Competenze scrittorie e dinamiche sociolinguistiche: problemi e metodi

a cura di
Simone CICCOLONE, Antonietta MARRA, Giuliano MION

Cagliari
UNICApress
2026

Sezione Ateneo
RESOCONTI/18

Competenze scrittorie e dinamiche sociolinguistiche:
problemi e metodi
a cura di Simone Ciccolone, Antonietta Marra, Giuliano Mion

In copertina: *Riscritture dinamiche*, foto di Giuliano Mion

Cagliari, UNICApress, 2026 (http://unicapress.unica.it)

ISBN 978-88-3312-233-5 (versione online)
DOI https://doi.org/10.13125/unicapress.978-88-3312-233-5

Indice

Sulle potenzialità di un approccio informatizzato all'analisi di testi notarili medievali

Giovanni Abete, Elisa D'Argenio, Mariafrancesca Giuliani[1]

1. Introduzione

In questo lavoro presenteremo gli strumenti e le procedure adottate all'interno del progetto *GeoDocuM* (Geografie Documentarie Meridionali) per l'estrazione e l'analisi del lessico documentario latino che caratterizza i testi delle prassi legali prodotti nel meridione italiano in epoca altomedievale. Il progetto, sviluppato da Mariafrancesca Giuliani, Giovanni Abete ed Elisa D'Argenio, si inserisce nell'ambito degli studi di sociolinguistica storica dell'unità di ricerca *Latinitas langobarda* diretta da Rosanna Sornicola (http://www.latinitaslangobarda.unina.it/) e si contraddistingue per l'adozione di una prospettiva d'indagine macroscopica e intertestuale incentrata sulla variazione lessicale che associa analisi statistiche, rappresentazioni cartografiche ed analisi qualitative di dettaglio.

1 Questo contributo è stato realizzato nell'ambito del progetto di ricerca *Writing expertise as a dynamic sociolinguistic force: the emergence and development of Italian communities of discourse in Late Antiquity and the Middle Ages and their impact on languages and societies*, P.I. Piera Molinelli, finanziato dal MIUR (PRIN 2017WLBK3Z) ed è frutto del lavoro congiunto dei tre autori. Per quanto riguarda la stesura del testo, i paragrafi sono da attribuire nella maniera che segue: §§ 1, 2 e 4.1 a Giovanni Abete; §§ 3, 4.2, 4.2.1 e 4.2.2 a Elisa D'Argenio.

In altre sedi abbiamo descritto obiettivi e metodi del progetto *GeoDocuM*, concentrandoci soprattutto sulle procedure di georeferenziazione e rappresentazione cartografica dei dati lessicali (Giuliani *et al.* 2023, 2024a). In questa sede intendiamo soffermarci invece sulle modalità di interrogazione del corpus e, nello specifico, sulle potenzialità di un approccio informatizzato all'analisi dei testi. Per tutte le fasi di lavoro di *GeoDocuM* (creazione del corpus, interrogazione, analisi statistica e rappresentazione cartografica dei dati lessicali), ci avvaliamo di procedure automatizzate attraverso l'uso di *script* nel linguaggio di programmazione *R* (R Core Team 2020). Per l'analisi dei testi, ci serviamo in particolare delle funzioni offerte da *Quanteda*, una libreria di *R* finalizzata alla creazione e all'analisi di corpora testuali (https://quanteda.io/; Welbers *et al.* 2017).

2. L'allestimento di un corpus di testi notarili medievali

La base dati del progetto è costituita da un corpus di testi digitalizzati che mira a includere, nel lungo periodo, tutte le edizioni affidabili di documenti notarili latini redatti tra l'VIII e l'XI secolo nell'Italia meridionale peninsulare. Al momento sono confluiti nel corpus i testi già digitalizzati e disponibili nell'archivio ALIM 2.0 (http://alim.unisi.it/; Ferrarini 2017; D'Angelo/Monella 2019) limitatamente all'area considerata e datati entro il 1100. Si tratta dei documenti contenuti nelle seguenti raccolte documentarie (cfr. Giuliani *et al.* 2024a, anche per i dettagli bibliografici delle fonti considerate):

- Codex Diplomaticus Cavensis (voll. 1-10) = CDCv;
- Regii Neapoletani Archivi Monumenta (voll. 1-4) = RNAM;
- Codice Diplomatico Verginiano (voll. 1-2) = CDV;
- Codice Diplomatico Barese (voll. 1, 3-5 e 7-10) = CDB;
- Codice Diplomatico Pugliese (voll. 20-21) = CDP.

Dal punto di vista tecnico, il corpus consta di migliaia di file di testo in formato .txt (uno per ciascun documento notarile) e di una tabella di metadati che associa ad ogni file di testo tutte le informazioni disponibili (raccolta, volume, numero del documento, data di redazione, luogo di riferimento, ecc.). Sia i file di testo che la tabella dei metadati sono stati realizzati automaticamente per mezzo di *script* di *R* a partire dai file di testo annotati in XML-TEI che sono liberamente scaricabili dall'archivio ALIM. Una volta opportunamente predisposti, questi materiali vengono

implementati in un corpus testuale gestito attraverso la libreria *Quanteda* di *R*, potendo così usufruire di tutti gli strumenti di analisi testuale offerti da questa libreria e delle ancor più ampie risorse di un linguaggio di programmazione come *R*.

Disporre di un ampio corpus di testi digitalizzati e accompagnati dagli opportuni metadati consente, tra le altre cose, di ottenere delle statistiche descrittive sulla consistenza e i limiti del corpus. A questo proposito, la libreria *Quanteda* offre delle funzioni utili a calcolare per ciascun documento e, secondariamente, per ciascuna raccolta documentaria, la lunghezza in parole, la lunghezza in caratteri e le forme uniche attestate[2]. Il corpus di *GeoDocuM* si compone dunque di 2.593 documenti, per un totale di 1.719.731 parole e 10.535.341 caratteri. In media ciascun documento contiene 663 parole. Le forme uniche attestate nell'intero corpus sono 727.884.

