LINGUAGGIO, MENTE E SOCIETÀ

a cura di
Ludovico Fulci e Enrico Sciubba

EUROMA

Indice

La chiave probabilistica delle lingue: teoria linguistica e applicazioni computazionali

di Isabella Chiari

Le lingue, i parlanti e i processi che sottostanno alla produzione e alla ricezione linguistica sono oggetti di studio di numerose discipline. Essi si incontrano e si scontrano oggi su un terreno applicativo e teorico nuovo, fornito dalla nascita negli anni Cinquanta del Novecento della linguistica computazionale. L'obiettivo di questo contributo è triplice: osservare alcune interazioni tra linguistica generale, linguistica dei corpora e linguistica computazionale; vedere come la linguistica e la statistica entrano nella costruzione di strumenti computazionali efficaci; riflettere sulle ricadute che l'incontro tra statistica e computazione ha sulla nostra rappresentazione delle lingue storico-naturali.

Il filo conduttore di questo percorso sarà il ruolo giocato dalla matematica, dalla statistica e dalla quantità – cose diverse -, sia usate per descrivere il linguaggio, sia interiorizzate dagli utenti nelle loro conoscenze implicite, nella linguistica computazionale e nella teoria del linguaggio. Quali sono le direzione della linguistica computazionale che tengono conto della dimensione probabilistica delle lingue? Quali aspetti sono centrali per i nuovi obiettivi che essa si pone? Quali caratteristiche del linguaggio sono messe in luce in modo da assumere nuovo significato per una teoria della produzione e della ricezione linguistica?

Esistono nella storia del pensiero linguistico linguistica diversi modi di considerare la dimensione quantitativa nella e

della lingua. Questo è stato l'oggetto di diverse discipline, ognuna con la propria etichetta, che sono sorte nel corso del Novecento: la statistica linguistica, la linguistica probabilistica, la linguistica matematica o linguistica quantitativa. Le differenze e le somiglianze sono spesso difficili da cogliere, sia dal punto di vista metodologico sia per il diverso apporto teorico dei rispettivi modelli linguistici proposti. Raggrupperemo entro il termine di «linguistica quantitativa» tutti i diversi approcci che si costituiscono all'intersezione della linguistica e della matematica e statistica dall'altra.

In campo puramente linguistico l'utilità degli strumenti matematici per lo studio del linguaggio è stata messa in luce già dai greci e dai romani che avevano individuato il comportamento peculiare delle parole cosiddette «di alto uso» e di quelle rare o uniche in un testo (*hapax legomena*). L'interesse per il modo in cui le parole ci si presentano con frequenze diverse, nei testi e all'apprendimento, è stato infatti oggetto di attenzione anche per crittografi, stenografi, insegnanti di lingue straniere: segno dell'evidente portata pratica derivante dall'osservazione degli aspetti quantitativi delle lingue. La vera svolta in questo campo, dal punto di vista storico, avvenne con la disponibilità dei calcolatori elettronici per automatizzare una buona parte del processo di spoglio, analisi e verifica dei dati.

1. Parole e numeri

La linguistica quantitativa ha assunto nella riflessione sul linguaggio diverse declinazioni, spesso anche in completa antitesi teorica. Vediamo brevemente in cosa consistono i principali tre filoni della linguistica quantitativa: (a) *approcci logico-matematici*, che mirano a fornire modelli matematici del funzionamento delle lingue; (b) *approcci descrittivi*, che mirano all'estrazione di regolarità statistiche da grandi quantità di raccolte testuali; (c) *approcci di tipo psicolinguistico* che intendo-

no sottolineare il ruolo dei processi probabilistici nell'apprendimento, nella produzione e nella ricezione linguistica.

Il primo approccio che, per varie ragioni, risulta anche quello più noto e più fortunato, è quello di tipo logico–matematico, in genere rappresentato meglio dalla dizione *linguistica matematica*, e si definisce con obiettivi di tipo modellistico e predittivo[1]. L'algebra e la teoria degli insiemi costituiscono il modello principale per la descrizione dei fatti linguistici in questa prospettiva, con l'elaborazione di apparati di assiomi, teoremi e corollari che rappresentano le strutture linguistiche a diversi livelli, dalla fonologia alla sintassi. Questo approccio si propone anche per i suoi aspetti predittivi, ossia per la capacità di prevedere, dato un modello, quali manifestazioni linguistiche possono o non possono occorrere.

