Premessa

Abbiamo progettato questo volume, discutendone con collaboratrici e collaboratori, nell'autunno del 2003. Il lavoro di coordinamento si è rivelato più complesso di quanto immaginavamo. Ora che è terminato, ringraziamo gli autori e le autrici per la attenzione e pazienza con cui hanno rispettato le indicazioni redazionali. Nella raccolta figurano alcuni contributi di cui noi due curatori siamo autori o coautori. Inoltre la stesura dell'introduzione generale al volume è di Tullio De Mauro e le introduzioni alle sezioni sono di Isabella Chiari.

i.c. e t.d.m.
Roma, 7 febbraio 2005

Gli studi statistici della forma e della sostanza dei suoni

Anche nel campo fonetico e fonologico, l'applicazione degli strumenti statistici, per l'analisi e per l'interpretazione dei fatti linguistici, è un'operazione che ha sedotto numerosi linguisti dagli anni Trenta del XX secolo in poi. Lo studio delle caratteristiche statistiche delle unità fonologiche e delle unità fonetiche in alcuni casi ha condotto a mere elencazioni senza particolari scopi interpretativi, quasi che i numeri da soli possano fornirci un'informazione utile sulle strutture linguistiche. Tale fascino dei numeri ha riguardato dagli anni Sessanta in poi numerose analisi a tutti i livelli linguistici, non solo quello fonologico, facendo spesso perdere di vista al lettore l'effettiva utilità delle ricerche quantitative come strumento euristico, interpretativo e di indagine linguistica. Si tratta di quella che Y. Lebrun (1966: 105) chiamava tendenza a "se laisser griser par le jeu subtil des nombres".

Non v'è dubbio che l'applicazione rigorosa degli strumenti che permettono la valutazione quantitativa dei fenomeni che riguardano il dominio fonologico e fonetico possa condurre a interessanti indicazioni sulla strutturazione, ma anche sul funzionamento dei processi di produzione e ricezione linguistica. Vi sono tuttavia molti modi di rispondere all'esigenza descrittiva segnalata da Gustav Herdan (1964: 66):

> Non sarebbe facile considerare completa la descrizione dei fonemi in una lingua, se sapessimo soltanto *quali* fonemi sono compresi nel suo sistema fonetico, senza sapere *quanto* è usato un particolare fonema, non solo nel dizionario, ma anche nell'uso dei membri della comunità linguistica, o, brevemente, senza conoscere il peso funzionale dei fonemi.

Escludendo osservazioni del tutto occasionali, si può dire che la prima vera e propria trattazione sistematica dei sistemi fonologici in termini quantitativi sia stata quella di G. K. Zipf dagli anni Trenta in poi.

Antecedenti rilevanti sono stati, con obiettivi radicalmente diversi, per lo più di tipo pratico, gli studi condotti sui sistemi stenografici e sulla crittografia, basati su analisi di tipo grafematico delle caratteristiche delle lingue naturali. Nel Cinquecento appaiono alcuni trattati di

crittografia interessanti ed esaustivi come quelli sul francese di Blaise de Vigenère, di Giambattista della Porta; nel Seicento quelli di John Wilkins, di Johannes Trithemius; nell'Ottocento di Arthur Joseph Hermann, del marchese de Viaris; nel Novecento di Luigi Sacco, di Mario Zanotti e molti altri. Accanto a questi emergono anche testi di steganografia come quelli di Daniel Schwenter, Ch.-Fr. Vesin, André Herman, Giovanni Piccoli. In questi lavori per la prima volta sono infatti discussi temi relativi alla frequenza delle lettere e delle parole di un testo, ma soprattutto sono svolte anche in modo sottile discussioni sulle proprietà peculiari dei testi che ne definiscono l'identità, la facilità o difficoltà ad essere trasmesso o nascosti, la struttura specifica dal punto di vista puramente formale.

La stenografia oramai quasi scomparsa (con l'importante eccezione delle trascrizioni giudiziarie e delle sedute pubbliche delle Camere), ancora oggetto di studio, ricerca e applicazione ulteriormente affinati è la crittografia, soprattutto connessa con lo sviluppo delle nuove e nuovissime tecnologie. Accanto a ciò si aggiunge oggi anche un settore di grande interesse linguistico, per alcuni importanti aspetti, che si occupa delle tecniche di compressione dei testi per la trasmissione. Come spesso nella storia della riflessione linguistica, molti spunti di indagine si sono avviati anche in questi casi da problemi eminentemente pratici, artigianali e concreti, conducendo a significativi sviluppi teorici e applicativi.

La fonologia statistica in una prospettiva specificatamente linguistica si può, senza temere troppo semplificazioni, far risalire alla scuola di Praga, e in particolare ai lavori di Mathesius, Trnka e Vachek, con notevoli contributi dagli anni Quaranta in poi anche in ambito sovietico. Dagli anni Sessanta in poi la prospettiva praghese incomincia ad essere sistematicamente integrata con metodologie e quadri di riferimento legati alle nozioni di entropia e ridondanza, sviluppate all'interno del paradigma della teoria dell'informazione (cfr. Chiari 2002: 35-64). L'attenzione si sposta dunque dal semplice computo delle frequenze assolute e relative delle classi fonematiche alla osservazione delle regole di restrizione e delle combinazioni possibili dei fonemi in sequenze (già in questa direzione va Mathesius nel suo studio sul ceco del 1911 e successivamente anche Kučera e Monroe 1968 sul russo, mentre sull'inglese spiccano i lavori di Dewey, 1923 e Denes, 1963). Applicazioni ispirate variamente ai lavori di Shannon le troviamo in Miller, Selfridge, 1950; Shannon, 1951; Miller, 1956; Do-

ležel, 1963; Petrova, 1968; Boguslavskaja, 1968; Novak, Piotrovskij, 1968; Chapanis, 1954; Newman e Gerstman, 1952, mentre per la lingua italiana abbiamo i dati forniti da Manfrino (1960).

