

GIULIO VACCARO

## Rappresentatività e bilanciamento in un corpus di italiano antico: appunti sul *Corpus TLIO*

Il contributo propone alcune riflessioni sull'uso del *Corpus TLIO* come strumento per l'analisi semantica e per la ricostruzione lessicale dell'italiano (antico). Gli incrementi realizzati a partire dal 2016 ne hanno modificato infatti profondamente la struttura e hanno sensibilmente mutato, conseguentemente, i risultati che la ricerca consente di ottenere. Si propongono qui alcune considerazioni sulla composizione di un corpus ai fini della ricostruzione complessiva del lessico dell'italiano antico, per cui sarebbe necessario 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.

*Parole chiave:* italiano antico, corpus TLIO, tradizioni discorsive, bilanciamento, rappresentatività.

### 1. *Il Corpus TLIO: nascita e sviluppo.*

Come è abbondantemente noto, gli studiosi della storia linguistica italiana godono di un privilegio pressoché unico tra i colleghi romanisti e, generalmente, studiosi di lingue moderne: quello di avere (e di avere avuto fin dalla fine degli anni Novanta) a disposizione in rete un corpus di enormi dimensioni per descrivere la fase antica della lingua. Si tratta ovviamente della galassia dei corpus realizzati dall'Opera del Vocabolario Italiano del CNR (OVI), un sistema che ha conosciuto, nel corso del quarto di secolo di servizio, varie fasi. Da ultimo è stata creata una separazione tra il *Corpus OVI dell'italiano antico*, destinato a contenere tutti i testi antichi editi in modo affidabile, e il *Corpus TLIO*, su cui si fonda il vocabolario. I due corpus sono stati fondati, fino al 2016, sulle stesse basi teoriche ma con una diversa opera di marcatura dei lemmi; a partire da questa data l'aggiornamento del *Corpus TLIO* è stato vincolato a alcuni criteri, esplicitati nel sito

dell'OVI (<[www.ovi.cnr.it](http://www.ovi.cnr.it)>): principalmente la datazione dei testi entro la fine del Duecento e l'«appartenenza ad aree linguistiche scarsamente documentate», ma anche l'«eccezionale rilevanza lessicale e/o culturale» e le scritture femminili<sup>1</sup>.

Le radici del *Corpus TLIO* (che è, ovviamente, il nocciolo del *Corpus OVI dell'italiano antico*) vanno ricercate negli anni Sessanta, dunque agli albori della linguistica computazionale e della linguistica dei corpus: ciò ha fatto sì che proprio all'interno dell'OVI si siano affrontate per la prima volta – in un'epoca in cui la riflessione teorica si trovava ancora a dover lottare con i limiti di operatività imposti dal *calcolatore* – questioni metodologiche relevantissime, dettate soprattutto dalla considerazione che il corpus non fosse uno strumento in sé concluso ma che esso dovesse servire come base per la realizzazione della prima sezione cronologica del *Vocabolario Storico Italiano*<sup>2</sup>. Insomma, il corpus non poteva “limitarsi” a essere grande (soprattutto per la nebulosità dei confini) ma – come sosteneva nell'autunno del 1974 l'allora direttore dell'OVI, d'Arco Silvio Avalle – doveva avere due caratteristiche principali e coesistenti per poter procedere alla redazione del vocabolario: la prima era la stabilità dell'insieme; la seconda la sua rappresentatività<sup>3</sup>. Le ragioni portate da Avalle erano molto semplici: solo un insieme stabile avrebbe infatti potuto garantire un vocabolario che descrivesse coerentemente uno stato di lingua; solo un insieme rappresentativo delle diverse realizzazioni testuali avrebbe fatto sì che il vocabolario non fosse afflitto da quel “morbo della letterarietà” che era stato il portato tradizionale della lessicografia fiorentina.

