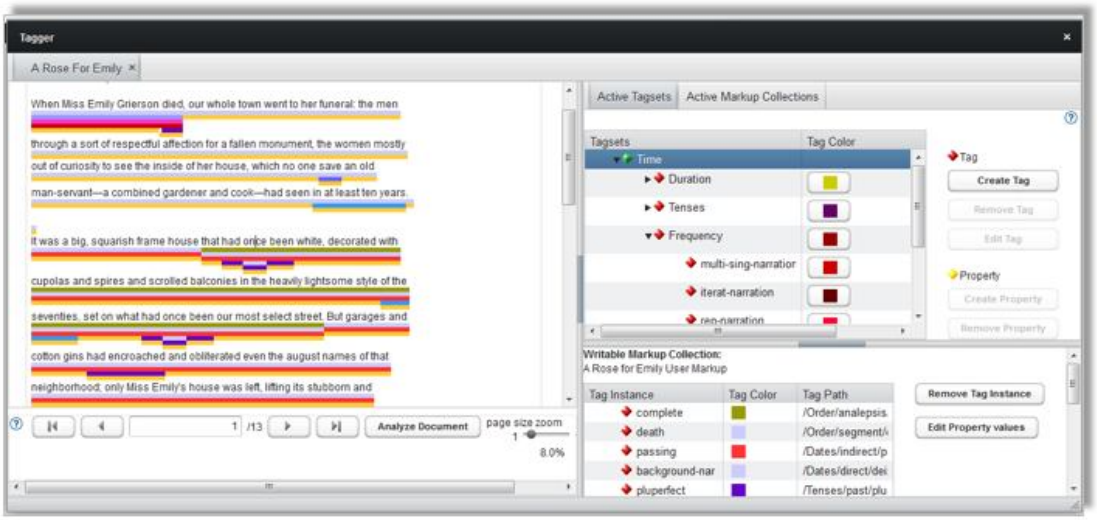
10. *Narratologia computazionale*

Questa – che distinguiamo dall'analisi statistica dei testi ma ne è un evidente sviluppo - è l'applicazione più nuova e ancora meno praticata delle Digital Humanities letterarie. Nel settore italianistico Giuseppe Gigliozzi (Gigliozzi 1993) vi aveva dedicato studio pioneristici, applicandoli soprattutto a testi con un livello di schematizzazione narrativa molto alto come le fiabe (Cenerentola e Pinocchio) o le novelle di Pirandello (*La vittoria delle formiche*). Questi esperimenti, che usavano per lo più software oggi obsoleti come *Sebnet*, hanno fatto scuola sul piano metodologico e nel 2002 Myriam Trevisan li ha applicati nel suo studio sul *Fu Mattia Pascal* che usa il procedimento dell'atomizzazione, cioè la suddivisione del testo in tutte le particelle (aggettivi, predicati) che definiscono il campo semantico di un nome, e soprattutto di un personaggio, come aveva tentato Alinei su Calvino già nel 1984, con macchine rudimentali. Si creano così i *cluster* che abbiamo visto alla base del topic modelling e altri tipi di analisi quanti-qualitative. Gli strumenti di analisi sono tabelle in cui ogni colonna è riferita a un ruolo del protagonista (bibliotecario, figlio, giovinetto, marito, giocatore, viaggiatore, ospite, redivivo) e i campi di ogni colonna includono gli elementi con cui si relaziona, gli attributi e le apposizioni correlate, gli stati d'animo, gli atti compiuti e i luoghi (Trevisan 2002: 199), documentando l'associazione costante di Mattia Pascal al tempo passato, come morto fra i vivi, e l'incomunicabilità fra i personaggi del romanzo.

