Verbum IV/2, pp. 271- 294.
1585-079X © Akadémiai Kiadó, Budapest, 2002

IL PROGETTO ITALANT E LA GRAMMATICA DEL CORPUS

LORENZO RENZI

Università di Padova
lorenzo.renzi@unipd.it

In the first part of this paper, the author presents the main concepts of the "corpus grammar" and he discusses it. He disagrees with the claim that the corpus grammar could compete with generative grammar and with other modern linguistic theories. The second part of the paper deals with the "Italant" project, that aims to constitute a grammar of Old Italian (i.e. of the Old Florentin linguistic veriety, written between the middle of the 13th till 1300, approximatly). The author argues that such a grammar has to be made basically using the same criteria which lead the linguist when he tries to describe a living language, although this could sound paradoxical.

La Grammatica dell'Italiano antico, diretta da Giampaolo Salvi e da me, che costituisce il progetto *Italant*, sarà basata su un *corpus*, quello costituito dalle scritture fiorentine dalle origini all'anno 1300 circa.

Si tratta del *corpus* dell'*OVI (Opera del vocabolario italiano* e in particolare il *TLIO, Tesoro della lingua italiana delle origini*, diretto da Pietro Beltrami) ora accessibile via Telnet attraverso la collaborazione con la University of Chicago e la University of Notre Dame negli Stati Uniti e l'Università di Reading in Inghilterra.

Da questo *corpus* deriva, compendiato e fornita di un nuovo sistema apposito di interrogazione, il *Padua Corpus*. Quando il nostro progetto è iniziato, il *Padua Corpus*, che è stato allestito per noi *dall'équipe* del CNR di Firenze che operava al TLIO, e in particolare dal direttore Pietro Beltrami e dall'informatico Domenico Iorio Fili, appariva come lo strumento essenziale di ricerca. Ma, con il potenziamento del *TLIO* conseguito attraverso la collaborazione delle due Università americane ricordate (grazie a Mark Olsen dell'Università di Chicago e Theodore J. Cachey e Christian Dupont dell'Università di Notre Dame, Indiana), e con la constatazione

che il *corpus* ridotto dava in molti casi (contro le nostre aspettative) troppo pochi dati, appare ora che il lavoro può essere condotto in modo ottimale sul l'OVI (in ogni caso tutti i collaboratori hanno a disposizione i due strumenti di ricerca, che, provenendo dalla stessa fonte, hanno il vantaggio di non contraddirsi mai). Si tratta di selezionare dall'OVI, che contiene testi e documenti volgari d'Italia dalle origini fino al 1375, i soli testi fiorentini limitandoli all'anno 1300 circa.

Sia l'OVI che il Padua corpus sono forniti di sistemi di interrogazione, strumenti essenziali per il reperimento delle forme. Pensati per le operazioni essenziali della lessicografia, queste operazioni, sfruttate ingegnosamente, possono essere utili in molti casi (ma probabilmente non in tutti) anche nella ricerca morfologica e sintattica. Per il Padua *corpus* è stato predisposto un sistema di interrogazione, detto *Gatto* (Gestione automatizzata del tesoro delle origini) particolarmente adattato appunto per la ricerca morfologica e sintattica.[1]

Cosa c'è di nuovo in tutto ciò?

La grammatica di una lingua antica è sempre il risultato dello spoglio e dello studio di un *corpus*: *corpus* cartaceo all'antica, o, da qualche tempo, *corpus* elettronico.[2] Nella necessità assoluta di basarsi su un *corpus*, lo studio di una lingua antica si differenzia dalla grammatica di una lingua moderna, che può essere basata sulla competenza del parlante (diretta o acquisita), la cosiddetta introspezione, sostenuta dall'orientamento generativista. (Vedremo in seguito come, a nostro parere, non ci si possa servire di un *corpus* senza una competenza acquisita da parte dello studioso).

Vorrei ora esaminare in che modo intendiamo servirci per la futura *Grammatica dell'italiano antico,* alla quale lavoriamo da anni, del *corpus*. Sarà forse una *grammatica del corpus?*

CORPUS LINGUISTICS

Si scrive molto oggi di „linguistica del corpus", *corpus linguistics,* e soprattutto se ne fa molta nei laboratori americani, europei (soprattutto inglesi) e di altri continenti. La linguistica del corpus (o dei *corpora*, come propone per l'italiano Rema Rossini Favretti, in ingl. *corpus linguistics)* erede della *computational linguistics* è probabilmente oggi la tendenza di maggior suc-

[1] Il sito dell'OVI si trova in http://www.csovi.fi.cnr.it oppure in http://www.vocabolario.org. Per il TLIO, l'OVI e il GATTO e la loro storia, vedi Squillacioti, P., Mosti, R., Larson p. 2001.

[2] Sull'unitarietà del concetto di *corpus*, vedi Caravedo 1999, cap.I.

cesso nella ricerca universitaria, tanto da sfidare ormai con successo non solo la linguistica storica e la filologia, ma perfino le varie forma di linguistica moderna, tra cui la grammatica generativa. E' interessante che alla luce della linguistica del corpus capiti di trovare la grammatica generativa classificata tra i vari approcci „tradizionali" alla lingua.

Ma cosa si intende, oggi, con *linguistica del corpus?*

A leggere la letteratura corrente, verrebbe da dire: molto *corpus*, e poca linguistica. La cosiddetta „corpus linguistics", è piuttosto un'elaborazione del *corpus* stesso, che non una teoria della lingua indotta dall'uso del *corpus*. Torniamo più tardi su questa precisazione.

Attingiamo per alcune osservazioni ad alcune delle fonti correnti per lo studio della *linguistica del corpus:* Sinclair, J. 1991, Oostdijk e de Haan 1994; Botley & McEnery 1996, McEnery e Wilson 1996, la recente raccolta di Rema Rossini Favretti 2000. E vediamo un po'.

La linguistica del corpus consiste essenzialmente nella costituzioni di *corpora* e nell'applicazione a questa di alcune tecniche. Le principali, tra queste, sono:

(1) le „concordanze" intese come raccolta di forme uguali, date assieme a una porzione del testo da cui sono state prelevate;
(2) la „lemmatizzazione", intesa come la pratica corrente per raccogliere delle voci in paradigmi;
(3) l'„etichettatura grammaticale", cioè l'attribuzione della categoria a ogni singola parola: nella letteratura *tagging;*
(4) l'analisi in costituenti immediati: *parsing*;
(5) la ricerca automatica di cooccorrenze di un dato elemento con altri: *Collocations* o *Cooccurence patterns*, *Corresponding Analysis*, o denominazioni simili.

Nella concezione della *linguistica del corpus* queste pratiche sono in scala ascendente di complessità, dalla più semplice e, apparentemente, meno redditizia alla più sofisticata. Ma il nostro parere, come vedremo, non è lo stesso. In tutti i casi la cosiddetta *linguistica del corpus* è, a nostro parere, un'elaborazione del *corpus* stesso, ma assolutamente non una teoria linguistica paragonabile a quante altre hanno meritato e meritano questo nome.

(1) Le *concordanze* non hanno bisogno di commento in quanto sono pratiche nate fuori dalla linguistica informatica, che questa ha assunto da tempo felicemente. In particolare lo strumento informatico permette di superare gli ovvi limiti di spazio connaturati alla forma cartacea, che aveva portato in passato spesso a rinunce e mutilazioni.[3] Inoltre i corpora in-

[3] Per es. nelle *Concordanze dei Promessi sposi*, a cura di Giorgio De Rienzo, Egidio Del Boca e Sandro Orlando, Milano, Banca del Monte di Milano–Mondadori, 1985, in

formatici possono essere messi in rete e diventano utilizzabili dagli studiosi anche prima che la raccolta sia terminata: è quanto avviene, tra gli altri casi, con l'OVI. Anche questo è un vantaggio enorme, che libera gli studiosi dalla necessità di attendere per anni i risultati di imprese laboriose, costose e complesse. Nell'attesa che la raccolta di dati sia finita, si può lavorare con raccolte provvisorie, spesso già, sempre come l'OVI, di dimensioni vastissime.