Tre sono le sue principali direzioni: 1) lo studio della struttura delle categorie grammaticali; 2) la definizione di classi e di relazioni tra oggetti linguistici (modelli analitici del linguaggio); 3) la definizione delle cosiddette grammatiche formali (cfr. Gladkij 2002). L'area della linguistica matematica che costituisce il suo nucleo è la teoria dei linguaggi formali e l'elaborazione di grammatiche formali. Anche la posizione di Noam Chomsky si richiama a questo approccio, soprattutto nelle sue prime declinazioni, in particolare nella centralità che attribuisce alla sintassi e alla nozione di regola trasformazionale. La grammatica generativa è infatti un particolare tipo di grammatica formale, intesa come esplicita descrizione di un linguaggio formale, e può essere vista come un apparato deduttivo di regole, dalle quali è possibile generare le frasi del linguaggio e le loro rappresentazioni strutturali. La possibilità di costruire macchine che producano comportamenti linguistici è fondata sulla presenza di un *model-*

[1] Si tratta di un approccio molto comune nell'ex-Unione Sovietica, nei paesi dell'Europa dell'est, e negli Stati Uniti dagli anni Sessanta del Novecento in poi. Esempi ne sono i lavori del rumeno Solomon Marcus, del russo Igor Mel'chuk, di Zellig S. Harris, e, per alcuni aspetti, anche di Noam Chomsky e di Maurice Gross.

lo, un quadro astratto e formale, che contempli in dettaglio tutte le possibilità dell'interazione. La linguistica computazionale ha dunque, tra i suoi compiti principali, quello di definire i modelli che rendano possibile una *performance* adeguata da parte della macchina.

Chomsky ha determinato un vigoroso passaggio della linguistica novecentesca verso l'approccio modellistico (cfr. Ferrari 2000, p. 16)[2], a sua volta alla base di un settore tra i più notevoli della linguistica computazionale chiamato *Natural Language Processing* (ossia trattamento automatico del linguaggio naturale), con il principale obiettivo di implementare regole generali, date le quali fosse possibile far produrre al programma frasi ben formate della lingua (il programma potrà, per esempio, produrre la frase *la chitarra è un poco scordata*, ma non **poco la un scordata è chitarra*); e, data una serie di frasi ben formate di una lingua, fornire un'analisi dal punto di vista morfologico, sintattico, ecc.

Diversi fenomeni linguistici tuttavia mettono a dura prova la formalizzabilità, matematizzazione, calcolabilità e rappresentabilità delle lingue mediante regole: l'apertura dell'insieme dei segni possibili, l'estensibilità dei significati e la presenza di diverse e imprevedibili accezioni, la presenza di sinonimie e omonimie, la cristallizzazione di polirematiche e collocazioni, la presenza di ambiguità sintattiche e di complesse articolazioni pragmatiche, il continuo rimando linguisticamente riflessivo delle lingue e il gioco inestricabile di linguistico ed extralinguistico, di conoscenza ed esperienza.

---

[2] Uno degli aspetti centrali assunti in questa prospettiva è il presupposto della *discretezza degli elementi linguistici*, ossia il fatto che «elementi discreti sono definiti come tagli in insiemi di eventi continui» (Harris 1968, p. 6). Se si concepisce la lingua come integrazione di piani, ognuno dei quali discretizzabile, allora avremo un terreno che si adatta ad essere trattato e descritto mediante modelli formali. Altri presupposti sono la linearizzazione (cioè la possibilità di rappresentare gli elementi discreti in sequenze lineari, anche se i rapporti interni tra gli elementi non sono sempre lineari), la finitezza ed enumerabilità delle combinazioni di elementi per formare frasi, la formalizzabilità, e la deduttività dell'apparato matematico.

C'è da sottolineare inoltre che l'approccio della linguistica matematica è *non-quantitativo*, esattamente come l'algebra, a differenza della prospettiva che diremo della statistica linguistica. Questo significa che la matematica è presa come modello per il formalismo che descrive le grammatiche, permettendo non solo di descrivere i testi delle lingue, ma il loro più interno strutturarsi in grammatiche, che definiscono ciò che è e ciò che non è lingua. A interessare è la struttura formale della *lingua*, ossia del sistema, non dei testi, (o in termini saussuriani "atti di *parole"*, atti linguistici concreti unici e irripetibili).