Di vera e propria narratologia computazionale si è cominciato a parlare soprattutto dal workshop di Istanbul 2012 *Computational Models of Narrative*, partendo dai precedenti lavori di formalizzazione degli schemi narrativi come la *Morfologia della fiaba* di Vladimir Propp (1928), l'indice dei motivi di Annti Aarne e Stith Thompson (Aarne-Thompson 1961, poi aggiornato da Uther 2004) e la taassonomia di *Discours du récit* di Gerard Genette (1972). Nel 2009 cominciò a essere testato il software CATMA (Computer Aided Textual Markup annd Analysis, catma.de), nati da una riscrittura del padre di tutti i software di analisi testuale, ossia TACT (di John Bradley e Ian Lancashire). CATMA lavora come tutti i programmi di questo tipo, cercando cooccorrenze, indici di frequenza ecc., ma richiede una marcatura XML-TEI fatta in parte a mano, perché molti fenomeni narratologici non sono riducibili a segni lessicali la cui ricerca sia automatizzabile. Questo però consente una elasticità e una modificabilità praticamente infinita che non è possibile

in software automatici. Ci si può, anzi si dovrebbe, lavorare in più persone, dato che l'interpretazione dei passi può essere controversa: Meister 2014 cita al proposito una versione collabortiva di CATMA denominata *CLEA* (Collaborative Literature Exploration and Annotation) che avrebbe anche esteso l'annotazione rendendola semiautomatica, ma il progetto, a conferma dell'estrema fragilità di questi strumenti, è fermo al 2016 e il sito di riferimento citato da Maister ctma.de/clea.html) già irraggiungibile dopo soli due anni.





11. Il Distant Reading

Oggi analisi di questo tipo si sono evolute in forme assai più sofisticate, la cui elaborazione si deve soprattutto al LitLab di Stanford e al metodo denominato "Distant reading", inventato da Franco Moretti, già autore dei volumi su *Il Romanzo* e sulle *Opere-mondo*.

Nell'omonimo volume del 2013 Moretti proponeva alcune applicazioni rivoluzionarie del metodo accanto a studi più tradizionali di tipo storico-culturale. Una riguardava la statistica relativa alla lunghezza dei titoli dei romanzi in rapporto all''incremento della loro produzione nell'Inghilterra del XVIII secolo («market expands, and titles contract»). Un altro (*Network Theory, Plot Analysis*), che riprendeva assunti del suo precedente lavoro (*Style, Inc.*) si interroga sulla necessità di una teoria del network (cioè dell'analisi delle