Dal punto di vista metodologico si sono presto configurati i principali strumenti matematici e statistici per lo studio della fonologia e della fonetica soprattutto la teoria della probabilità, la statistica descrittiva (per lo studio delle unità: fonemi, foni, sillabe) e, per lo studio della fonotassi, soprattutto il calcolo combinatorio.

Come si può vedere nei saggi contenuti in questa sezione di fonetica e fonologia del volume, anche qui i metodi variano: dal computo e confronto di occorrenze come nel saggio di Albano Leoni e Clemente sulla ridondanza fonematica, in quello di Chiari e Castagna sui nessi consonantici, o quello di Koesters Gensini sul rapporto tra lunghezza delle parole e frequenza, fino a diversi strumenti probabilistici, matematici e informazionali più articolati come si può osservare nel saggio sul rendimento funzionale.

Concretamente un'analisi di tipo quantitativo richiede obbligatoriamente la predisposizione di un corpus, costruito e pianificato in modo da rendere i risultati rappresentativi ed estensibili a simili porzioni di lingua (o testi). Osservata da questo punto di vista la linguistica quantitativa si configura come una branca metodologica intrecciata alla linguistica dei corpora. In questo caso occorrerebbe distinguere veri e propri *corpora testuali* (autentici e composti di *running words*) rispetto a *repertori* (detti a volte, forse impropriamente, corpora lessicali o lessici).

Se si tentasse di rilevare una serie di nodi di riflessione teorica e metodologica che emergono dal tentativo di analizzare con strumenti quantitativi il dominio fonetico e fonologico, anche solamente osservando i saggi contenuti in questa sezione, si potrebbero segnalare almeno cinque questioni:

I) Il problema dell'*unità di analisi* (fono, fonema, sillaba)
II) Il problema dell'*inventario*
III) Il problema dell'*interfaccia fonetica/fonologia*
IV) Il problema del *rapporto tra fonologia sincronica e diacronica* e della sua valutazione quantitativa
V) Il problema dell'*interazione tra livello fonologico e livelli più ampi* (morfologico, testuale)

Il problema dell'*unità di analisi* discusso variamente sia nel saggio introduttivo su foni e fonemi che in quello del rendimento funzionale, è tipico di ogni analisi linguistica, sia quantitativa che qualitativa, non occorre dunque qui soffermarvisi. Il problema dell'*inventario*, anch'esso centrale per qualunque discussione, implica: a) l'individuazione delle unità e b) la valutazione del contesto (posizionalità, restrizioni, ecc.). Entrambe le questioni sono spinose nel dominio fonologico: vi si intrecciano dibattiti nel merito della definizione stessa di unità e anche numerosi problemi di trattamento matematico, più intricato ogni qual volta non si abbia a che fare con frequenze assolute prese singolarmente.

Il problema dell'*interfaccia fonetica/fonologia*. L'integrazione di considerazioni che riguardano la composizione statistica e/o quantitativa di corpora e repertori è particolarmente utile per indagare le dinamiche della produzione linguistica, e soprattutto il circolo che, saussurianamente, si istituisce tra *langue* e *paroles*. Tema questo particolarmente cruciale nella riflessione teorica attuale sui rapporti tra fonetica e fonologia (il problema dell'interfaccia) illustrato in alcuni suoi punti critici nel contributo di Albano Leoni e Clemente. Il problema della dinamicità e parziale artificialità della distinzione tra *fonologia diacronica e sincronica* e della relazione tra i due punti di vista, tanto fondamentale e delicato a ogni livello dell'analisi linguistica, e illustrato qui nel lavoro sul rendimento funzionale. Il problema dell'*interazione tra livello fonologico e livelli più ampi* (morfologico, testuale) riguarda lo scarto che si crea tra composizione fonologica osservata, per così dire, al microscopio (per sequenze di fonemi o per sillabe) e composizione fonologica così come si dispiega e configura nella costituzione dei morfi e del lessemi e, ancora maggiormente, nella costituzione di testi. Si tratta di un tema particolarmente complesso centrale alla discussione sullo statuto teorico della nozione di rendimento funzionale e altrettanto dominante nella problematizzazione del rapporto, notoriamente esplicitato come tendenza generale da Zipf, tra lunghezza e frequenza delle parole nei testi, ripreso e discusso da Sabine Koesters Gensini nel suo saggio in questo volume.

Una questione inoltre attraversa le cinque appena elencate ossia la discussione del modello teorico-linguistico che sottostà a tutte le specifiche scelte e che per esempio affronti dello statuto complesso e incerto della nozione di rumore, di ottimalità, di minimo sforzo, e anche di contesto e di variazione. Questione questa che tocca in modo cru-

ciale non solo il livello fonologico, ma soprattutto il livello stilistico-testuale delle lingue, come emerge in altri luoghi di questo volume e in particolare nell'ultima sezione su *Testi e usi*.