All'opposto del precedente si trova l'approccio della *statistica linguistica*, che è invece focalizzato sui testi (scritti e parlati) per individuare e descrivere le regolarità statistiche mostrate dalle diverse unità testuali, con una particolare attenzione al lessico. Si tratta di una tradizione piuttosto antica. La crittografia (*kryptos* "nascosto", e *graphein* "scrivere") dal Cinquecento con Giambattista della Porta fino al Novecento, la steganografia (*stèganos* "nascosto") di Johannes Trithemius, e la stenografia (*stenós* "stretto, breve"), che si fa risalire a Senofonte, affrontano i temi relativi alla frequenza con la quale le lettere e le parole di un testo occorrono, oltre a quelli che interessano le caratteristiche che aiutano a definire la facilità o difficoltà di trasmissione di un messaggio o la sua possibilità di essere celato. Oggi la stenotipia raccoglie tali eredità con la costruzione di macchine per le trascrizioni, usate nei tribunali o in Parlamento, insieme alla crittografia informatica e alle tecniche di compressione dei testi per la trasmissione.

La linguistica ha raccolto queste eredità a partire dalla Scuola di Praga negli anni Trenta del Novecento con studi sulla fonologia statistica e sul lessico[3]. In questa prospettiva l'obiettivo è l'approssimazione alle concrete produzioni testuali, anche letterarie, con l'ambizione di coglierne il profi-

[3] Emergono, per le ricadute teoriche, soprattutto i lavori di George K. Zipf, Benoit Mandelbrot, Pierre Guiraud, Charles Muller e di Gustav Herdan.

lo linguistico e stilistico. Il nucleo è l'individuazione di tendenze e regolarità, in modo induttivo e quantitativo, e non regole (deduttive).

Per capire in maniera intuitiva il ruolo giocato dalle frequenze degli elementi linguistici basta osservare alcuni semplici esempi. Viaggia su internet la mail con il testo seguente: *sceondo uno sutdio dell'uvinesrita di Cmabrigde, l'odrine delle ltteree in una praloa non ipmrota, 'uinca csoa che h ipmrotatne h che la pirma e l'utlima saino al psoto guisto. il rseto puo eresse in un dsiodrine ttoale e ptotete smepre lggeree sneza porlbemi. E preche il crevlleo uamno non lggee ongi ltterea da sloa, ma la proala cmoe un isneime.*

Un qualunque italiano madrelingua non ha difficoltà a leggere questo testo, anche se le lettere sono state mescolate in maniera più o meno casuale, lasciando solo la prima e ultima nella posizione originaria. La capacità di ricostruzione del messaggio dipende dal fatto che vi è una notevole quantità di predicibilità e ridondanza nei testi, dovuta alle note differenze nelle frequenze delle lettere (e dei fonemi). Una 'Z' è molto meno frequente di una 'E' o di una 'R', ecc.

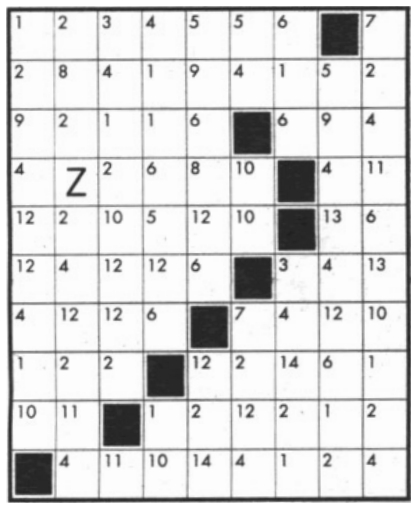

Fig. 1 Parole crociate crittografate

Ognuno di noi, seppure intuitivamente, conosce tali differenze, tanto da essere capace di svolgere il gioco enigmistico delle *parole crociate crittografate* (vedi Fig. 1).

Questo tipo di gioco enigmistico, tra i più semplici contenuti nella *Settimana enigmistica*, mostra quanto siano prevedibili e interiorizzate le frequenze delle lettere alfabetiche (e indirettamente dei nostri sistemi fonologici).

Nello schema infatti sono presenti caselle vuote ciascuna delle quali contiene un numeretto (da 1 a 26), che rappresenta univocamente una lettera dell'alfabeto. Non ci sono definizioni. È fornita solamente una piccola chiave, costituita da una parola di tre o quattro lettere (che associa i primi tre o quattro numeri). Il solutore deve ricostruire l'intero schema sulla base delle corrispondenze, senza alcuna indicazione sul significato delle parole che deve ricostruire. Come è possibile risolvere questo gioco se non appoggiandosi alle conoscenze che abbiamo, spesso senza sapere di avere, sul fatto che una E è certo più frequente di una Q, una I più di una Z, ecc.

