How to reference this article
Vaccaro, G. (2020). Corpus di letteratura come corpus di lingua: il caso del Medioevo. *Italica Wratislaviensia*, 11(1), 143–165.
DOI: http://dx.doi.org/10.15804/IW.2020.11.1.06

Giulio Vaccaro
(Opera del Vocabolario Italiano-CNR, Firenze)
vaccaro@ovi.cnr.it
ORCID ID: 0000-0002-8087-9910

CORPUS DI LETTERATURA COME CORPUS DI LINGUA: IL CASO DEL MEDIOEVO

CORPORA OF ITALIAN LITERATURE AS CORPORA OF ITALIAN LANGUAGE: THE CASE OF THE MIDDLE AGE

Abstract: The implementation of a corpus of a historically determined variety of a language poses very relevant methodological problems. The principle one: is it possible to create a corpus actually representative of a linguistic variety of which we do not know, nor can we know, the extension, and in which the weight of literary texts necessarily is more conspicuous than in a real language? The answer from the point of view of *corpus* linguistics is certainly negative. But numerous tools are available today (first of all the *Corpus TLIO* and the *Corpus OVI dell'Italiano antico*) that would seem to show the opposite.

Starting from the OVI experience, some methodological reflections on the creation of corpus of historical varieties of language will be proposed, showing how, although it is impossible to apply criteria of balance and representativeness *stricto sensu*, a series of corrections can be applied (in in particular by selecting the texts on the basis of the discursive traditions) which allow to have tools that effectively meet the needs of linguistic research.

Keywords: Mediaeval Literature; Discourse traditions; *Corpus OVI dell'italiano antico*; Corpus Lingustic; Italian Lexicology.

Received: 03/02/2020; Accepted: 15/06/2020; Published: 31/07/2020
ISSN 2084-4514 e-ISSN 2450-5943

1. STORIA E GEOGRAFIA DELL'ITALIANO ANTICO

La realizzazione e lo sviluppo di corpus di una varietà storica di una lingua – quali sono il *Corpus TLIO* e il *Corpus OVI dell'italiano antico*, oggetti principali di questo lavoro[1] – pongono problemi metodologici rilevantissimi, e principalmente uno: è possibile realizzare un corpus effettivamente rappresentativo di una varietà linguistica di cui non conosciamo, né possiamo conoscere, l'estensione e in cui, soprattutto, il peso dei testi letterari è, per forza di cose, assai più cospicuo di quello che sarebbe in un corpus di lingua contemporanea (e di converso il peso dei testi spontanei è minore, quando non trascurabile o assente)? Conseguentemente, è possibile per una fase storica di una lingua, e in generale per un corpus diacronico, essere caratterizzato non solo dalla finitezza dell'insieme, ma anche dalla selettività e dalla rappresentatività?

Si tratta di questioni che per l'italiano antico sono ulteriormente complicate da dati intrinseci. Il primo, tutt'altro che banale, è quale sia l'estensione geografica dell'italiano antico: se esso vada inteso, dunque, come la fase antica dell'italiano contemporaneo (e coincida, dunque, *grosso modo* col fiorentino antico: lo intende in questi termini la *Grammatica dell'italiano antico* di Salvi e Renzi (cf. Salvi & Renzi, 2010) oppure se esso vada inteso, «prima della codificazione della lingua nazionale, come un insieme di varietà che vanno dal piemontese al siciliano, da trattarsi unitariamente» (Beltrami, 2008, p. 34; è questa la prospettiva del *TLIO*).

[1] Il *Corpus TLIO* contiene 2439 testi per complessive 22.399.150 occorrenze di 464.488 forme grafiche distinte ed è lemmatizzato esaustivamente rispetto alle forme; il *Corpus OVI dell'italiano antico* (che include al suo interno integralmente il *Corpus TLIO*) contiene 2570 testi per complessive 24.310.040 occorrenze di 486.361 forme grafiche distinte. Prescindendo da discorsi più specifici sui confini geografici e cronologici, preciso che intendo con l'etichetta di *italiano antico* la lingua che emerge dall'insieme dei testi prodotti prima del 1375 in un qualunque volgare del dominio italo-romanzo.

Il secondo, ancor più complesso, è l'estensione cronologica di questa fase[2]. Com'è noto, il *TLIO* ha dichiarato come estremo cronologico più recente la morte del Boccaccio (1375): questa data fu assunta dopo un'ampia e articolata discussione tra quanti vedevano il momento di passaggio agli inizi del Trecento (ossia al 1321, intendendo con questa data l'inizio di una tradizione linguistica del fiorentino volta progressivamente allo schiacciamento degli altri volgari) e quanti, invece, lo posponevano fino al Quattrocento inoltrato, se non anche al 1525 (e dunque fino alla canonizzazione bembiana del modello linguistico delle Tre Corone) o all'ultima edizione rivista dell'*Orlando furioso* (1532)[3]. Mancano, però, delle indicazioni in positivo della scelta del 1375 come data limite: essa è, certamente, una data simbolica, visto che si può ben dire che la morte dell'ultima delle Tre Corone rappresenti, di fatto, la fine della fase di creazione di quell'idea di italiano (antico) che sarà poi assunta da Bembo e divulgata dalla Crusca e che coinciderà di fatto con la nozione stessa di italiano scritto almeno fino all'Ottocento inoltrato. Il taglio al 1375 implicava che i testi più recenti fossero scritti da persone nate ancora nella prima metà del secolo, il che consentiva, almeno in linea tendenziale, di minimizzare l'impatto di quella prima grande

[2] Non è questa la sede per una disamina delle proposte di periodizzazione dell'italiano, che oscillano da quella tradizionale per secoli (che è poi, nei fatti, quella che si è imposta anche nel *TLIO*) a quelle che si fondano su criteri interni (penso soprattutto alle proposte di Tesi, 2007): per un'analisi dei termini della questione si veda Serianni, 2015, pp. 18–21.

[3] La questione era ben presente nel dibattito che precedette l'inizio dei lavori dell'Opera del Vocabolario Italiano: «si pone anzitutto il *problema della frontiera* tra il Tesoro e il Vocabolario, al quale si possono dare tre soluzioni: 1° o limitare il Tesoro ai primi due secoli, Duecento e Trecento; 2° o estendere il Tesoro fino alla fase unitaria della lingua letteraria, cioè fino al 1500, o addirittura fino al 1525, anno della pubblicazione delle *Prose della volgar lingua* di Pietro Bembo; 3° o spingere il Tesoro fin dentro il Quattrocento, ma "a pettine" cioè limitando la schedatura ai testi tuttora prevalentemente dialettali» (Vaccaro, 2013, pp. 310–311). Nella discussione si manifestarono due opposte tendenze: una (sostenuta da Mario Fubini e Gerhard Rohlfs) era per l'inclusione complessiva del Quattrocento; l'altra (sostenuta da Bruno Migliorini, Gianfranco Contini, Vittorio Santoli, Carlo Tagliavini e Arrigo Castellani) era orientata – sia pur con proposte diverse – per una soluzione più restrittiva, che portasse la frontiera ben dentro il Trecento.