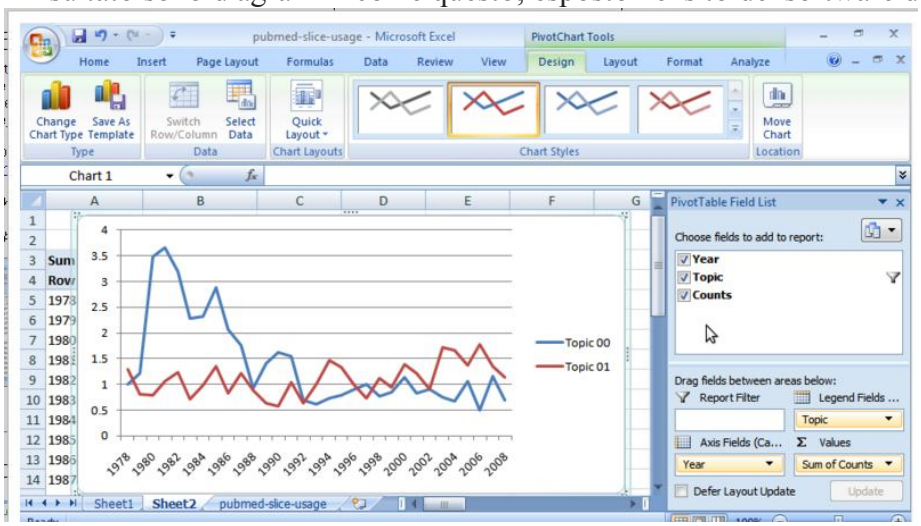
relazioni, ad esempio fra personaggi) e dimostra che, pur non essendo necessario teorizzare il metodo, è stato necessario effettuare statistiche per capire le relazioni fra i personaggi di Amleto per capire alcuni aspetti del dramma che fino ad allora non erano emersi, apprezzando l'emergere di prove quantitative in letteratura: «oggi possiamo replicare in pochi minuti ricerche che richiesero a un gigante come Leo Spitzer mesi e anni di lavoro. Se ci interessiamo di fenomeni linguistici o sstilistici possiamo effettuare operazioni che le generazoini precedenti potevano solo sognare». La *Network theory* è definita così come «una teoria che studia le connessioni (usualmente chiamate edges) all'interno di un ampio gruppo di soggetti (chiamati nodes o vertices)», la cui analisi rivela caratteristiche inattese di grandi sistemi. Il metodo prevede che i vettori colleghino i personaggi se si sono scambiati parole e documenta da una parte dati che si aspettavano (Amleto è il centro del network), ma senza di lui si intravede una bipartizione fra la corte da una parte e la zona col fantasma e Fortebraccio dall'altra, e soprattutto si scopre che in termini quantitativi Claudio è centrale quasi quanto Amleto (distanza dal centro 1,62 contro 1,45), ma in termini strutturali no, perché la sua figura interessa soprattutto personaggi periferici. Orazio invece è poco meno importante di Claudio in termini quantitativi ma molto di più in termini strutturali, perché è il punto cardinale di personaggi meno connessi, che «puntano al mondo oltre Elsinore», collegandosi al subplot inglese, norvegese e francese, o al mondo dei morti, creando l'idea che Elsinore sia la punta di un iceberg, la cui dimensione nascosta viene intuita attraverso questi personaggi.
L'applicazione a romanzo funziona meno perché nei romanzi l'azione non è dettata sempre dai personaggi, ma Moretti riesce a rendere significativi i risultati inserendo negli schemi direzione, peso e semantica e a collegarli a una teoria della diffusione del romanzo legata o paragonabile alla diffusione del sistema capitalista e all'evoluzione delle specie animali, con l'emergere di una costante: "l'intreccio dal centro, lo stile dalle periferie".
Più recentemente Moretti ha raffinato la sua analisi nel saggio *Operationalizing. On the function of Measurement in Literary Theory* (Moretti 2013b), applicando il metodo alla *Phèdre* di Racine, chiedendosi perché la massima parte dello spazio di parola è occupato non dagli scambi fra la protagonista e il marito Teseo o l'aspirante amante Ippolito ma con la "confidente" Enone: un risultato inatteso, scoperto solo grazie alle statistiche informatiche, che Moretti collega alla poetica neoclassica di Racine. Altre applicazioni riguardano il *Macbeth, Othello,* e soprattutto l'*Antigone* di Sofocle, dimostrando l'infondatezza testuale dell'interpretazione hegeliana della tragedia e trovando tracce di tendenze, ad esempio, nelle quali il protagonista non è una realtà fondamentale della costruzione drammatica ma solo una "istanza" speciale della più generale categoria di "centralità". Questo conferma che il nuovo approccio, scoprendo dati invisibili ai nostri sistemi percettivi e alle nostre capacità di gestirli, possono cambiare la storia della letteratura, oltre che la teoria. Secondo Moretti, e secondo noi, «il potere sperimentale senza precedenti degli strumenti e degli archivi digitali offre una possibilità unica di ripensare le categorie dello studio letterario» (Moretti 22013b: 119, trad. nostra). A suo avviso «le digital humanities prbabilmente non hanno ancora cambiato il territorio dello storico della letteratura o la lettura di testi singoli; ma l'operazionalizzazione (la formalizzazione, direbbe Orlandi) ha certamente cambiato o radicalizzato la nostra relzione con i concetti: ha suscitato le nostre aspettative trasformando concetti in formule magiche che possono chiamare alla vita un intero mondo di dati empirici; e ha affinato il nostro scetticismo perché se i dati si ribellano contro il loro creatore, allora il concetto è realmente in pericolo. É diventato immaginabile un programma di ricerca orientato dai dati e ricco di dati, che mette alla prova e, se necessario, falsifica [smentisce] la conoscenza ricevuta dello studio letterario» (ivi).
L'approccio di Moretti, che ha avuto eco mondiale e riscosso premi e riconoscimenti, ha suscitato prevedibilmente anche critiche, in parte fondate (specie quelle sulla scarsa

trasparenza dei dati e sull'assenza o bassa definizione della loro "pesatura", indispensabile per misurare il loro grado di significatività, in parte molto naif o talvolta pretestuose, come sempre accade quando una nuova tendenza teorica si affaccia all'orizzonte e gli specialisti abituati ad altri sistemi di analisi si sentono spiazzati o superati dalla novità. Ne rendiamo conto nel dibattito pubblicato su *The Mechanic Reader* (Ciotti-Stella 2015).