Come mostra il grafico in figura 1, la grande maggioranza dei documenti proviene dalle raccolte documentarie campane, e in particolare dal *Codex Diplomaticus Cavensis* (64%) e dai *Regii Neapoletani Archivi Monumenta* (20%). I documenti del *Codice Diplomatico Verginiano* costituiscono solo il 4% del totale, mentre i documenti delle raccolte pugliesi (*Codice Diplomatico Barese* più *Codice Diplomatico Pugliese*) assommano al 12% del totale. Ad integrazione del grafico, sono riportati in tabella 1, per ciascuna raccolta documentaria, il numero di documenti, il numero di parole e il numero di forme uniche attestate.

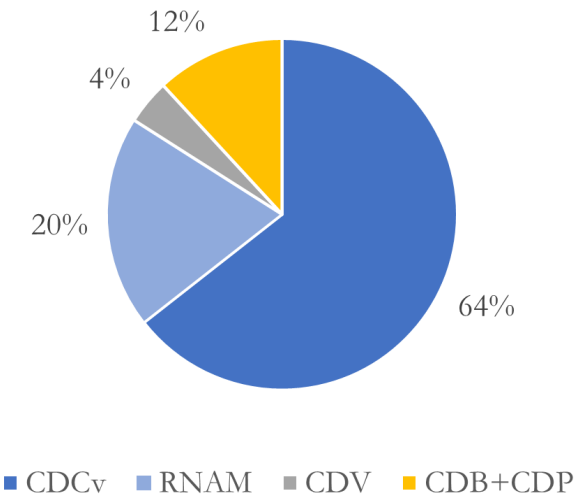

Fig. 1 – Percentuale di documenti presenti nel corpus per raccolta documentaria.

2 Rendiamo con forme uniche l'inglese unique forms, espressione con la quale si indica nella letteratura sul *text-mining* l'elenco di tutte le forme attestate, privo di eventuali doppioni. Una forma unica è dunque una sequenza esatta di caratteri, che occorre una o (solitamente) più volte in un dato corpus.

Tab. 1 – Numero di documenti, di parole e di forme uniche distinti per raccolta documentaria.

Raccolta	Documenti	Parole	Forme uniche
CDCv	1.671	6.501.247	11.394
RNAM	507	2.267.359	3.402
CDV	107	401.797	1.206
CDB	209	903.843	3.510
CDP	99	461.095	1.225

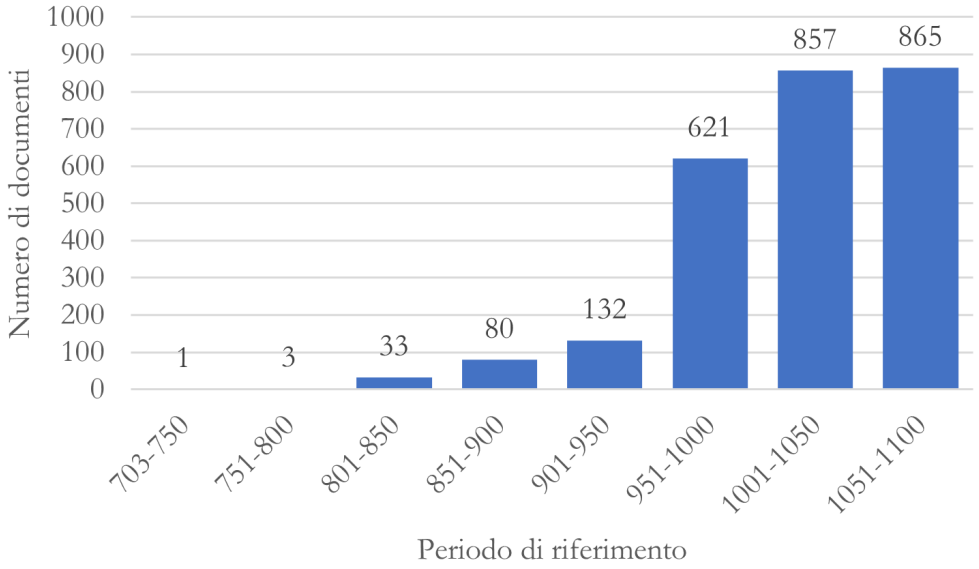

Fig. 2 – Distribuzione cronologica dei documenti del corpus.

Altre preziose informazioni sulle caratteristiche del corpus si ottengono dai metadati e specificatamente da data e luogo di riferimento dei documenti (v. nota 3). Per quanto riguarda la distribuzione cronologica dei testi, il grafico in figura 2 mostra la consistenza numerica dei documenti per ogni mezzo secolo: sebbene i più antichi testi del corpus risalgano all'VIII secolo, si tratta in realtà di pochissimi esemplari, mentre è solo a partire dal IX secolo che la documentazione si fa un po' più cospicua, per poi subire un netto aumento a partire dalla seconda metà del X secolo e attestarsi su valori molto alti per tutto il secolo XI (risalgono a questo secolo 1.722 documenti, il 66% del totale).

Tra i metadati, il luogo di riferimento del documento riveste una speciale importanza per il nostro progetto, che ha tra i suoi obiettivi quello di sviluppare rappresentazioni

cartografiche della variazione lessicale. Per agevolare le operazioni di georeferenziazione, i luoghi di riferimento indicati nei documenti sono stati ricondotti a un comune moderno corrispondente. Ciò consente di sfruttare per le operazioni di cartografazione le coordinate geografiche dei comuni moderni, che sono facilmente reperibili[3]. I documenti di *GeoDocuM* ricadono in 134 comuni moderni, ai quali ci riferiremo convenzionalmente come ai "punti" del corpus.