La struttura statistica del lessico e del vocabolario

Che cos'è il vocabolario fondamentale di una lingua? Che cos'è il vocabolario di base? Che cosa rappresentano dell'uso linguistico? Quanto differiscono nella composizione qualitativa e quantitativa rispetto a lessici più ampi? Differisce sotto qualche aspetto il vocabolario di base usato da un parlante madrelingua da quello usato da un parlante di italiano come seconda lingua? Sono tutte domande centrali per la comprensione del funzionamento del lessico di una lingua e soprattutto dell'uso che in circostanze diverse del lessico di questa lingua si fa.

Dal punto di vista quantitativo e statistico il lessico è solitamente osservato per sciogliere e spiegare principalmente tre tipi di problemi (cfr. Těšitelova, 1992: 69): a) la discussione su quale sia l'unità di popolazione o unità di analisi del lessico e quale, in relazione a ciò, debba essere l'ampiezza del corpus perché esso possa essere rappresentativo degli usi lessicali di una lingua o di un testo; b) la costruzione, analisi e discussione delle caratteristiche del vocabolario di una lingua in relazione alla frequenza d'uso delle parole; c) la valutazione articolata della cosiddetta 'ricchezza del vocabolario' di un testo, di un autore, ecc.

Il problema metodologico dell'unità e della rappresentatività attraversa, affrontato in diversi modi, tutto il presente volume, dal livello fonetico/fonologico fino a quello più propriamente stilistico. Il problema della ricchezza del vocabolario è invece più volte discusso nei saggi che compongono l'ultima sezione (*Testi e usi*), in particolare nei contributi di Emanuela Piemontese e di Rita Plantera, come si dirà più avanti. La questione che invece tocca più da vicino questa sezione è quella segnalata al punto b), ossia la discussione e l'ordinamento del lessico di una lingua in relazione alla frequenza d'uso.

Come più volte viene detto nei saggi che fanno parte di questa sezione, la nozione di vocabolario di base e quella sua antecedente storica di vocabolario fondamentale, sono nozioni tutt'altro che scontate. La selezione ed analisi dei lessemi presi in conto risente di volta in volta degli scopi che, nel corso soprattutto dell'ultimo secolo, hanno motivato la costruzione delle liste. Pedagogisti, psicologi e ancora una volta stenografi, molto prima che linguisti, si sono occupati infatti del-

la predisposizione di tali repertori. Al lavoro di impianto manuale sul tedesco di Kaeding (1897), uno dei più ampi dal punto di vista dell'estensione del corpus tra i dizionari di frequenza di prima generazione, sono seguiti quelli sull'inglese di Thorndike (1921, 1931-32), di Vander Beke (1930) sul francese, di Buchanan (1927) sullo spagnolo. Diversi lavori poco conosciuti di questo tipo sono stati svolti nell'URSS anche sulle lingue europee e soprattutto a fini pedagogici e per l'insegnamento delle lingue straniere (cfr. Alekseev, 1973).

Tra i dizionari di seconda generazione invece troviamo i lavori di Josselson (1953) e Zasorina (1977) sul russo, il famosissimo dizionario dell'inglese americano di Kučera e Francis (1967), la serie di dizionari di Juilland et alii (1970) sulle lingue romanze a partire da rumeno, spagnolo e francese, e il LIF (1971) per l'italiano e il più recente LIP (1993). Alla terza generazione appartengono invece i dizionari di frequenza costruiti a partire dai grandi corpora di riferimento delle lingue europee, come ad esempio il Leech et alii (2001) basato sul *British National Corpus*.

I dizionari di frequenza sono stati il naturale antecedente dell'elaborazione dei vocabolari fondamentali, almeno nell'accezione che oggi si dà al termine in ambito linguistico. I vocabolari fondamentali infatti si identificano oggi con la zona di alta o di massima frequenza d'uso tra le fasce in cui si può suddividere il lessico di una lingua. Accanto a questi si sviluppa sin dagli anni Sessanta il filone che potremmo dire opposto, anche qui soprattutto nei paesi dell'ex-URSS, sulla frequenza di testi specialistici e dei linguaggi settoriali (giornalismo, fisica, chimica, ecc.).

Nella prima metà del Novecento si incominciarono a produrre anche i cosiddetti *word books* contenenti le parole del vocabolario di alto uso, partendo dall'inglese (cfr. Hunter, 1931; Thorndike, 1932; Palmer e Hornby, 1937; Faucett e Maki, 1940; Fries e Traver, 1940; Thorndike e Lorge, 1944; Roberts, 1965), e successivamente prodotte per il francese (cfr. Henmon, 1924), per lo spagnolo (cfr. Keniston, 1933) per il tedesco (Schinnerer e Wendt, 1933; Pfeffer, 1964) e per altre lingue come svedese, russo, portoghese, mentre per la lingua italiana la prima lista di parole di alta frequenza fu curata da Knease (1931) sulla base di un corpus di tipo letterario.