Altri esempi di attività simili possono essere i diversi tipi di *cloze*, ossia di quel procedimento (usato soprattutto nei test di comprensione e negli esercizi di verifica delle competenze in una lingua straniera), per cui alcune parole - o lettere - di un testo vengono cancellate e devono essere ricostruite. Un esempio con le lettere possono èi in Esempio 1, mentre esempi di cloze sul vocabolario sono l'Esempio 2 e l'Esempio 3.

Esempio 1

_L / L_NG_ _GG_ _ / _ / G_V_RN_T_ / D_LL_ / PR_B_B_L_T_ (italiano)

Esempio 2

Dimenticavo di dire che (1)_____ signora Teresa ha avuto (2)_____ bella idea di presentarmi (3)_____ suoi parenti, facendomi passare (4)_____ un suo nipote "ospite (5)_____ di lei per un (6)_____ periodo di convalescenza," e (7)_____, colto di sorpresa, non (8)_____ la prontezza di contraddirla, (9)_____ dato il via a (10)_____ reazione a catena di (11)_____. Ben presto mi ritrovo (12)_____ in un intrico crescente (13)_____ parentele (Chiari 2002, p. 476).

Esempio 3

I am a _____ elderly man. The nature of _____ avocations for the last thirty _____ has brought me into more _____ ordinary contact with what would seem _____ interesting and somewhat singular set _____ men, of whom as yet nothing that I know _____ has ever been written (testo tratto da Melville, *Bartleby the scrivener*).

---

[4] Soluzione Esempio 1: il linguaggio è governato dalla probabilità.

A livello lessicale, ad esempio, già dagli anni Cinquanta, è stato osservato come le parole si distribuiscono nei testi in modo che poche vengano usate spessissimo, e tantissime parole presenti nei nostri vocabolari invece siano invece usate pochissimo. Il vocabolario di base della lingua italiana (cfr. De Mauro 1980), ad esempio, contiene nella sua fascia più interna, il vocabolario fondamentale, 2.000 parole che coprono circa il 90% delle occorrenze di un qualunque testo scritto o discorso parlato. La stratificazione statistica del lessico, simile in tutte le lingue, privilegia dunque alcune unità. Oltre alle parole grammaticali che costituiscono la tessitura ineliminabile dei testi, anche un nucleo di sostantivi, verbi, aggettivi, avverbi popola la stragrande maggioranza dei nostri testi, mentre altre parole vengono usate solo in testi specialistici o per determinati propositi stilistici. La mera registrazione di una parola nel vocabolario di base tuttavia non significa che si tratti di una parola semplice, quanto invece, come osservato da G. K. Zipf agli inizi del Novecento, che essa possiede una certa generalità semantica e pluralità di accezioni.

Anche lo studio sull'acquisizione lessicale dei bambini si è giovata dell'approccio quantitativo. Si tratta in questi casi di studi di carattere descrittivo che sottolineano la natura incrementale dell'apprendimento del lessico, mostrando anche la complessità della valutazione e dell'interpretazione dei dati relativi ai processi di acquisizione lessicale (cfr. Ferreri, 2005).

La statistica linguistica, pur avendo notevolmente privilegiato lo studio del lessico, ha compiuto progressi importanti nell'analisi delle caratteristiche quantitative dei sistemi fonologici, della sostanza dei suoni, in morfologia e in sintassi quanto nell'individuazione delle proprietà generali dei testi. Su questioni più prettamente testuali si sono misurati numerosi linguisti affrontando temi quali la scelta lessicale e la sua omogeneità, la valutazione della leggibilità, la relazione tra lunghezza delle parole o delle frasi e quella testo, concentrazione delle frequenze lessicali, la comparazione tipologica tra lingue diverse in testi simili, stilometria forense (per una ras-

segna bibliografica omnicomprensiva della letteratura su questo e gli altri settori della linguistica quantitativa cfr. Köhler, 1995). La valutazione delle caratteristiche linguistiche proprie di un testo (letterario e non) in particolare ha riscosso molto interesse a partire dallo studio di Pierre Guiraud del 1954, Sull'italiano, oltre ai lavori pionieristici di padre Roberto Busa, i primi lavori di statistica testuale sono di Heilmann (1961), Rosiello (1965), De Mauro (1966), Alinei (1971), Prosdocimi (1977), Cortelazzo (1984) e ancora prima la ricerca degli statistici Boldrini, Faleschini e Lonstergo (1948). Altri temi classici in quest'area sono le questioni relative alla misura e valutazione della ricchezza del vocabolario (si veda per tutti la monografia di Cossette, 1994), alla valutazione della correttezza teorica e statistica della nota legge armonica di Zipf sul rapporto tra rango e frequenza.