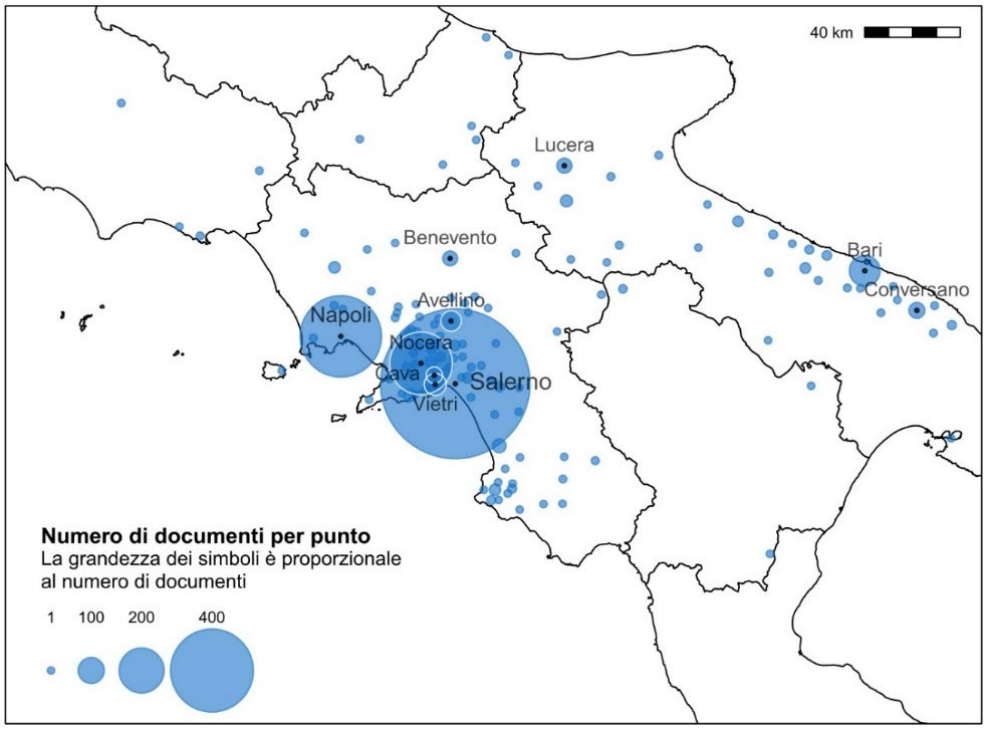

Fig. 3 – Distribuzione geografica dei documenti del corpus.

3 Per una discussione esaustiva sui problemi e i metodi dell'attribuzione e del trattamento delle localizzazioni geografiche in *GeoDocuM*, rinviamo a Giuliani et al. (2024a). In estrema sintesi, l'informazione d'elezione per la localizzazione dei documenti è per noi il luogo di redazione riportato nel testo dal rogatario, in quanto associa idealmente il luogo di formazione del documento all'area in cui il notaio esercitava la propria professione e alle relative tradizioni linguistiche e scriptologiche. In assenza dell'informazione sul luogo di redazione facciamo riferimento ad altri luoghi menzionati nel testo, come quello in cui è avvenuta l'azione giuridica o è collocato il bene oggetto del negozio. Dal punto di vista operativo, in tutti i casi in cui in ALIM il metadato relativo al luogo risultava mancante, rappresentato dall'indicazione generica di un'area (es. "Salerno, territorio di") o da un toponimo medievale, abbiamo proceduto, attraverso la lettura di prima mano dei documenti e la consultazione di letteratura pertinente, ad integrare, quando possibile, l'informazione, individuando un comune moderno di riferimento.

Come mostra la carta in figura 3, solo alcuni punti sono rappresentati in maniera consistente: Salerno (752 documenti), Napoli (392 documenti), Nocera (293), Bari (125), Vietri sul Mare (78), Avellino (70), Cava de' Tirreni (48), Conversano (46), Lucera (41), Benevento (40). Di contro, bisogna anche rilevare che molti punti sono caratterizzati da una scarsa documentazione: 96 punti compaiono infatti in meno di 5 documenti, e di questi 52 (più di un terzo del totale) sono rappresentati da un solo documento (cfr. Giuliani *et al.* 2023: 1111).

3. Caratteristiche del corpus e formulaicità

Come evidente dalle informazioni appena fornite, il nostro corpus risulta non bilanciato dal punto di vista sia della distribuzione diatopica sia di quella cronologica. Questa condizione di partenza, inevitabile per un corpus storico, potrebbe porre in dubbio l'opportunità dell'uso di tecniche di analisi quantitativa e statistica e la rappresentatività dei dati che restituiscono. Ciò che però sostiene, a nostro avviso, la validità degli obiettivi di *GeoDocuM* è costituito, oltre che dalla combinazione delle analisi quantitative con le analisi qualitative e dai correttivi adottati, in particolare, per le rappresentazioni cartografiche (v. § 4.1)[4], dalla stessa tipologia testuale indagata e dal contesto storico-culturale in cui fu prodotta.

Fino all'avvento dei Normanni, l'Italia meridionale peninsulare si caratterizzava per un dualismo che affondava le proprie radici nelle alterne vicende delle contese tra Longobardi e Bizantini e nella divisione politica, con risvolti etnici, linguistici e culturali, tra vaste aree, per lo più interne, che ricadevano sotto il dominio dei primi e zone, prevalentemente costiere, che erano sotto l'egida, più o meno nominale, dei secondi[5]. Tali circostanze non hanno mancato di produrre riflessi anche nella documentazione notarile. Su uno strato di elementi condivisi che originano, da un lato, dalle strutture e dalle formule documentarie di epoca tardo-antica e, dall'altro, da simili esigenze nella gestione delle azioni giuridiche e della loro rappresentazione testuale, si modellarono schemi formulari parzialmente differenziati (cfr. Pratesi 1992). I tipi documentari si consolidarono quindi nelle diverse prassi notarili attraverso la reiterazione di formule, costruzioni sintattiche ed elementi lessicali, che mostrano sia caratteri di conservatività, in ossequio a modelli espressivi consacrati dalla tradizione giuridica,

4 Per una discussione più ampia su questo punto, rinviamo a Giuliani et al. (2023: 1110-1114).

5 I risvolti culturali e linguistici di questo dualismo sono stati discussi in dettaglio da Giuliani (2007) e Sornicola (2012).

sia segni di innovatività nell'accoglimento di termini della lingua quotidiana. Proprio sull'esistenza di similarità e contrapposizioni *GeoDocuM* fonda i presupposti delle proprie ricerche, mirando a sondare il "peso" e la profondità di solidarietà lessicali su diversa scala (municipale, areale, regionale), così come di selezioni lessicali esclusive di alcuni centri di scrittura.