(2) Poco c'è da dire anche sulla *lemmatizzazione*. Questa operazione, ben nota alla lessicografia, separa le voci omonime ma dal significato diverso, e raccoglie quelle che presentano mere differenze fonetiche o grafiche. Per fare degli esempi semplici in italiano, si tratta di distinguere tra, mettiamo, *muto* „che non parla" e *muto* „io cambio" o tra *ratto* „grosso topo" *ratto* „rapido" (arcaico) (si tratta di casi di *omonimia)*, di raccogliere insieme parole come *obiettivo* e *obbiettivo*. In italiano antico si tratta anche di raccogliere insieme forme uguali ma rese con grafie diverse come, per es., *cane*, *chane* e *kane*. La lemmatizzazione costituisce inoltre dei paradigmi mettendo assieme le forme singolari e plurali dei nomi, singolari e plurali maschili femminili (eventualmente neutre) degli aggettivi, modi, tempi e persone dei verbi, ecc. ecc. Tutte queste forme vengono raccolte sotto una sola voce, il lemma.

Nei protocolli generali della linguistica del corpus questa seconda operazione appare spesso come la prima, dato che il lavoro che porta alle concordanze viene dato per presupposto (Sinclair 1991, 41–42; McEnery e Wilson 1996, 42).

La lemmatizzazione è un'operazione estremamente utile per chi consulta un *corpus*, ma può contenere degli errori e delle omissioni. Questo pericolo si accentua nell'operazione successiva.

(3) Il *tagging* consiste nell'assegnazione ad ogni parola del *corpus* di un'etichetta grammaticale, per es. *a, di*, *con* saranno etichettate Preposizione, *mano* o *piedi* nome, il primo singolare e il secondo plurale, ecc. ecc.

L'individuazione e l'attribuzione dell'etichetta viene fatta manualmente o automaticamente, o con una mescolanza delle due tecniche. Oggi il caso più comune è che l'etichettatura sia mista, venga fatta cioè automaticamente grazie a una procedura complessa, ma siccome si sa che non tutti i casi saranno risolti automaticamente, un controllo manuale provvederà a sistemare i casi residui. Questo controllo deve essere affidato naturalmente a un „etichettatore", cioè a un linguista esperto, molto attento e

---

cui le parole grammaticali sono state soppresse per esigenze editoriali (vedi anche la recensione di Gianfranco Folena *Misure mazoniane* nell'„Indice", III, 1986, n. 5.

naturalmente ben retribuito: il suo lavoro è lungo, e nonostante le apparenze, molto difficile.

Come si sa, infatti, l'appartenenza di una data parola a una categoria grammaticale è chiara in certi casi, meno in altri. Ci sono dei casi semplici, centrali nell'ordinamento grammaticale di una lingua, di facile individuazione, e sono certamente la maggioranza, ma ci sono anche casi più difficili, periferici nella grammatica, che presentano delle resistenze non solo a un'analisi automatica, ma anche al lavoro di un linguista accorto. Per fare un esempio nel nostro dominio, quello dell'italiano antico, nella sequenza „mangiava di frutte", *di* è preposizione, o articolo partitivo, o cos'altro? Dare le etichette alle parole presupporrebbe di aver risolto proprio quei problemi che costituiscono alle volte il tema stesso della ricerca. Qui il *tagging* non può certo aiutarci, visto che è il *tagging* stesso che deve ricevere informazioni dal lavoro del linguista.

Nonostante questo limite, il *tagging* è un'operazione di indubbia utilità, visto che potenzia e semplifica la possibilità di interrogazione del *corpus*. Avendo a disposizione un *corpus* annotato in questo modo *(tagged corpus)*, potrò, per es., cercare se ci sono nel *corpus* coppie di preposizioni, del tipo *(in su*, *di per* ecc.) senza bisogno di esplicitare tutta la combinatoria possibile, potrò cercare delle sequenze come, per es., quella di *verbo avverbio*, ecc.ecc. Un altro caso, reale, è il seguente: per documentare in italiano antico il tipo della comparativa corrispondente al tipo dell'ital. mod. *più buono che ricco*, l'interrogazione per via lessicale risulta pesantissima. Bisogna cercare la cooccorrenza di due elementi di altissima frequenza come *più* o *che*, che si trovano nella vicinanza l'uno dell'altro per le più svariate ragioni (o per nessuna ragione): nell'OVI, selezionando il solo fiorentino fino al 1300, ci sono più di mille occorrenze. La ricerca diventerebbe facilissima potendo interrogare la sequenza *più* Aggettivo *che* Aggettivo (e *più* Agg. *di* Agg.).

Tuttavia il *tagging* comporta anche dei rischi: se, per tornare all'esempio di prima, tutte le occorrenze di *di* seguite da SN sono state indicate come preposizioni, è difficile che a qualcuno venga in mente che possano essere articoli partitivi, e comunque, se questa possibilità gli viene in mente, dovrà disfare il *tagging* e rifare tutto da sé.

(4) Veniamo alla successiva operazione, il *parsing* (Mc Enery e Wilson 1996, 42–49), cioè la divisione del testo in costituenti immediati, corrente nella tradizione linguistica americana almeno da Bloomfield, e accettata anche da Chomsky in *Syntactic Structures* come parte del suo modello.

Anche il *parsing* può essere manuale o automatico Il genere di problemi che si è posto per il *tagging*, si pone in modo molto più grave per il *parsing*. Non sono in grado di valutare i problemi di un *parsing* automatico, che a quanto pare viene comunque praticato ancora raramente. Le proce-

dure per assegnare delle etichette sintattiche a un testo devono essere estrememente sofisticate. Quanto al *parsing* manuale, se penso a qualcuno al quale venga affidato il compito di eseguire un *parsing* manuale, mi assalgono fortissimi dubbi. Se ci fosse davvero qualcuno capace di individuare correttamente tutte le frasi e assegnare loro il vero indicatore sintagmatico, ebbene allora quello sarebbe un perfetto linguista, per il quale la sintassi non avrebbe più segreti. Di fronte alla sua prova non ci sarebbe più materia per la ricerca. Mentre, si sa, la ricerca è infinita.

In realtà, come il *tagging*, ma più di questo, il *parsing* può svolgere solo un lavoro grosso, approssimativo. Ma mentre sono convinto dell'utilità del *tagging*, sono scettico sull'utilità di avere un corpus provvisto di *parsing*. Questo non solo per il rapporto tra costi e benefici, che immagino che possa essere difficilmente positivo, ma anche perché non vedo a cosa possa servire avere in un *corpus* indicazioni di soggetti, oggetti, predicati, ecc. ecc. Diversamente dalle categorie grammaticali, quelle sintattiche sono ridotte in numero, e ognuno può trovare rapidamente da sé tutti i soggetti, tutti gli oggetti, i predicati e gli altri elementi di una frase (salvi i dubbi di cui dicevo). E cioè: quasi ogni frase ha un soggetto visibile, molte frasi hanno un oggetto, tutte un predicato: non vedo in che modo il lavoro del linguista possa essere favorito dal semilavorato che gli può fornire un testo annotato.

(5) *Collocations* (o *Cooccurence patterns* o *Corresponding Analysis)*. Si tratta della più ambiziosa delle tecniche di *linguistica del corpus*, in quanto costituisce una vera e propria tecnica di „procedura di scoperta", nel senso di Chomsky 1957. Nel Cap.VI di *Syntactic Structures*, Chomsky descriveva questo procedimento come quello che pretenderebbe, dato un corpus, di ricavarne direttamente la grammatica, e lo riteneva un requisito troppo forte, cioè irrealistico, per la costituzione di una grammatica

Ma, per quanto possa parere sorprendente, proprio di una tale pretesa si tratta qui. Se ne può avere un'idea dagli esempi di procedimenti riportati da McEnery e Wilson 1996, 71 ss. Un saggio di prima mano si ha nell'analisi del verbo inglese *to budge* „scostare, smuovere" in Sinclair 1997.

Il lettore italiano se può fare un'idea sulla propria lingua dall'articolo di Rema Rossini Favretti (2001). L'autrice mostra come, con tecniche adeguate, si possano fare emergere da un *corpus* di italiano scritto (il CORIS di Bologna) le forme *so* e *conosco*, e il profilo caratteristico dei contesti in cui queste appaiono. I contesti sono molto vari, ma alcune costanti sembrano profilarsi, seppure con molta approssimazione. Ambedue le forme, *so* e *conosco*, coocoorrono spesso con il pronome *lo*, ma si differenziano in altri casi: *so* predilige verbi all'infinito, frasi introdotte da *che* e altri elementi subordinanti, mentre *conosco* si accompagna spesso a un nome (sarà l'oggetto: *Conosco un tuo amico…)*. A differenza di *so*, che presenta

solo la cooccorrenza *lo so*, *conosco* accanto a *lo*, presenta anche *la*, *li*, *le*. Per quanto incerti e provvisori, questi tentativi vanno chiaramente nella direzione del procedimento di scoperta, la loro ambizione non potendo essere altra di quella di una grammatica che, passo dopo passo, si faccia da sé.