La posizione della statistica linguistica è generalmente di tipo descrittivo, ossia consiste nel mirare a individuare regolarità nella composizione testuale, che possono non essere evidenti durante la produzione o ricezione di un testo, ma che possono servire ad esempio per fornire un profilo statistico dell'uso lessicale di un autore, di un testo o di una data tipologia testuale. Poiché la statistica linguistica estrae regolarità dai testi, essa ha bisogno di poter accedere a vastissime quantità di materiale testuale e si associa quindi, pur non identificandovisi con la *linguistica dei corpora*[5].

Attraverso l'individuazione del profilo statistico di un testo è possibile, ad esempio, avere elementi per attribuirlo a un determinato autore, o classificarlo automaticamente come appartenente a un dato genere testuale. La statistica linguistica riveste inoltre anche un ruolo predittivo nel momento in

[5] La *linguistica dei corpora* si occupa dello studio, della gestione e dello sfruttamento di corpora testuali. Un corpus può essere definito come: «A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language» (Eagles, 1996:3) e un corpus elettronico in particolare è «a corpus which is encoded in a standardized and homogeneous way for open-ended retrieval tasks» (Eagles, 1996: 3).

cui, sulla base di ampie osservazioni campionarie, è possibile fornire un modello che definisca le caratteristiche linguistiche e statistiche di tutti i testi che appartenenti alla stessa categoria. Nonostante le metodologie della statistica linguistica siano molteplici e anche molto complesse e siano in fondo basate su una serie di ipotesi linguistiche sul comportamento dei testi, si può dire che il loro carattere sia essenzialmente di tipo induttivo e sperimentale, a differenza del metodo della linguistica matematica.

Dal punto di vista computazionale le informazioni statistiche possono essere usate sia per la costruzione di simulazioni matematiche, sia per lo sviluppo di strumenti tradizionali della linguistica applicata, come traduttori, dizionari elettronici, analizzatori grammaticali, lemmatizzatori, ecc. Nonostante l'attenzione dei linguisti per la dimensione quantitativa delle lingue sia stata continua durante tutto il Novecento, la linguistica computazionale, solo in tempi relativamente recenti, ha rivalutato l'utilità di tenere in considerazione anche questi aspetti nel disegno di applicazioni computazionali con il risultato di un arricchimento di prospettiva sia specificatamente connesso al successo dei metodi probabilistici sia alle possibili ricadute teoriche di tale successo.

L'approccio psicolinguistico diversamente da quelli fin qui presi in esame considera i fattori di frequenza come interni alle strategie di acquisizione linguistica, di produzione e di ricezione in contesti ordinari. Il presupposto teorico di questa prospettiva si fonda sull'idea che se l'apprendimento linguistico è basato essenzialmente sull'esperienza e l'uso, anche la competenza relativa alla frequenza degli elementi che si trovano a tutti i livelli linguistici viene interiorizzata e trasformata, mediante diverse forme di astrazione, in competenza sui costrutti generali e in strategie di produzione e ricezione linguistica. Si sottolinea dunque la centralità dell'interiorizzazione dei fattori statistici nella performance linguistica (produttiva e ricettiva), in fo-

nologia, fonotassi, accesso al lessico, ma anche nei meccanismi di lettura e scrittura e nell'apprendimento della lingua materna e delle seconde lingue. La centralità di questa dimensione è efficacemente riassunta da Nick Ellis: «Frequency is thus a key determinant of acquisition because "rules" of language, at all levels of analysis (from phonology, through syntax, to discourse), are structural regularities that emerge from learners' lifetime analysis of the distributional characteristics of the language input» (cfr. Ellis 2002a: p. 144).