4. Strumenti per l'analisi dei testi

4.1. Individuare le occorrenze di un tipo lessicale

L'analisi linguistica di un tipo lessicale presuppone la capacità di individuarne tutte le occorrenze all'interno del corpus di riferimento. Per fare ciò è necessario definire l'elenco delle forme uniche di un tipo lessicale, ossia l'elenco completo delle combinazioni di caratteri con cui quel tipo compare nei testi. Nella fattispecie di un corpus di documenti notarili latini medievali, questa operazione è complicata dal forte polimorfismo che caratterizza i testi e dalla scarsa standardizzazione ortografica. Pertanto, risulta spesso impossibile prevedere in anticipo tutte le varianti formali con cui un tipo lessicale si presenterà nei testi. Per ovviare a questo problema, esistono degli strumenti informatici che aiutano a individuare le forme uniche di un tipo lessicale in maniera veloce e parzialmente automatica. A questo proposito, due metodologie si sono rivelate particolarmente utili in *GeoDocuM*, da usare eventualmente in combinazione: una basata sulla distanza fono-ortografica dall'idealtipo, l'altra sull'uso delle espressioni regolari.

Per quanto riguarda il primo metodo, questo consiste nel definire un idealtipo e nell'individuare, attraverso un algoritmo, le forme abbastanza simili all'idealtipo. Per farlo, utilizziamo la distanza di Levenshtein (dL), ossia la somma degli inserimenti di carattere, cancellazioni di carattere o sostituzioni di carattere necessari per giungere da una forma A ad una forma B (il calcolo può essere effettuato in *R* attraverso la funzione "adist"). Poniamo ad esempio di voler cercare nel corpus tutte le occorrenze del tipo *anditus*, un termine che compare frequentemente nei documenti ad indicare un accesso o un diritto di accesso (cfr. Giuliani *et al.* 2023: 1112-1114). Il tipo presenta un notevole numero di varianti formali, che non è possibile conoscere preliminarmente alla ricerca. Una volta definito un idealtipo da usare come punto di riferimento,

ad esempio *anditu*[6], con un linguaggio di programmazione come *R* è possibile ottenere tutte le parole del corpus che presentano rispetto all'idealtipo una dL uguale o inferiore a una certa soglia impostata dall'utente. Nel caso specifico, le parole che presentano una dL ≤ 2 rispetto alla sequenza *anditu* sono le seguenti: *adit*, *aditi*, *aditum*, *ambitu*, *andat*, *andica*, *andita*, *anditam*, *anditas*, *andite*, *anditi*, *anditis*, *andito*, *anditu*, *anditum*, *anditus*, *angilu*, *anticu*, *audit*, *audita*, *audito*, *auditu*, *auditum*, *aunitu*, *binditu*, *cannitu*, *esditu*, *handita*, *handitum*, *indita*, *reditu*, *seditu*, *tandiu*. Come si può vedere, l'elenco contiene anche forme che non hanno nulla a che fare con *anditus*, ma il ricercatore esperto non avrà difficoltà a rimuoverle, ottenendo così l'elenco delle seguenti forme: *andita*, *anditam*, *anditas*, *andite*, *anditi*, *anditis*, *andito*, *anditu*, *anditum*, *anditus*, *handita*, *handitum*. L'elenco "sfrondato" potrebbe a questo punto dirsi completo, ma non possiamo esserne del tutto certi perché con una dL relativamente bassa (in questo caso ≤ 2) c'è il rischio di non catturare forme che risultino significativamente più lunghe dell'idealtipo. Ad esempio, se fosse attestata una forma come **anditibus* questa sfuggirebbe alla ricerca, in quanto caratterizzata rispetto all'idealtipo *anditu* da una dL = 4 (1 sostituzione + 3 inserimenti di carattere). Per evitare questo rischio, è certo possibile aumentare il valore soglia, tuttavia bisogna tener presente che valori troppo alti possono generare elenchi molto lunghi di forme, difficili da gestire manualmente. Ad esempio, nel caso in questione, una dL ≤ 4 produrrebbe un elenco di ben 2.157 forme uniche, che richiederebbero uno spoglio manuale lungo e tedioso.

Una soluzione più agile al problema appena esposto consiste nell'uso delle espressioni regolari. Un'espressione regolare è una sequenza di caratteri che specifica un preciso *pattern* da ricercare nel testo. In riferimento al tipo lessicale *anditus*, le forme che siamo già stati in grado di individuare condividono un *pattern* piuttosto semplice: includono tutte la sequenza "andit". Utilizzando un'espressione regolare[7], possiamo quindi individuare con *R* tutte le parole del corpus che presentano questo *pattern*: la ricerca produce l'elenco esatto di forme già individuate con il metodo basato sulla distanza di Levenshtein, con in aggiunta la forma derivata *anditellum*, che non eravamo stati in grado di individuare[8].

L'esempio appena considerato mostra come le espressioni regolari siano uno strumento potente e preciso; tuttavia, la correttezza di questo metodo di ricerca dipende in maniera cruciale dal modo in cui viene formulata l'espressione regolare e, in ultima

6 In base alla nostra esperienza, è preferibile scegliere come idealtipo una forma priva di consonante finale. Ciò consente, a parità di soglia impostata per la distanza di Levenshtein, di catturare un maggior numero di forme.

7 In questo caso l'espressione regolare è molto elementare e corrisponde semplicemente a "andit".

8 Si tratta di un hapax che compare in un documento cavense: Et ipsum anditellum per suprascriptas fines et mensuras quam commune dimiserunt, CDCv VIII, 124.

analisi, dalla correttezza delle predizioni formulate dal ricercatore. Se ad esempio, ignorando l'esistenza di forme inizianti per "h", avessimo ricercato le forme che iniziano per "andit"[9] anziché quelle che semplicemente contengono questa sequenza, avremmo mancato le forme *handita* e *handitum*. Poiché, come abbiamo già osservato, in testi come i nostri è difficile avere un'idea precisa di quali forme aspettarsi, può essere preferibile, soprattutto in fase esplorativa, un metodo di ricerca più libero e meno "guidato", quale è effettivamente il metodo basato sulla distanza di Levenshtein, al quale affiancare l'uso di espressioni regolari per ricerche più mirate.