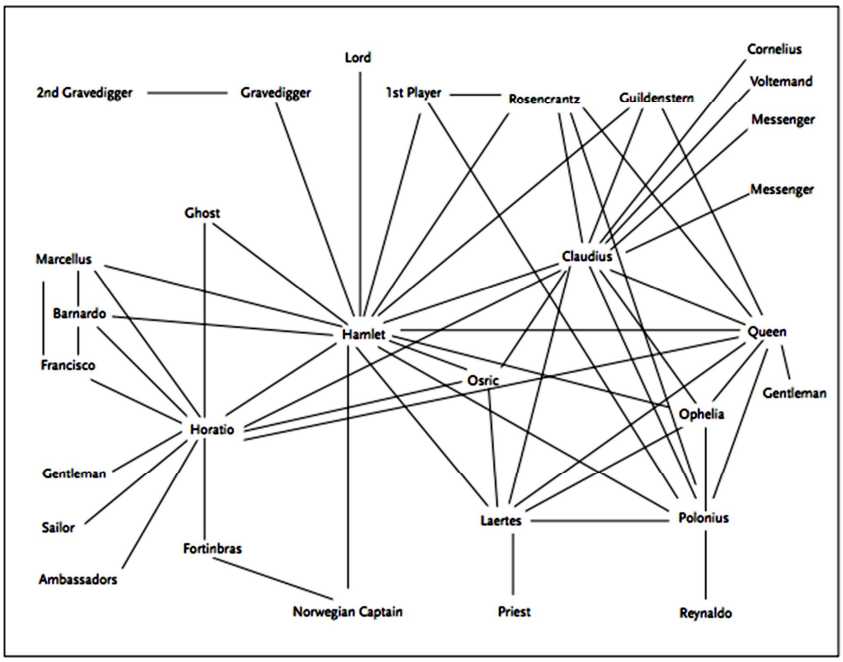
12. Metodi critici e ricerche storiche: topic modelling, network analysis, sentiment analysis

Il laboratorio di Stanford e dunque Franco Moretti e Matt Jockers hanno reso relativamente popolari e spesso credibili anche metodi che non producono statistiche da analizzare ma elaborazioni delle medesime mirate a un determinato scopo. Una di queste è il *topic modelling*, che potremmo tradurre con 'configurazione tematica', che individua appunto gruppi di semantemi riferibili a motivi o temi presenti nei testi esaminati. Lo strumento più comunemente (si fa per dire: in Italia è quasi ignorato o esercitato con un alto tasso di approssimazione) è la *Latent Dirichlet Allocation,* il cui software più diffuso è lo Stanford Toolkit (nlp.stanford.edu/software/tmt/tmt-0.4/). Naturalmente qui tutto dipende da come sono preparate le "librerie" (cioè le parole che concorrono a formare un topic, un tema) e da quali e quuanti topic si cercano in un testo: un sistema molto usato per sostenere che questi strumenti non funzionano è testare un'opera su topic che chiaramente le sono incompatibili (ad es. temi di saluto ai morti tipici degli epitafi in un corpus di poesia d'amore) e poi mostrare che il software non dà risultati. Oppure cercare in un testo temi che ovviamente devono essere presenti (ad es. l'amore nella poesia d'amore) per poi affermare che lo strumento non ha portato risultati nuovi. Come sempre, tutto dipende dall'operatore e dalla sua buona fede: la LDA mira infatti soprattutto a individuare topiche latenti ma presenti.

Il risultato sono diagrammi come questo, esposto nel sito del software di Stanford:



Analogamente la *Network Analysis* è un metodo usato nelle scienze sociali che individua reti di relazioni sviluppatee intorno a nodi (nodes) legati da ties o links: in letteratura i suoi usi più spettacolari sono legati ai capitoli che Franco Moretti ha dedicato all'*Amleto* di Shakespeare e all'*Antigone-Eumenidi* di Sofocle, individuando agglomerati di relazioni (il gruppo intorno a Claudio e quello intorno a Orazio) che l'analisi letteraria "tradizionale" non aveva messo in luce e forse non poteva mettere in luce. Vediamo qui di seguito il suo celebre diagramma:



