Qualche riflessione per la costituzione di un *corpus* di latino tardo[1]

Pierluigi Cuzzolin

(Università di Bergamo)

Abstract

This paper aims to present some observations on the methods used to assemble and organise collections of texts, i.e. the corpora of what are wrongly known as the *dead languages.* Although more technical aspects such as how to format or check the corpus are not dealt with here, representativeness, adequate corpus size and effectiveness are three essential features that must be taken into consideration for any late Latin corpus.

Key Words – Late Latin; Corpus linguistics; representativeness; adequate corpus size; effectiveness

Questo contributo intende presentare alcune riflessioni sui metodi che regolano la costruzione e l'organizzazione di raccolte di testi come, ad esempio, i corpora delle (erroneamente definite) lingue morte. Senza qui trattare gli aspetti più tecnici, legati per esempio al formato e al controllo del corpus, vi sono in ogni caso tre caratteristiche indispensabili e necessarie per un corpus di latino tardo: rappresentatività, dimensione adeguata, efficacia.

Parole chiave – latino tardo; Corpus linguistics; rappresentatività; grandezza del corpus; efficacia

[1] Il presente lavoro costituisce la redazione scritta e rielaborata della relazione presentata al congresso *Latin Vulgaire - Latin Tardif XIII* tenutosi a Budapest dal 3 al 7 settembre 2018. Ringrazio ancora amici e colleghi per le stimolanti osservazioni rivolte alla presentazione, Rosanna Sornicola per le sue puntuali osservazioni e due referee anonimi per gli utili commenti. Come sempre errori ed eventuali cattive idee espresse sono attribuibili solo a me.

1. Premessa

Le riflessioni del presente lavoro nascono dall'insoddisfazione che chi scrive ha provato in passato, e continua spesso a provare, nei confronti delle selezioni di testi sulle quali si basano alcuni contributi che intendono argomentare una qualche tesi di linguistica, a qualunque livello di analisi (ma in prevalenza si tratta di lavori sintattici), in prospettiva sia diacronica sia sincronica. Tale insoddisfazione è generalizzata ed è dovuta al fatto che chi utilizza un *corpus* creato *ad hoc*, posto che esso sia soddisfacente, non esplicita quasi mai i criteri in base ai quali tale corpus è costituito. Anche chi scrive ha talvolta utilizzato in passato piccoli *corpora* il cui presupposto era di mostrare che assemblare testi in maniera che si direbbe *random* poteva comunque dare risultati interessanti e affidabili. Credo sia evidente però la necessità ormai di dare un fondamento più stabile ed epistemologicamente consapevole alla pratica di basarsi su una selezione di testi per indagini su fasi storiche di una lingua, nel nostro caso il latino.

Riflessioni in questo senso non sono ricorrenti nel nostro ambito di studi: il lavoro più importante e recente, che ha il merito di discutere esplicitamente i criteri per la costituzione di un corpus di latino tardo, specificamente del latino merovingico, è quello scritto a più mani, da Selig, Eufe e Linzmeier (2017), un progetto di lavoro con il quale il presente articolo condivide molti dei presupposti dell'impostazione generale e mostra parecchi punti di convergenza. Un lavoro su cui tornerò più avanti.

Almeno negli ultimi tre decenni è venuta affermandosi con sempre maggiore successo la tendenza a mettere a disposizione degli studiosi di vari settori della linguistica i corpora; questo vale anche per un settore non ovvio, almeno per certi aspetti, come quello della linguistica storica (a proposito si vedano una informata e particolareggiata rassegna contenuta in Kytö 2011, che risulta ancora assai utile e l'altrettanto utile lavoro di Rissanen 2008).

I problemi che la costituzione di un corpus pone sono però numerosi e la facilità di trovare la loro soluzione, o almeno una soluzione, varia in misura considerevole perché i criteri per la sua costituzione devono essere molteplici e adatti alla lingua, e più precisamente alla varietà di essa che si vuole indagare. Non sarà inopportuno ricordare che quando si tratta di varietà di lingua, questa è definibile utilizzando criteri diastratici e diafasici, criteri troppo spesso sacrificati a descrizioni di sistema, e dunque considerati, piuttosto ingenuamente, neutrali rispetto a questa prospettiva sociolinguistica.

Nel presente lavoro cercherò di mettere in rilievo solo alcuni dei problemi specifici che presenta la costituzione di un corpus di latino tardo. Si tratta di una rassegna delle considerazioni di metodo su ciò che implica una impresa del genere più che di suggerimenti o proposte operative su ciò che si deve fare, anche perché conviene essere chiari su un punto: a mio parere, qualunque corpus su una cosiddetta *lingua morta*, etichetta che non a caso oggi si tende a sostituire con quella più appropriata di *lingua a corpus chiuso*, non potrà alla fin fine che coincidere necessariamente con l'intero corpus restante di tale lingua, il quale, per quanto possa sembrare un paradosso, è già costituito, ed è stato costituito in massima parte dal caso. Quindi l'operazione di costituire un corpus di latino tardo, a rigore, consisterà nella scelta, compiuta secondo un qualche criterio, di alcune opere dall'insieme di tutte le opere che vengono attribuite al latino tardo, definito cronologicamente in base a qualche criterio, per giungere a un corpus che sia comprensivo di tutte le opere dalle quali l'operazione per costituire il corpus ha preso avvio. Per chi sia dell'opinione espressa da chi scrive, un corpus di una determinata epoca è completo solo quando viene a contenere tutte le opere che ci sono giunte di quella determinata epoca. In fondo è questo il paradosso della costituzione di un corpus quando si tratta di lingue a corpus chiuso.

Per questa ragione, il punto centrale delle mie riflessioni verterà sull'aspetto epistemologico dell'operazione di costituire un corpus transitorio e parziale di una lingua a corpus chiuso, mentre non comprenderà quegli aspetti tecnici che pure sono spesso intimamente legati alla costituzione di un corpus e non ne costituiscono solo un momento esecutivo (illustra egregiamente il punto McGillivray 2014).