Un ulteriore esempio aiuterà a chiarire la complementarità dei due metodi. La ricerca delle forme con una dL ≤ 2 rispetto all'idealtipo *barbanu*, che, come è noto, è una denominazione dello zio paterno (cfr. Giuliani *et al.* 2023: 1108-1109), consente di individuare la maggior parte delle varianti formali di questo tipo, come ad esempio *barbanem*, *barbaneo*, *varbano*, *varbanum*. Aumentando la soglia a ≤ 3, riusciamo a catturare anche una forma come *barvaneo*, con "v" postconsonantica, che precedentemente era sfuggita. È evidente come la varietà di forme dipenda in questo caso dalla possibile presenza del betacismo, oltre che dal polimorfismo della terminazione. Tenendo conto di ciò, è possibile formulare un'espressione regolare che catturi le forme inizianti per "barban" includendo anche le varianti con betacismo e tutte le possibili terminazioni, ossia "^(b|v)ar(b|v)an". La formula consente di individuare le seguenti forme (in ordine alfabetico): *barbane*, *barbanei*, *barbanem*, *barbaneo*, *barbaneoque*, *barbanes*, *barbaneum*, *barbaneus*, *barbani*, *barbano*, *barbanoque*, *barbanum*, *barbanus*, *barvaneo*, *varbaneo*, *varbaneoque*, *varbani*, *varbanis*, *varbano*, *varbanum*. Sono state, cioè, individuate tutte le forme uniche del tipo lessicale *barbanus*, incluse forme con *-que* enclitico quali *barbanoque*, *barbaneoque*, *varbaneoque*, che non era stato possibile scovare con il metodo basato sulla distanza di Levenshtein.

Una volta definito l'elenco delle forme uniche di un tipo lessicale, attraverso la funzione "kwic" (*keyword in context*) di *Quanteda* è possibile ricercarne in blocco le occorrenze nel corpus, e realizzare una tabella che includa per ogni occorrenza una certa porzione di cotesto, l'indicazione del documento in cui si trova, e tutti i relativi metadati.

A titolo esemplificativo, la tabella 2 mostra le prime cinque occorrenze nel corpus del tipo *barbanus*. L'ampiezza del cotesto, che normalmente fissiamo a dieci parole prima e dopo la parola chiave, è qui estremamente ridotta per ragioni di spazio. Per lo stesso motivo non sono riportati i metadati che accompagnano ciascun record.

9 L'espressione regolare sarebbe stata in tal caso "^andit", con il simbolo ^ a specificare l'inizio di parola.

Tab. 2 – Prime cinque occorrenze di barbanus *nel corpus (tot. 109) individuate con la funzione "kwic"*

Cotesto precendente	Keyword	Cotesto seguente
propinquo casam que fuit gaidonis	**Barbani**	mei in quibus, ut
alio capite abet fine landemari	**Varbanis**	meis, in ipso capite
et abbatis, qui fuit	**Barbanem**	meum, per hanc cartula
ipsius iohanni presbiteri et abbati	**Barbani**	meo, et ipse iohannes
rebus ipsa ipsius iohanni presbiteri	**Barbani**	meo datum fuit, sic

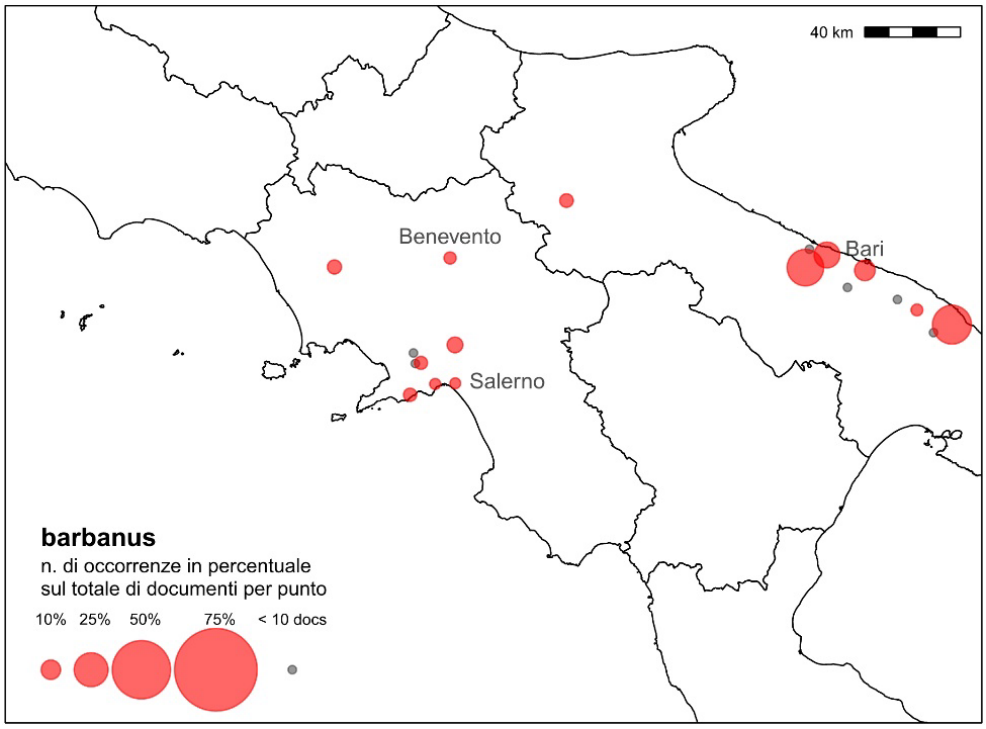

Fig. 4 – Distribuzione geografica del tipo lessicale barbanus.

La carta in figura 4 mostra, ad esempio, la distribuzione geografica del tipo lessicale *barbanus*: la grandezza delle "bolle" rende conto della frequenza di occorrenza di questo tipo in percentuale rispetto al numero di documenti disponibili per ciascun punto del corpus[10]; i pallini in grigio, di dimensione fissa, indicano attestazioni in punti che

10 Se uno stesso documento presenta più attestazioni di uno stesso tipo lessicale, queste vengono contate come una sola attestazione (cfr. Giuliani et al. 2023: 150).

sono rappresentati da meno di dieci documenti e per i quali il riferimento alla presenza in percentuale rischierebbe di essere inaffidabile. Con questi accorgimenti, a fianco delle attestazioni dei territori longobardi campani, vengono valorizzate anche le attestazioni di area pugliese, dove il tipo è infatti molto caratteristico, nonostante il minor numero di documenti disponibili per quest'area. Si tratta pertanto di soluzioni che riescono a fornire una rappresentazione quantitativamente adeguata della distribuzione geografica del tipo lessicale indagato, tenendo conto della disomogeneità di documentazione che inevitabilmente caratterizza un corpus storico come il nostro (cfr. §§ 2-3 e Fig. 3; per una più ampia esemplificazione rinviamo a Giuliani *et al.* 2023, 2024a).