Con il primo argomento, tra l'altro, Avalle individuava un problema che è oggi infinitamente più complesso vista l'estrema facilità (tecnica) per ciò che riguarda la possibilità di aggiornamento degli strumenti *online*. Se infatti è vero che far crescere un corpus è relativamente semplice – basti pensare che il primo corpus si cui si fondava il *TLIO* era di circa 15 milioni di occorrenze e quello attuale di oltre 23 milioni –, l'altra faccia della medaglia è che, relativamente alla redazione di un vocabolario di uno stato di lingua, un corpus incre-

<sup>1</sup> Per un panorama generale sugli strumenti si vedano gli interventi raccolti in Leonardi & Squillacioti 2019.

<sup>2</sup> Per la storia dell'Opera del Vocabolario Italiano dalla fondazione fino al 1992 (anno di inizio della direzione di Pietro Beltrami), cfr. Vaccaro 2013.

<sup>3</sup> Le considerazioni di Avalle si leggono in Vaccaro 2013: 355-363.

mentato rispetto agli inizi di circa il 53% (o, reciprocamente, in cui la parte su cui sono state redatte le prime voci è pari al 65% dell'attuale) darà inevitabilmente risposte diverse. La rilevanza di questa criticità non vale tanto per il più epidermico dei problemi, ossia l'aggiunta, sempre possibile con facilità in uno strumento *online*, di singole voci nel vocabolario. Se si guarda agli ultimi incrementi lessicali si noterà infatti che essi pertengono in misura quasi esclusiva all'aspetto glossaristico ineliminabile nel *TLIO*, che redige una voce per ciascun lemma attestato nel *Corpus TLIO*: di certo però voci come *ammanifestare*, *balievole* o *canesco* rappresentano prima ancora formazioni desultorie che lessemi periferici nella lingua.

L'aspetto invece più rilevante è che questo 53% aggiunto fatalmente documenterà in modo diverso la distribuzione del lessico innanzitutto sotto il profilo delle tradizioni discorsive (che è la cosa più rilevante sul piano lessicografico), e in subordine anche sull'asse della diacronia (e varrebbe qui la pena aprire una riflessione sul concetto di *ultima attestazione* proposto da D'Achille 2020) e – ancorché in misura minore e meno rilevante – della diatopia.

È insomma concreto il rischio, già additato da Guadagnini (2016), che nel passaggio dall'attestazione all'analisi semantica un corpus avente per obiettivo precipuo la descrizione semantica (cosa ovvia per il *Corpus TLIO*) 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 alcun peso nella storia dell'italiano (antico), che vanno ulteriormente ad ampliare quella quota già altissima di *hapax* o di lessemi monotestuali che costituisce già oggi quasi i due quinti del lessico descritto nel *TLIO*: per avere un dato numerico basti considerare che nelle 42.006 voci pubblicate nel *TLIO* gli *hapax* attestati nel corpus sono circa 12.000 e i lessemi documentati solo fuori corpus (per la gran parte anch'essi monoattestati) circa 5000. Ciò vuol dire che circa il 40% del *TLIO* è costituito da lessemi episodici nella storia dell'italiano antico.

Questa prima questione si intreccia strettamente con la seconda che poneva Avalue, quella della rappresentatività, che incrocia tuttavia un tema di portata maggiore. La domanda di base è infatti: è possibile realizzare un corpus rappresentativo di una varietà linguistica di cui

non conosciamo l'estensione e in cui il peso dei testi letterari è soverchiante? Conseguentemente, è possibile per un corpus sincronico rispetto a una fase antica di lingua o per un corpus diacronico caratterizzarsi non solo per la finitezza dell'insieme, ma anche per selettività e rappresentatività?

## 2. *Distorsione e bilanciamento: il problema della diatopia*

Ritornando al *Corpus TLIO*, Avalle riteneva che quanto raccolto rappresentasse ormai in modo adeguato le tradizioni testuali e la parte omessa fosse trascurabile o «inerte» (cfr. Vaccaro 2013: 361). Ciò implica che qualunque ampliamento, qualunque *rimeditazione* (questo il termine che usa a De Robertis 1985: 445) non può che partire da una riflessione preliminare sulla qualità di quel lavoro e sulle eventuali necessità e sull'opportunità di rimodularne il bilanciamento all'interno delle singole tradizioni discorsive.