Accanto ai *word books* sempre a scopi didattici si diffusero i dizionari fondamentali, che raccoglievano definizioni ed esempi delle parole con maggior frequenza, usando quelle stesse parole per definire tut-

te le altre. Per l'inglese è molto noto il *Basic English* (1944) di Ogden. In Unione Sovietica la costruizione di cosiddetti *vocabolari minimi* per le principali lingue europee riscuote particolare successo con il lavoro di Rakhmanov, 1947). Per la lingua francese, si registra il caso più noto di dizionario fondamentale creato con forti basi statistiche ad opera di Gougenheim (1958), il quale è anche uno dei primi a occuparsi del problema della parole di alta disponibilità, che faranno il loro ingresso nei vocabolari di base. In un secondo tempo compaiono dizionari fondamentali anche per il tedesco (Pfeffer, 1970) e le altre lingue europee. Per l'italiano compaiono il volume di Bruno Migliorini (1943), che tuttavia non ha una solida costruzione statistica, e soprattutto i lavori di De Mauro che sono confluiti nella costituzione del Vocabolario di base (De Mauro, 1980), la cui genesi e composizione è esposta in dettaglio nel presente volume nel saggio di Giuliani, Iacobini e Thornton.

In relazione alla definizione quantitativa e qualitativa del vocabolario di base i saggi contenuti in questa sezione discutono appunto questioni di tipo specificatamente linguistico e questioni di tipo metodologico. In particolare emerge una rappresentazione del vocabolario di base come un'entità dotata di sue proprie caratteristiche (qualitative e quantitative appunto), ben distinte da quelle esibite dal vocabolario generale di una lingua. Entità che ha una sua fisionomia diacronica e sincronica precisa, stratificata in modo proprio, composta da in modo molto stabile da secoli (con lessemi entrati in italiano soprattutto nel XIII, XIV e XVI secolo), e con poche innovazioni recenti (cfr. Giuliani, Iacobini, Thornton in questo volume).

Diverse considerazioni, correlate alla natura e al momento storico in cui il vocabolario di base della lingua italiana è stato elaborato, spingono a rivederne alcune caratteristiche non radicali: da una parte, l'inclusione di osservazioni sulla lingua parlata, oggi disponibili con maggior facilità di venticinque anni fa, almeno parzialmente permette di riformulare (quantomeno dal punto di vista dell'uso) una parte dei lemmi inclusi o esclusi; dall'altra, poche o pochissime innovazioni e altrettanto poche parole cadute in disuso per motivi contingenti (si pensi al destino della *lira*). Ma la fascia che per sua natura solleva maggiori questioni rimane quella dell'alta disponibilità, discussa per alcune caratteristiche teorico-metodologiche ed empiriche nel saggio di Domenico Russo e nel merito dei lemmi inclusi in quello di Francesco De Renzo. Quest'ultimo ricostruisce per molte parole il passaggio

da una fascia all'altra nelle diverse versioni del vocabolario di base che si sono seguite negli anni, notando proprio la relativa chiusura dei lemmi del VdB, ossia il fatto che le differenze registrate ad esempio per il vocabolario di alta disponibilità non sono tanto nuove entrate, ma passaggi soprattutto dalla fascia di alto uso all'alta disponibilità. Questo tipo di considerazioni stimolano a osservare meglio proprio la definizione dei confini tra le fasce e anche a provare a trovare qualche tipo di giustificazione o spiegazione (socio-culturale e/o linguistica) per tali passaggi, così come anche per la relativa autonomia complessiva della porzione di lingua identificata così nettamente dal vocabolario di base.

Il saggio di De Renzo, come per altri versi anche quello di Carloni e Vedovelli, riportano inoltre il vocabolario di base alla sua genesi teorico-pratica, i *word books* per l'apprendimento delle lingue, cui si è accennato poco sopra. Nel primo caso il legame con la prospettiva pedagogica è svolto attraverso un confronto con il lessico conosciuto e usato dai bambini della scuola elementare (tema svolto in maniera più estesa nel capitolo di Silvana Ferreri nella sezione *Conoscenza e usi delle parole* di questo volume), utile per formare un'immagine delle diverse fasi dell'acquisizione della propria lingua materna.

Nel caso del lavoro di Carloni e Vedovelli, prima riflessione sui dati emersi da un ampio progetto di ricerca interuniversitario su lessico e sintassi dell'italiano parlato da apprendenti stranieri, il cardine dell'indagine risiede nella valutazione della dinamicità del percorso di apprendimento dell'italiano come lingua seconda e nella peculiarità delle caratteristiche del parlato dei non nativi nelle fasce lessicali di alta frequenza rispetto alle stesse fasce osservate nei testi prodotti da parlanti nativi. La dinamicità caratterizzante del percorso di apprendimento di una lingua straniera (dinamicità che tuttavia contraddistingue anche la competenza di ogni parlante nativo) è, quasi paradossalmente, particolarmente adatta a essere rappresentata mediante strumenti quantitativi, come sottolinea appunto Vedovelli nella parte teorico-metodologica del saggio, proprio per la sua facoltà di cogliere la variabilità e variazione (sincronica e diacronica) degli usi linguistici.

Tutti i temi fin qui sintetizzati ritornano sotto una luce diversa anche nel contributo di Tommaso Russo dedicato a un problema delicato e nuovo connesso con le questioni metodologiche e teoriche che presiedono alla costituzione di un lessico di frequenza per la lingua italiana dei segni, in cui vengono affrontate alcune questioni relative alla

costruzione, analisi e interrogazione di corpora di trascrizioni di testi della LIS insieme a questioni insidiose e spinose anche per l'analisi delle lingue verbali.