4.2. Potenzialità delle analisi statistiche automatiche

Sebbene la casistica finora indagata nell'ambito di *GeoDocuM* abbia riguardato serie lessicali di interesse a noi già noto sulla base della letteratura o della conoscenza diretta dei testi, la formulaicità che contraddistingue i documenti notarili (cfr. § 3) è un dato che apre il campo a risultati non attesi, prodotti per il tramite di procedure automatiche fondate sull'analisi statistica. In particolare, discuteremo ed esemplificheremo di seguito le potenzialità di studio offerte dall'estrazione automatica delle collocazioni di due parole e delle forme uniche in specifici sottocorpora. Nella prospettiva di indagine adottata da *GeoDocuM* tali procedure si rivelano particolarmente utili, combinate con l'analisi della distribuzione diatopica, per individuare *pattern* lessicali che interessano in maniera maggioritaria o esclusiva tradizioni scriptologiche di specifiche aree.

4.2.1. Analisi delle collocazioni di due parole

Con il termine "collocazione" ci si riferisce al fatto che «certain lexical items tend to co-occur more frequently in natural language use than syntax and semantics alone would dictate» (Krishnamurthy 2006: 596). Si tratta quindi di combinazioni di parole che tendono a presentarsi contigue in un testo più spesso di quanto ci si aspetterebbe sulla base della pura casualità. L'identificazione e l'analisi delle collocazioni può essere d'aiuto per approfondire le associazioni semantiche e le relazioni concettuali presenti in un corpus e identificare usi formulari peculiari di gruppi di testi e, quindi, nell'ottica di *GeoDocuM*, di aree o località.

Per l'individuazione delle collocazioni nel nostro corpus ci siamo avvalsi della funzione "textstat_collocations" della libreria *Quanteda*. Applicando la funzione ad un corpus è possibile ottenere un elenco di collocazioni personalizzato in base a parametri impostati dall'utente (numero massimo di parole da includere nella collocazione, frequenza minima, ordinamento crescente o decrescente sulla base di un parametro scelto, ecc.). In questa sede presenteremo alcune delle collocazioni del nostro

corpus ottenute impostando una ricerca che restituisse le prime 100 combinazioni di due parole ordinate per valori di *lambda*[11] decrescente. Semplificando, il valore *lambda* misura la forza di associazione tra le parole in un testo sulla base di un calcolo che determina quanto la loro occorrenza insieme è superiore o inferiore rispetto alla frequenza attesa se le due parole fossero indipendenti. Un valore alto di *lambda* indica dunque una collocazione potenzialmente significativa e informativa del corpus, poiché data da una combinazione di parole che tendono a comparire contigue più spesso di quanto ci si aspetterebbe. Di fronte a un corpus come il nostro, però, è necessario tenere presente che un valore alto di *lambda* non è un indizio sicuro della significatività di una combinazione e, viceversa, combinazioni significative potrebbero non avere un valore alto di *lambda*: il calcolo della forza dell'associazione potrebbe essere infatti "disturbato" dal pervasivo e imprevedibile polimorfismo ed è pertanto sempre necessaria un'analisi di dettaglio dei contesti da parte del ricercatore.

Una volta ottenuto l'elenco delle collocazioni, abbiamo proceduto per ciascuna combinazione di parole alla ricerca di tutte le possibili varianti formali (secondo le procedure indicate nel § 4.1) e alla visualizzazione in forma tabellare dei contesti di occorrenza associati ai relativi metadati. Presentiamo di seguito alcune delle collocazioni che si sono rivelate di maggiore interesse.

Alla prima posizione dell'elenco, cioè con il valore di *lambda* più alto, troviamo la combinazione "christe fave" (48 occorrenze). Questa associazione di parole non presenta nel corpus alcuna variante formale. Il vaglio dei contesti di occorrenza e della loro distribuzione ha consentito di constatare che si tratta di un sintagma che costituisce l'invocazione verbale presente nella parte iniziale del protocollo di documenti essenzialmente napoletani (45 occorrenze localizzate a Napoli, 3 occorrenze a Pozzuoli).

Alla terza posizione occorre la combinazione "iugalium personarum" (63 occorrenze). La ricerca delle varianti formali ha consentito di individuare altre forme del sintagma (*iugalibus personarum*, *iugalium personam*) per un totale di 72 occorrenze complessive, tutte localizzate a Napoli. Il sintagma è adoperato sempre nel medesimo contesto d'uso, ossia per indicare che due persone menzionate come genitori di una donna sono coniugi uniti in matrimonio (*Certum est me eufimia filia quidam domini theodori.*

11 Il valore di lambda, calcolato dalla funzione "textstat_collocations", è una misura di associazione statistica utilizzata per quantificare la forza di coesione delle espressioni multiparola. Tecnicamente, per un'espressione target di numero K parole (specificato dall'utente attraverso il parametro size), lambda rappresenta il coefficiente del parametro di interazione K-dimensionale in un modello log-lineare saturo. Tale modello è adattato alle frequenze dei termini che costituiscono l'insieme delle espressioni multiparola ammissibili (cfr. la documentazione della funzione, a cui si rimanda anche per ulteriori dettagli tecnici e per i riferimenti bibliografici: https://quanteda.io/reference/textstat_collocations.html#).

et quidam domina theodonanda iugalium personarum, RNAM, I.1, 19, p. 70, rr. 4-5, Napoli, 934).

Rispettivamente alla quinta e alla tredicesima posizione occorrono le combinazioni "honeste femine" (49 occorrenze) e "honesta femina" (229 occorrenze). La forza della combinazione è rimarcata proprio dal fatto che il sintagma compare con due varianti formali nell'elenco. Il sintagma registra 296 occorrenze totali e risulta caratteristico della documentazione prodotta nei domini bizantini, poiché 288 occorrenze sono localizzate a Napoli e le restanti 8 in due documenti redatti rispettivamente a Gaeta e a Sorrento.