Di fatto, la massiccia inclusione nel *Corpus TLIO* di testi aggiunti sulla base di una selezione in cui diacronia e diatopia fanno aggio su qualunque altra valutazione ha cessato dichiaratamente di fare del *Corpus TLIO* un corpus stabile ma non ne ha fatto e 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, il nuovo *Corpus TLIO* che prevede solamente aggiunte casuali qualitativamente può essere solamente un glossario, per quanto ampio, di un conglomerato, per quanto numeroso, di testi: la questione, ovviamente, non si pone sul dipolo della presenza/assenza del singolo testo, bensì su una valutazione qualitativa complessiva che investe l'idea stessa di descrizione lessicografica di una lingua. Preliminare rispetto alla "rimeditazione" di un corpus che voglia essere di base per una descrizione lessicale (sia essa dell'italiano antico o di qualunque varietà storica di una lingua) dovrebbe essere una riflessione di fondo: un qualunque corpus che si prefigga di andare al di là del mero dato di attestazione non può 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 deve 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 interne (tradizioni discorsive, tipologie testuali) e esterne alla lingua (diacronia e diatopia). Il solo aggiungere testi al fine di incrementare numericamente un 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 esplicitato con estrema chiarezza da Burgassi & Guadagnini (2017), 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” (p. 11).

Poco rileva o almeno poco dovrebbe rilevare, quindi, quanto un'edizione sia commentata o quanto essa sia affidabile dal punto di vista fonologico (e in subordine morfologico) rispetto al manoscritto o ai manoscritti di base: una banca dati testuale, come qualunque opera umana, è costruita e progettata per una finalità specifica, e dunque la sua qualità e la sua efficacia vanno misurate rispetto a quella medesima finalità. Per quanto dirlo possa parere lapalissiano, il *Corpus OVI* è stato realizzato per essere la base su cui costruire il vocabolario dell'italiano antico. Dunque il principio interno che ne guida la costituzione 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.

Nel passaggio dal dato dell'attestazione a quello dell'analisi, un corpus destinato all'analisi semantica che si arricchisca di testi scelti in base a criteri diversi da quello lessicale (compreso un corpus esaustivo, e a prescindere dal fatto che sia esaustivo rispetto all'edito, al noto o al conservato) o sposterà una prospettiva glossaristica o, mantenendo la prospettiva del vocabolario, necessiterà di correttivi nella redazione delle voci (esclusione degli *hapax*, vaglio dell'attestazione, ecc.). Per contro, un corpus destinato alla redazione di un vocabolario in cui i testi siano aggiunti con criteri casuali dal punto di vista del lessico rappresentato finirà per fornire una descrizione casuale della lingua, in cui sono solo i grandi numeri a garantire la plausibilità dei risultati.

Al contrario, solo in un corpus bilanciato tra tutte le variabili è possibile 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 le parole di Cosimo Burgassi e Elisa 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: 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 non può che essere, banalmente, quanto il testo contribuisca al dettaglio di determinate zone lessicali in rapporto alla documentazione già nota. Per fare un solo esempio, la versione italiana del *Lancelot en prose* (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 mai usciti dal manoscritto in cui erano contenuti. Al contrario testi più tardi e certamente meno interessanti sotto il profilo culturale, come per esempio il *Trattato dell'arte del vetro* di Benedetto di Baldassarre Obriachi (Milanesi 1864: 69-109), sono latori di documentazione altrimenti inattuabile 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 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) sotto il profilo della diatopia.

Il dato della distribuzione linguistica dei testi, infatti, 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 storiche – molto di più si è edito; il secondo, contingente, è che la pubblicazione di singoli documenti locali è stata spesso delegata (altrove più che in Toscana) a eruditi del luogo in pubblicazioni a diffusione estremamente marginale 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. È d’altronde un dato evidente che la massima diversificazione nell’analisi semantica del lessico avviene sulla base delle tipologie testuali, delle tradizioni discorsive e dell’uso linguistico (ovvero in diafasia e in diastratia), in misura minore in diacronia (ma tale variabilità tende a scemare in una descrizione in sincronia quale è quella del *TLIO*) e solo in minima parte in diatopia.