mutazione che si verificò dopo la peste del 1348. Non si può non intravedere, tuttavia, che questa partizione trova un suo senso più nella prima delle due prospettive in cui s'intende l'italiano antico (ossia quella di Salvi & Renzi, 2010) che nella seconda, visto che la data rappresenta certamente uno spartiacque per il fiorentino antico e per alcune varietà toscane, molto meno per altre varietà non toscane: significativamente, tra l'altro, il 1375 è il confine della prima sezione del *Corpus MIDIA*, che aderisce (nel complesso) alla nozione di italiano antico presente in Salvi & Renzi, 2010 e contiene difatti, per l'epoca antica, esclusivamente testi toscani (o toscanizzati, come nel caso dei poeti della Scuola siciliana) (Cimaglia, 2017). Fin dagli inizi, nel *Corpus TLIO* si operò per l'inclusione di testi anche successivi alla data limite del 1375 che fossero stati prodotti da autori che avessero già cominciato a scrivere prima di questa data (è per esempio il caso di Antonio Pucci) e di testi prodotti sicuramente nel Trecento, ma di cui non fosse possibile una più precisa collocazione; via via con maggiore larghezza (ma senza il vincolo dell'esaustività rispetto alle edizioni affidabili) si sono immessi nel corpus anche «alcuni testi con le stesse caratteristiche di edizione e di lingua, databili con precisione o genericamente tra la fine del secolo XIV e l'inizio del XV» (Beltrami, 2012, p. 3). Negli ultimi anni[4] è stata invece creata una separazione tra il *Corpus OVI dell'italiano antico*, che conterrà tutti i testi editi in modo affidabile composti entro il 1400, e il *Corpus TLIO*, il cui aggiornamento è stato vincolato a criteri diversi: la datazione entro la fine del Duecento, l'«appartenenza ad aree linguistiche scarsamente documentate», l'«eccezionale rilevanza lessicale e/o culturale», le scritture femminili[5].

In generale, il prevalere della logica della documentazione della variabilità in diatopia rispetto a quella della qualità lessicale ha portato a un sempre più massiccio ingresso di testi tardo-trecenteschi all'interno dei corpus dell'OVI, appiattendo di fatto la periodizzazione dell'italiano

[4] All'interno del progetto finanziato *CoVo. Il corpus del vocabolario italiano delle origini: aggiornamento filologico e interoperabilità»* (PRIN 2015), coordinato da Lino Leonardi presso l'Università di Siena.

[5] Cf. <http://www.ovi.cnr.it/images/pdf/Criteriperlaggiornamento.pdf>. Per l'ultima tipologia non è stata tuttavia considerata Margherita Datini.

antico sulla tradizionale partizione per secoli che era stata propria degli studi di Migliorini. Resta vero, tuttavia, che anche in area toscana molteplici e non facilmente districabili sono i fili che annodano il secondo Trecento (e a maggior ragione il tardo Trecento) con il secolo successivo, almeno nella sua componente non umanistica. Basti pensare a quei frutti che nel primo Quattrocento arrivano a maturazione e che più saranno duraturi e informeranno la cultura dei secoli successivi, come per esempio la perdurante fortuna dei *Reali di Francia* di Andrea da Barberino (che insieme al *Leggendario de' Santi* e al *Guerrin meschino* rappresenteranno a lungo la cifra culturale delle classi popolari, come mostra bene il sarto manzoniano: «un uomo che sapeva leggere, che aveva letto in fatti più d'una volta il *Leggendario de' Santi*, il *Guerrin meschino* e i *Reali di Francia*, e passava, in quelle parti, per un uomo di talento e di scienza»). I caratteri peculiari dell'italiano antico – inteso nel senso complessivo di 'volgari dell'area italo-romanza' – paiono, insomma, cominciare a sfrangiarsi molto più in là della fine del Trecento, almeno alla metà del secolo successivo. A voler proprio indicare una data simbolica da cui princìpi la diffrazione la si potrebbe indicare nel 1454, quando la pace di Lodi sancì l'affermazione politica e culturale dei Medici: l'attività di Cosimo, Piero e Lorenzo e la restaurazione del volgare fiorentino in funzione della politica medicea portano sia a un'ampia affermazione del volgare anche in domini tradizionalmente inesplorati sia a una poderosa e ideologica espansione del volgare fiorentino – pur nella sua declinazione umanistica – sugli altri. L'affermazione ovviamente non fu né rapida né senza resistenze, e d'altronde nella stessa Firenze rimasero autori legati a una tipologia linguistica fortemente arcaica. Due casi per tutti: Matteo Franco e Nicolò Machiavelli[6].

2. RIMEDITAZIONI: TRA ESAUSTIVITÀ E RAPPRESENTATIVITÀ

Nello specifico, il cambio dei criteri di inclusione nello strumento su cui si fonda la redazione del vocabolario cessa dichiaratamente di

[6] Su cui si vedano rispettivamente Frosini, 1990 e 2014.

fare del *Corpus TLIO* un corpus esaustivo ma non ne fa in automatico un corpus bilanciato: in ultima analisi se il vecchio *Corpus TLIO* poteva fondare una descrizione lessicografica che fosse un *vocabolario della lingua* e non solo il glossario del corpus stesso, un *Corpus TLIO* che preveda solamente aggiunte parziali numericamente e casuali qualitativamente (mi riferisco ovviamente alla qualità lessicale) può essere solamente un glossario di un conglomerato, per quanto numeroso, di testi.

Preliminare rispetto alla «rimeditazione» (come la definiva già nel convegno per il centenario della Crusca nel 1983 Domenico De Robertis: si veda De Robertis, 1985, p. 445) del *Corpus TLIO* (o di un qualunque corpus voglia essere di base per una descrizione lessicale dell'italiano antico) dovrebbe essere una riflessione di fondo:

> la documentazione disponibile, che la si intenda come l'insieme dei documenti editi, l'insieme dei documenti noti o l'insieme dei documenti conservati, è il risultato dell'azione aleatoria del tempo e della storia: qualunque corpus che documenti una fase antica di lingua è di necessità un corpus non campionato, poiché la popolazione madre è ormai inattingibile (Guadagnini, 2016, p. 764).

La conseguenza ovvia è che un qualunque corpus non possa essere una semplice somma di testi (anche là dove – proiettando la questione su un piano meramente teorico – la somma dei testi inclusi coincida col totale dei testi editi, e quest'ultimo coincida a sua volta con il totale dei testi conservati) ma si debba fondare, comunque, su un principio di valutazione qualitativa delle testimonianze, e dunque sulla *selezione* di quelle ritenute significative ai fini della determinazione dell'insieme e del bilanciamento di queste testimonianze sull'asse delle variazioni esterne alla lingua (diacronia e diatopia) e interne alla lingua (tradizioni discorsive, tipologie testuali). Il solo aggiungere testi al fine di incrementare numericamente il corpus, senza una seria riflessione preliminare che investa il piano complessivo del lessico dell'italiano antico e della sua rappresentazione, porta come unico risultato quello di mettere a disposizione un corpus più grande sotto il profilo quantitativo, ma non migliore qualitativamente.