All'ottava posizione dell'elenco occorre il sintagma "christianissimis viris" (29 occorrenze complessive, tenuto conto delle varianti formali). La combinazione lessicale è usata esclusivamente in documenti redatti a Napoli e cooccorre in 28 casi con il verbo *adpretiare* nelle formule di sanzione per indicare gli uomini che saranno chiamati a fare la valutazione economica di un bene (*Si vero legitimi det memoratus filius meus eiusque heredibus medietate pretium quantum predicta terra appretiata fuerit a christianissimis viris*, RNAM, I.1, 19, p. 62, rr. 3-5, Napoli, 932). La ricerca dei contesti di occorrenza del tipo *adpretiare* nel corpus permette di rilevare che nelle località che invece ricadono nei domini longobardi gli aggettivi con cui sono qualificati questi *homines* sono *doctus*, *bonus* e *sapiens*, anche in combinazione (*per bonis et doctis hominibus* in un documento beneventano, *per doctos et sapientes homines* in un documento cavense).

Menzioniamo, da ultima, la collocazione che compare alla venticinquesima posizione dell'elenco, "scriptoris discipulo", poiché ci consente di mettere in luce un aspetto rilevante per valutare l'efficacia dell'interazione tra dato lessicale e dato geografico in *GeoDocuM*. Delle 32 occorrenze rilevate, 31 sono localizzate a Napoli ed una soltanto a Melfi, secondo i metadati forniti da ALIM. Chiaramente, il sintagma testimonia la presenza di apprendisti all'interno dell'organizzazione gerarchica dei notai laici della Curia napoletana (Martin 2011: 67-72). In realtà, infatti, la localizzazione lucana è un errore, poiché la data topica riportata nel protocollo del documento è *Neapolis* (CDCv I, 143): le solidarietà lessicali e le specificità storico-culturali di un'area possono consentire anche di individuare e correggere una localizzazione inesatta.

Le collocazioni di due parole con i valori di *lambda* più alti mettono dunque in luce, in particolar modo, combinazioni lessicali in uso esclusivamente a Napoli. Se ciò non desta di certo meraviglia, poiché «Napoli è stata la capitale di un piccolo ducato, contraddistinto da un notevole "particolarismo", rispetto ai contermini territori longobardi» (Cuozzo e Martin, 1995, p. 8) – "particolarismo" che ha interessato anche la

cultura grafica, linguistica e giuridica dei notai napoletani[12] –, nondimeno risulta interessante che la procedura automatica corrobori quanto già noto e contribuisca a meglio delinearlo, confermando i dati con l'immediatezza e l'esaustività delle sue ricerche e fornendone di ulteriori, meno indagati o attesi.

4.2.2 Sottocorpora e forme uniche

La possibilità di lavorare con sottocorpora definiti sulla base di parametri scelti dall'utente si dimostra particolarmente utile per identificare lessemi caratteristici o esclusivi di specifiche aree o località[13]. A titolo esemplificativo, illustreremo di seguito alcuni dei risultati ottenuti relativamente alle forme uniche esclusive di Napoli, Amalfi e Gaeta (nell'ottica di individuare usi linguistici peculiari dei domini bizantini *vs* domini longobardi) e di quelle pugliesi (al fine di identificare lessemi tipici delle fonti documentarie provenienti dalla Puglia *vs* Campania).

Dopo aver creato i diversi sottocorpora con la funzione "corpus_subset" della libreria *Quanteda* sulla base del metadato relativo al comune o alla regione e aver estratto le relative forme uniche da mettere a confronto, sfruttiamo la funzione di base "setdiff" di *R* che consente di ottenere la lista degli elementi di un insieme che non sono presenti in un altro insieme. I dati estratti tramite questa procedura necessitano naturalmente di essere raffinati, sia perché la presenza di una forma nell'elenco non garantisce, date le possibili varianti formali, che il lessema non sia attestato anche altrove[14], sia perché la numerosità delle forme uniche restituite (per il sottocorpus di Napoli, Gaeta e Amalfi, ad esempio, ben 4.911) richiede un intervento manuale di scrematura da parte del ricercatore sulla base delle proprie conoscenze e della propria esperienza. Nonostante questi limiti, e sebbene diversi dati siano già noti in letteratura, riteniamo che l'esplorazione dell'elenco di forme ottenute possa essere comunque di un certo interesse, soprattutto nella prospettiva – che *GeoDocuM* sperimenta – di sondare la variazione lessicale e di raffigurarne le configurazioni di distribuzione.

12 La questione del "particolarismo" dei curiali napoletani è discussa in prospettiva linguistica in Sornicola (2012), Giuliani (2012) e Giuliani et al. (2024b) a cui si rimanda anche per ulteriori riferimenti bibliografici.

13 Le ricerche automatiche che sfruttano i confronti tra sottocorpora sono precipuamente orientate a dare risalto alle differenze formali e lessicali e possono dunque contribuire a mettere a fuoco alcune specificità diatopiche. Per un'approfondita discussione sull'utilità di questo tipo di indagini e sui risultati a cui può condurre rimandiamo a Giuliani (2023), che applica il metodo dell'esplorazione dei "lemmari esclusivi" a specifici testi del Corpus TLIO (Tesoro della Lingua Italiana delle Origini), al fine di isolarne possibili localismi e diatopismi.

14 Anche in questo caso, come visto in precedenza, procediamo quindi alla ricerca di tutte le forme del lessema nel corpus e alla visualizzazione dei contesti di occorrenza con i relativi metadati.

Ci limitiamo qui a riportare a titolo d'esempio alcuni lessemi significativamente non attestati altrove, testimonianza di una tradizione scriptologica e formulare, di usi linguistici conservativi o a connotazione locale. Per il sottocorpus di documenti di Napoli, Gaeta e Amalfi segnaliamo: i grecismi *apothecella* 'piccola dispensa', *exadelfus*, *-a* 'cugino, -a' (cfr. Sornicola 2012: 38), spesso nell'espressione *exadelf- german-*, *ipotheca* 'magazzino' e *egripus* 'canale, solco di confine' (lessema in uso esclusivamente a Napoli)[15]; il già citato termine di parentela *iugalis* 'coniuge'[16]; *triclinium* 'camera da letto' (cfr. Sornicola 2015: 246); *spurcitia* 'immondizia'. Tra i lessemi esclusivi dei documenti pugliesi citiamo: *cabea* 'baldacchino' (cfr. Ferrari 2023: 208-209)[17]; i lessemi, relativi alla qualità della terra, *cocibelina* 'fertile, che produce legumi cottoi, che si cuociono facilmente' e il suo antonimo *crudia* 'poco fertile, che produce legumi difficili da cuocere' (cfr. D'Argenio 2018: 218); *orreata* (*casa*) «casa a più piani o […] provvista di soffitta o di una loggia coperta» (Gelao 1981: 18).