Il diagramma è ottenuto inserendo in un programma il numero delle battute del testo in cui il dialogo è fra due personaggi: queste battute vengono rappresentate attraverso vettori grafici fra i medesimi, scoprendo così aggregazioni e polarizzazioni inattese.
Anche in questo caso uno degli argomenti polemici più infantili che viene utilizzato per svalorizzare queste ricerche è che non ci sarebbe stato bisogno del software per individuare le relazioni: di fatto però nessuno, prima che fosse disponibile il software, si è preso la briga di contare a mano le battute divise per personaggi e di riflettere criticamente sui risultati, e la cosa è stata possibile solo grazie alla *network analysis* e alle competenze letterarie di Moretti. Se mai si potrebbe obiettare che il diagramma non rappresenta la dimensione quantitativa (cioè la numerosità di occorrenze di una relazione) perché tutti i vettori hanno uguale spessore. Ma di questo rendono conto altri diagrammi. I programmi che è possibile usare sono molti (basta dare un'occhiata all'articolo *Social Network Analysis* di Wikipedia) ma uno di quelli più diffusi si trova nel pacchetto "R", che contiene anche altri strumenti stilometrici.

Un altro uso di R consente quella che è chiamata la *sentiment analysis,* cioè l'analisi dell'atteggiamento soggettivo di un parlante (nel caso di linguaggio naturale) o di un personaggio (nel caso di opere letterarie) o di un mittente (nel caso di epistole) basato su una sorta di varietà del topic modelling che isola nuclei semantici legati alle emozioni: a questo scopo è stato costituito perfino un vocabolario multilingue di lessico delle emozioni (che può essere usato sia per indagini di mercato sia per studi letterari o psicologici) come *EmoLex*. Di solito questo tipo di ricerca, su cui si sono sollevati molti dubbi per l'inevitabile grosslanità di una statistica meccanica su sfumature così delicate dell'espressione letteraria, si basa sull'estrapolazione di dati dei social media (Facebook, Twitter ecc.), ma esistono anche tool appositi come il celebre *Suyzeth Package*, sempre all'interno del pacchetto R, frequentemente usato nel LitLab di Stanford, anche se l'autorità in questo campo è Mohammad. L'analisi, che ha avuto un certo successo su Joyce, è in corso di applicazione all'epistolario di Svevo nell'*Archivio Digitale Italo Svevo* (vd. Fenu 2017). Ne riproduciamo qui alcuni diagrammi.

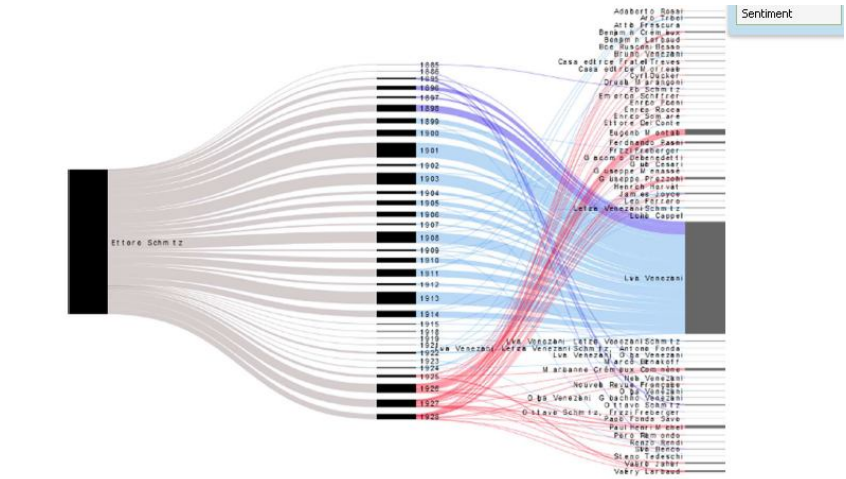


fig. 1. diagramma alluvionale del carteggio di Italo Svevo ottenuto da *data frame csv* tramite *tool* RAW