La valutazione dei testi si dovrà basare, dunque, non su piani contingenti (lingua del testo, tipologia dell'edizione, ecc.), bensì sul piano della storia della lingua e su un'analisi dei dati fondata su una "filologia dei grandi numeri". È quanto esplicitano con estrema chiarezza Burgassi & Guadagnini (2017) in un volume che rappresenta il più ampio e intelligente uso di macrodati ricavabili dal *Corpus OVI*:

> riteniamo che una corretta interpretazione dei dati restituiti dal *Corpus OVI* risulti dall'applicazione di una filologia 'dei grandi numeri': con questa denominazione indichiamo un tipo di analisi (ma prima ancora un punto di vista) che si attua su un piano supra-testuale e che consiste nell'osservare la testimonianza lessicale in prospettiva contrastiva, vale a dire interpretando la documentazione alla luce della caratterizzazione diatopica e diastratica, e delle diverse tipologie di documento e di tradizioni discorsive (Burgassi & Guadagnini, 2017, p. 11).

Poco rileva, o almeno poco dovrebbe rilevare, quindi, quanto un'edizione sia commentata, se sia accompagnata da un glossario o quanto essa sia affidabile dal punto di vista fonologico (e in subordine morfologico) rispetto al manoscritto o ai manoscritti di base[7]: una banca dati testuale, come qualunque opera umana, è costruita e progettata eminentemente per una finalità specifica, e dunque la sua qualità e la sua efficacia vanno misurate rispetto a quella medesima finalità, il che non esclude che uno strumento possa venire incontro *anche* a esigenze diverse: è lo stesso motivo per cui *Google books* può essere profittevolmente usato per le ricerche in ambito lessicale, con opportune cautele, eppure non essere un corpus lessicografico (Gomez Gane, 2008 e soprattutto Maconi, 2016). Per quanto dirlo possa parere lapalissiano, il *Corpus TLIO* è (stato) realizzato per essere la base su cui costruire il vocabolario dell'italiano antico. Dunque il principio interno che ne guida la costituzione

[7] Singolarmente, tra l'altro, il pregiudizio verso i testi "raddrizzati" dal punto di vista fono-morfologico si appunta esclusivamente sui testi toscani ma assai di rado su quei testi non toscani in cui pure la *facies* linguistica è oggetto di ricostruzione editoriale: penso qui a casi come l'edizione di Bonvesin de la Riva (Contini, 1941, che nel *Corpus TLIO* porta anche la marca di «Testo significativo») o dell'Anonimo romano (Porta, 1979).

e su cui se ne deve verificare la tenuta e la qualità è l'affidabilità del lessico testimoniato dalle edizioni. Ciò, ovviamente, non esclude che il corpus possa essere usato anche per altre ricerche su piani linguistici diversi dal lessico, e dunque sulla grafia, sulla fonologia, sulla morfologia e sulla sintassi; ma tutte queste ricerche richiedono un alto grado di attenzione, di selezione e di (pre-)analisi dei dati da parte dell'utente[8]. D'altronde proprio l'incentrare l'attenzione sul lessico giustifica – in tutti i corpus gestiti dall'OVI – il deliberato uso di edizioni di qualità assai eterogenea. Per dare solamente un'idea della delicatezza e della complessità delle questioni sottese, basti pensare che l'edizione più antica oggi inclusa nel *Corpus TLIO* è quella dei trattati di Ugo Panziera, stampata a Firenze nel 1492[9], la più recente quella del libricciolo di conti del pistoiese Rustichello de' Lazzàri (Francesconi, Frosini & Zamponi, 2018). Un arco cronologico di oltre cinque secoli si giustifica solo guardando l'opzione di fondo del corpus: privilegiare il dato lessicale, ritenuto sostanziale ai fini dell'inclusione (o no), rispetto a quello fonomorfologico (per non parlare di quello grafico), che è invece, almeno dal punto di vista lessicografico, ininfluente: il corpus dunque non può che essere uno strumento banausico e ancillare, là dove il lavoro scientifico di interpretazione dei dati è – e non può che essere – competenza di chi rediga le singole voci del *TLIO* o di chi compia ricerche a ampio spettro

[8] È uno dei rischi additati da Burgassi-Guadagnini, 2014, p. 22: «la circolarità ineliminabile fra qualità delle edizioni e qualità (vale a dire affidabilità) dei *corpora* testuali e degli studi che ne derivano è una tara che inficia la scientificità dei risultati soltanto per chi abbia la feticistica presunzione di estrapolare – dai *corpora* e dalle edizioni – dei dati di verità: la consapevolezza che qualunque testo restituito da qualunque tipo di edizione è di per sé un testo "ricostruito" consente invece, a nostro avviso, di preservare il valore, ma ancora prima il senso e la legittimità, di strumenti o di analisi che coprano vasti insiemi di materiali in una prospettiva ampiamente comparatistica, al netto del margine di oscillazione, di variabilità, di potenziale cambiamento nella lezione o nell'interpretazione, che è sempre postulabile per ogni dato testuale».

[9] [*Incominciano alcuni singulari tractati di frate Ugo Panziera de' frati minori*], impresso in Firenze, per Antonio Mischomini, MCCCCLXXXXII adi VIIII di giugno.

sull'italiano (antico), e in particolare sulla semantica[10]. Naturalmente quello delle edizioni sette e ottocentesche (ma il problema è più sensibile per le seconde che per le prime) e – più in generale – del grado di affidabilità delle edizioni ai fini dell'inclusione in un corpus destinato a studi di tipo lessicale è un problema delicato all'atto della costituzione di una base di dati e di un corpus di italiano antico. Tuttavia, come detto, l'opzione elettivamente lessicografica, e dunque centrata sulla qualità dell'attestazione dei singoli lessemi e non sull'aderenza a tratti grafico-fonetico-morfologici, ha consentito di includere anche edizioni di testi fondamentali pure nel caso in cui esse fossero al di sotto (o addirittura molto al di sotto) di una soglia di accettabilità:[11] «il fatto che sul corpus si debba costruire il vocabolario impone di utilizzare anche testi editi male, senza i quali il vocabolario presenterebbe buchi vistosi» (Beltrami, 2016, p. 76). È il caso del volgarizzamento del *Tesoro*, che si legge in gran parte ancora nell'edizione curata da Luigi Gaiter (Gaiter, 1878–1883), che interviene sistematicamente riscrivendo il testo sulla base dell'edizione Chabaille del testo francese, o del volgarizzamento dell'*Egidio Romano* senese, per cui si dispone ora dell'ottima edizione Papi (2016).

Nel complesso, solo un tipo di approccio che tiene in considerazione tutte le possibili variabili (in primo luogo comunicative e sociali; in subordine per il caso italiano, vista la distribuzione marcatamente sbilanciata della documentazione, geografiche) in cui un lessema compare può consentire di ricostruire i caratteri del lessico italiano antico, distinguendone il nucleo (il lessico fondamentale), la grande fascia intermedia (il lessico comune), la periferia (il lessico tecnico, le parole proprie di determinati ambienti culturali o di determinate situazioni comunicative) e di discriminare i punti che si pongono oltre la periferia della lingua

[10] In subordine il corpus mostra un tasso abbastanza ampio di affidabilità anche sul piano sulla sintassi, che è (anche a causa dell'alto tasso di continuità tra italiano antico e italiano ottocentesco) uno dei punti meno esposti a modifiche nelle edizioni.