Alla prova dei fatti, tuttavia, i risultati ottenuti dal *corpus* per via induttiva sono del tutto parziali, e sono paragonabili, direi, ai primi dati che uno studente inesperto può tirar fuori da se stesso per via introspettiva. Gli basterà un solo istante in più per ricordarsi che anche *sapere* può avere un oggetto *(so la strada*, *non so la geografia)*., esempi casualmente assenti nel suo corpus. Gli basterà un po' di pratica grammaticale scolastica per generalizzare le osservazioni fatte dicendo che sia *sapere* che *conoscere* sono verbi transitivi. Gli basterà un po' di esperienza per ricordarsi che, per avere informazioni precise su temi come questi, basta aprire un buon vocabolario, anche un vocabolario modesto, ma che ha alle spalle il poderoso, secolare lavoro. Questo lavoro sulla nostra lingua è stato fatto secoli fa dagli Accademici della Crusca, poi da quel gigante della lessicografia che è stato da Niccolò Tommaseo. Questo lavoro è stato fatto sull'introspezione *e* sul *corpus* dell'italiano scritto, un *corpus* che per secoli, prima dell'elettronica, si trasportava artigianalmente su schede.

La questione è se la ricerca automatica delle corrispondenze, o cooccorrenze, possa col tempo, raffinando i propri metodi, andare oltre i risultati ottenuti con i metodi ormai collaudati che abbiamo ricordato. Sarà possibile che un giorno una voce di vocabolario fatta da un calcolatore, secondo la tecnica qui presentata delle cooccorrenze, sia più precisa e più ricca di quella fatta da un lessicografo o da un grammatico esperto? Potrà sostituirlo? potrà batterlo, come il computer ormai batte il più esperto rivale in una partita a scacchi? Questi primi balbettamenti preludono alla costituzione di una grammatica che, grazie a un procedimento di scoperta, si farà da sé? Sinclair (1991) preconizzava questo momento. Ma dai primi risultati che ci presenta Rema Rossini Favretti, dagli esempi suggeriti dalle analisi di Biber 1993a (in Mc Enery e Wilson 1996 75–76), non oseremmo davvero pensarlo. Più felici sono indubbiamente le analisi di *decline*, *yield*, *set*, *of, second* e *back* in Sinclair 1991, di *to budge* in Sinclair 1997. Ma qui è la indubbia abilità dello studioso nel tirare le fila che salva la situazione, studioso che ammette che da un certo punto in là bisogna procedere „largely on a subjective basis" (Sinclair, 1991, 106).

Ma la linguistica del corpus non ci aveva promesso di fare il più possibile a meno del fattore umano?

Prima di chiudere questa rassegna, dobbiamo dedicare qualche parola all'approccio *quantitativo* della linguistica computazionale, approccio che viene opposto orgogliosamente a quello meramente qualitativo delle varie

linguistiche teoriche „tradizionali". Noto, in primo luogo, che la *linguistica del corpus* sembra avere ormai dimenticato del tutto il lavoro prezioso, seppur limitato, svolto dalla statistica linguista degli Anni Sessanta, di cui ho già lamentato, in altre occasioni, la precoce messa in oblio. L'acquisizione principale era stata La legge di Zipf, universale, che mette in rapporto inverso la frequenza di una parola, il suo rango e la sua lunghezza. Era stato possibile stabilire per ogni lingua il lessico di base, un elemento indispensabile per l'insegnamento delle lingue straniere, per la preparazione di grammatiche e metodi di insegnamento, ecc. Suggestive, anche se laboriose statistiche lessicali, erano state messo al servizio della stilistica, alla ricerca dello „scarto" tra lo stile di un autore, o, meglio, di un genere letterario rispetto alla media linguistica. Queste osservazioni si leggevano e si possono leggere ancora con utilità nelle opere di G.K. Zipf ancora del 1935 e dell'altro americano Mandelbrot, nei libri dei francesi Pierre Guiraud e Charles Muller, dell'inglese Gustav Herdan, nel romeno Salomon Marcus (tramite anche per l'Italia di diversi autori russi). Questo genere di ricerche, legato nella impostazione originaria americana, alla psicologia empirista e al comportamentismo, è caduto sotto i colpi della rivoluzione di Chomsky. Questa si apre, ricordiamolo, proprio con una critica serrata al versante linguistico dell'opera B.F.Skinner, il più influente psicologo comportamentista americano (Chomsky recensiva polemicamente *Verbal Behaviour* (1957) di Skinner nel 1959).

Tuttavia il patrimonio costituito da questi studi non è andato del tutto dimenticato. In Europa la sua incompatibilità con la nascente grammatica generativa non veniva in genere drammatizzata, anzi erano sono spesso gli stessi studiosi, quelli capaci di concepire la linguistica fuori dal quadro più tradizionalmente umanistico e „filologico", a mostrare un pari interesse per questi diversi approcci. Così alcuni principi della linguistica quantitativa di quegli anni sono in realtà definitivamente acquisito a diversi rami della linguistica applicata. Da noi, Tullio De Mauro ne ha rappresentato una versione etorodossa ma estremente suggestiva, nel suo carattere militante, nella sua *Guida all'uso delle parole* (1980), e anche il recente *LIP (Lessico dell'italiano parlato*, con Federico Mancini, Massimo Vedovelli e Miriam Voghera, 1993) sarebbe inconcepibile senza lo sfondo degli studi citati prima.

Ma veniamo alla dimensione quantitativa così come è concepita oggi. Prima di tutto la recente *linguistica del corpus* rivendica le sue potenzialità per il fatto di poter disporre, tramite le tecniche informatiche moderne, di una massa quantitativamente molto superiore al passato: i moderni *corpora* consistono di milioni di parole rispetto alle poche migliaia di quelli del passato. Quanto gli obiettivi, quello che la recente *linguistica del corpus* sembra proporsi è effettivamente tutt'altra cosa dagli obiettivi della

vecchia statistica linguistica. Prendiamo come esempio la rassegna di studi informatici sull'anafora di Biber riportata in Botley e McEnery (1996, particolarmente 29–33). I rapporti anaforici tra l'elemento detto *antecedente* e la *ripresa* vengono prima identificati attraverso etichettatura (mista, dato che un'etichettatura solo automatica sarebbe insufficiente). Si individuano così le diverse categorie grammaticali che possono costituire antecedente e ripresa: per es. quest'ultima può consistere di un pronome personale, o di un pronome dimostrativo, ecc. Poi si contano i diversi tipi, se ne stabilisce la consistenza statistica nel *corpus*. Si conta la *distanza* media che separa l'antecedente dalla ripresa. Possiamo fermarci qui, e dare il nostro parere. Se i primi calcoli mi sembravano probabilmente superflui (anche se non posso escludere del tutto che un bravo linguista sappia tirarne frutto), l'ultimo mi pare francamente assurdo: la distanza tra antecedente e ripresa non può essere una distanza lineare, calcolabile in numero di parole. E questo perché la struttura della frase ha un'architettura sintattica, e, se parliamo di distanza, questa non può essere calcolata in linea d'aria passando sopra mari e montagne. E' da un pezzo che sappiamo che si deve andare oltre la linearità del significante. Gli stessi linguisti del *corpus*, visto che ammettono l'analisi in costituenti immediati, ammettono a quanto pare che la linearità nasconde una gerarchia. Se troveremo una distanza media, questa sarà una media insignificante, perché ignora i fattori che favoriscono o impediscono il rapporto anaforico. In conclusione, un approccio statistico di questo tipo mi sembra o irrilevante o sbagliato.

Per concludere, quale sarebbe, secondo i suoi rappresentanti, lo status della *linguistica del corpus*? secondo le affermazioni di alcuni studiosi la *linguistica del corpus* sarebbe una metodologia *(methodology)*, avrebbe cioè uno status inferiore a quello di teoria *(theory)*. Lo stesso status viene sottinteso, direi, dal più preciso termine *approach*, nella misura in cui, per es, sarebbe possibile affrontare un problema linguistico, per es. quello semantico e sintattico della anafora, per via informatica assumendo un quadro teorico esterno (come quelli per es. della linguistica di Halliday), sempre secondo una proposta di Botley e McEnery (1996).