Le tendenze statistiche nella conoscenza di lessemi, significati e usi

Dato l'immenso interesse che, sin dagli esordi della linguistica quantitativa, ha risvegliato lo studio del lessico, due sezioni diverse di questo volume sono state dedicate alla sua esplorazione. Nella sezione precedente si sono visti gli aspetti che permettono di elaborare ed estrarre dal lessico della lingua il nucleo delle parole che costituiscono il vocabolario di base, in questa sezione invece sono raccolti contributi che si focalizzano soprattutto sulle competenze lessicali, la loro acquisizione (Ferreri), la loro stabilizzazione in uno stato di lingua rappresentato dai dizionari (De Mauro e Ferreri), la loro dinamicità come mostrata dal rinnovamento e invecchiamento lessicale (Bolasco). È affrontato inoltre un tema 'classico' della statistica linguistica, ossia la relazione, individuata da Zipf, tra numero delle accezioni di una parola e frequenza (Carloni) che rimanda a un settore della linguistica quantitativa poco battuto, ossia la semantica statistica.

Come è stato osservato nella sezione precedente, il lessico dal punto di vista storiografico è il dominio maggiormente trattato nel corso del XX secolo con metodi quantitativi. Riassumere i contributi più significativi in questo campo sarebbe impresa pressoché impossibile. Basti dunque citare alcuni nomi, a complemento di quelli fatti in precedenza, e tralasciando l'amplissimo tema della valutazione della ricchezza del vocabolario inclusa nella sezione di questo volume dedicata al testo: al pluri-menzionato Zipf, è necessario aggiungere soprattutto in area francese gli studi sul vocabolario di Pierre Guiraud (che saranno ripresi anche nell'ultima sezione di questo volume), quelli di Charles Muller, di Gustav Herdan, notevoli non solo per gli aspetti empirici e applicativi ma soprattutto per le proposte teoriche e metodologiche. Per quanto riguarda invece i temi connessi allo sviluppo (quantitativo) del linguaggio e le dinamiche diacroniche negli usi linguistici si ricordano in particolare: sui temi svolti in questo volume da Bolasco, in una prospettiva specificatamente linguistica, l'essenziale saggio di Mańczak (1966), lo sviluppo di una vera e propria disciplina, la *glottocronologia*, fondata da Swadesh (1952; 1955), centrata appunto sulla valutazione statistica del mutamento linguistico, approc-

cio particolarmente fortunato negli anni Sessanta e Settanta in Unione Sovietica e continuato dalla linguista canadese Sheila Embleton (in particolare 1986); e, sempre in area slava, il contributo di Golovin (1965) sul russo e quello dello stesso Herdan (1964). Per quanto riguarda invece i temi connessi con l'acquisizione del lessico, si possono menzionare gli studi di Deuszing (1919; 1921) e di Bakonyi (1933; 1934; 1939) sul tedesco, e soprattutto il contributo teorico e metodologico di J. E. Hall (1954) oltre ai riferimenti riportati da Ferreri nel suo saggio.

Come si è osservato per tutti i livelli dell'analisi linguistica, anche per il lessico, o potremmo dire, soprattutto per il lessico, si rende primaria la discussione e la scelta dell'unità di analisi: parola testuale, lessema, lemma, parola grafica, ecc. Il tema è essenziale sia dal punto di vista teorico che applicativo ed è presentato in entrambe le vesti nel contributo di Ferreri, di De Mauro e Ferreri, e indirettamente, nei suoi risvolti applicativi, in quello di Bolasco.

Tra le questioni cruciali per la valutazione quantitativa delle competenze vi è quella, trattata esplicitamente nel saggio di Silvana Ferreri, della asimmetria tra piano della produzione e piano della ricezione, osservata soprattutto nelle prime fasi di acquisizione della lingua e, anche, attraverso il problema summenzionato dell'unità di analisi: com'è immagazzinato il lessico mentale di un individuo? Come avviene l'accesso in ricezione e in produzione?

Particolarmente interessante sarebbe inoltre l'osservazione comparativa dei dati sulla prevalenza di nomi e verbi nelle fasi di acquisizione illustrata da Ferreri su lingue diverse, con le considerazioni che riguardano la manifestazione testuale (scritta e parlata) di queste due categorie nell'italiano contemporaneo, come spia di diverse strategie di elaborazione mostrata nel saggio di Voghera, contenuto nella seconda sezione di questo volume (*La grammatica*), da indagare eventualmente anche in una prospettiva cross-linguistica.

Per quanto riguarda invece la valutazione della dinamicità intrinseca del lessico di una lingua e della sua correlazione a fatti, circostanze ed eventi extralinguistici, emerge con particolare chiarezza l'inscindibilità delle considerazioni sulla natura mobile del lessico, mostrata da neologismi e obsolescenti, con i processi socioculturali della comunità linguistica che parla tale lingua. Bolasco, nel suo saggio sulle tendenze diacroniche più recenti dell'italiano, propone un indice che segnala gli andamenti anomali nell'uso delle parole di un

corpus, in modo da fornire un utile strumento statistico per la valutazione (qualitativa) delle tendenze linguistiche in un arco di tempo dato.

Un ponte inoltre va gettato, in relazione a tutti i temi citati (unità di analisi, competenze in produzione e comprensione, caratteristiche statistiche del lessico, costruzione della testualità e uso di verbi e nomi, ecc.), tra considerazioni quantitative e qualitative che riguardano l'acquisizione della lingua materna, come illustrate nel saggio di Ferreri, e le parallele considerazioni applicate al parlato di apprendenti l'italiano come lingua straniera (cfr. Carloni e Vedovelli in questo volume).