[11] Sulla questione ha più volte richiamato l'attenzione Pietro Beltrami: cf. almeno Beltrami, 2010.

(tipicamente prestiti, adattati e no, *hapax* o comunque lessemi a attestazione monotestuale)[12].

La compresenza di tutte le variabili consente di individuare la posizione di un lessema (o – per essere più precisi – la posizione di ciascuno dei significati di un lessema) in una delle fasce lessicali. Poco (se non nulla) si può derivare dalla sola presenza di un alto numero di attestazioni, che è una condizione necessaria, ma sicuramente non sufficiente. Per usare ancora le parole di Burgassi e Guadagnini «una stima puramente numerica [...] non renderebbe giustizia dell'essenza del lessema, che acquista spessore dalla sinergia dei numeri e dalla valutazione dei contesti nei quali i numeri si collocano» (Burgassi & Guadagnini, 2017, p. 69).

Il punto di riflessione necessario e preliminare all'inclusione (o no) di un testo in un corpus che abbia come obiettivo precipuo il lessico (quali sono, per l'appunto, il *Corpus TLIO* e il *Corpus OVI*) non può che essere, banalmente, quanto il testo contribuisca al dettaglio di determinate zone lessicali in rapporto alla documentazione già nota. Per fare solo un esempio, la versione italiana del *Lancelot en prose* rocambolescamente scoperta (e ora ottimamente edita) da Cadioli (2016), rappresenta senza dubbio uno snodo fondamentale per comprendere lo sviluppo della diffusione e della ricezione della materia arturiana in Italia, ma lessicalmente testimonia esclusivamente un'ampia gamma di gallicismi che non sono tuttavia, almeno in base ai dati in nostro possesso e a quanto si può inferire dalla nostra conoscenza della prassi traduttoria nel Due e Trecento, mai usciti dal manoscritto in cui erano contenuti. Al contrario testi magari anche più tardi e meno interessanti sotto il profilo culturale, come per esempio il *Trattato dell'arte del vetro* di Benedetto di Baldassarre Obriachi (Milanesi, 1864, pp. 69–109), sono latori di documentazione altrimenti inattingibile di "lessico di bottega", quindi di lessico tecnico-specialistico, ma anche di tutto quel lessico materiale che è tipicamente composto di quella parte della lingua fatta di

[12] Si tratta dell'articolazione proposta già da De Mauro, 1980 e poi alla base del *GraDIt*. Il modello è ben applicabile all'italiano antico, come dimostrano i vari lavori di Cosimo Burgassi e Elisa Guadagnini (e per tutti si veda ora Burgassi & Guadagnini, 2017).

parole che si usano ma, tendenzialmente, non si scrivono o si scrivono poco: in ultima analisi, quello che – in un corpus di lingua contemporanea – definiremmo lessico "ad alta disponibilità". Proprio la tendenziale scarsa attestazione di questi vocaboli, che sono perlopiù quelli maggiormente soggetti alla variazione geografica, rende, di fatto, inapplicabile un'analisi del lessico (e, conseguentemente, poco utile un incremento del corpus) esclusivamente sotto il profilo della diatopia[13].

Inoltre, il dato della distribuzione linguistica dei testi si scontra con due problemi: il primo, storico, è che in Toscana, nel Medioevo, si è scritto in volgare più che altrove e, soprattutto, molto più si è conservato, e – per ragioni culturali – molto di più si è edito; il secondo, contingente, è che l'edizione di singoli documenti locali è stata spesso delegata (altrove più che in Toscana) a eruditi del luogo in pubblicazioni a diffusione estremamente marginale (si veda per esempio il caso delle Marche, per le quali si dispone oggi dell'accurata bibliografia di Aprea, 2018) e in edizioni spesso tutt'altro che impeccabili.

Infine, l'immissione di testi sulla base di considerazioni di ordine diatopico può essere un criterio solo e esclusivamente nell'ottica di documentare il numero più elevato possibile di "punti" in cui un lessema è diffuso. L'enorme squilibrio dei dati numerici tra l'area toscana (da cui proviene sì "solo" il 54% de testi, ma ben l'80% delle occorrenze) e il resto del dominio italo-romanzo non consente infatti – se non in numeratissimi casi (perlopiù in cui il referente sia già geograficamente referenziato: per esempio i nomi di pesi e misure o di imposte) –, partendo dall'incrocio dei dati sulla presenza/assenza di un lessema in una data area, di inferire nulla sulla distribuzione di quella determinata parola in italiano antico, e dunque sul tasso di "regionalità" di un singolo vocabolo[14].

---

[13] Infatti anche nella costituzione del *Corpus TLIO* in origine, e fino agli inizi della direzione di Avalle, la marcatura dell'area linguistica di un testo non era neppure prevista.

[14] Un dato che va tenuto in conto è che, se si confrontano i citati rispetto al perimetro dei "citabili", il rapporto di copertura per l'area toscana è sensibilmente inferiore rispetto a quello delle altre aree. Non si può dunque fare un discorso che parta da un principio strettamente statistico di "rappresentazione", perché per molte aree non

Che la presenza di un determinato lessema in un'area non sia segno di regionalità vale in modo massimo per la Toscana (ossia: un termine attestato solo in Toscana non è di necessità un toscanismo), ma vale anche in senso inverso, e vale anche nel caso in cui vi sia una comunanza di più aree contro la Toscana. Come hanno già rilevato Burgassi e Guadagnini, quando si verifica quest'ultima condizione «tale lessema è nella maggioranza dei casi un forte latinismo lessicale, e ciò descrive un fatto culturale più che linguistico» (Burgassi & Guadagnini, 2017, p. 22). Un fatto culturale da intendere, però, come l'opposizione tra una Toscana linguisticamente più matura e con un modello culturale e linguistico forte e una periferia che di questo modello è invece priva e dunque è fatalmente più esposta a soluzioni estemporanee, prima tra tutti il latinismo, la cui ragione andrà ricercata nella

> particolare fisionomia linguistica dell'uomo medievale: quella condizione di contatto culturale che è alla base del volgarizzare si dà in fondo, seppure in termini variabili, per qualsiasi testo romanzo. La tensione tra latino e volgare accompagna il costituirsi delle pratiche scrittorie, l'assurgere dei volgari a lingue di cultura, nonché il loro differenziarsi in registri e codici. Ne consegue che qualsiasi scrivente medievale si muove, seppure a vari livelli e in diversa misura, in un orizzonte linguistico pervaso dal latino (o dai latini) (De Roberto, 2017, p. 234).