Le informazioni sulle frequenze (dei fonemi, delle sillabe, dei lessemi, ma anche a livello sintattico) sono costantemente usate dagli utenti di una lingua. Gli esempi di pratiche risolutorie di giochi o test che abbiamo presentato nel paragrafo precedente, la loro relativa facilità è già una prima testimonianza della capacità dell'individuo di interiorizzare e, all'occasione, mobilitare le proprie conoscenze sulle frequenze. Ma tali conoscenze non emergono solamente in pratiche così marginali (e certamente, per molti versi, di tipo eminentemente metalinguistico), esse sono infatti messe in atto anche in maniera inconsapevole per la soluzione di quotidiani possibili fraintendimenti e di ordinaria comprensione.

Nella lettura, ad esempio, parole molto frequenti e con pronunzia regolare sono lette più rapidamente di parole più rare o imprevedibili nella conversione grafema-fonema (cfr. Coltheart 1978). Nell'ascolto il riconoscimento uditivo di una parola funziona meglio e più rapidamente se la parola è molto frequente (cfr. Luce 1986; Kirsner 1994), nella produzione lessicale vengono processate più rapidamente sequenze di parole frequenti, idiomatiche, e formule fraseologiche, soprattutto se si è sotto pressione (è il caso del commento o cronaca radiofonici o televisivi di eventi sportivi in cui l'azione è rapida, come mostra Kruiper, 1996). Anche nella ricezione e comprensione dei testi l'esperienza permette ai locutori di riconoscere e prevedere le forme più fre-

quenti di occorrenza dei verbi, le loro collocazioni, la struttura statistica e informativa delle parole che compongono un testo o un discorso.

L'approccio psicolinguistico si focalizza sui momenti di acquisizione e apprendimento e sulle fasi di processamento del materiale linguistico in produzione e ricezione. A differenza degli orientamenti precedentemente descritti è centrato più direttamente sull'utente (parlante e ascoltatore) come soggetto che usa in maniera più o meno consapevole informazioni sulle frequenze ricavate attraverso la sua esperienza. Si tratta di una prospettiva soprattutto indagata da psicologi del linguaggio, linguisti cognitivi, neuropsichiatri mediante osservazioni di tipo sperimentale[6].

2. Le direzioni della linguistica computazionale

Intorno agli anni Sessanta del Novecento la linguistica trova un terreno fertile di applicazione e nuovi impulsi teorici dal congiungersi di due diversi fattori. Il primo fattore dipende dal grande successo che riscuote nel mondo la teoria generativa e trasformazionale di Noam Chomsky, che propone una visione delle lingue fondata sulla *competenza linguistica,* intesa come capacità di produrre frasi ben formate e di esprimere intuitivamente giudizi di grammaticalità (dire se una data frase è grammaticale o no, ossia se appartiene alle frasi generabili con la grammatica di una data lingua) e afferma il ruolo centrale della nozione di *regola*, come insieme di condizioni necessarie e sufficienti a specificare la possibilità o impossibilità di generare un numero potenzialmente infinito di frasi ben formate, con un approccio radicalmente deduttivo. Tale orientamento è detto *approccio basato su regole* (o *modellistico*).

---

[6] Per una panoramica su questo tipo di approccio si vedano Ellis (2002a, 2002b) e Saffran, (2003).

Come si è detto, tale posizione, per la sua evidente formalizzabilità, è stata assunta a modello teorico per lo sviluppo di teorie linguistiche di tipo matematico o algebrico, le quali a loro volta hanno svolto la funzione di guidare l'implementazione delle grammatiche prodotte su sistemi informatici. Una conseguenza, non logicamente diretta, della posizione chomskyana è un deciso discredito della possibilità di costruire modelli del linguaggio fondati sulla teoria della probabilità e su materiale testuale prodotto di atti di *performance* linguistica (i *corpora*). Chomsky infatti afferma: «Evidently, one's ability to produce and recognize grammatical utterances is not based on notions of statistical approximation and the like» (Chomsky, 1957: pp.15-16). Si apre così una cesura radicale tra metodi (e applicazioni computazionali) di tipo basato su regole o grammaticali e metodi (e applicazioni) di tipo probabilistico e *corpus-based*.

Come si può vedere, l'avversione chomskyana allo studio dei corpora si riversa su due piani: la possibilità di descrivere il funzionamento della lingua mediante modelli probabilistico-statistici (mentre si afferma al contrario la "matematizzabilità" della grammatica), e la necessità di osservare la lingua attraverso le sue realizzazioni concrete, ossia i testi scritti o parlati (decretando di conseguenza la relativa inutilità della linguistica dei corpora per la comprensione dei fenomeni di produzione e ricezione linguistica).