Ma per alcuni autori, a quanto pare, la *linguistica del corpus* si può già candidare ad essere una vera e propria „teoria". „Teoria" è, dagli anni Sessanta, il termine che dà piena dignità a un indirizzo di studio. Molti studiosi sottolineano invece il carattere iniziale della *linguistica del corpus*, e questo richiamo è destinato a giustificare preventivamente i difetti delle ricerche. Tuttavia, anche con queste riserve, più di uno studioso ritiene che la *linguistica del corpus* sia già in grado di portare una sfida al più importante rivale, la grammatica generativa. Secondo Geoffry Leech (1992, cfr. Botley Mc Enery 1999, 23–25), la *linguistica del corpus* si opporrebbe alla grammatica generativa perché si basa sulla *performance* e non sulla *compe-*

tence; b) perché si pone scopi descrittivi anziché esplicativi; c) perché mira a una visione quantitativa anziché qualitativa dei fatti del linguaggio; d) perché di ispirazione empirista anziché razionalista.

Si potrebbe aggiungere: e) che la *linguistica del corpus* si applica a una sola lingua, anziché mirare a stabilire degli universali linguistici.

Si tratta in realtà, a mio parere, tranne che per il punto d), più di arretramenti, di ridimensionamenti degli obbiettivi, che di veri e propri obiettivi alternativi. Vediamo punto per punto (escludendo per il momento l'ultimo punto, la cui discussione ci porterebbe lontani dal nostro tema):

(a) la *competence* si estrae, non sempre agevolmente, dal mare magnum, dal magma, della *performance*. Rinunciare a questa operazione può essere un sollievo, ma momentaneo, perché prima o dopo il caos è destinato ad aggredirci, e avergli fatto buon viso prima non ci esimerà dal doverlo affrontare dopo.

(b) la descrizione è la prima fase alla quale può seguire la seconda fase, più difficile e talvolta ardua o, qualche volta, addirittura inaccessibile, la spiegazione: cosa di più comodo di dichiarare che la spiegazione è inutile?

(c) l'ideale sarebbe che a una visione qualitativa corrispondessero realtà quantitative diverse: la vecchia statistica linguistica, che ho ricordato prima, aveva ottenuto qualche successo, pur limitato, su questo piano, quando per es. aveva stabilito che c'è un rapporto definibile matematicamente tra la lunghezza di una parola, la sua frequenza e il suo contenuto di informazione (legge di Zipf). Ora, che senso ha mettere l'accento su aspetti quantitativi *anziché* qualitativi? Quello che avrebbe senso sarebbe portare avanti il limite posto da Zipf, stabilire cioè altri nessi tra i molti principi qualitativi che la migliore ricerca linguistica ha messo in luce finora, e le poche regolarità statistiche stabilite. Non ha senso invece stabilire un'opposizione tra le due, né tanto meno proclamare una superiorità dell'elemento quantitativo sul qualitativo.

(d) la questione dell'empirismo e del razionalismo è più complessa. Nel I capitolo della loro *Corpus* Linguistics, Mc Enery e Wilson discutono in modo polemico, ma anche, bisogna riconoscerlo, con una certa ricerca di equilibro, *what Chomsky said*, a partire da *Syntactic Structures* del 1957, e le prospettive che, a loro parere, si aprono con la rivoluzione metodologica proposta dalla c*orpus linguistics*. Da questo appassionato confronto, sarebbe impossibile ricavare, a mio parere, che il quasi mezzo secolo passato abbia portato a rivedere in modo decisivo alcuni dei postulati di Chomsky. In particolare, non ve-

> do perché dovremmo mettere in questione il fatto che il linguaggio umano (in tutte le sue forme storiche), faccia un uso infinito di mezzi finiti. Così è, e ne consegue che il più gigantesco dei *corpora* è destinato a rappresentarlo in modo insufficiente, come abbiamo già visto. Il che non vuol dire che i *corpora* siano inutili, come vedremo dopo. Ma l'introspezione resta il modo privilegiato di interrogare i dati della nostra lingua: aggiungiamo che è *necessaria* per discernere nel caos della *performance* e per estrarne ciò che è precisamente *competence*, o, come diceva Saussure, *langue*. La *linguistica del corpus* non ci farà tornare indietro a prima di Chomsky, che sarebbe in realtà un retrocedere alle spalle di Chomsky, di Saussure e forse di Dionisio Trace e Apollonio Discolo verso il caos primitivo.

Aggiungo che lo studio delle lingue straniere, o delle lingue del passato, riesce nella misura in cui riusciamo a formarci una competenza sui generis di quelle lingue, una competenza imperfetta, ma pur sempre un dominio sui dati bruti, infinitamente superiore di quanto sarebbe un'impossibile acquisizione di dati della performance. Ci chiediamo: quali sono veramente le possibilità di un corpus, anche esteso, estesissimo, di sostituirsi alla *competence*? Vediamo un esempio. Nonostante sia composto da 50 milioni di parole, il corpus che sta alla base di *Wörterbuch der italienischen Verben*, di Peter Blumenthal e Giovanni Rovere (Stuttgart, Klett, 1998), opera, sia detto a scanso di equivoci, quanto mai meritoria e che non è un esempio di linguistica del corpus, non contiene nessun esempio di *ringraziare per qualcosa*, ma solo di *ringraziare di qualcosa*. Qualsiasi parlante italiano, a meno che non sia afflitto da amnesia, sa, grazie alla sua competenza, che si dice: *Ti ringrazio di avermi aiutato* ma anche *Ti ringrazio per avermi aiutato*. Ma un corpus di 50 milioni può mancare di questa seconda forma. Se ne conclude che anche il corpus più esteso non può gareggiare, nemmeno nella completezza dei dati, con la competenza di un nativo.

Tutto ciò non vuol dire che un eccesso di introspezione non abbia portato nel generativismo a casi curiosi di solipsismo (ricordiamo le affermazioni precedute dall'espressione: *In my dialect…)*. Ma non si può certo giudicare una teoria dalle sue degenerazioni.

Tornando al nostro soggetto, e concludendo, il problema è se, accanto all'introspezione, che resta la prima fonte di dati, i *corpora* possano svolgere un ruolo positivo. La nostra riposta è sì. Tutto sta nel modo in cui sapremo interrogare i *corpora* per estrarne dei dati, e come sapremo servirci di questi. E' sbagliato invece voler attribuire ai *corpora* compiti ai

quali questi non possono assolvere. Le operazioni proposte dal *linguistica del corpus* o sono utili ma modeste, o ambiziose ma irrealizzabili.

Nelle pagine che seguono ricorderò più di una volta che, anche prima che Chomsky la teorizzasse e la mettesse al centro della ricerca linguistica, il ruolo centrale nella ricerca linguistica era svolto dall'introspezione. Questa, oltre a fornirci i dati linguistici, ci fornisce degli elementi essenziali per la costituzione di una grammatica: ci dice se questa o quella frase è grammaticale o no, se due frasi siano sinonime, se due parole sono omonime, ecc. ecc. Rinunciare a queste operazioni, o introdurle surrettiziamente, come sembra pretendere la *linguistica del corpus* con la sua rinuncia al „razionalismo" a favore dell'„empirismo", non può essere di nessuna utilità.