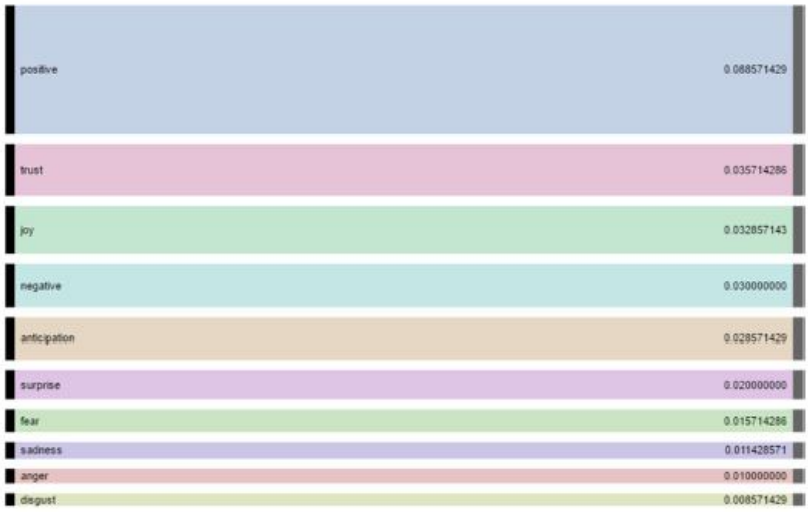


fig. 3: rappresentazione del *sentiment* della lettera di Svevo a Montale datata 17 febbraio 1926 tramite *tool* RAW

12. Studi sul genere letterario

Studi sugli aspetti linguistici del concetto di genere letterario hanno cominciato a diffondersi soprattutto negli ultimi anni, a partire da Swales 1990. Nel '99 Maria Wolters e Mathias Kirsten hanno presentato un esperimento di categorizzazione del genere sulla base della presenza di determinate parti del discorso[1], applicate al corpus LIMAS di 500 testi tedeschi contemporanei di circa 2000 parole ciascuno, basato su classi di genere estremamente ampie, come testi politici, testi giuridici, testi economici, testi narrativi ecc., che noi definiremmo piuttosto classi tematiche e non generi letterari. Criteri seguiti sono: TTR, Sentence Lenght (Words per Sentence), Word Lenght (Characters per Words) e POS frequency (Wolter-Kirsten 1999). Il risultato è esposto in una serie di algoritmi difficili da comprendere per non esperti di statistica ma abbastanza vaghi da decretare il fallimento di analisi fondate su categorie linguistiche troppo generiche, ma lascia la certezza che la scelta del campione pesa sugli esiti, che le function words, le forme finite dei verbi, così come i nomi, sono meno frequenti in determinate categorie testuali come i testi accademici. Nel 2007 Karlgren ed Eriksson hanno presentato un contributo su *Authors,*

[1] *Exploring the Use of Linguistic Features in Domain and Genre Classification*, in «Proceedings of EACL '99» pp. 142-9. Criteri seguiti sono: TTR (Type/Token Ratio), Sentence Lenght (Words per Sentence), Word Lenght (Characters per Words) e POS frequency,

Genre, and Linguistic Convention che muoveva dallo stesso problema riscontrato da noi nelle ricerche sugli autori delle *Epistolae duorum amantium*: separare le caratteristiche linguistiche d'autore da quelle del genere per verificare se, ad esempio, la lingua o almeno il lessico di un autore presenta variazioni da un genere all'altro maggiori o minori di quelle che si riscontrano fra un autore e l'altro. Insomma, la lingua di Cicerone oratore è diversa da quella di Cicerone epistolografo più o meno di quanto la lingua media di Cicerone sia diversa da quella di Livio o Virgilio? Ha più implicazioni linguistiche il genere o l'autore o l'epoca? Karlgren ed Erikkson si esercitavano nella misurazione di avverbi e tipi di frasi nei vari generi di articolo di giornale di un'annata del «Glasgow Herald», concludendo che gli avverbi si trovano in ogni sottogenere ma la loro collocazione è individuale, e giungendo anch'essi a esiti interlocutorii una loro problematizzazione del concetto di genere.

Le basi più solide del metodo restano dunque, oltre ai lavori seminali di Sinclair[2], l'esplorazione effettuata da Douglas Biber nel 1991 in *Variation across speech and writing*, dove elenca le frequenze di alcune decine di dati, dalla ratio type/tokens, cioè il rapporto fra forme e occorrenze, alla presenza di determinati tempi verbali, o congiunzioni, o proposizioni divise per alcune decine di generi più specificamente riconoscibili: tra queste una tabella è dedicata alle *personal letters*. Fra gli studi più recenti il reading antologico curato da Tony McEnery e Richard Xiao-Yukio *Corpus-based Language Studies* (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 2001 e Bybee-Hopper 2001.