Il secondo fattore che ha contribuito allo sviluppo della linguistica computazionale ai suoi albori fu la diffusione dei primi calcolatori potenti, in grado - opportunamente programmati - di svolgere complessi calcoli in tempi brevissimi. Le possibilità tecniche offerte dall'informatica congiunte a una certa potenza descrittiva del modello formale generativo hanno determinato la nascita di una serie di programmi di ricerca legati alla possibilità di sviluppare un connubio fertile tra informatica e linguistica. Nasce così nel 1962 l'Association of Computational Linguistics (ACL). La direzione che prende la linguistica computazionale negli anni a

seguire fu tutt'altro che rettilinea. Da punti di vista anche radicalmente diversi si dipanarono quelli che oggi sono i principali terreni di applicazione della disciplina.

La traduzione automatica (*machine translation*), ipotizzata alla fine degli anni Quaranta dall'ingegnere e matematico Warren Weaver, diventa realtà - controversa – nelle sue varie declinazioni di *traduzione assistita dal computer* (*computer-aided translation*, CAT), *traduzione automatica assistita* (*human-aided machine translation*), di *translator's workstation*, con l'impulso economico e di ricerca degli organismi internazionali come la Ue, la Nato, l'Onu e delle grandi industrie multinazionali, che tutt'oggi ne forniscono linfa vitale. Temi cari alla linguistica teorica come la complessità traduttiva della polisemia, delle collocazioni e polirematiche, dell'omonimia, dell'ambiguità sintattica diventano altrettante questioni applicative per cui necessita specifico trattamento con evidenti conseguenze sulla bontà del prodotto della traduzione.

L'elaborazione o trattamento del linguaggio naturale (*Natural Language Processing* – NLP - e *Understanding*), proprio a partire dalle posizioni chomskyane con l'influenza dei temi e delle applicazioni dell'intelligenza artificiale, assume un ruolo principe nella disciplina andando spesso ad identificarsi con essa. Tenendo come cardine della ricerca il livello sintattico, il NLP sviluppa strumenti di analisi sintattica come il *parsing*, o l'etichettatura grammaticale come il *Part-of-Speech tagging*. Anche in questo campo emergono questioni legate al funzionamento effettivo delle lingue naturali che possiedono ambiguità sintattiche (in *il medico visita il paziente con gli occhiali*: gli occhiali di chi sono?), semantiche (nella frase *Il calcio l'ha mandato in prigione*, il calcio è il gioco o l'atto di dare una pedata?), anaforiche (*l'uomo ha discusso con la ragazza del suo problema*, "suo" dell'uomo o della ragazza?), oltre a una manifestazione parlata molto diversa da quella scritta, che spesso presenta caratteristiche tutt'altro che ben formate grammaticalmente, ellittiche, parziali e incomplete (Esempio 5).

Esempio 4 - Frammento di dialogo dal LIP (De Mauro et alii, 1993: RC6)

C: Nicola siccome non c'ero pochi minuti fa vorrei sapere se_ quali quali persone
A: non c'eri?
C: no non c'ero quando avete parlato delle associazioni
A: ah va be'
C: eh quali persone possono aderire siccome noi abbiamo l'apporto del nostro laboratorio
A: questo
C: anche perche'
eh i fondi_ insomma le cinquemila lire di di non contrari
A: ma di gente di ruolo o di gente non di ruolo?
C: no no
A: eh? # ah no allora sia chiaro una volta
C: ah quello volevo dire

La lessicografia computazionale si occupa della elaborazione, ristrutturazione e utilizzazione dei dizionari tradizionali e della compilazione di *dizionari-macchina* (che contengono informazioni di carattere linguistico relative a specifici ambiti di applicazione - fonetico, sintattico, semantico, ecc. - da utilizzare in altre applicazioni computazionali, come quelle dell'NLP) e *dizionari informatizzati*, costruiti sulla base di strumenti di *Computer-Aided Traditional Lexicography* pur corrispondendo agli usi e alle funzioni principali dei dizionari a stampa tradizionali.