Questa fisionomia generale si intreccia, nel caso dei volgarizzamenti con altre due questioni: «l'intersecarsi di una generalizzata condizione di diglossia che diviene bilinguismo culturale in determinati àmbiti e la

---

avrebbe alcun senso: dire, per esempio, che il *Corpus OVI* ha per i testi calabresi un rapporto di copertura del 100% non nasconde il dato che si stia parlando di due testi; e dire che questo rapporto è per i testi pugliesi dell'87,5% equivale comunque a dire sette su otto. Anzi, porre il discorso su questo piano porta invariabilmente fuori strada: a ben vedere, infatti, tra le aree con un maggior numero di "citabili" (ho considerato per questo conto il numero limite di almeno 30 testi "citabili") supera un rapporto di copertura del 70% solamente il Veneto; l'Emilia si attesta intorno al 65%; la Sicilia e la Toscana sono sotto il 60%. Per la Toscana, poi, il rapporto è molto alto per le aree periferiche (come per esempio l'Amiata), supera il 70% per Pisa e il 65% per Siena, mentre è di poco inferiore al 60% per Firenze e addirittura sotto il 40% per i testi che non sono precisamente ascrivibili a una determinata area della regione.

vicinanza tra la lingua madre dei volgarizzatori e quella della loro fonte» (De Roberto, 2017, p. 234). Proprio la condizione di chi nel Medioevo scrive (e volgarizza) fa sì che il latinismo sporadico sia dovuto a un semplice trascinamento, e una presenza molteplice in singoli punti non collegati sia da ascrivere piuttosto a fenomeni di neologia ricorsiva che a «una soluzione che è possibile predicare come marcata alla luce della documentazione italiana antica» (Dotto, 2017, p. 324)[15].

3. UN CORPUS PER TRADIZIONI DISCORSIVE

Viceversa, il dato essenziale per un corpus di italiano antico è la sua capacità di documentare adeguatamente, e possibilmente in modo bilanciato, le diverse tipologie lessicali, in modo tale da rappresentare il lessico riducendo, per quanto possibile, il rischio di sovrarappresentare determinate aree a detrimento di altre. Si tratta di un problema in realtà comune nella lessicografia italiana (storica e – per quanto possa sembrare paradossale – anche dell'uso), in cui si largheggia spesso nella documentazione di forme della lingua poetica minore e minima del Due e Trecento, ossia di uno dei settori più largamente indagati dal punto di vista filologico e letterario, mentre intere quote di lessico rimangono marginali. È il caso, per esempio, dei suffissati in *-anza* tipici della lirica

[15] L'interpretazione è, a mio avviso, distorta dall'intendere i due volgarizzamenti posti in parallelo (il Valerio Massimo di Accurso da Cremona – per cui non si può escludere tra l'altro un tramite catalano – e l'anonimo volgarizzamento veneto dell'*Ars amandi* di Ovidio) come «due punte avanzate della ricostruzione dell'antico fuori Toscana» (Dotto, 2017, p. 323). Tuttavia, parlare di «ricostruzione dell'antico» fuori dalla Toscana, ma anche solo fuori di Firenze (e mi spingerei ancora oltre: al di fuori di una peculiare tradizione discorsiva tutta interna alla cultura fiorentina; già il volgarizzare domenicano a Firenze è cosa diversa), mi pare strada impervia: l'istituzione progressiva di un rapporto di profondità storica col mondo antico, e dunque il passaggio da una visione di antico e moderno che procede lungo un *continuum* (quello che è – insomma – «il volgarizzamento totale e radicale, della parola e della realtà»: Tanturli, 1986, p. 883) a quella di due mondi che sono percepiti chiaramente come autonomi e sono dunque soggetto/oggetto di storicizzazione (Tanturli, 1986, pp. 873–874) è elemento che mi pare difficilmente riscontrabile nelle altre aree d'Italia, dove il rapporto col mondo classico è piuttosto di tipo archeologico.

duecentesca e primo trecentesca: su 359 voci contenute nel *TLIO* circa i due quinti (153) sono costituiti di attestazioni uniche di lessemi che sono, evidentemente, tipici di una tradizione discorsiva ma scarsamente rappresentativi della lingua e della sua evoluzione.

Vi sono, invece, alcune tipologie testuali e tradizioni discorsive che più di altre rispondono ai criteri di redazione di un vocabolario. È certamente il caso dei glossari,[16] che forniscono una sorta di "vocabolario in sincronia", anche se le testimonianze sono spesso difficilmente interpretabili nel momento in cui si passi dal campo dell'attestazione a quello dell'interpretazione del significato. Fatalmente, del resto, la presenza del termine volgare come glossa secca di un termine latino polarizza l'orientamento del significato su quest'ultimo; una questione di polarizzazione del significato che, d'altronde, finisce per essere decisiva anche nei volgarizzamenti, in particolare nei casi in cui il significato di un termine non sia immediatamente perspicuo, come accade – d'altronde – anche per alcuni xenismi (spesso individuati come tali già dagli scriventi antichi) che nulla hanno a che vedere con il lessico dell'italiano (antico) e la cui inclusione nel *TLIO* è giustificata esclusivamente dalla tensione all'esaustività nella descrizione della documentazione: si tratta, in ogni caso, di una lessicografia che punta – e non può essere altrimenti – alla mera decodificazione del dato (ossia: all'aspetto glossaristico: cosa vuol dire la determinata parola in quel determinato contesto) più che alla sua codificazione (ossia quale spazio semantico occupi un lessema nella lingua)[17].

Ben più rilevante, per estensione e per impegno culturale, è il caso dei commenti danteschi, la cui presenza – anche per l'esistenza oggi di un *Vocabolario dantesco* (*VD*) – sarebbe invece da incrementare an-

---

[16] Per un bilancio sulla pubblicazione e sugli studi di glossari negli ultimi decenni, cf. Aresti, 2017.

[17] Si veda per esempio il caso di *atirotipa*, in cui l'allogenia è dichiarata nel testo («Una altra mainira se truova che se chiama atyrotypa in arabico e in latim arnicara»). Non del tutto diverso il caso di parole come *tabarzet*, disponibili solo in locuzioni nominali (*zucchero tabarzet*, *miele tabarzet*) in cui il termine rimane però semanticamente vuoto.

che a fronte di edizioni filologicamente non impeccabili[18]. I commenti sono infatti testimoni al contempo della ricezione e della tradizione di una lingua d'autore spesso nient'affatto semplice, come dimostrano i casi in cui le interpretazioni del significato di una parola divergano, anche sensibilmente, tra commentatore e commentatore: casi per certi versi estremi di quest'incertezza sono parole come *cagnazzo* o *chioccio* o *chiappa* o *stormo* o addirittura *cerchia* (cf. *VD*, s.vv). La testimonianza dei commenti, dunque, è doppiamente fondamentale: da un lato per comprendere, su un piano più banale, il significato delle parole (anche se i commenti che richiamano un basso livello culturale sono pochissimi; per il Trecento le *Chiose Selmi* e le *Chiose Cagliaritane*[19]), dall'altro per capire quali fossero i "dantismi forti", ossia quelle parole e accezioni più tipiche di Dante e di più difficile interpretazione (e in questi casi maggiore è la dispersione nell'indicazione dei significati, maggiore è il grado di specializzazione dantesca del significato). In fondo, solo la presenza dei commenti danteschi consente di percepire il livello di lingua e di comprensibilità del testo dantesco e il fatto che «Dante non fu certo mai, neppure a Firenze, il poeta dei bottegai» (Dionisotti, 1965, p. 334). Per di più i commenti fiorentini (l'*Ottimo*, Lancia, il volgarizzamento del Bambaglioli) sono spesso ricettori di una lunga tradizione di testi – principalmente di volgarizzamenti, ma anche di altri commenti – che vengono direttamente dal tardo Duecento o dai primi del Trecento: basti pensare al ruolo che hanno avuto per l'*Ottimo* i volgarizzamenti di Orosio o dei *Fatti dei Romani* o delle *Metamorfosi* del Simintendi.