Lo studio statistico della morfologia e della sintassi

La morfologia e la sintassi non sono state tra i primi domini della linguistica sottoposti a indagini di tipo quantitativo. L'analisi dei diversi aspetti della grammatica è infatti un fatto relativamente recente. L'attenzione incomincia a cadere sulla statistica grammaticale solamente a partire dagli anni Sessanta (cfr. Těšitelova 1992: 109)[1], iniziando dalla quantificazione della presenza delle diverse parti del discorso in varie tipologie testuali, affrontando in seguito specifici aspetti dell'analisi morfologica del verbo e del nome. In sintassi, sempre dagli anni Sessanta, si incominciano ad affrontare temi quali la lunghezza della frase[2] e successivamente si giunge all'analisi quantitativa dei diversi tipi di sintagmi e dell'ordine delle parole con i primi lavori di G. U. Yule e di Herdan sull'inglese[3]. Ma come può essere definita la statistica grammaticale, entro cui si fanno confluire la statistica morfologica e sintattica? Una possibile definizione è la seguente:

> Grammatical statistics, one of the basic domains of quantitative linguistics, studies the frequency, distribution, and relations of units, i.e., grammatical categories, grammatical phenomena (features from the point of view of statistical methods), and on the basis of statistical data attempts to model them. It tries to explain how grammatical phenomena depend on and condition each other when functioning in a text, etc. (Těšitelova 1992: 100)

Tra i lavori più significativi, a puro titolo indicativo, si segnalano per quanto riguarda la morfologia, ancora una volta numerosi i contributi

---

[1] Tra i primi a occuparsi di questioni di statistica sintattica anche uno dei curatori di questo volume (cfr. De Mauro, 1959; 1960).

[2] Il problema della lunghezza delle frasi era affrontato primariamente in relazione all'attenzione da subito posta sui problemi di stilo-statistica o statistica testuale di cui si parlerà nell'ultima sezione di questo volume.

[3] In campo morfologico e sintattico inoltre le differenze tipologiche tra le lingue incominciano a farsi sentire con maggiore nettezza richiedendo dunque un trattamento sia teorico che metodologico e statistico radicalmente diverso per gruppi linguistici diversi (si pensi al diverso trattamento della sintassi in lingue flessive e isolanti, o alla scelta dell'unità di analisi della morfologia nelle stesse).

provenienti dall'URSS. In particolare sono alquanto rilevanti i lavori di Andreeva sui tipi morfologici del russo e di Bartkov sull'inglese; i numerosi lavori di Mel'čuk e quello di Moreau (1963) sulla morfologia del francese; il lavoro di Greenberg (1950) sulle lingue semitiche; sulla produttività morfologica rimandiamo direttamente ai riferimenti proposti nel saggio di Gaeta e Ricca in questo volume (cfr. soprattutto i lavori di Baayen). Per quanto riguarda la sintassi invece, solamente per dare un'idea dei lavori più ambiziosi, si ricordano sulla quantificazione delle categorie grammaticali la monografia di Barth (1961) su inglese, francese e spagnolo, il lavoro di Těšitelova (1973); sulla frequenza dei sintagmi il lavoro di Beebe (1980) sull'inglese; su diversi aspetti della sintassi dell'inglese Grammon (1963), di Keniston (1937) sullo spagnolo, di Uhlířová (1969; 1973) sul ceco.

Per quanto riguarda specifici problemi linguistici e metodologici centrali per le analisi quantitative nel campo morfologico e sintattico, come emerge chiaramente anche dai saggi contenuti in questa sezione, esistono alcuni temi dominanti: la definizione dell'unità di analisi e della relativa estensione del corpus, il problema teorico ed empirico della valutazione della produttività, la determinazione della relazione tra prospettiva diacronica e sincronica dei fenomeni sintattici in relazione alla dimensione testuale, e soprattutto della relazione tra lingua scritta e lingua parlata.

La principale difficoltà che si riscontra nella considerazione quantitativa dei fatti grammaticali dipende da una sorta di paradosso statistico. Tale paradosso dipende dal fatto che, selezionando come unità di analisi una categoria grammaticale o un tipo di sintagma, ci si ritrova, anche utilizzando corpora relativamente piccoli, con grandi valori di frequenza, poiché le categorie grammaticali o i tipi di sintagmi sono relativamente pochi rispetto agli elementi appartenenti ad altri domini linguistici (come quello lessicale o fonologico). Poiché tuttavia l'ampiezza reale dell'unità è ben più ampia (essendo principalmente costituita dall'enunziato o dalla frase), per avere risultati effettivamente rappresentativi dal punto di vista statistico, tenendo conto dei contesti sintagmatici in cui l'unità occorre, è necessario ricorrere a corpora di estensione significativamente più ampia rispetto ai corpora che vengono sottoposti ad analisi fonologiche o lessicali. Nella statistica

grammaticale è dunque capitale tenere in conto dei contesti di occorrenza, più che della semplice registrazione del dato di frequenza (anch'essa tuttavia importante e indicativa)[4].