Studi sugli aspetti linguistici del concetto di genere letterario hanno cominciato a diffondersi soprattutto negli ultimi anni, a partire da Swales 1990.

Nessuno di questi studi prende in esame *corpora* di testi antichi, e fra i pochi precedenti in questo campo possiamo citare solo il lavoro di Luon- Mellet 2003 sul corpus degli storici latini che dimostra come l'uso di determinate caratteristiche sintattiche si distribuisce nel *corpus Caesarianum* negli altri storiografi latini: tempi e modi verbali, parti del discorso, uso dei casi separano nettamente Cesare e i suoi epigoni da Sallustio, Tacito e Curzio Rufo, individuando sottodistinzioni, come, all'interno del corpo cesariano, quella fra il *Bellum Hispanicum* e le altre opere.

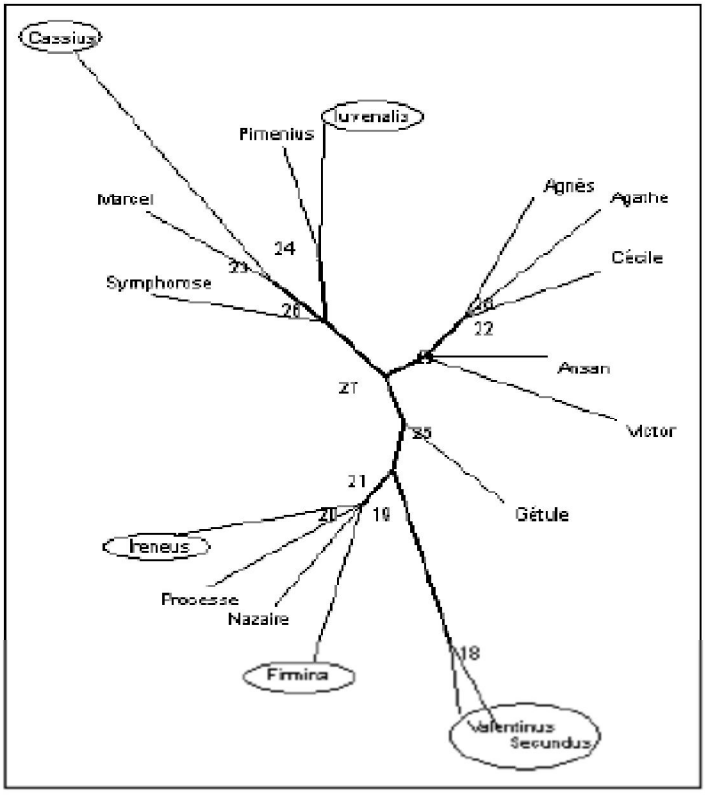
L'esperimento da noi tentato (Stella 2012, Stella 2015) sul linguaggio dell'epistola latina da Cicerone a Petrarca allarga ulteriormente, rispetto al modello di Luon-Mellet, l'arco cronologico considerato, ponendo anche la questione della eventuale continuità di genere fra latino classico, imperiale, medievale e umanistico, ed entra più specificamente in caratteri individuali come le scelte lessicali, affrontando in questo modo anche la questione dei rapporti di dipendenza stilistica o lessicale fra autori o della maggiore prossimità lessicale fra un autore umanistico e uno medievale o classico.

L'esperimento si basava su cinque tipi di analisi: le parole con frequenza più alta, le parole vuote, la TTR (Type/Token Ratio), le parole proprie a ogni autore e la quota di condivisione delle forme lessicali (token overlap), quest'ultima ricavata sia sulle forme singole che sui clusters (che il programma chiama *locuzioni*). Ogni tipo ha fornito risultati utili, come la maggiore frequenza del pronome di seconda persona (*te*) in epistolari meno teorici e intellettualizzati, l'emersione nei testi medievali di preposizioni probabilmente "spinte in alto" dal loro uso nella lingua quotidiana e vernacolare, la differenziazione di stili (uno stile *enim*, caratteristicco dei classici e di Petrarca, e uno stile *autem* tipici di Agostino, Alcuino e Abelardo), l'alta varietà lessicale in Plinio e Petrarca rispetto agli altri, la tendenza a una maggiore condivisione del lessico con l'autore immediatamente precedente (che individua una sorta di legge della contiguità diacronica e riconosce