Su un piano diversissimo rispetto a quello dei commenti a Dante si collocano invece i testi agiografici, a partire dal testo che più di tutti permea la lingua successiva, le *Vite dei santi Padri* di Domenico Cavalca, le varie agiografie dei santi sono più o meno organizzate in strut-

---

[18] I commenti inclusi nel *Corpus TLIO* sono a oggi solamente Jacopo Alighieri, Jacopo della Lana, l'*Ottimo*, la *Declaratio* di Guido da Pisa, le *Chiose Selmi*, Maramauro, Boccaccio e lo pseudo-Boccaccio, Francesco da Buti, nonché i capitoli ternari di Bosone da Gubbio e Mino di Vanni. Per nessuno dei testi è stata recepita una delle nuove edizioni. A oggi la ricerca lessicale sul lessico dantesco e sui dantismi è fortemente debitrice dell'ampio corpus del *DDP*.

[19] Per le edizioni si vedano rispettivamente: Avalle, 1900 e Carrara, 1902.

ture testuali (tipicamente la *Legenda aurea*, ma non mancano casi di strutture dedicate a singoli santi o a gruppi di santi), più o meno ampie, più o meno rimaneggiate da copisti-autori che si muovono su un labile confine:

> il basso quoziente di coesione interna dei singoli capitoli, formati da sezioni giustapposte e tendenzialmente autonome, favorisce gli interventi di destrutturazione del testo, rendendo difficile distinguere cadute e omissioni come lacune ereditarie o, al contrario, come tagli effettuati in modo indipendente dai singoli copisti, per lo più su digressioni o commenti esegetici, citazioni da *auctoritates* e unità narrative nelle seriazioni di aneddoti miracolistici. La tendenza, evidente soprattutto nei testimoni quattrocenteschi, ad adattare o ammodernare il testo per facilitarne la fruizione attraverso operazioni di riassetto linguistico-stilistico o strutturale, rende inoltre in molti casi inutilizzabili i riscontri della collazione e impone in definitiva l'esclusione delle convergenze a livello lessicale o stilistico, e in parte anche a livello strutturale, come elementi congiuntivi (Cerullo, 2015, p. 258).

Per di più i testi della devozione medievale rimangono vitali assai a lungo nella tradizione popolare (un caso per tutti: i *Fioretti di san Francesco*). Quello agiografico rimane, dal punto di vista linguistico e lessicale, un campo di fatto inesplorato, anche per l'insufficienza filologica dell'indagine sui testi.

Ma anche altre tradizioni discorsive pur già ben rappresentate nel corpus sarebbero da ampliare: è il caso, in particolare, dei testi storici cittadini, che hanno conosciuto – negli ultimi anni – una nuova fase di studio e rianalisi, concentrato tuttavia più sul fronte degli studi storici che su quello testuale, che consente oggi di inquadrare pienamente le cronache all'interno di una realtà storica e linguistica comunale[20] (si pensi per esempio alle cronache fiorentine sul tumulto dei Ciompi, e in particolare alla cosiddetta «Cronaca dello squittinatore», che rappresenta l'unico documento "interno" alla rivolta[21] o testi falsamente antichizzati prodotti per una determinata cerchia sociale, come la *Cro-*

[20] Si veda per tutti Miglio-Francesconi, 2017.

[21] Per l'edizione cf. Scaramella, 1900, pp. 73–102.

nica malispiniana[22]), municipale (si pensi alla *Cronaca di Partenope*[23]) o al momento di passaggio verso la signoria (si pensi per esempio alla monumentale *Polyhistoria* di Niccolò da Ferrara, ancora oggi quasi del tutto inedita[24]).

Pare dunque evidente che ai fini della ricostruzione complessiva del lessico dell'italiano antico sia necessario un corpus con un triplice bilanciamento. Il principale è quello dall'appartenenza a singoli generi (anche molto ampi) o a tradizioni discorsive insufficientemente rappresentate (si pensi ai manuali di medicina, ai ricettari di cucina, ai libri di viaggio...), incrociata con i dati geografici e con quelli cronologici. Una rappresentazione dei testi provenienti da tutte le aree italiane, infatti, è sicuramente indispensabile, là dove si tenga presente però che non basta che un testo provenga da una determinata area per farne, *sic et simpliciter*, un testo lessicalmente affidabile per quell'area. Il dato è noto e lo notavano per esempio già Ferdinando Gabotto e Delfino Orsi pubblicando, nel 1891, il primo (e purtroppo unico) volume dedicato ai laudari piemontesi: «rilevare l'influsso umbro in queste laude pare perfino ozioso» (Gabotto & Orsi, 1891, p. XII). Il punto è esattamente questo: i laudari di Bra e di Carmagnola rappresentano senza dubbio dei monumenti della letteratura piemontese del Quattrocento, ma ciò non implica (o almeno: non implica *necessariamente*) che essi siano anche dei monumenti del volgare piemontese del Quattrocento. Ciò non vuol dire, naturalmente, proporre il mito della genuinità del testo "di carattere pratico" contro l'artificio del testo letterario: come mostrano i numerosi studi di Federigo Bambi (si vedano almeno Bambi, 2018 e 2019) anche il volgare dei notai è per sua natura esposto all'interferenza del latino. Rimanendo in ambito piemontese i *Parlamenti* canavesi due/trecenteschi editi da Bertoni (1910) rappresentano senza dubbio un testo lessicalmente rilevante, ancorché letterario. Del pari è necessario rap-

[22] La collocazione della *Cronica* malispiniana al tardo Trecento è un dato acquisito agli studi grazie a Porta, 1991.

[23] Per l'edizione della *Cronaca di Partenope*, cf. Kelly, 2001. Per comprendere la genesi del testo risultano tuttavia fondamentali le messe a punto di De Caprio, 2012, pp. 18–28 e Montuori, 2017.

[24] L'unica porzione di testo edita è in Muratori, 1738.

presentare in modo adeguato anche la variazione (almeno: la possibile variazione) nella microdiacronia: privilegiare i testi duecenteschi – rimasti salvi nell'Ottocento dalla filologia che perseguiva il mito del secolo d'oro della lingua – vuol dire comunque sottorappresentare la prima grande età della scrittura in volgare, in cui – soprattutto per Firenze e la Toscana – si scriveva effettivamente di tutto in volgare. L'incrocio di questi tre dati, tuttavia, non può prescindere da un quarto: quello dell'analisi critica preliminare del lessico di un testo rimane lo strumento fondamentale nelle mani dello studioso.