2. Presupposti teorici e metodologici per la costituzione di un corpus in linguistica storica

2.1. Prima di passare a illustrare che valore possa assumere la parola corpus nell'ambito di studi sul latino tardo, e, almeno in linea di principio, per ogni fase di una lingua documentata, conviene richiamare un fatto spesso trascurato: ovvero, che l'idea di costituire un corpus in linguistica trova un termine di confronto particolarmente istruttivo con il cosiddetto *blood screening*. Come è noto, per analizzare i valori del sangue di una persona al fine di ottenere dati validi e utili dal punto di vista medico per il suo valore diagnostico da tempo si ricorre all'analisi di un campione di sangue, senza che sia necessario analizzare l'intera quantità di sangue contenuta nel corpo della persona, un'operazione che sarebbe improponibile e tecnicamente complessa, a dir poco. Il presupposto corretto che sta alla base del *test* è che il campione di sangue prescelto riproduca esattamente la distribuzione degli elementi che lo compongono. Insomma: fare l'analisi di un campione sanguigno dà gli stessi risultati di una analisi condotta su tutta la quantità di sangue, di cui il campione costituirebbe una rappresentazione soltanto percentuale.

Il secondo presupposto per il *test* del sangue è dunque che il campione sia rappresentativo dell'insieme, consenta cioè di fare analisi e di trarre conclusioni a fini medici assolutamente corrette e affidabili perché l'insieme di cui rappresenta un campione è considerato biologicamente e fisiologicamente *omogeneo*.

La domanda che si pone immediatamente nel caso di una campionatura linguistica è scontata: è possibile adottare il criterio utilizzato per il test del sangue, cioè mediante una campionatura, per costituire un corpus linguistico che abbia finalità storiche? Rispondere a questa domanda è lo scopo del presente lavoro, ma una risposta esauriente, e dunque convincente, richiede parecchie precisazioni e distinzioni.

Innanzitutto, se uno dei presupposti basilari del test del *blood screening* è quello dell'omogeneità del materiale organico che viene analizzato, l'idea che metodo così efficace possa essere trasferito per analogia dal campo medico a quello linguistico può mantenere una propria validità con l'unica condizione che il corpus sia sincronico; ma anche in questo caso non mancano riserve e *caveat*. La differenza decisiva sta nel fatto che nei campioni di sangue viene presupposta un'omogeneità che manca invece nei campioni tratti dalle scienze umane e sociali: nel primo caso si parla di popolazione univariata, nel secondo di popolazione multivariata (si veda il classico volume di Blalock 1984).

Se invece la trasposizione del metodo ha come fine la costituzione di un corpus diacronico, e in particolare quello di epoche tarde di una lingua a corpus chiuso, con ambizioni di ricavarne informazioni di tipo diacronico, o più propriamente storico, allora scopo e metodo si rivelano abbastanza illusori. Nessuna meraviglia, dunque, se nell'impresa ambiziosa e utilissima di costituire un corpus di latino merovingico, quindi molto simile alla prospettiva di cui trattano le pagine presenti, Selig, Eufe e Linzmeier hanno adottato consapevolmente la soluzione più radicale, così sintetizzata dagli studiosi:

> Le linguistique (sic!) diachronicien doit donc se résigner à l'idée que, pour beaucoup de formes, il ne dispose que d'une documentation éparse et que, pour s'approcher des

> réalités communicatives situées en dehors de la scripturalité, il est obligé de se servir d'une documentation indirecte […] Ces restrictions quantitatives et qualitatives et la partialité de la documentation écrite historique nous ont induits à abandonner toute idée de représentativité propagée par la linguistique du corpus (Selig et al. 2017: 731-732).

Quanto è stato fin qui precisato impone allora di formulare le domande essenziali in modo diverso da quanto normalmente si è fatto e capire piuttosto che cosa si *possa intendere* per corpus in linguistica storica.

È opportuno precisare che la gran parte di ciò che viene preso come punto di riferimento nel presente lavoro è tratto dal recente *Handbook* dedicato alla *corpus linguistics* (O'Keeffe e McCarthy 2010). Questa scelta è voluta perché, a parere di chi scrive, tale *handbook* costituisce a tutt'oggi, oltre che un ottimo e assai consapevole strumento di lavoro, e un sicuro punto di riferimento per la disciplina in questione, anche un repertorio di conoscenze condivise su che cosa si debba intendere per *corpus linguistics* e come si debba praticare tale disciplina. Questo insieme di nozioni, condivise e accettate, può valere come temine di confronto per comprendere più a fondo quale sia la specificità di un corpus di linguistica storica per un'epoca particolarmente complessa come quella latina tarda (anche se ogni epoca ha peculiarità sue proprie che la rendono complessa); e conseguentemente quali metodi siano i più appropriati per la sua costituzione.

Tre sono i criteri grazie ai quali, almeno a mio parere, la validità di un tale corpus potrà essere controllata e sottoposta a verifica, e tutti e tre sono intimamente connessi fra loro, tanto che su alcuni punti dell'argomentazione si sovrappongono, almeno in parte: la sua rappresentatività, la sua grandezza, la sua efficacia. A ciascuno di questi criteri verranno dedicate alcune riflessioni che mi paiono essenziali.

2.2. Come sempre, però, quando si cerca di definire qualche concetto o qualche strumento concettualmente fondato diventano necessarie alcune riflessioni preliminari: nel caso in questione, che cosa si intenda con il termine corpus e che cosa ne consegua.

Per definire che cosa si intenda con corpus riporto la definizione classica offerta da David Crystal (1992: 85):

> A collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. The main purpose of a corpus is to verify a hypothesis about language – for example, to determine how the usage of a particular sound, word, or syntactic construction varies. Corpus linguistics deals with the principles and practice of using corpora in language study.