L’enorme squilibrio dei dati numerici tra l’area toscana (da cui proviene sì “solo” il 54% dei testi, ma ben l’80% delle occorrenze) e il resto del dominio italo-romanzo non consente infatti, partendo dall’incrocio dei dati sulla presenza/assenza di un lessema in una data area – 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) – di inferire nulla sulla distribuzione di quella determinata parola in italiano antico, e dunque sul tasso di “regionalità” di un singolo vocabolo. Il solo Leonardi (2019: 26) si è soffermato sulla questione, proprio partendo da questi dati:

“Se andiamo a verificare nel vocabolario la presenza di voci fondate su documentazione esclusivamente non toscana, cioè fondata su quel 20% di testi [*rectius*: occorrenze] (...), il dato appare compatibile con quella dimensione, attestandosi attorno al 17%, a fronte

dell'83% corrispondente alle voci per le quali sussiste una documentazione toscana (...).”

“Ma la documentazione non toscana è presente in un numero significativo di queste voci, attestate *anche* in Toscana: queste voci a “Documentazione mista” (...) sono pari a un 28% del totale, per cui le voci ad attestazione solo toscana (...) nel *TLIO* risultano pari al 55% del totale, una percentuale ben più ridotta rispetto al peso dei testi [*rectius*: occorrenze] toscani sull'insieme del corpus. È un dato che può sembrare irrilevante, nella sua astrattezza, ma è un dato da cui si può partire per un'analisi del lessico italiano delle origini che tenti di valutare l'effettiva consistenza e il peso relativo delle varietà regionali dell'italiano scritto”.

Se la comparazione tra le percentuali dei *token* delle singole aree linguistiche e quella dell'attestazione dei singoli lessemi non finisce di convincermi (vista l'ovvia considerazione che la gran parte delle occorrenze in ciascuna area sarà riconducibile ai lessemi costituenti il lessico fondamentale e che tale lessico è privo di variazioni sostanziali in diatopia), i dati che emergono paiono additare piuttosto la difficoltà di un'analisi delle varietà regionali dell'italiano antico. La struttura stessa delle voci del *TLIO*, in cui il lemma è individuato sulla base dell'etimo indipendentemente dalle realizzazioni fonetiche, ci consente di dire solamente che (almeno) un 28% dei lessemi diffusi in più di un'area linguistica dell'Italia medievale ha un etimo comune: un dato che, in numeri assoluti, è pienamente sovrapponibile in una comparazione tra le diverse lingue romanze.

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 & Guadagnini (2017: 22), 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».

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 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.

### 3. *Un triplice bilanciamento*

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 quella specifica area. La constatazione è banale e la facevano per esempio già Ferdinando Gabotto e Delfino Orsi pubblicando il primo (e purtroppo unico) volume dedicato ai laudari piemontesi: «rilevare l'influsso umbro in queste laude pare perfino ozioso» (Gabotto & Orsi 1891: XII). I laudari di Bra e di Carmagnola rappresentano infatti 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, poiché essi rientrano pienamente in un modello letterario che è a un tempo anche un modello linguistico. Ciò non vuol dire, naturalmente, riproporre il mito della genuinità del testo di carattere pratico contro l'artificiosità del testo letterario, visto che anche il volgare notarile è per sua natura esposto all'interferenza, in particolare del modello latino, e anche la scrittura dei mercanti riflette al suo interno le dinamiche del contatto linguistico con le zone in cui si commercia (si pensi a quanto accade

nei documenti conservati nell'archivio del mercante pratese Francesco di Marco Datini) o con le aree con cui si ha contatto.

Del pari è necessario rappresentare 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.