[2] Soprattutto *Corpus, Concordance, Collocation* (1991), *Reading Concordances* (2003) e *Trust the Text* (2004).

maggior forza a questa continuità, sia essa di uso o di imitazione, rispetto a quella del genere letterario). All'interno di questa sequenza si evidenziano poi prossimità di tipo storico-sociale, come il lessico professionalmente intellettuale e "cittadino" che collega Abelardo, filosofo parigino, a Petrarca umanista combattuto fra spazio urbano e agreste. I clusters evidenziano l'uso frequente di locuzioni che introducono citazioni o memorie testuali, mentre le parole "individuali" caratterizzano il mondo privato di ogni autore, o almeno la sua rappresentazione epistolare, cioè amicale (ad esempio *paupertas, querela, fastidium, calamus, Ital-* in Petrarca, ognuno a capo di un plesso tematico ben noto agli studiosi del poeta aretino).

Un esperimento successivo ha tentato il medievista Edoardo D'Angelo in collaborazione con la medievista "computazionale" Cathérine Philippart de Foy, collaboratrice del LASLA (il laboratorio di analisi informatizzata delle lingue antiche a Liegi, uno dei più antichi dopo quello di padre Busa), che in nel 2013 hanno analizzato un corpus agiografico umbro misurando la distanza intertestuale di parametri come il POS (le parti del discorso), che ha confermato il raggruppamento dei testi per sottogenere (da una parte le *Vitae*, dall'altra le *Passiones* più dialogate). Questo è u esempio dei risultati, rappresentati graficamente in quello che si chiama diagramma arboriale (ad albero):



Coniugandosi con i dati storico-filologici i risultati dell'analisi informatica portano all'identificazione di gruppi di testi provenienti dalla stessa area geografica e dagli stessi centri di produzione letteraria e individua con relativa certezza parentele fra i testi, prima evidenziando le aggregazioni legate al genere letterario, poi i raggruppamenti interni a ogni genere.

13. *Strumenti per lo studio di formule e clausole*

Nel campo delle lingue classiche, latine in particolare, troviamo gli strumenti più sofisticati per la produzione di dati utilizzabili in analisi stilistiche. Fra i più antichi *Poesis* di Mastandrea e Tessarolo, poi evoluto nel magnifico portale *Musisque deoque* (http://www.mqdq.it/public/) che consente di effettuare letture e ricerche su tutta la poesia latina dalle origini al VI secolo, inclusa quella epigrafica, in edizione critica e talora anche sulle varianti, con possiblità di cercare qualsiasi parola, serie di parole, specie in clausola, cioè alla fine del verso, selezionando se necessario un determinato schema metrica o periodo temporale o gruppo di autori.
Questo serve a una vasta serie di obiettivi di ricerca: a comprendere meglio lo stile di un autore, a individuare imitazioni, fonti e riprese, a identificare autori di frammenti non riconosciuti, a ricostruire catene di reimpieghi formulari, a sviluppare ricerche tematiche ecc., ma il sistema ha potenzialità ancora inesplorate: un esempio inatteso lo ha fornito Paolo Mastandrea nel contributo a *The Mechanic Reader* (2015), dimostrando imitazioni di Marziale da Catullo guidate più dalla fonetica che dalla semantica o dalla metrica, e dunque individuabili solo da un software. Vi mostriamo qui un esempio di risultato di query su una clausola virgiliana, che "scopre" la sua fonte lucreziana e le sue imitazioni successive, comprese le reinterpretazioni cristiane:

Altri software di questo tipo sono *Tesserae* (tesserae.caset.buffalo.edu/), un motore di analisi molto potente che ha la caratteristica di comparare i testi individuando in pochi secondi quanto "Catullo" c'è in "Virgilio", verso per verso, con lemmatizzatore interno e non ha un database di testi proprio ma attinge a quelli della biblioteca digital *Perseus*, molto ampia eppure molto limitata al latino classico. É in grado di effettuare anche analisi di frequenza, statistiche sui tri-grammi e perfino ricerche tematiche semantizzate, cioè non basate solo su elementi lessicali del tema in oggetto. Un nuovo tool (Tracer) è in via di costruzione da parte di Greta Franzini e lavora su qualsiasi set di testi preparati, indipendentemente da altre biblioteche disponibili (www.etrap.eu).