Nel passaggio dall'attestazione all'analisi semantica, un corpus lessicale che si arricchisca di testi scelti in base a criteri diversi da quello lessicale finirà per arricchirsi di *hapax* provenienti dal novero degli occasionalismi, delle autoschediastiche formazioni deverbali o denominali, di latinismi marginali; di una serie di parole, dunque, che non hanno avuto peso nella storia dell'italiano. Quindi, o lo strumento lessicografico che se ne produrrà privilegerà una prospettiva glossaristica o, mantenendo la prospettiva del vocabolario, necessiterà di correttivi nella redazione delle voci (esclusione degli *hapax*, vaglio dell'attestazione, ecc.). La mera registrazione del dato di fatto (la presenza di una parola, a prescindere dalla sua episodicità, della sua effettiva consistenza e, in ultima analisi, dal suo valore storico) rientra in quella «disponibilità a discutere della lingua di testi che, grazie alla possibilità di essere *interrogati*, offrono il dubbio vantaggio di poter rimanere *non letti*» (Tomasin, s.d., p. 16).

SIGLE

Corpus MIDIA = Morfologia dell'Italiano in DIAcronia, coordinato da P. D'Achille, consultabile online all'indirizzo *www.corpusmidia.unito.it*.
Corpus TLIO = Corpus TLIO (vers. 10 agosto 2019), consultabile online all'indirizzo *http://tlioweb.ovi.cnr.it*.
Corpus OVI dell'italiano antico = Corpus OVI dell'italiano antico (vers. 10 agosto 2019), diretto da P. Larson & E. Artale, consultabile online all'indirizzo *http://gattoweb.ovi.cnr.it*.

DDP = Dartmouth Dante Project, consultabile online all'indirizzo *https://Dante.Dartmouth.EDU*.
GraDIt = De Mauro, T. (Ed.). (2000). *Grande dizionario italiano dell'uso*. Torino: UTET.
TLIO = Tesoro della Lingua Italiana delle Origini, fondato e da P.G. Beltrami, consultable online all'indirizzo *http://tlio.ovi.cnr.it/tlio*.
VD = Vocabolario dantesco, diretto da P. Manni, consultabile online all'indirizzo *www.vocabolariodantesco.it*.

BIBLIOGRAFIA

Fonti

Avalle, G. (1900). *Le Antiche chiose anonime all'Inferno di Dante secondo il testo Marciano*. Città di Castello: Lapi.
Bertoni, G. (2010). Note e correzioni all'antico testo piemontese dei 'Parlamenti ed epistole'. *Romania*, *XXXIV/155–156*, 305–314.
Cadioli, L. (Ed.). (2016). *Lancellotto. Versione italiana inedita del «Lancelot en prose»*. Firenze: Edizioni del Galluzzo per la Fondazione Ezio Franceschini.
Carrara, E. (Ed.). (1902). *Le chiose cagliaritane*. Città di Castello: Lapi.
Contini, G. (Ed.). (1941). *Le opere volgari di Bonvesin da la Riva*. Roma: Società Filologica Romana.
Francesconi, G., Frosini, G., & Zamponi, S. (Eds.). (2018). *Libricciolo di conti di Rustichello de' Lazzàri (1326–1337)*. Pistoia: Brigata del Leoncino.
Franco, M. (1990). *Lettere*. (edited by G. Frosini). Firenze: Accademia della Crusca.
Gabotto, F., & Orsi, D. (Eds.). (1891). *Le laudi del Piemonte*. Bologna: Romagnoli-Dell'Acqua.
Gaiter, L. (Ed.). (1878–1883). *Il Tesoro di Brunetto Latini volgarizzato da Bono Giamboni, raffrontato col testo autentico francese edito da P. Chabaille*. Bologna: Romagnoli.
Kelly, S. (2011). *The Cronaca di Partenope. An introduction to and critical edition of the first vernacular history of Naples, c. 1350*. Leiden: Brill.
Milanesi, G. (Ed.). (1864). *Dell'arte del vetro per musaico. Tre trattatelli dei secoli XIV e XV*. Bologna: Romagnoli.

Muratori, L.A. (Ed.). (1738). *Polyhistoria... ab anno MCCLXXXVII usque ad MCCCLXVII, italice conscripta*. In *Rerum Italicarum Scriptores*, XXIV, Milano, ex typographia Societatis Palatinae, coll. 699–848.

Papi, F. (Ed.). (2016). *Il 'Livro del governamento dei re e dei principi' secondo il codice BNCF II.IV.129*, I, *Introduzione e testo critico*. Pisa: ETS.

Porta, G. (Ed.). (1979). Anonimo Romano. *Cronica*. Milano: Adelphi.

Scaramella, G. (Ed.). (1900). *Il tumulto dei Ciompi. Cronache e memorie*. In *Rerum Italicarum Scriptores*, vol. XVIII/3. Bologna: Zanichelli.

Studi

Aprea, F. (2018). *Bibliografia dei testi volgari marchigiani dalle Origini al 1550*. Roma: Aracne.

Aresti, A. (2017). L'edizione di glossari latino-volgari prima e dopo Baldelli. Una rassegna degli studi e alcuni glossarietti inediti. *Studi di lessicografia italiana*, *XXXIV*, 35–82.

Asperti, S. (1998). *I Vangeli in volgare italiano*. In L. Leonardi (Ed.), *La Bibbia in italiano tra Medioevo e Rinascimento*. Atti del convegno internazionale (Firenze, 8–9 novembre 1996). (pp. 119–144). Firenze: SISMEL-Edizioni del Galluzzo.

Bambi, F. (2018). *Scrivere in latino, leggere in volgare. Glossario dei testi notarili bilingui tra Due e Trecento*. Milano: Giuffrè.

Bambi, F. (2019). *Qualche postilla sulla lingua dei notai del medioevo*. In J. Visconti (Ed.). *Parole nostre. Le diverse voci dell'italiano specialistico e settoriale* (pp. 125–140). Bologna: il Mulino.

Beltrami, P.G. (2008). La nuova lessicografia dell'italiano antico: il 'Tesoro della Lingua Italiana delle Origini'. *Bollettino dell'Atlante linguistico degli antichi volgari italiani*, *I*, 33–52.

Beltrami, P.G. (2010). *Lessicografia e filologia in un dizionario storico dell'italiano antico*. In C. Ciociola (Ed.). *Storia della lingua e filologia*. Atti del VII Convegno ASLI (Pisa-Firenze, 18–20 dicembre 2008). (pp. 235––248). Firenze: Franco Cesati.

Beltrami, P.G. (Ed.). (2012). *Norme per la redazione del 'Tesoro della Lingua Italiana delle Origini'*. Retrieved from *http://tlio.ovi.cnr.it/TLIO/NormeTLIO.pdf*.

Beltrami, P.G. (2016). *Il Tesoro della Lingua Italiana delle Origini: Caratteristiche, problemi, futuro*. In R. Coluccia, J.M. Brincat, F. & Möhren (Eds.). *Actes du XXVIIe Congrès international de linguistique et de philologie romanes* (Nancy, 15–20 juillet 2013). *Section 5: Lexicologie,*

phraséologie, lexicographie (pp. 71–78). Nancy: ATILF/SLR. Retrieved from *www.atilf.fr/cilpr2013/actes/section-5.html*.