In questa definizione, solo in apparenza priva di punti che necessitano di un approfondimento, sono in particolare due i punti che suscitano dibattito: il primo punto è costituito dal fatto che manca – forse perché dato per presupposto, ma certo non esplicitato – un aspetto che invece è stato riconosciuto come caratterizzante, ovvero la rappresentatività del corpus scelto[2].

[2] Mette conto di riportare anche la definizione che di *corpus linguistics* offre *Wikipedia*, perché, oltre a essere diventata la sintesi ideale della conoscenza condivisa nelle voci trattate, in essa si trovano alcuni spunti di riflessione che sono invece assenti nella più concisa definizione di Crystal riportata sopra, e mostra una maggiore consapevolezza della teorizzazione che è alla base della costituzione di un corpus: «*Corpus linguistics* is the study of language as expressed in corpora (samples) of 'real world' text. *Corpus linguistics proposes that reliable language analysis is more feasible with* corpora *collected in the field in its natural context ("realia"), and with minimal experimental-interference.* […] The text-corpus method is a digestive approach that derives a set of abstract rules that govern a natural language from texts in that language, and explores how that language relates to other languages. Originally derived manually, corpora now are

Il secondo punto è il fatto che la principale funzione di un corpus viene identificata nella sua capacità di verificare una qualche ipotesi intorno al linguaggio, e più precisamente intorno al suo uso (anche se non rimane chiaro in che rapporto stiano funzione e struttura). Poiché sono in particolare le condizioni d'uso e la frequenza con cui un fenomeno ricorre ad essere considerati tratti fondanti della disciplina, è inevitabile trarre la conclusione che la costituzione di un corpus di latino tardo dovrà basarsi su presupposti epistemologici diversi da quelli appena menzionati. Nel caso del latino, e non solo della sua fase tarda, l'ipotesi che sta a fondamento della costituzione di un corpus con valenza storica, ipotesi che viene identificata non completamente a ragione con il suo scopo, è che esso possa documentare in modo affidabile le eventuali tendenze che caratterizzano quella particolare fase linguistica, cioè che si possano controllare linee di sviluppo o di mutamento con metodo verificabile.

Tuttavia, come sarà specificato sotto, non meno appropriata risulta anche la registrazione delle variazioni degli stati di lingua nei testi dei quali si costruisce il corpus, laddove l'arco temporale dovesse risultare piuttosto ristretto od omogeneo, secondo almeno uno dei parametri sociolinguistici rilevanti come la diafasia o la diastratia. È proprio il tratto inerente di eterogeneità, però, che rende ogni corpus possibile difficilmente comparabile al *blood test* citato sopra.

È dunque abbastanza comprensibile che la linguistica dei corpora presenti caratteristiche costitutive che poco si adattano alle esigenze di indagini su fasi antiche di una lingua, quale che essa sia. Passo all'illustrazione partita dei tre criteri in base ai quali valutare la validità di un corpus.

3. La rappresentatività

3.1. Se la rappresentatività del corpus è una caratteristica fondamentale e fondante della stessa disciplina della *corpus linguistics*, ciò corrisponde a sostenere che il campione sul quale si basa il corpus sarà tanto più rappresentativo quanto maggiore sarà la sua capacità di contenere tutti gli elementi che possono essere sottoposti a ricerca col numero minore, o comunque con un numero adeguato di testi, che sia però inferiore al numero complessivo di testi dell'epoca che intende rappresentare. Si noti che questa affermazione non è in contrasto con l'assunto fondamentale del presente lavoro, e che si trova formulato poco sopra, e cioè che per lingue a corpus chiuso, un qualunque corpus non potrà rappresentare altro che un momento transitorio nella costituzione dell'unico corpus possibile.

Questo è, a mio parere, il punto cruciale, che ho già esplicitato sopra e a proposito del quale bisogna essere chiari: mentre la costituzione di un corpus, nel caso di lingue vive, parte dal presupposto che la base da cui attingere il campione è illimitata, nel caso delle lingue come il latino, o una sua parte, come il latino tardo, la base da cui attingere è già esso stesso un campione, selezionato da fattori esterni ed estranei alla volontà di chi tale

automatically derived from source texts» (*Wikipedia*, <https://en.wikipedia.org/wiki/Corpus_linguistics> [ultimo accesso 15/09/2019]; il corsivo è mio). È interessante osservare che nella voce *Corpus Linguistics* consultata il 24/8/2018 la parola corpora era tradotta tra parentesi letteralmente dalla parola *bodies* e non, come nella redazione corrente del lemma, dalla parola *samples*. Così pure alla voce *Text corpus*, tratta anch'essa da *Wikipedia*, si legge: «In linguistics, a corpus […] or text corpus is a large and structured set of texts (nowadays usually electronically stored and processed). Text corpora are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory» (*Wikipedia*, <https://en.wikipedia.org/wiki/Text_corpus> [ultimo accesso 15/09/2019]). Anche in questo caso è interessante notare che *l'incipit* della voce inglese *Text corpus* presente su Wikipedia all'altezza cronologica del 24/8/2018 era il seguente: «A text corpus is a large and unstructured set of texts».

corpus usa: scelte culturali che hanno cancellato parte di ciò che era stato prodotto, danni prodotti da temperie storiche violente, a volte accidenti di pura casualità. Dato questo presupposto, dunque, la costituzione di un qualunque corpus sarà in realtà sempre quella di un sotto-corpus.

Ma questo dato fa capire un altro punto cruciale che distingue la costituzione di un corpus di lingua moderna da uno di lingua a corpus chiuso: la competenza linguistica del linguista che costruisce il corpus coincide, almeno in linea di principio, con ciò che viene elicitato, sia nel caso di corpora orali sia di corpora scritti. Si potrebbe anche sostenere che il linguista sa quali fenomeni può trovare ma non in quale percentuale statistica si presentano. Nel caso di una lingua a corpus chiuso, la competenza del linguista che costruisce il corpus coincide, almeno in linea di principio, con le nozioni acquisite con l'apprendimento della grammatica della lingua in questione e dal numero e dal tipo di testi dei quali ha conoscenza. Questo comporta che nel primo caso, siamo in presenza di un metodo di analisi *corpus-based*, nel secondo, di un metodo di analisi *corpus-driven*[3].