Al termine di questa breve rassegna, speriamo di aver mostrato come le modalità di ricerca automatica che stiamo sperimentando in *GeoDocuM*, unite ad una riflessione critica sulle loro potenzialità e limiti, possano rappresentare un valido strumento di supporto per il ricercatore nelle indagini volte a mettere in relazione dati lessicali e dati geografici di un corpus storico, in una prospettiva in cui gli studi di dettaglio possano giovarsi di quadri d'insieme di riferimento, analizzati anche alla luce di indagini socio-storiche.

Riferimenti bibliografici

CDCv = Morcaldi, M. / Schiani, M. / De Stefano, S. (a cura di) (1873-1893), *Codex Diplomaticus Cavensis*, I: Petrus Piazzi, Napoli; II-VIII: Hoepli, Milano-Pisa-Napoli.

Cuozzo, E. / Martin, J.-M. (1995), "Il particolarismo napoletano altomedievale", *Mélanges de l'Ecole française de Rome. Moyen-Age, Temps Modernes*, 107/1: 7-16.

15 Su quest'ultimo termine come studio di caso approfondito nell'ambito di GeoDocuM, si veda Giuliani et al. (2024b).

16 Il termine non è mai impiegato nella documentazione di area longobarda, neanche settentrionale, ma ha significativo riscontro nei papiri ravennati.

17 Rimandiamo a Ferrari (2023) per una approfondita disamina dei termini della cultura materiale che occorrono nelle fonti documentarie medievali dell'Italia meridionale.

D'Angelo, E. / Monella, P. (2019), "ALIM (Archivio della Latinità Italiana del Medioevo). Storia, attualità, prospettive di una banca-dati di testi mediolatini", in Canettieri, P. / Santini, G. / Tinaburri, R. / Gamberini, R. (a cura di), *La Filologia Medievale. Comparatistica, critica del testo e attualità.* Atti del Convegno (Viterbo, 26-28 settembre 2018), L'Erma di Bretschneider, Roma-Bristol, 203-225.

D'Argenio, E. (2018) [2019], "Novità lessicali nel *Codice Diplomatico Barese* (secoli X-XIII)", *Archivum Latinitatis Medii Aevi*, 76: 209-222.

Ferrari, V. (2023), *Il lessico della cultura materiale nei documenti medievali dell'Italia meridionale (IX-XII secolo)*, Giannini, Napoli.

Ferrarini, E. (2017), "ALIM ieri e oggi", *Umanistica digitale*, 1: 7-17.

Gelao, C. (1981), *Itinerari per Bari medievale*, Edipuglia, Bari.

Giuliani, M. (2007), *Saggi di stratigrafia linguistica dell'Italia meridional*e, PLUS, Pisa.

Giuliani, M. (2012), "Il policentrismo campano alla luce della documentazione medievale", in Sornicola, R. / Greco, P. (a cura di), *I documenti notarili altomedievali di area campana: bilancio degli studi e prospettive di ricerca.* Atti della giornata di studio (Napoli, 3 dicembre 2009), Tavolario Edizioni, Cimitile (NA), 191-213.

Giuliani, M. (2023), "Variazione e omogeneità nel più antico repertorio lessicale italiano", *Bollettino dell'Atlante lessicale degli antichi volgari italiani*, 11: 9-44.

Giuliani, M. / Abete, G. / D'Argenio, E. (2023), "I sondaggi, i metodi e le analisi del progetto *GeoDocuM*. Alla ricerca delle tendenze locali e sovralocali del latino documentale dell'Italia meridionale", *Zeitschrift für romanische Philologie*, 139/4: 1101-1130.

Giuliani, M. / Abete, G. / D'Argenio, E. (2024a), "Carte d'archivio meridionali (secc. VIII-XI). Lavori in corso per una mappatura dei dati lessicali", in Proietti, D. / Valente, S. (a cura di), *Carte altomedievali e centri di documentazione. Ricerche storico-linguistiche, dati e considerazioni teorico-metodologiche*, Aracne, Roma, 133-181.

Giuliani, M. / Abete, G. / D'Argenio, E. (2024b), "Fenomeni di coesione e particolarismi alla luce delle indagini di *GeoDocuM*: il caso del lat. mediev. napol. *egripus*", in Consani, C. / Guazzelli, F. / Perta, C. (a cura di), *Gruppi professionali come fattore di innovazione linguistica. Evidenze documentarie in Europa tra Tarda Antichità e Medioevo*, Alessandria, Edizioni dell'Orso, 29-53.

Krishnamurthy, R. (2006), "Collocations", in Brown, K. (ed.), *Encyclopedia of Language and Linguistics.* Vol. II, Elsevier, Boston-Oxford, 596-600.

Martin, J.-M. (2011), "Les documents de Naples, Amalfi, Gaète (IXe-XIIe siècle): écrìture, diplomatique, notariat", in Martin, J.-M. / Peters-Custot, A. / Prigent, V. (dir.), *L'héritage byzantin en Italie (VIIIe-XIIe siècle). I. La fabrique documentaire*, École française de Rome, Roma, 51-85.

Pratesi, A. (1992), "L'eredità longobarda nel documento latino di età normanno-sveva", in *Id.*, *Tra carte e notai. Saggi di diplomatica dal 1951 al 1991*, Società Romana di Storia Patria, Roma, 439-448.

R Core Team (2020), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, https://www.R-project.org/.

RNAM = Spinelli, A. (a cura di) (1845-1861), *Regii Neapolitani Archivi Monumenta*, 6 voll., ex Regia typographia, Neapoli.

Sornicola, R. (2012), *Bilinguismo e diglossia dei territori bizantini e longobardi del Mezzogiorno: le testimonianze dei documenti del IX e X secolo*, Giannini, Napoli.

Sornicola, R. (2015), "*Curiales*, *notarii*, *presbyteri* nella Campania alto-medievale. Alcuni problemi di sociolinguistica storica, con particolare riguardo alla morfo-sintassi", in Consani, C. (a cura di), *Contatto interlinguistico tra presente e passato*, LED, Milano, 237-281.

Welbers, K. / Van Atteveldt, W. / Benoit K. (2017), "Text analysis in R", *Communication Methods and Measures*, 11/4: 245-265.