Seppure trattato in modo indiretto, questo tema è capitale nel contributo di Annibale Elia a questo volume. Il quadro illuminato nella prospettiva lessico-grammaticale fa infatti emergere con evidenza quanto fenomeni che apparentemente ricondurremmo alle stesse classificazioni sintattiche risultino a un'osservazione più fine radicalmente diversi, o come dice Elia si osserva "una quasi individualità del comportamento sintattico di ogni uso verbale". Questo fatto ci induce a due considerazioni: dal punto di vista metodologico spinge verso un'analisi più fine (appunto come Elia la propone) degli usi sintattici e dunque, paradossalmente, ad ampliare l'estensione dei corpora di riferimento per lo studio della sintassi (rispetto al lessico), da un punto di vista empirico e quantitativo suggerisce la auspicabilità di una applicazione dell'analisi lessico-grammaticale a grandi corpora testuali per valutare l'incidenza quantitativa dei diversi usi "individuali" cui sottoponiamo di volta in volta le stesse tipologie verbali.

Il primo dei problemi summenzionati, ossia quello dell'unità di analisi (e di questa in relazione all'estensione del corpus) attraversa tutti i saggi della sezione, ma in particolare emerge esplicitamente in Voghera e in Gaeta e Ricca nelle scelte sulla valutazione delle forme di parola e delle loro specifiche occorrenze e distribuzioni.

Il saggio di Gaeta e Ricca, in particolare, mette in primo piano la discussione metodologica sulla valutazione quantitativa della produttività (morfologica), nozione questa spesso ambigua e indefinita in modo simile alla nozione di rendimento funzionale discussa nella prima sezione del volume.

Centrale invece al saggio di Miriam Voghera sulle categorie sintattiche e a quello di Policarpi e Rombi sulla sintassi dell'italiano contemporaneo è l'uso di strumenti quantitativi per la valutazione, da una parte, della differenza tra scritto e parlato (in particolare in Voghera), dall'altra, in entrambi i saggi della relazione tra fenomeni sintattici e

---

[4] Tra i lavori specificatamente dedicati a questo settore vi è in particolare la monografia di Těšitelova (1980), che affronta in modo sistematico i problemi del campionamento e del trattamento delle unità di analisi in statistica morfologica e statistica sintattica.

testualità nel complesso. Nel caso della relativa differenza tra scritto e parlato il dato quantitativo è usato come evidenza per mostrare come diverse strategie testuali possano sottostare alle differenti modalità espressive, in relazione da una parte alle diverse strutturazioni testuali tipiche di scritto e parlato, ma soprattutto in relazione alle diverse prassi di progettazione linguistica che vengono messe in atto dal locutore. Un simile sforzo, applicato su un singolo problema – quello degli usi del *che* nell'italiano contemporaneo – è svolto da Aureli con un'analisi sul corpus del LIP al fine di associare gli usi del *che* alle diverse tipologie testuali ivi rappresentate.

Nel caso invece del saggio di Policarpi e Rombi dominante è la valutazione della variazione diacronica della sintassi dell'italiano correlata a una esigenza di chiarificazione anche quantitativa dell'immagine globale che ci si fa di una lingua in un suo determinato stato, immagine fortemente influenzata da quella che si potrebbe chiamare una selezione arbitraria e non bilanciata dei testi che appartengono al nostro corpus mnemonico dell'italiano. Per non incorrere dunque nell'errore di farsi fuorviare da tale corpus mnemonico (che può dunque essere tutt'altro che rappresentativo della lingua in generale) è necessario costruirne uno, appropriato e bilanciato appunto, e interrogabile e analizzabile con strumenti quantitativi.

La struttura statistica dei testi e la sua analisi

Il testo come unità comunicativa definita in quanto equilibrio tra caratteristiche imposte dalla lingua stessa e caratteristiche che il singolo produttore del messaggio conferisce al suo prodotto individuale è diventato molto presto uno degli interessi principali della linguistica quantitativa, se non quello dominante. Uno dei testi cardine di questa disciplina non a caso ha il titolo *Language as choice and chance* (1956). Come già più volte rilevato in questo volume, il lessico dei testi, la sua valutazione in relazione alla lingua e in relazione al suo autore sono stati i primi settori indagati con una gran moltitudine di analisi, metodologie proposte e risultati elaborati. Le caratteristiche del testo sottoposte e sottoponibili a indagine quantitativa sono pressoché infinite: scelta lessicale e sua omogeneità, valutazione della leggibilità, relazione tra lunghezza delle parole o delle frasi e testo, concentrazione delle frequenze lessicali, comparazione tipologica tra lingue diverse in testi simili, stilometria forense, sono solo alcuni dei problemi analizzati in questo settore (per una rassegna bibliografica omnicomprensiva della letteratura su questo e gli altri settori cfr. Köhler, 1995).

I primi testi ad essere analizzati sono stati testi di tipo letterario, su cui si sono esercitati numerosi studiosi con obiettivi diversi: a) l'individuazione delle peculiarità che risiedono nelle scelte lessicali di un autore in un testo; b) la determinazione delle caratteristiche linguistiche (lessicali, macroscopiche, morfo-sintattiche) che differenziano e/o accomunano testi di tipologie diverse; c) la predisposizione di strumenti di stilo-statistica che permettano di risolvere problemi di attribuzione della paternità di un'opera a un autore o a un altro (cfr. in particolare, oltre al noto lavoro di Yule 1944, si vedano le ricerche di Mosteller).