Uno dei problemi cruciali della costituzione del corpus che garantisca la sua rappresentatività, e anzi se ne pone a fondamento, è l'arco cronologico che lo delimita e all'interno del quale si trovano i testi utilizzati: del corpus costituisce una caratteristica decisiva per la scelta dei testi che lo formano. Sulla periodizzazione del latino chi scrive ha proposto insieme con una dottissima collega una possibile cronologia (Cuzzolin e Haverling 2009); ma ne sono state proposte altre, basate su criteri differenti. È tuttavia chiaro che la periodizzazione, a mano a mano che dall'epoca classica si passa a quella tarda diventa sempre meno precisa anche per l'aumento della documentazione, non solo letteraria. E dunque la necessità di distinguere con maggiore accuratezza diverse epoche, delimitandone l'arco temporale, è ormai una condizione irrinunciabile. Risulta assolutamente condivisibile, dunque, la scelta dei colleghi Selig, Eufe e Linzmeier di costituire un corpus di latino limitato al solo latino merovingico, per la relativa facilità con cui è possibile circoscrivere questa varietà non è solo linguisticamente ma anche culturalmente. Riprenderò queste considerazioni alla Sezione 4.1.

3.2. Se consideriamo quest'ultimo punto, la costituzione di un corpus di testi latini tardi, che è il caso che qui interessa, non può essere fatta per verificare alcuna ipotesi sulla lingua per le seguenti ragioni: innanzitutto, mentre, nel caso di una lingua moderna, la rappresentatività del corpus – una condizione imprescindibile perché il corpus sia considerato scientificamente valido – costituisce una condizione preliminare, un *prius*, nel caso di una lingua non più parlata, e dunque a corpus chiuso, è semmai una condizione identificabile a posteriori, un *posterius*. Insomma, solo una volta che il corpus sia stabilito si può verificare se esso è rappresentativo.

E il concetto stesso di rappresentatività nel caso di lingue a corpus chiuso, come il latino per l'appunto, è inerentemente problematico. Nel caso del latino tardo l'opportuna distinzione tra la rappresentatività statistica contrapposta alla rappresentatività testuale, e dunque storica, perde una parte consistente del suo rilievo dal momento che il corpus si basa su testi, che in quanto tali, sono storicamente e culturalmente connotati (utili riflessioni in Gard e McGillivray 2017, in particolare il cap. 2).

L'idea, insomma, che il corpus sia un insieme di testi selezionati organizzato in modo da poter soddisfare criteri specifici che rendono tali testi funzionali alle analisi

[3] Riporto le chiare definizioni che di questi due concetti ha offerto Biber: «Corpus-based research assumes the validity of linguistic forms and structures derived from linguistic theory. The primary goal of research is to analyse the systematic patterns of variation and use for those pre-defined linguistic features. Corpus-driven research is more inductive, so that the linguistic constructs themselves emerge from analysis of a corpus» (Biber 2012: 1).

linguistiche, in linea di principio ragionevole, non può essere applicato in modo meccanico alle lingue come il latino tardo. Queste osservazioni valgono per tutte le lingue o loro varietà, e dunque anche quelle moderne, quando il campione si basi su testi.

Il problema metodologico di un *prius* rispetto a un *posterius* nella costituzione del corpus come si è accennato sopra è stato ed è oggetto di recenti riflessioni ed è sintetizzabile nella domanda se le analisi linguistiche debbano essere *corpus-based* o *corpus-driven*. Inutile ricordare che le nostre analisi linguistiche e le conclusioni che ne traiamo sono di necessità *corpus-driven* ma a partire da un campione molto più ampio *corpus-based*. Si tratta di un paradosso ma anche in questo caso il paradosso ha qualche conseguenza. La questione è complessa e merita qualche precisazione.

Il punto, che si rivela ineludibile, è che l'indagine linguistica condotta su un corpus può guidarne la sua costituzione ma allo stesso tempo ne viene anche necessariamente condizionata. In realtà, la logica che sta dietro ai numerosi corpora, che possono essere di varia grandezza (ma si veda la Sezione 4.), è quasi sempre (e le eccezioni sono poche) di tipo quantitativo: inserire cioè nel corpus opere del numero più grande possibile di autori considerati significativi. Ovvero: il corpus è costituito selezionando quei testi che riportano la lingua di alcuni autori in particolare, e che si presuppone possano servire da banca dati essenziale per ottenere informazioni generalizzate su un particolare fenomeno, quale che sia il suo livello di analisi. Non è senza ragione, dunque, sostenere che in questi casi la rappresentatività verrebbe così ottenuta per mezzo della quantità. Ma come è ben noto, in un periodo come quello tardo, e anche sull'arco cronologico del latino tardo, come si accennava, le proposte avanzate dagli studiosi non sono identiche, e bisogna tenere presente che comincia a essere ampiamente documentata una produzione che non sarebbe corretto cercare di ricondurre alla norma classica, e che anzi se ne distacca progressivamente. Va ovviamente tenuto presente che non si tratta solo della ricchissima produzione degli autori cristiani, per i quali l'idea di costituire un corpus linguistico si è già concretata in alcune iniziative preliminari importanti, soprattutto sotto forma di lessici e concordanze di numerosi autori, ma anche di quegli scritti, pur quantitativamente meno numerosi, di carattere tecnico come quelli della medicina, della veterinaria o di arte militare, solo per citare alcuni fra i casi più noti e discussi, la cui descrizione così spesso pone problemi di carattere interpretativo.