CORPORA DI LINGUE ANTICHE

E' il momento di dire qualcosa dei *corpora* di lingue antiche e dei *corpora* che offrono dati storici su lingue vive

Diciamo subito che la presenza di *corpora* simili ci sembra aprire prospettive nuove e importanti negli studi, ma non per le ragioni che vengono presentate dalla *linguistica del corpus*. Come vedremo, i *corpora* elettronici ci offrono oggi una quantità di dati raccolti insieme come mai è stato in passato, quando lo studioso poteva far ricorso unicamente ai testi e agli studi che gli offrivano le biblioteche e alla propria memoria: tutte cose che restano peraltro indispensabile anche oggi. I sistemi di interrogazione, che il ricercatore deve saper usare in modo ingegnoso, sono un formidabile aiuto alla ricerca. Ma, come credo di avere mostrato, non c'è niente di nuovo e di buono nel metodo, o nella teoria che ispira la ricerca, che possa venire dall'uso di *corpora* elettronici. Anzi ne possono venire solo dei difetti. Questi sono principalmente due: la tendenza a sostituire l'interrogazione dei testi alla loro conoscenza diretta, e l'uso precoce della statistica. Nel primo caso, la pigrizia che induce a preferire la ricerca dei dati nel *corpus* automatico anziché attraverso una conoscenza diretta, esimendosi alla fine da questa, è stata già denunciata da Rissanen (1989 cit. in McEnery e Wilson 1996108, 108). Mc Enery e Wilson obiettano non senza ragione che si tratta di un caso di cattivo uso dell'informatica e non di un difetto della *linguistica del corpus:* è vero, ma il pericolo esiste. Nel secondo caso, il discorso è più delicato: lo studio accurato dei dati porta a raggrupparli e suddividerli. Ma più i raggruppamenti si fanno precisi e minuti, e quindi più numerosi, più diventa difficile, come ha notato di nuovo Rissanen (1989), che ci siano le condizioni necessarie per delle in-

dagini statistiche. Perché queste siano possibili, infatti, è necessario che i campioni indagati siano statisticamente rappresentativi, e abbiano quindi una certa consistenza numerica. Lo studioso cosciente di questo rischio dovrà quindi rinunciare in molti casi a classificazioni e raggruppamenti troppo fini, oppure rinunciare alla quantificazione. In realtà, a mio parere, l'effettiva necessità di tali statistiche è tutta da dimostrare, e temo che in molti casi l'insistenza sulle statistiche di molti rappresentanti della linguistica del *corpus* dipenda dal fatto che queste sono facilmente acquisibili attraverso gli strumenti informatici, mentre altri dati, che sarebbero più importanti, non lo sono. Nello studio della lingua la superiorità del quantitativo sul qualitativo non è stata affatto dimostrata, fino ad oggi, e tantomeno è stata dimostrata la possibilità di trasformare il qualitativo in quantitativo. Si può deprecare questo stato di cose, ma è un fatto che lo stato della ricerca in linguistica è molto diverso da quello di altre scienze che si sono costituite come tali acquisendo uno status matematico. Forse questo, sia detto tra parentesi, dovrebbe far dubitare della definizione della linguistica come scienza naturale, proposto, questa volta, proprio da Chomsky, del quale non vogliamo essere dei seguaci pedissequi e degli ammiratori incondizionati (ma questa è un'altra storia).

IL RUOLO DEL CORPUS NEL PROGETTO ITALANT

Il progetto Italant ha lavorato finora su un *corpus* prevalentemente non etichettato *(plain or row corpus)*, anche se in parte lemmatizzato, come abbiamo già ricordato, sempre a cura dell'OVI. Inoltre un Gruppo di Torino rappresentato da Carla Marello e Manuel Barbera sta predisponendo, nel quadro della stessa ricerca cofinanziata, l'etichettatura del *Padua Corpus*, trasformandola in *Corpus Taurinense*. Le parti non ancora ultimate dell'impresa potranno usufruire, credo, di questo *corpus* etichettato grammaticalmente e trarne grande vantaggio.

Come abbiamo già detto, *tagging* e lemmatizzazione possono comportare degli errori, cosicché lo studioso non potrà di norma esimersi dal controllare il lavoro già fatto. Ma anche se sarà così, l'aiuto sarà stato grandissimo. Qualche volta dovrà rifare il lavoro perché quello che gli interessa è stato ignorato per qualche ragione (svista, ignoranza del fenomeno, suo peso eccessivo, ecc.) da chi ha elaborato i dati. Inoltre lo studioso avrà bisogno spesso di servirsi di categorie grammaticali di grana più fine di quelle di cui è stato provvisto il *corpus*: il *tagging* non prevede, per es., di notare i verbi inaccusativi, i pronomi espletivi, ecc. ecc. Nonostante tutto questo, non si può certo sottovalutare il vantaggio di

possedere un *corpus* etichettato, che ci permette di avere a disposizione alcuni risultati bell'e pronti, almeno per i fenomeni più chiari ed univoci.

Il lavoro vero e proprio del linguista resta da fare.

Un esempio. Supponiamo di avere il nostro corpus annotato per funzioni grammaticali. Giampaolo Salvi, nel suo capitolo in preparazione sull'Impersonale, nota che in it. ant. il tipo di impersonale rappresentato da una frase come: *là si trovava sempre più ribaldi che in niun'altra terra* (Novellino, 85.3) non va interpretato come in it. mod., dove *ribaldi* sarebbe oggetto, ma come un esempio di mancato accordo del soggetto postverbale delle costruzioni inaccusative. Questo lavoro può essere piuttosto ostacolato che agevolato da un *parsing* già condotto sul *corpus* in cui *ribaldi* sia dato come oggetto. Lo studioso, infondo, non deve chiedere al *corpus* e al suo sistema di interrogazione altro che gli risparmi la fatica di cercarsi da sé tutti i *si* e, possibilmente, di dividere in due gruppi il *si* impersonale dal riflessivo. Poi dovrà lavorare da solo.

Aggiungiamo che l'etichettatura *(tagging)* non affronta, perché non lo può fare, problemi morfologici: *prefissi*, *suffissi* e tanto più fatti di *morfologia flessiva* sono al di fuori della sua portata. Non c'è manuale di linguistica che non avverta il principiante che la „parola" non è un primum linguistico, ma solo una prima approssimazione morfologica, sintattica e semantica. Ora il *tagging* classifica proprio „parole", con tutta la grossolanità di questo concetto.

In questa ottica, il *corpus* più utile al linguista sarà quello fornito di un buon sistema di interrogazione, comprendente la ricerca di singole di parole grafiche, di sequenze di due o più parole, anche non adiacenti (cosiddette cooccorrenze), che fornisca contesti adeguati (allargabili su richiesta a porzioni più vaste). Il programma *Gatto*, predisposto per un tipo di ricerca sintattico, rende possibile anche l'interrogazione del punto interrogativo e dell'esclamativo, che in genere fanno parte invece dei segni che servono all'interrogazione *(wildcard characters):* lo scopo di questa estensione è evidentemente quello di facilitare la ricerca sull'interrogazione, l'esclamazione e altri fenomeni connessi. Su questo fronte ci sono stati inizialmente dei problemi, ma l'ultima versione li ha risolti. Serve poi che il corpus contenga referenze precise ai vari testi e all'edizione utilizzata, alla loro data e localizzazione. Tutte queste caratteristiche si ritrovano nell'OVI.

Ma una volta dati un corpus e un buon sistema di interrogazione, le armi più formidabili per la ricerca sono l'ingegnosità dello studioso nell'interrogare il corpus e la sua capacità di sfuggire alle trappole e alle tentazioni dello strumento.

IL NOSTRO LAVORO

Come interrogare un *corpus* in vista della descrizione morfologica e sintattica?

Un fenomeno sintattico può venire interrogato attraverso le sue manifestazioni visibili di carattere lessicale. Così per es. l'interrogazione viene studiata a partire dagli introduttori interrogativi *(chi*, *come*, *perché* ecc.): l'identificazione di questi elementi resta un compito del ricercatore, che deve avere delle conoscenze previe della lingua che sta indagando, nel nostro caso l'italiano antico, o deve farsele nel frattempo (o altrimenti si dimenticherà, per es., di *chente* „quale"). E per questo non c'è che leggere e studiare i testi. Lo studio, morfologico, sintattico, semantico, dell'articolo definito si baserà sull'osservazione dei contesti in cui si trovano *lo, il, 'l, el, li, gli, i, le* ed alcune altre forme (un problema è posto dalle delle preposizioni articolate, scritte unite o separate). Un'eventuale lemmatizzazione dovrà raccolto utilmente queste forme sotto una singola voce, che dovrebbe essere ancora *lo*, ma chi usa la lemmatizzatizzazione non sarà giustificato se trascurerà alcune forme o esempi che il lemmatizzatore ha dimenticato: la responsabilità, in questo e in altri casi simili, resta naturalmente sua.