Proprio quest'aspetto è quello più significativo: fare un corpus di una fase storica di una lingua non è un qualcosa di meccanico, di banale o di ancillare. L'auspicio è che gli aspetti testuali e lessicali vengano sviluppati in un nuovo *Corpus TLIO* che risponda in prima battuta a criteri di bilanciamento (che potrebbero utilmente essere quelli già previsti per *MIDIA*). Questo non tanto per contribuire alla realizzazione, ormai prossima, del *TLIO*, quanto per poter legare un corpus di lingua antica agli stati di lingua successivi, il cui pilastro finale potrebbe auspicabilmente essere il *VoDIM* e il cui primo stato potrebbe essere il *TLIQ* (*Tesoro della lingua italiana del Quattrocento*) suggerito in occasione del convegno per il trentennale dell'OVI da Tullio De Mauro (Leonardi & Maggiore 2016: 256).

### *Riferimenti bibliografici*

- Burgassi, Cosimo & Guadagnini, Elisa. 2017. *La tradizione delle parole: Sondaggi di lessicografia storica*, Strasbourg: ELiPhi.
- Cadioli, Luca (a cura di). 2016. *Lancellotto. Versione italiana inedita del «Lancelot en prose»*. Firenze: Edizioni del Galluzzo per la Fondazione Ezio Franceschini.
- Corpus TLIO per il vocabolario*, diretto da Larson, Pär, Artale, Elena & Dotto, Diego, consultabile online all'indirizzo <<http://tlioweb.ovi.cnr.it>>.
- Corpus OVI dell'italiano antico*, diretto da Larson, Pär, Artale, Elena & Dotto, Diego, consultabile online all'indirizzo <<http://gattoweb.ovi.cnr.it>>.

- D'Achille, Paolo. 2020. «A te l'estremo addio»? Il problema dell'ultima attestazione nella linguistica e nella lessicografia italiana. *Studi di lessicografia italiana* 37. 333–355.
- De Robertis, Domenico. 1985. L'Ufficio filologico dell'Opera del Vocabolario, il suo impianto, il suo lavoro. In *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)*, 444–451. Firenze: presso l'Accademia.
- Gabotto, Ferdinando & Orsi, Delfino (a cura di). 1891. *Le laudi del Piemonte*. Bologna: Romagnoli.
- Guadagnini, Elisa. 2016. Lessicografia, filologia e «corpora» digitali: qualche considerazione dalla parte dell'OVI. *Zeitschrift für romanische Philologie* 132(3). 755–792.
- Leonardi, Lino. 2019. Filologia e lessicografia digitali: l'Opera del Vocabolario Italiano a quota 40.000. *Bollettino dell'Opera del Vocabolario italiano. Supplementi* 7. 15-31.
- Leonardi, Lino & Maggiore, Marco (a cura di). 2016. *Attorno a Dante, Petrarca, Boccaccio: la lingua italiana. I primi trent'anni dell'Istituto CNR Opera del Vocabolario Italiano. Convegno internazionale (Firenze, 16-17 dicembre 2015)*. Alessandria: Edizioni dell'Orso.
- Leonardi, Lino & Squillacioti, Paolo (a cura di). 2019. *Italiano antico, italiano plurale. Testi e lessico del Medioevo nel mondo digitale. Atti del convegno internazionale in occasione delle 40.000 voci del TLIO (Firenze, 13-14 settembre 2018)*. Alessandria: Edizioni dell'Orso.
- MIDIA. *Morfologia dell'Italiano in DIACronia*, coordinato da D'Achile Paolo, consultabile online all'indirizzo <<https://www.corpusmidia.unito.it/>>.
- Milanesi, Gaetano (a cura di). 1864. *Dell'arte del vetro per mosaico. Tre trattatelli dei secoli XIV e XV*. Bologna: Romagnoli.
- Vaccaro, Giulio. 2013. Veniamo da molto lontano e andiamo molto lontano. Documenti per la storia dell'Opera del Vocabolario Italiano dalle origini al 1992. *Bollettino dell'Opera del Vocabolario Italiano* 18. 277–390.
- VoDIM. *Vocabolario Dinamico dell'Italiano Moderno*, consultabile online all'indirizzo <<https://vodim.accademidellacrusca.org/>>.