E va ulteriormente messo in conto anche un altro dato di fatto, e cioè che certi fenomeni, data la loro specificità, possono essere messi in luce, ovvero emergono come documentabili, solo quando il corpus possiede una certa grandezza (su questo aspetto di progressiva *emersione* del dato raro, e dunque a volte particolarmente prezioso, si veda il capitolo di Nelson 2010).

3.3. Il fatto decisivo è che per il linguista storico, qui inteso in senso lato, cioè anche come colui che studia stadi sincronici di fasi antiche di lingue vive o estinte (estinto qui vale come sinonimo di *a corpus chiuso*[4]), un corpus, ovvero una selezione che si è sempre supposto essere rappresentativa, è da sempre uno strumento di lavoro insostituibile e, si potrebbe anche sostenere, necessario. Per citare sull'argomento, fra i tanti, un testo recente di linguistica computazionale applicata al latino:

> […] The best option for defining Latin seems to be to overcome any dichotomy and consider Latin as the outcome of a series of particular historical, geographical and cultural circumstances, necessarily leading to an inhomogeneous linguistic system

[4] Si tratta di una precisazione importante, se si tiene conto del fatto che, come già osservato, scopo del corpus, nel caso del latino, è quello di individuare tendenze o variazioni all'interno di stati di lingua.

> where elements from different areas and registers met and were only partially transmitted by the sources (McGillivray 2014: 15).

Dati questi presupposti è conseguente la conclusione che ne deriva: «Of necessity, historical linguistics has always been corpus-based since by far the principal evidence of language change and evolution is found in collections *of texts of different periods and locations*» (Tognini Bonelli 2010: 14; il corsivo è mio).

Se dunque un'ipotesi iniziale da verificare non può essere la ragione per la costituzione di un corpus, la ragione andrà individuata in altro, ovvero almeno nell'individuazione di tendenze. «The linguist aims to describe language use rather than identify linguistic universals. *The quantitative element (frequency of occurrence) is considered very significant* and, depending on the specific approach, is taken to determine the categories of description» (Tognini Bonelli 2010: 15 [n. 6]; il corsivo è mio).

Per quanto da quest'ultima affermazione si possa anche in parte dissentire – per una teoria generale del linguaggio umano cercare di identificare universali linguistici è un compito non meno importante e centrale della descrizione di una lingua e dell'uso che ne viene fatto – per il ricercatore che lavori sul latino tardo il parametro della frequenza è importante anche in senso negativo, non solo positivo: gli elementi che ricorrono meno frequentemente, sia come *types* sia come *tokens*, hanno pari importanza, se non addirittura maggiore, degli elementi che sono più frequenti per ricavare informazioni che consentono generalizzazioni sullo sviluppo della lingua in questione, e tali da rendere possibile qualche previsione sulle dinamiche del cambiamento linguistico. La frequenza è argomento troppo pervasivo nella linguistica dei corpora perché non si possa accennarne, mostrandone gli aspetti problematici che le sono inerenti.

La conclusione fondamentale è che, data l'impostazione stessa della *corpus linguistics*, un corpus servirà primariamente a dare un quadro delle tendenze strutturali di una varietà di lingua per un certo periodo, o di quella lingua in generale, ma non potrà rivolgersi ad esso chi abbia come scopo primo quello di trovare deviazioni dalla norma perché l'elicitazione di tali dati è a dir poco problematica, quando siano documentati.

4. Grandezza del corpus

4.1. Ovviamente la grandezza del corpus costituisce uno dei criteri fondamentali in base ai quali esso viene costruito. Si tratta tuttavia di un principio molto generale, a cui è possibile guardare come la risultante dell'intersezione di altri parametri di analisi più sottili. Non si tratta solo di decidere se il corpus deve comprendere i diversi aspetti ricavabili dalla sua analisi a livello diafasico, diatopico, diastratico: prima di ogni altra cosa a determinare la grandezza del corpus sarà l'arco temporale all'interno del quale si colloca il campione dei documenti scelti.

Suddividere in periodi linguisticamente oltreché storicamente significativi una qualsiasi lingua è assai problematico; nel caso del latino tardo lo è in modo particolare; e poco sopra si è accennato al fatto che la periodizzazione del latino diventa sempre meno affinata a mano a mano che la cronologia si abbassa. Quanti e quali significati siano stati dati all'aggettivo *tardo* meriterebbe uno studio a sé stante, anche se una lettura obbligatoria in questo senso rimane il volume di Einar Löfstedt pubblicato nel 1959, intitolato per l'appunto *Late Latin*, che ha dato un'impronta innovativa alla trattazione del problema. In questo contributo, pur fondamentale, all'aspetto cronologico viene dato il giusto rilievo, ovviamente, ma l'interesse dichiarato è rivolto principalmente alla questione di

determinare con la maggior precisione possibile quando il latino tardo cessò di essere parlato, dando così luogo al romanzo, piuttosto che determinare il momento in cui il latino può cominciare a essere definito tardo. Ha qualche rilievo il fatto che Löfstedt colloca genericamente gli inizi del latino tardo intorno al 200:

> Whether we are to make Late Latin start with Apuleius, Gellius, and Fronto, or – perhaps more plausibly – to refer it to the age of Tertullian and the earliest martyrologies, that is, around or shortly before 200, is a question of terminology rather than of substance. In the world of language there are no sudden transitions (Löfstedt 1959: 1).

Löfstedt avanza questa proposta, formulata come se l'argomento rimanesse sostanzialmente periferico nella discussione, soggiacendo all'idea che il modo migliore e più sicuro per identificare le epoche in cui suddividere la storia di una lingua sia quello di rifarsi a osservazioni di carattere grammaticale e stilistico. Né la questione ha ricevuto da allora trattazioni più approfondite nella pur vasta bibliografia sull'argomento (che qui volutamente tralascio di citare): il perno fondamentale intorno al quale ruotava il problema è sempre parso la fine del latino tardo e la sua transizione verso il romanzo, piuttosto che il momento del suo inizio.