Andiamo avanti: per chi studia l'articolo la sua assenza in un Sintagma Nominale è altrettanto interessante della sua presenza. Naturalmente l'assenza come tale non può essere etichettata. Ma per studiare l'assenza dell'articolo, o articolo zero, si potrà far ricorso all'etichettatura dei nomi e degli aggettivi che, in quanto primi elementi di un SN, dovrebbero essere preceduti dall'articolo. Se invece ne sono privi, si deve cercare di capire perché. Ma devo dire che il numero dei SN in una sola pagina e in ogni riga di un testo è talmente alto, e i casi sia di presenza che di assenza di articolo sono così numerosi, che non vale la pena di ricorrere a questo espediente: basta scorrere un testo e gli esempi si presentano naturalmente in folla. Il difficile è capire la logica dell'assenza di articolo, che dipende da fattori diversi che devono esser rigorosamente distinti gli uni dagli altri. E a questo l'etichettatura servo ben poco, servono invece ipotesi teoriche.

Mettiamo ora il caso, proseguendo i sondaggi sull'articolo, di volere studiare il partitivo. Qui anzi dobbiamo stabilire se l'italiano antico aveva la possibilità di esprimere il partitivo come fa l'ital.mod. quando dice Dei *bambini giocavano, Ho visto* dei *bambini* o *ho spalmato* della *marmellata sul pane,* ecc., e cioè utilizzando *di* + art.def. rispettivamente in posizione di soggetto (come nel primo caso) e di oggetto (nel secondo e nel terzo). Sta-

bilito che di occorrenze simili non ce ne sono nel Duecento, bisognerà aver l'idea di cercare di senza preposizione, come si trova in toscano mod. e nel Manzoni *(si videro di gran novità…)* (Rohlfs, *Morfologia,* par. 424): è inutile notare che non ci sono procedimenti meccanici che possano aiutarci, si tratta solo di avere l'idea (un'idea che ci può essere tutt'al più suggerita da conoscenze previe sulla storia della lingua italiana, o dall'analogia con il francese o da un'altra fonte estrinseca di ispirazione- o infine dalla lettura attenta dei testi italiani antichi). Questa volta nella gran folla dei di soccorrono subito alcuni esempi. Una parte di questi si rivelano illusori: per es. *in tant'ha di segnoria* (Chiaro Davanzati, canz. 30, p.112) *di segnoria dipenderera da tanto: tanto di segnoria*. Ma altri casi sono senz'altro buoni, come per es. *tu hai di belle femmine* (Novell. ed. Favati, 36, p.211). Ma bisognerà chiedersi: si tratta di un vero partitivo, o non sarà piuttosto che il verbo avere può reggere un sintagma introdotto da di? e cioè: in ital.ant. il verbo avere poteva reggere accanto alla reggenza di un oggetto anche quella di di +SN? Questo problema si pone per tutti i casi di di retto da verbi come dire, dare, fare, chiedere, prendere e forse qualcun altro oltre ad avere. Dal dubbio si uscirebbe se ci fossero casi di soggetti partitivi, ma nel Duecento non ce ne sono. Ne appariranno più tardi, e nel Quattrocento (quando la preposizione appare articolata), l'uso del partitivo nel soggetto sarà largamente documentato anche se limitato al caso di soggetto di costruzioni inaccusative. Questa restrizione è significativa, perché ci fa vedere che storicamente il partitivo si sviluppa a partire dall'oggetto, e se si estende poi al soggetto lo fa passando per quella posizione intermedia che è quella del soggetto inaccusativo, che unisce proprietà dell'oggetto e del soggetto. Non c'è dubbio che i casi duecenteschi come quello citato dal *Novellino* rappresentano il primo nucleo da cui si sviluppa il partitivo moderno, ma resta da decidere se, nella sincronia duecentesca, si debba parlare già di partitivo o si sia ancora nell'ambito delle reggenze dei verbi. Questa è la logica che sto seguendo nella ricerca che ho in corso per il capitolo sull'*Articolo* per la Grammatica dell'Italiano Antico, una logica abbastanza complessa, si vede.

La descrizione degli espedienti da usare nell'interrogazione del *corpus*, e del limite che anche questi espedienti hanno, sono, credo, la prova migliore del fatto che il nostro lavoro linguistico si pone al di fuori di quella che viene comunemente chiamata *linguistica del corpus*. La gran parte del lavoro, e tutto il lavoro che ha una qualche rilevanza *linguistica* in senso forte, è fatta dallo studioso e non dalla macchina.

Dopo avere illustrato brevemente la tecnica di lavoro sul *corpus*, vorrei adesso mostrare a quali pericoli sia soggetta. Preciso che si tratta di pericoli, e non di difetti insiti nel testo, come abbiamo già visto a proposito delle osservazioni di Rissanen: accettiamo per una volta le contro-

obiezioni di McEnery e Wilson. Il pericolo è che, invece della ricerca di generalità (regole, sottoregole, ecc.), lo studioso si limiti a commentare uno per uno agli esempi presenti nel corpus.

Non credo che una tale metodologia di lavoro sia stata mai teorizzata, ma di fatto si riscontra spesso in lavori condotti sul *corpus* anziché sull'introspezione. La mia idea è che, anche in una grammatica basata su un corpus, si debba tendere a dare lo stesso genere di studio grammaticale che si darebbe in una grammatica basata sull'introspezione. Nonostante le limitazioni della documentazione, nonostante sia impossibile eseguire dei *test* di grammaticalità ed elicitare dati come si fa da una lingua viva, miriamo ad avvicinarci allo stesso ideale descrittivo e esplicativo che ci guida nello studio delle lingue vive. E, almeno in molti casi, questo è possibile.

Quali sono le caratteristiche di una lingua viva? Una lingua viva possiede *regole discrete*, per cui tale o tale forma o costruzione è *grammaticale* o *agrammaticale*, distinzione qualitativamente diversa da quella di *attestato* o *non-attestato*.

Si potrebbe pensare che ci siano due vie e due metodologie diverse: quella propriamente linguistica basata sulla prima delle due dicotomie date sopra *(grammaticale: agrammaticale)*, e quella filologica basata sulla seconda *(attestato: non-attestato)* Ma si farebbe torto alla filologia, perché già la buona filologia ottocentesca, invece di accontentarsi della registrazione di forme documentate, aveva ricavato dallo studio dei testi alcuni regole, o leggi, *Gesetze* come si diceva allora in tedesco, che era allora la lingua della scienza. Queste regole avevano carattere predittivo, come la legge del Gröber, del 1877 sulla forma dell'articolo italiano, o la legge Tobler-Mussafia, formulata per l'italiano dal secondo dei due studiosi nel 1886, sulla posizione dei pronomi clitici rispetto al verbo, leggi che riguardano proprio l'italiano antico. Queste ed altre regole delle lingue antiche, regole di carattere fonologico, morfologico o sintattico, stabilite da studiosi dell'Otto- e del Novecento, non erano diverse nella forma da quelle che si stabiliscono nelle lingue vive. In chiave moderna, generativista, diciamo per es. che le regole dell'italiano antico dovevono sottostare a quelle restrizioni che si sono provate come universali, o come altamente probabili. Questo criterio, che si potrebbe chiamare di *verosimiglianza*, deve guidarci nello stabilire la grammatica di una lingua antica.

Si può anche porre la questione: si deve ammettere la *variazione* in italiano antico? Certamente, perché la variazione è stata osservata in tutte le lingue vive. Con *variazione* intendo il fatto che in una lingua ci sono certe parti che presentano trattamenti alternativi, uno dei quali, in genere, appare connesso a un registro più alto, ed è in genere conservativa, l'altra è un'innovazione. Questa variazione è, a mio modo di vedere, niente

meno che la chiave del cambiamento linguistico (voglio dire del modo in cui avviene il cambiamento linguistico, come ho cercato di mostrare in un mio lavoro recente (Renzi 2000)).

Come esempio di variazione linguistica, porto il caso delle forme in italiano antico del pronome deittico di III persona, in it.mod. *quello*. Ci interessa la forma antica del soggetto singolare, che è *quelli*, che è in alternanza fonologica con *quegli*, *quei*, così come nel pronome pers. *elli* alternava con *egli*, *ei*, *e'*. La variazione di cui vogliamo parlare è quella che riguarda la funzione sintattica dei pronomi.

In fior. ant. abbiamo al nominativo sing. *quelli (quegli*, *quei)*, per es.:

> ben è ragione che nullo omo mi pianga,
> ch'io sono ben come *quei* che si vide
> ne l'agua infino a' denti (Chiaro Davanzati, Canz. 9, vv.15–17, p.39).