Il primo punto, riguardante la valutazione delle caratteristiche linguistiche proprie di un testo (letterario e non), ha attratto speciale attenzione a partire dallo studio accurato e per molti versi pionieristico di Pierre Guiraud del 1954, seguito da Mistrík, 1967; Doležel e Bailey, 1969; Dugast, 1979), fino a condurre all'esigenza di trovare misure e strumenti per la valutazione dell'apporto linguistico individuale, come mostra il lavoro di Piemontese sull'intersezione tra aspetti pro-

duttivi e ricettivi del testo in una prospettiva quantitativa. Anche del secondo aspetto tratta il lavoro citato sugli stili verbali individuali, mettendo in luce anche gli aspetti più delicati della valutazione linguistica di tipologie testuali diverse, soprattutto in relazione ai problemi della valutazione della comprensione e leggibilità.

Anche sulla rilevazione degli aspetti quantitativi che differenziano testi diversi negli anni Sessanta e Settanta sono comparsi numerosi lavori. Tra i più rilevanti sono da segnalare gli studi di Alekseev (in particolare 1977), Juhan Tuldava e M. V. Arapov. Per quanto riguarda l'analisi lessico-statistica di testi italiani sono da citare, tra gli altri, oltre a padre Roberto Busa, almeno Heilmann (1961), Rosiello (1965), De Mauro (1966), Alinei (1971), Albano Leoni (1970-72), Prosdocimi (1977), Cortelazzo (1984) e ancora prima i lavori degli statistici Boldrini, Faleschini e Lonstergo (1948).

Gli studi sul testo sono stati presto catalizzati intorno alla questione della misurazione e valutazione della ricchezza del vocabolario (si veda per tutti la monografia di Cossette, 1994 in cui gli indici di Yule, Guiraud, Herdan, Muller, Brunet, Uber e Dugast vengono attentamente analizzati e confrontati), questione legata alla valutazione della correttezza teorica e statistica della nota legge armonica di Zipf sul rapporto tra rango e frequenza, riformulata con obiettivi di maggior accuratezza da Benoit Mandelbrot in diversi lavori. Su alcuni di questi aspetti si sofferma, con attenzione anche al versante della leggibilità, il lavoro di Plantera sulla temperatura informazionale.

Gli ultimi due saggi di questa sezione sono invece una piccola incursione in quella ampia area di intersezione, oggi in particolare espansione, che riguarda la costruzione di strumenti informatizzati di analisi e sintesi linguistica, ossia la linguistica computazionale. La relazione tra statistica linguistica in senso tradizionale e linguistica computazionale è sempre stata infatti molto stretta. Ancora una volta nell'ambito di ricerca sovietico, sin dagli anni Trenta, lo studio delle caratteristiche quantitative delle lingue naturali ha fornito il primo impulso per l'ideazione di strumenti computazionali (perfino prima della nascita dei veri e propri calcolatori). Si veda come esempio il modello di traduzione automatica basato sull'esperanto come interlingua e proposto dal russo Pëtr Trojanskij. La statistica linguistica ha sempre fornito strumenti e risultati fondamentali per lo sviluppo delle applicazioni della linguistica computazionale. Vi sono almeno due aspetti particolarmente rilevanti per illustrare tale contributo. Da una parte la

statistica linguistica ha spinto alla costruzione dei grandi corpora di riferimento delle lingue europee, che oggi costituiscono una preziosissima base di informazioni sulle caratteristiche effettive di uso della lingua scritta e della lingua parlata di cui si servono le applicazioni computazionali (si pensi ai dizionari elettronici basati su corpus, alla traduzione automatica *example-based*, alle *translation memories*). Dall'altra la pura applicazione di alcuni metodi di tipo statistico ai dati linguistico-testuali (applicazione da alcuni definita guidata da principi "non linguistici") sono derivati alcuni degli strumenti più potenti della linguistica computazionale attuale (si pensi allo *statistical natural language processing*, ai sistemi di estrazione della terminologia e delle collocazioni, alla *statistical machine translation*). Il rapporto tra le due discipline non è sempre stato facile (si ricordino le prime posizioni adottate dai sostenitori dell'approccio *rule-based* al trattamento automatico del linguaggio di ispirazione chomskyana), ma è attualmente imprescindibile, se si vuole superare una certa impasse nel successo di analisi e di copertura di grandi corpora di dati testuali, senza ricorrere al dispendioso intervento di integrazione umana al lavoro svolto in automatico.

In questo volume, volutamente, si è cercato di concentrare l'attenzione su procedimenti e problemi legati all'uso dei metodi quantitativi in linguistica, evitando di trattare specificatamente le questioni applicative e teoriche legate all'incontro con la linguistica computazionale. Tale scelta è dipesa da svariate ragioni, tra cui, non ultimo, il fatto che aprire uno spazio specifico alla linguistica computazionale avrebbe significato aggiungere un secondo volume a quello attuale, già piuttosto corposo.

Per la loro stretta relazione con alcune questioni dibattute in diversi saggi di questo volume si sono inclusi, per contro, due contributi che illustrano problemi applicativi di questo tipo: il saggio di Mastidoro e Amizzoni in cui vengono descritti alcuni strumenti per l'analisi e la gestione di materiale testuale (controparte applicativa di alcune questioni discusse soprattutto da Piemontese), e il saggio di Simonetta Vietri (la cui lettura va utilmente abbinata al lavoro di Elia incluso nella seconda sezione di questo volume) su alcune delle potenzialità funzionali del sistema integrato INTEX, efficace, non solamente per la costruzione di dizionari e grammatiche a stati finiti, ma anche, ed è qui quel che più interessa, per estrarre una gran quantità di informa-

zione linguistiche (morfo-sintattiche) fini per le successive elaborazioni statistiche.