4.2. Come è noto periodizzare una lingua in base a criteri interni di sviluppo, ovvero strutturali e di sistema, è, almeno fino a oggi, molto complesso se non addirittura impossibile, per più ragioni, che non è qui il caso di ricordare. L'unica operazione che risulta ragionevole e concretamente attuabile è quella di basarsi su criteri esterni, e dunque sociali, politici, o culturali. Nella storia del latino la data da cui far cominciare, *almeno convenzionalmente*, il periodo tardo della lingua latina dovrebbe essere individuato perché da quel momento in poi si dovrebbe cominciare a constatare, a un qualche livello di analisi, un mutamento decisivo anche nella grammatica del latino. Un esempio può aiutare a rendere il discorso meno astratto. Se si potesse accertare che la cosiddetta *Constitutio Antoniniana*[5], proprio per i suoi presupposti sociali e culturali, ebbe delle conseguenze decisive, con le quali rendere ragione dei successivi mutamenti linguistici, allora il 212 potrebbe essere l'anno adatto allo scopo. Purtroppo, per quanto noi sappiamo, la *Costitutio Antoniniana* non risponde al criterio suggerito e dunque l'anno della sua emanazione può essere indicato come data convenzionale dalla quale far cominciare il latino tardo solo enfatizzandone la sua convenzionalità. Se mai dunque si trovasse una data sulla quale si potrà concordare, tale data dovrà essere legata a un fattore esterno necessario per rendere ragione dei mutamenti che da quella data in poi si verificarono nella storia della lingua latina.

Un secondo, essenziale parametro che incide significativamente è quello del tipo di corpus che si intende costruire, ovvero se il corpus è di tipo generale, che intende cioè essere rappresentativo di tutti i tipi di testualità che la documentazione rimasta ci testimonia, o se invece intende essere un corpus *specialistico*. In quest'ultimo caso il corpus potrà essere di dimensioni comunque inferiori, e comunque, almeno per le lingue vive, un corpus specialistico è generalmente considerato di piccole dimensioni quando non supera le 250.000 parole, e per corpora orali di lingue vive anche la cifra di 1.000.000 di parole oggigiorno è considerata piccola.

[5] Come è noto, si tratta del decreto col quale nel 212 l'imperatore Caracalla concesse la cittadinanza a tutti i residenti all'interno dei confini dell'impero, tranne ai *dediticii* (qualunque gruppo sociale ci fosse dietro a questa denominazione). Sui complessi problemi che pone questo importante documento si veda Purpura (2012).

Come si diceva comunque nelle pagine precedenti, il vero punto di riflessione sarà dunque ancora una volta sul metodo da adottare per trasformare il punto di partenza, ovvero tutta la produzione tardolatina rimasta, in punto di arrivo.

Il problema di fondo a cui queste riflessioni rimandano è quello della campionatura statistica. Per le lingue moderne si è avuto un significativo cambiamento nell'approccio al problema:

> A movement then grew in the 1990s that was more concerned with corpus exploitation than corpus exploration […]. This movement saw the value of smaller corpora and stressed their pedagogical purpose over their lexicographical potential. Small corpora, it was held, can be very useful, providing they can offer a 'balanced' and 'representative' picture of a specific area of the language. This recognition of a need for smaller, more specialised corpora increased (Nelson 2010: 55).

Questo approccio potrebbe, e forse dovrebbe, essere opportunamente adottato anche per le lingue antiche, o per varietà di esse. Ovviamente per la costruzione di un corpus in qualche lingua moderna sono stati elaborati metodi raffinati, e addirittura sofisticati, per avere, tramite una campionatura bilanciata e affidabile, su cui non è qui il caso di addentrarsi (si vedano le pagine molto chiare di McEnery e Wilson 2001: 81-85), corpora in cui il dato qualitativo è integrato dal dato quantitativo; e, molto verisimilmente, tutto questo potrà avere conseguenze differenti se si tratta di lessico contrapposto ad altri fattori come quelli morfologici o sintattici.

Nonostante sembri quasi lapalissiano il sostenerlo, è opportuno ribadire che una procedura del genere, nel caso di lingue a corpus chiuso, non può essere applicata se non a rischio di ottenere risultati statisticamente e percentualmente distorti. Come si è ricordato sopra, il linguista che voglia costituire un corpus di una lingua estinta come il latino si trova a disporre di una documentazione eterogenea e non bilanciabile, scelta per noi da circostanze in gran parte fortuite o culturalmente definite. Purtroppo sono i limiti intrinseci del campione di partenza che impediscono una ulteriore campionatura, convincente oltre ogni ragionevole dubbio.

5. Efficacia

Questo principio equivale a quello che nella letteratura specialistica viene identificato anche con l'etichetta di *Praticità*. Conviene chiarire subito il punto in questione: il termine praticità, ancora prima di quello di efficienza, rimanda a un approccio in cui, fra i tratti costitutivi della struttura di un corpus, ci sia la possibilità di dare a eventuali interrogazioni una risposta in termini di risparmio temporale. Per chi però è alle prese con un corpus di lingua estinta, ovvero a corpus chiuso, una simile condizione non è prioritaria. Questo significa che i criteri per la costituzione del corpus differiscono da, o meglio: non coincidono completamente con, quelli squisitamente tecnici, legati alle capacità dello strumento informatico utilizzato per la messa in rete o su altro supporto dell'intero campione.

Ma un punto di particolare rilievo, che spesso risulta fra le cause esplicite proprio di quella insoddisfazione per i vari corpora di lingue antiche, strettamente connesso all'efficacia, è il fatto che l'interrogabilità e l'analisi del corpus devono essere condotte in modo empirico e non introspettivo. Come è stato esplicitamente sostenuto:

> *The analyses must be empirical – rather than introspective* – since language users often are not consciously aware of their most typical choices. The analyses must cover numerous data in order to tell which language choices are widespread, which occur

> predictably although under rare circumstances, and which are more idiosyncratic (Conrad 2010: 227; il corsivo è mio).