L'obliquo, cioè la forma usata per l'accusativo e gli altri casi diversi da quello del soggetto era *quello*, *quel*. Ora abbiamo casi in cui questa forma occorre dove ci saremmo aspettato la forma del soggetto *quelli*:

> per ch'un nasce Solone e altro Serse
> altro Melchisedec e altro *quello*
> che, volando per l'aere, il figlio perse
> (Dante, Par. VIII, 124–126)

dove *quello* (che designa Dedalo, che perse il figlio Icaro) è in posizione predicativa del soggetto. *Quello,* soggetto**,** regge il verbo, ma non da posizione adiacente in Brunetto Latininell'esempio seguente:

> …che quello che avea loquenzia congiunta con sapienzia, avenìa che, per giudicio di moltitudine di gente e di sé medesimo, paresse essere degno di reggiere le publiche cose. (Rettorica p. 30)

Infine ci sono anche esempi di *quello* in posizione canonica:

> …*quello* è nimico di sè medesimo, che prolunga la vita al suo nimico (Tesoro di Brunetto Latini volgarizzato da Bono Giamboni, p. C368)

Ci sono d'altra parte anche dei casi in cui la forma in *-i* è oggetto, come in Dante Inf.II, 104:

> Che non soccorri *quei* che t'amò tanto?

analogamente a quanto avviene con il dimostrativo di I pers.: Inf. I, 103: *Questi* (oggetto) non ciberà terra né peltro (soggetto)…

Qual era allora il sistema del pronome dimostrativo di 3.a persona? Tenendo conto della grande massa dei dati, e considerando che l'origine, per analogia, della forma in *-i* è il sing. lat. *qui(s)* (Renzi 1998), il sistema doveva essere questo:

Sistema A	+Umano	-Umano	
	quegli	quello	SOGG.
	quello	quello	OBLIQUO

Ma c'erano, come abbiamo visto, delle eccezioni. Cosa vuol dire delle eccezioni? Vuol dire che premeva alle porte un altro sistema, che è semplicemente questo:

Sistema B quello SOGG. e OBLIQUO

in cui, semplicemente, *quegli* sparisce e c'è una sola forma per soggetto e obliquo, cioè in realtà questa distinzione non c'è più. E siccome era la forma del soggetto che distingueva l'umano dal neutro, anche questa distinzione cade: il sistema si semplifica radicalemente. Questi cambiamenti, già in atto nel Duecento, portano direttamente al sistema moderno, in cui *quegli* non c'è più, o, se si preferisce è un innocuo arcaismo: la forma *quegli* è oggi relegata al margine alto della lingua, là dove le forme agonizzanti spesso di rifugiano, nelle penne dei letterati (cfr. Renzi 1998).

Questo tema è morfologico, ma in realtà ha implicazioni sintattiche perché riguarda lo status e i limiti dei casi in italiano antico. Nel mio articolo del 1998, appena citato, ho dimostrato che, nonostante la loro origine etimologica, un sistema a tre per cui:

quegli	SOGG:
colui	GENIT/DAT
quello	OGG:

l'assetto casuale, da quando abbiamo documentazione, è stato sempre in italiano, del tipo a due casi:

SOGGETTO contro OBLIQUO (cioè GENIT/DAT + OGG)

E cioè nel nostro caso:

Quegli SOGG. contro *colui* e *quello* OBLIQUO

in attesa che, come abbiamo visto sopra, anche questa distinzione sia obliterata (anche *colui* verrà usato non solo come obliquo ma anche come soggetto).

Torno al tema dei rapporti tra attestato e possibile. Bisogna reagire alla tentazione semplicistica di tracciare un'equazione:

non attestato = *impossibile*
attestato = *possibile*

Vanno infatti aggiunte altre due possibilità:

non attestato ma *possibile*
attestato ma *impossibile*

Quest'ultimo caso può sembrare il più improbabile, o paradossale, ma si legga quello che scriveva un linguista storico come Lucian Foulet nel lontano 1919 (3.a ed. 1958, par.53), quando notava che in francese antico l'ordine Complemento-Soggetto-Verbo è si attestato qualche volta „mais il viole une des règles les plus solidement établies de la syntaxe de l'ancien français", quella dell'„inversione" tra soggetto e verbo, per cui dopo un Complemento ci aspettiamo Verbo-Soggetto e non Soggetto-Verbo (oggi si parlerebbe di *Verb-Second)*. Vediamo che qui un filologo dell'inizio del Novecento esprime impavidamente il parere che ci siano esempi che violano delle regole. E ha ragione.

Certo resta il problema dei casi attestati ma teoricamente impossibili. Come risolverlo? Ci sono varie possibilità: alcuni esempi possono essere reinterpretati, per es. in un caso come quello appena dato può darsi che la sequenza sia interrotta: il Complemento è fuori frase, segue la coppia normale Soggetto-Verbo. O la frase può essere respinta: i testi scritti, come la lingua parlata contengono degli errori, come tutti i filologi e tutti i linguisti sanno.[4] Questo con buona pace della linguistica del *corpus* che si propone di privilegiare la *performance* alla *competence*, impegnandosi così, come ho già detto, in una lotta con i dati bruti che non potrà non perdere.

Quanto all'ultimo caso: *non attestato ma possibile*, mi limiterò a un caso molto semplice, quello dei pronomi (personali liberi e clitici, dimostrativi, ecc.) e del genere e numero. Dato un corpus sufficientemente vario, come è quello di Italant, la 1.a, 2.a e 3.a persona sono tutte sufficientemente rappresentate (favorite le prime due dalla lirica, la 3.a. dalle cronache e dalle opere narrative), ma quanto al genere e al numero la situazione è diversa. In generale il Singolare è più frequente del Plurale, e in generale, e in italiano in particolare, il Maschile è più frequente del Femminile. E' possibile, per es., per fare un esempio *fictum*, ma chiaro, che non troviamo documentate alcune forme in certi contesti, per es. che io abbia *trovatolo*, *trovatala*, *trovatili* ma non *trovatele*. Chiedersi, sapendo che tutti gli altri pronomi clitici seguono il participio passato, se per caso quello femminile di III plurale possa fare eccezione, non ha alcun senso. In mancanza di indizi negativi, non possiamo che estendere la regola anche al caso che ci manca, anche se lo scrupolo ci porterà a notare che c'è una lacuna nella documentazione di una forma.

Giunti a questo punto, è a portata di mano la conclusione alla quale miro: una grammatica di una lingua antica basata su questi criteri, potrà portare la sua somiglianza con la grammatica di una lingua moderna fino al punto di fare uso dell'asterisco, cioè di stabilire la agrammaticalità di questa o quella forma o costruzione. Italant ha già degli esempi di questo

[4] Vedi la mia breve nota: Renzi 1993.

genere. Nel Cap. sui verbi *essere* e *avere* „presentativi", Giampaolo Salvi nota che in ital. ant. questi verbi non sono accompagnati dai clitici *vi*, *ci:*

Dinanzi alla casa *avea* una fossa (Novellino, 38.10)

Se *vi, ci* sono presenti, hanno valore anaforico, cioè si riferiscono a un luogo già citato nel testo precedente, come nell'esempio seguente:

„…*sopra capo* di quel Signore, che ha?" […] „*Havi* un capello" (Novellino, 29.10)

È una situazione, scrive Salvi, „diversa dall' it. mod. in cui il *ci* del verbo presentativo, non essendo un complemento di luogo, ma solo un indicatore del valore presentativo, può cooccorere con un complemento di luogo: *In casa di Mario ci sono molti scarafaggi / Ci sono molti scarafaggi in casa di Mario.* In it. ant., questo non avviene mai perché è *vi/ci* che esprime il complemento di luogo". Se riprendiamo ora l'esempio citato sopra possiamo concludere che in ital.ant. non potremmo avere:

* Sopra capo vi ha un capello

Infatti, se *sopra capo* fosse fuori dalla frase, emarginato all'inizio, per effetto della legge Tobler-Mussafia avremmo l'enclisi della particella clitica al verbo *(havvi*, o, come nell'esempio sopra *havi)*. Allora *sopra capo* dovrà essere interno alla frase, ma allora *vi* non può essere presentativo, come abbiamo stabilito indipendentemente, né anaforico riferendosi a un locativo contenuto nella stessa frase. Quindi la frase è agrammaticale. Abbiamo perciò il diritto di asteriscarla, anche se non possiamo fare ricorso, come nel caso di una lingua viva, alla nostra introspezione o a quella di un informatore.