Burgassi, C., & Guadagnini, E. (2014). Prima dell'«indole». Latinismi latenti dell'italiano. *Studi di lessicografia italiana*, *XXXI*, 5–43.

Burgassi, C., & Guadagnini, E. (2017). *La tradizione delle parole. Sondaggi di lessicografia storica*. Strasbourg: ELiPhi.

Cerullo, S. (2015). Il volgarizzamento toscano trecentesco della 'Legenda aurea'. Appunti e prolegomeni per un'edizione critica. In: *Studi di filologia italiana*, LXXII, pp. 233–298.

Cimaglia, R. (2017). *La costituzione del Corpus MIDIA*. In P. D'Achille, & M. Grossmann (Eds.), *Per la storia della formazione delle parole in italiano. Un nuovo corpus in rete (MIDIA) e nuove prospettive di studio* (pp. 53–62). Firenze: Franco Cesati.

De Caprio, C. (2012). *Scrivere la storia a Napoli tra Medioevo e prima età moderna*. Roma: Salerno Ed.

De Mauro, T. (1980). *Guida all'uso delle parole*. Roma: Editori Riuniti.

De Robertis, D. (1985). L'Ufficio filologico dell'Opera del Vocabolario, il suo impianto, il suo lavoro. In I. Calabresi (Ed.), *La Crusca nella tradizione letteraria e linguistica italiana*. Atti del Congresso Internazionale per il IV centenario dell'Accademia della Crusca (Firenze, 29 settembre––2 ottobre 1983) (pp. 444–451). Firenze: Accademia della Crusca.

De Roberto, E. (2017). *Sintassi e volgarizzamenti*. In S. Cerullo, & L. Leonardi (Eds.). *Tradurre dal latino nel Medioevo italiano. Translatio studii e procedure linguistiche.* Atti del convegno (Firenze, Fondazione Ezio Franceschini, 16–17 dicembre 2014) (pp. 227–293). Firenze: SISMEL-Edizioni del Galluzzo.

Dionisotti, C. (1965). *Dante nel Quattrocento*. In *Atti del Congresso internazionale di Studi danteschi*, a cura della Società Dantesca Italiana e dell'Associazione Internazionale per gli Studi di lingua e letteratura italiana sotto il patrocinio dei comuni di Firenze, Verona e Ravenna (20–27 aprile 1965) (pp. 333–337). Firenze: Sansoni.

Dotto, D. (2017). Dal Veneto alla Sicilia: escursioni lessicali fuori dalla bottega dei volgarizzatori dei classici. *Bollettino dell'Opera del vocabolario italiano*, *XXII*, 317–379.

Frosini, G. (2014). *Lingua*. In: G. Sasso (Ed.). *Enciclopedia Machiavelliana* (pp. 720–732). Roma: Istituto della Enciclopedia Italiana, II.

Gomez Gane, Y. (2008). Google ricerca libri e la linguistica italiana: vademecum per l'uso di un nuovo strumento di lavoro. *Studi linguistici italiani*, *24*, 260–78.

Guadagnini, E. (2016). Lessicografia, filologia e corpora digitali: qualche considerazione dalla parte dell'OVI. *Zeitschrift für romanische Philologie*, *132*, 755–792.

Lorenzi Biondi, C., & Vaccaro, G. (2017). Firme e copie. I volgarizzamenti del secondo Trecento. In E. Guadagnini, & G. Vaccaro (Eds.), *Rem tene, verba sequentur. Latinità e medioevo romanzo: testi e lingue in contatto* (pp. 179–232). Alessandria: Edizioni dell'Orso.

Maconi, L. (2016). *Retrodatazioni lessicali con Google Libri: opportunità e inganni della Rete*. In C. Marazzini, & L. Maconi (Eds.), *L'italiano elettronico. Vocabolari, corpora, archivi testuali e sonori*. Atti della Piazza delle lingue 2014 (6–8 novembre). (pp. 73–93). Firenze: Accademia della Crusca.

Miglio, M., & Francesconi, G. (Eds.). (2017). *Le cronache volgari in Italia.* Atti della VI Settimana di studi medievali (Roma, 13–15 maggio 2015). Roma: Istituto Storico Italiano per il Medioevo.

Montuori, F. (2017). *Come "si costruisce" una cronaca*. In: M. Miglio, & G. Francesconi (Eds.), *Le cronache volgari in Italia*. Atti della VI Settimana di studi medievali (Roma, 13–15 maggio 2015) (pp. 31–88). Roma: Istituto Storico Italiano per il Medioevo.

Porta, G. (1991). *Le varianti redazionali come strumento di verifica dell'autenticità dei testi: Villani e Malispini*. In: S. Guida, & F. Latella (Eds.), *La filologia romanza e i codici*. Atti del Convegno (Messina, Università degli Studi, Facoltà di Lettere e Filosofia, 19–22 dicembre 1991) (vol. II, pp. 481–529). Messina: Sicania.

Salvi, G., & Renzi, L. (Eds.). (2010). *Grammatica dell'italiano antico*. Bologna: il Mulino.

Serianni, L. (2015). *Prima lezione di storia della lingua italiana*. Roma–Bari: Laterza.

Tanturli, G. (1986). Volgarizzamenti e ricostruzione dell'antico. I casi della terza e quarta Deca di Livio e di Valerio Massimo, la parte del Boccaccio (a proposito di un'attribuzione). *Studi medievali*, *s. III, XXVII*, 811–888.

Tesi, R. (2007). *Storia dell'italiano. La formazione della lingua comune dalle fasi iniziali al Rinascimento*. Bologna: Zanichelli.

Tomasin, L. (n.d.). *Che cos'è l'italiano antico?* Retrieved from *www.academia.edu*.

Vaccaro, G. (2013). Veniamo da molto lontano e andiamo molto lontano. L'Opera del Vocabolario Italiano dalle Origini al 1992. *Bollettino dell'Opera del vocabolario italiano*, *XVIII*, 277–390.

Riassunto: La realizzazione di un *corpus* di una varietà storicamente determinata di una lingua pone problemi metodologici rilevantissimi, e principalmente uno: è possibile realizzare un *corpus* effettivamente rappresentativo di una varietà linguistica di cui non conosciamo, né possiamo conoscere, l'estensione e in cui, soprattutto, il peso dei testi letterari è per forza di cosa assai più cospicuo che in una lingua reale? Se la risposta, dal punto di vista della linguistica dei *corpus* è senz'altro negativa, sono oggi disponibili numerosi strumenti (primo tra tutti il *Corpus OVI dell'italiano antico*) che parrebbero mostrare il contrario.

Partendo proprio dall'esperienza del *Corpus OVI*, si proporranno alcune riflessioni di tipo metodologico sulla realizzazione di *corpus* di varietà storiche di lingua, mostrando come, pur essendo impossibile applicare criteri di bilanciamento e rappresentatività *stricto sensu*, si possano applicare una serie di correttivi (in particolare selezionando i testi sulla base delle tradizioni discorsive) che consentano di avere strumenti effettivamente rispondenti alle necessità della ricerca linguistica.

Parole chiave: Letteratura medievale; Tradizioni discorsive; Corpus OVI dell'italiano antico; Linguistica dei corpus; Lessicologia dell'italiano.