Così per questo aspetto, dunque, calza a pennello quanto aveva ricordato ormai alcuni anni fa Anna Morpurgo Davies a proposito di questo problema:

> Per le lingue classiche ci si è domandato spesso fino a che punto la mancanza del parlante e della possibilità di introspezione possa essere decisiva e ci si è risposto in genere che, entro certi limiti, un'analisi moderna era possibile data la vastità del corpus […]. D'altra parte, il dubbio rimane, e con esso rimane una consapevolezza dei limiti entro cui si deve operare, consapevolezza che si rafforza quando si pensa non solo a questione di distribuzione e di sintassi, ma anche appunto a quei fattori sociolinguistici, etnolinguistici etc., a cui si accennava sopra (Morpurgo Davies 1992: 68).

Il lungo lavoro della Morpurgo Davies, una lettura più che caldamente raccomandabile a chiunque sia interessato ad avere una visione articolata del rapporto tra linguistica e linguistica storica, correttamente, a mio modo di vedere, sottolinea l'interazione tra la prospettiva diacronica e quella sincronica:

> Quello che la linguistica storica può fare per l'inglese è di introdurre una dimensione diacronica in uno studio che, dal punto di vista sincronico, dispone di mezzi e dati adeguati; quello che deve fare innanzitutto per il greco e il latino è di facilitare e talvolta permettere una descrizione sincronica; in aggiunta, (ma si tratta di un problema diverso) dovrà introdurre anche qui la dimensione diacronica (Morpurgo Davies 1992: 68).

Come si è sottolineato all'inizio di queste pagine, il corpus deve essere costruito per avere rappresentatività innanzitutto delle tendenze generali della lingua, almeno come documentate dai testi selezionati. Non mi pare inopportuno riportare una citazione significativa dal lavoro di Reppen, che accenna anche a livelli di analisi, come quello dell'intonazione, che richiederebbero ulteriori riflessioni:

> […] *a corpus can serve as a useful tool for discovering many aspects of language use that otherwise may go unnoticed*. Unlike straightforward grammaticality judgements, when we are asked to reflect on language use, our recall and intuitions about language often are not accurate. Therefore, a corpus is essential when exploring issues or questions related to language use. The wide range of questions related to language use that can be addressed through a corpus is a strength of this approach. Questions that range from the level of words and intonation to how constellations of linguistic features work together in discourse can all be explored through the lens of corpus linguistics. Questions related to aspects of how language use varies by situation, or over time, are also ideal areas to explore through corpus research (Reppen 2010: 31; il corsivo è mio).

6. Che cosa può dirci della lingua un corpus?

Una volta che il corpus sia costruito, resta da chiedersi se sia adeguato agli scopi per i quali è stato creato. Ovviamente un corpus può essere interrogato a tutti i diversi livelli di analisi, indipendentemente dagli aspetti tecnici di interrogabilità, che possono essere tenuti sufficientemente distinti dai criteri della sua costituzione e che dei quali si è già brevemente accennato. Va però osservato che, poiché il corpus è primariamente

organizzato secondo i diversi tipi di testualità presenti nella documentazione, quando si tratta di documentazione scritta, allora i dati ricavabili dalle analisi saranno condizionati dal tipo di testualità che li conserva. Questo è un altro punto chiave che non può essere eluso e da cui dipendono le risposte che può offrire l'analisi del corpus. Un esempio che è già stato ricordato sopra ed è stato ripetutamente discusso nella letteratura è il peso, non solo statistico, che si deve dare al cosiddetto latino dei cristiani. Il problema della valutazione di questa varietà di latino, come è ben noto, si pone non solo quando si voglia costituire un corpus, ma è un problema generale di più vasta portata. Se è vero che la scuola di Nimega, e in particolare i due studiosi di maggiore spicco, Schrijnen e Mohrman, hanno offerto un'interpretazione del latino dei cristiani come linguaggio specialistico, la maggioranza degli studiosi ha ridimensionato in modo sostanziale questa analisi. Tuttavia, di qualunque varietà che possa essere caratterizzata come un socioletto a sé stante, grazie all'essenziale omogeneità dell'ambiente culturale e sociale nel cui ambito esso veniva impiegato, si può opportunamente costruire un corpus, giustificabile proprio con le ragioni con le quali si è costruito il corpus del latino merovingico citato sopra.

Come si è già ribadito, il corpus può offrire dati che presentano un'idea delle tendenze generali del periodo preso in esame, meno può soccorrere il linguista che cerchi la forma anomala, la costruzione tipologicamente rara, l'arcaismo o fenomeni analoghi.

E, almeno per la situazione presente, ci sono molti aspetti per i quali la comunità scientifica non dispone di corpora adeguati oppure per i quali non sono ancora state elaborate le riflessioni metodologiche necessarie. Fra questi aspetti ci sono tutti quelli che riguardano la pragmatica, per esempio, ma non solo. A questo proposito mi paiono particolarmente adatte due citazioni tratte dal lavoro di Rühlemann, nelle quali la complessità di ricavare informazioni pragmatiche da corpora di lingue moderne, e dunque maggiormente controllabili, è ben messa in rilievo:

> If we take pragmatics as the study of meaning of text in context, it becomes clear that the relationship between pragmatics and corpus linguistics is *not unproblematic*. The reason is simple: *corpora record text, not meaning, and they record context only crudely* (Rühlemann 2010: 289; il corsivo è mio).

E ancora:

> Where there is a complete form-function mismatch, as in cases of conversational implicature, *a quantitative corpus study will be useless: what listeners take to be implicated in an utterance cannot be retrieved exhaustively from a corpus*; it can only, with varying degrees of confidence, be inferred *post hoc* (Rühlemann 2010: 290; il corsivo è mio).