Possiamo anche stabilire possibilità alternative, visto che, come abbiamo già detto, il fiorentino antico doveva avere al suo interno delle variazioni, come ogni altra lingua. L'esempio riguarda l'ordine delle parole, e questa volta non è limitato al toscano. In tutte la varietà italiane il verbo aveva capacità pronominale, nel senso che era sufficiente la morfologia verbale a indicare la persona che fungeva da soggetto, ma questa capacità si esercitava solo quando il verbo precedeva il soggetto (Vanelli, Renzi, Benincà (1985), ora in Benincà 1994). Di qui viene che nel caso della sequenza VS, se S era rappresentato da un pronome, il soggetto poteva apparire o no: in fior. possiamo avere *volete voi*… (Brunetto, Pro Ligario, p. 57), ma anche:

> E me, come conoscesti essere figliolo di pistore? (Novellino, II, p. 128, r. 71)

Che il pronome soggetto dopo il verbo fosse facoltativo si può ricavare da questo esempio da Dante:

…mi rivolsi loro e parla' io (Dante, D. C. Inf. 5, 115)

in cui il pronome soggetto appare dopo il secondo verbo ma non dopo il primo.

Lo stesso vale, per es., in romanesco antico, dove abbiamo *comenzo io* e dove *so'io venuto?* (Cronaca, XXIII), ma anche:

da quale novitate comenzeraio?

– in cui *-aio* rappresenta lat. HABEO, come in fiorentino e in italiano *-ò*, dunque „comincerò". Ne concludiamo che dove abbiamo *conoscesti* avrebbe potuto esserci *conoscesti tu*, e dove c'è *comenzeraio* potrebbe seguire *io*.

Costruendo una grammatica dell'ital.ant., basata su un corpus, possiamo, e, io direi, dobbiamo costruire delle frasi agrammaticali, accanto alla documentazione di quelli grammaticali, e possiamo e dobbiamo postulare usi alternativi. Procediamo così con la grammatica di una lingua antica come con quella di una lingua viva: e questo perché per avere una buona descrizione di una lingua dobbiamo dire ciò che può esserci, ma anche ciò che non può esserci. Chiarito questo punto, passerò a discutere un altro punto, che ha pure una poratata teorica maggiore di quella che potrebbe sembrare. Chi lavora su una lingua antica è portato spesso a considerazione di ordine apparentemente statistico, ma in realtà ingenuamente numeriche. C'è la tendenza a scrivere espressioni come: *la tale forma, o costruzione, è rara, la tale è più frequente di un'altra,* e simili.

Come antidoto all'ingenuità di tale asserzioni, si potrebbero leggere questa volta con utilità le osservazioni della linguistica del corpus, che non sono in realtà altro che applicazioni alla lingua delle norme generali della statistica come scienza. Cosa vogliono dire parole come *raro*, *frequente*, o espressioni come *più frequente, più raro?* Delle nozioni come quelle di frequenza non possono più essere usate ingenuamente dopo che a definirle è nata, e non da poco, un'intera scienza. Questa scienza, la statistica, ci dice per es. che se di un fenomeno (nel nostro caso di una forma, di una costruzione, ecc.) si trovano più esempi che di un'altra, bisogna osservare lo scarto, e poi vedere, attraverso una metodologia complessa, se lo scarto è sufficiente a ritenere la superiorità definitiva. Altrimenti può essere che, continuando lo spoglio, i risultati potrebbero rovesciarsi. Se poi il materiale da spogliare fosse esaurito senza che lo scarto possa essere giudicato rilevante in modo definitivo, dobbiamo rassegnarci all'idea che una valutazione sicura non è possibile. Una valutazione non si può fare a occhio, come in Italia credono molti linguisti. Se no si potrebbero chiudere i corsi e le Facoltà di Statistica.

Questo non toglie che il buon senso può guidare a giudizi certi in caso di sproporzioni macroscopiche: 100 contro 1 vuol dire certamente 100 contro 1 (ma invece se su 10 occorrenze, abbiamo 6 contro 4, questo non vuol dire proprio niente, come sa benissimo chi stia giocando a testa e croce: nel giro di pochi tiri il risultato potrebbe capovolgersi). Ma soprattutto il buon senso dovrebbe aiutarci a capire che, per esempio, riprendendo l'esempio di prima, se di un certo di caso di 3.a pl. femminile ci sono pochi esempi non è che quel caso fosse *raro*. Chi scrive una grammatica si trova ad ogni passo di fronte a casi del genere. Riporto come esempio un caso, questa volta, lessicale, quello delle parole *mamma* e *babbo*. Prima di Dante *mamma* è testimoniata due volte (in un documento del 1211), *babbo* mai. Dante stesso le usa solo nella *Divina Commedia*, quattro volte *mamma*, una volta *babbo*. Sono, erano delle parole rare? *Babbo* era più raro di *mamma?* niente affatto, erano parole estranee ai generi letterari e ai soggetti trattati dalle opere del tempo, perfino in carte private si preferiva scrivere *madre* e *padre*, testimoniate ambedue con centinaia di occorrenze. Ci voleva il genio di Dante per documentarcele. Diverso sarebbe il caso, per esempio, di *ragazzo* e *ragazza* che certamente nel Duecento e nel Trecento non c'erano ancora, almeno a Firenze e in Toscana. Qui zero occorrenze vuol dire davvero zero: cioè che la forma non c'è e non può esserci. Tutto ciò la semplice interrogazione del corpus non ce lo può dire. Non è che i *corpora* siano inutili, spero che nessuno si sia fatto questa idea leggendo quanto ho scritto. Ma ci sono tante altre cose che dobbiamo leggere e sapere per lavorare bene su una lingua antica, come ce ne sono per studiare una lingua viva.

BIBLIOGRAFIA

Botley, S. & McEnery, A.M., 1996, *Corpus-based and Computational Approaches to Discourse Anaphora,* Amsterdam, Benjamins.

De Mauro, Tullio, (1980), *Guida all'uso delle parole*, Roma, Editori riuniti, 12a ed. 1997.

LIP (Lessico dell'italiano parlato, 1993, di T. De Mauro, F. Mancini, M. Vedovelli, M. Voghera, Milano, Etaslibri.

Foulet, L. 1919, *Petite syntaxe de l'ancien français*, Paris, Champion.

McEnery, T. & Wilson, A. 1996, *Corpus Linguistics*, Edinburgh, Edinburgh University Press.

Oostdijk, N. e de Haan, 1994, edd., *Corpus-based Research Into Language. In honour of Jan Aarts*, Amstedam, Rodopi.

Renzi, L. 1993, intervento alla Tavola Rotonda in *La filologia romanza e i codici*, a cura di Saverio Guida e Fortunata Latella, Messina, Sicania, 1993, vol.II, pp. 781–782.
Renzi, L.1998, *La discendenza di* QUI, „Studi di grammatica italiana", 17, 1998, 5–36.
Renzi, L. 2000, *Le tendenze dell'italiano contemporaneo. Note sul cambiamento linguistico nel breve periodo*, „Studi di lessicografia italiana", XVII, 279–319.
Rissanen M. 1989 *Three problems connected with the use of diachronic corpora*, ICAME Journal, 13, 16–19.
Rossini Favretti, R. 2000, a cura di, *Linguistica e informatica. Corpora, multimedialità e percorsi di apprendimento*, Roma, Bulzoni.
Rossini Favretti, R. 2001, *La linguistica dei corpora in Europa: prospettive e analisi*, Lingua e stile, 2, 2001, 367–357.
Sinclair, J. 1991, *Corpus, Concordance, Collocations*, Oxford, Oxford University Press.
Sinclair, J. 1997, *The Lexical Item*, Ms.University of Birmingham.
Squillacioti, P., Mosti, R., Larson P. 2001, *Il tesoro della lingua italiana delle origini, in Opera del vocabolario. Bollettino, La lessicografia storica e i grandi dizionari delle lingue europee*, Alessandria, Edizioni dell'Orso, pp. 43–75.
Vanelli, L., Renzi, L., Benincà P. (1985), *Tipologia dei pronomi soggetto nelle lingue romanze*, ora in Benincà (1994) *La variazione sintattica. Studi di dialettologia romanza*, Bologna, Il Mulino, pp. 195–211.