A maggior ragione, dunque, tutte le informazioni che riguardano aspetti come registro, marche di discorso, o anche la rilevanza degli effetti conseguenti all'attività traduttiva (su cui ha riflettuto Panagl 2003) richiederanno che i futuri corpora rispondano a criteri di adeguatezza ancora più consapevoli e articolati.

7. Conclusioni

Come si era fatto presente all'inizio del presente lavoro, il lettore che abbia avuto la pazienza di seguire quanto sostenuto fin qua ha probabilmente trovato in queste pagine più

riflessioni sui problemi che la costituzione di un corpus pone e conseguenti inviti alla cautela metodologica che non risposte di carattere prescrittivo. L'idea che per costruire dei corpora di lingue antiche, e dunque estinte e a corpus chiuso, e il latino tardo è solo una di queste, si possano utilizzare *sic et simpliciter* i metodi e le prospettive che sono normalmente impiegate nella costituzione di corpora di lingue moderne, vive, con corpus aperto, appare quanto meno ingenua. In realtà, come si è cercato di mostrare, la costruzione di un corpus di lingue antiche richiede che si risponda consapevolmente ad alcune domande preliminari che concernono sia il metodo sia gli scopi della sua costruzione, tenendo ben a mente che tale corpus è condizionato, implicitamente e in maniera decisiva, dalla natura dei testi che lo costituiscono e che lo limitano, altrettanto implicitamente, nelle sue potenzialità.

Poiché, come si è sostenuto nelle pagine precedenti, tutti i possibili corpora non sono e non possono rappresentare altro che fasi intermedie in vista di un corpus ultimo che verrebbe a coincidere con tutta la documentazione testuale che ci è rimasta, sarebbe forse più utile e realizzabile la costruzione di corpora che siano rappresentativi di un qualche aspetto particolare del latino tardo. Ma si tratta di un'operazione che a tutt'oggi occorre ripensare nella sua globalità, perché i problemi aperti rimangono ancora numerosi.

Riferimenti bibliografici

Biber, Douglas (2012), 'Corpus-Based and Corpus-driven Analyses of Language Variation and Use', in Heine, Bernd; Narrog, Heiko (eds.), *The Oxford Handbook of Linguistic Analysis* Online Publication, < https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199677078.001.0001/oxfordhb-9780199677078> (ultimo accesso 15/09/2019).

Blalock, Hubert M. (1984), *Statistica per la ricerca sociale*, Bologna, Il Mulino.

Conrad, Susan (2010), 'What can a corpus tell us about grammar?', in O'Keefe, Anne; McCarthy, Michael (eds.), *The Routledge Handbook of Corpus Linguistics*, London-New York, Routledge, 227-240.

Cuzzolin, Pierluigi; Haverling, Gerd (2009), 'Syntax, sociolinguistics, and literary genres', in Baldi, Philip; Cuzzolin, Pierluigi (eds.), *New Perspectives on Historical Latin Syntax*, vol. 1: *Syntax of the Sentence*, Berlin-New York, Mouton de Gruyter: 19-64.

Crystal, David (1992), *An Encyclopedic Dictionary of Language and Languages*, Oxford, Oxford University Press.

Gard, Jenset B.; McGillivray, Barbara (2017), *Quantitative historical linguistics*, Oxford, Oxford University Press.

Kytö, Merja (2011), 'Corpora and historical linguistics', *Revista Brasileira de Linguística Aplicada* 11(2), 417-457.

McGillivray, Barbara (2014), *Methods in Latin Computational Linguistics*, Leiden, Brill.

Morpurgo Davies, Anna (1992), 'Il significato della linguistica storica nell'analisi delle lingue classiche', in Aa.Vv., *Atti dell'Accademia dei Lincei. La posizione attuale della linguistica storica nell'ambito delle discipline linguistiche* (Roma, 26-28 marzo 1991), Roma, Accademia Nazionale dei Lincei, 65-86.

O'Keeffe, Anne; McCarthy, Michael (eds). 2010, *The Routledge Handbook of Corpus Linguistics*, London-New York, Routledge.

Panagl, Oswald (2003), 'Danke ja! und danke nein! im Lateinischen', in Held, Gudrun (ed.), *Partikeln und Höflichkeit*, Bern, Peter Lang, 238-246.

Purpura Gianfranco (2012), '*Constitutio Antoniniana de civitate*', in Purpura, Granfranco (ed.), *Revisione ed integrazione dei Fontes Iuris Romani Anteiustiniani* (FIRA). *Studi preparatori I. Leges*, Torino, Giappichelli, 695-732.

Reppen, Randi (2010), 'Building a corpus. What are the key considerations?', in O'Keefe, Anne; McCarthy, Michael (eds.), *The Routledge Handbook of Corpus Linguistics*, London-New York, Routledge, 31-37.

Rissanen, Matti (2008), 'Corpus linguistics and historical linguistics', in Lüdeling, Anke; Kytö, Merja (eds.), *Corpus Linguistics. An International Handbook*: vol. 1, Berlin – New York, de Gruyter, 53-68.

Rühlemann, Christoph (2010), 'What can a corpus tell us about pragmatics?' in O'Keefe, Anne; McCarthy, Michael (eds.), *The Routledge Handbook of Corpus Linguistics*, London-New York, Routledge, 288-301.

Selig, Maria; Eufe, Rembert; Linzmeier, Laura (2017), 'CoLaMer (Corpus du latin mérovingien)', in García Leal, Alfonso; Prieto Entrialgo, Clara Elena (eds.), *Latin vulgaire latin tardif XI*, Hildesheim-Zürich-New York, Olms-Weidmann, 730-741.

Tognini Bonelli (2010), 'Theoretical overview of the evolution of corpus linguistics', in O'Keefe, Anne; McCarthy, Michael (eds.), *The Routledge Handbook of Corpus Linguistics*, London-New York, Routledge, 14-27.

Pierluigi Cuzzolin
Università di Bergamo (Italy)
pierluigi.cuzzolin@unibg.it