Le tecnologie della lingua parlata sono tra le applicazioni della linguistica computazionale che oggi sono massimamente al centro dell'attenzione. Basti pensare ai sistemi automatici che rispondono alle nostre richieste sull'orario ferroviario, ai sistemi di dettatura, alle voci sintetiche che ci forniscono dati e servizi telefonici. Le due aree che tradizionalmente costituiscono il fondamento delle tecnologie più avanzate sono quelle legate alla generazione o sintesi del parlato (*speech synthesis*) e al riconoscimento automatico di testi prodotti in parlato spontaneo (*speech recognition*). Accanto a tali aree oggi si collocano gli studi sul parlato multimodale (che mira a rappresentare insieme al se-

gnale fonico anche espressioni del viso, manifestazioni di emozione nella voce e nel corpo, movimenti della bocca) che permettono di creare simulazioni tridimensionali di teste parlanti (August, http://www.speech.kth.se/august/, Holger, Baldi, ma anche alcune versioni italiane come Lucia, sviluppata all'Istituto di Scienze e Tecnologie de lla Cognizione del CNR di Padova, soprattutto dal lavoro di Emanuela Magno Caldognetto e Piero Cosi, http://www.pd.istc.cnr.it/LUCIA). Ancora più all'avanguardia nel settore, per l'integrazione di diverse aree critiche della linguistica computazionale, si trova il «dialogo uomo-macchina», che serve a produrre sistemi che permettano a un utente non addestrato di entrare in relazione con una macchina usando una lingua naturale. Tali sistemi richiedono l'integrazione delle tecnologie del parlato con moduli di NLP, di carattere sintattico, semantico e moduli di intelligenza artificiale come il *problem solving* e modelli di gestione della interazione dialogica al fine di permettere operazioni quali insegnamento e *training* (per cui la macchina può tutorare diversi tipi di abilità come i *computer-assisted tutoring systems*), informazione (risposte a richieste degli utenti), comando (quando il dialogo è usato per esempio per guidare le azioni di un robot), assistenza (all'utente per prendere decisioni su problemi complessi).

E ancora tra le applicazioni computazionali di interesse generale vi sono l'*indicizzazione automatica,* che serve a produrre delle analisi rapide dei testi raccolti (per esempio sul web) attraverso la individuazione delle parole-chiave di un testo (*keyword extraction*); l'*information retrieval* e l'*information extraction*, che sono settori in fortissima espansione, proprio come conseguenza del fatto che quantità enormi di dati testuali sono immagazzinate elettronicamente per essere successivamente utilizzate e interrogate; il *text mining*, programma di ricerca che permette la categorizzazione e classificazione dei documenti, la tematizzazione, l'estrazione di relazioni tra dati e il suo riversamento sotto forma di database.

Applicazioni di *text mining* sono tipiche nei settori delle relazioni con il pubblico (smistamento di posta elettronica, *customer survey*), nell'analisi di specifiche tipologie testuali (analisi finanziaria, valutazione delle cartelle cliniche, ricerche su basi documentali giuridiche), nell'editoria e nelle telecomunicazioni, e nei sistemi di monitoraggio a fini di sicurezza (come nel caso della famigerata rete Echelon) in abbinamento ai cosiddetti sistemi esperti elaborati nell'intelligenza artificiale; la *summarization*, che consente di generare automaticamente riassunti di testi, rapporti estratti da dati strutturati e testi che estraggano informazioni rilevanti o pertinenti a partire da una base dati testuale.

Come si è visto non solo gli scopi e gli obiettivi della linguistica computazionale sono molteplici, ma anche le metodologie, le teorie e i modelli linguistici che sottostanno alle varie direzioni di ricerca sono diversi, a volte completamente contrastanti. Se pure il primo stimolo teorico venne dalla teoria generativa, non bisogna pensare che ad essa ci si è fermati. Convivono oggi infatti approcci assai diversi, e il predominio teorico chomskyano ha largamente lasciato spazio a posizioni spesso del tutto antitetiche nella teoria e negli interessi.

3. I testi nella linguistica computazionale

In che modo il paradigma *rule-based* è stato sfidato dal suo rivale approccio probabilistico e *corpus-based*? L'idea che la linguistica computazionale di stampo tradizionale sia fondata su una rappresentazione del funzionamento della lingua centrata sulla sintassi e radicata nella grammatica generativa è palesemente parziale e scorretta, anche se per molto tempo è stata identificata con tale paradigma. Fatto sta che le metodologie fondate *solamente* su un approccio modellistico si sono rivelate radicalmente fallimentari. Yorick Wilks (2006: p. 14) afferma in maniera perentoria che:

> […] it was generally assumed that the knowledge of the world and of language that a machine intelligence required could be programmed in directly, the content being provided by the researcher's intuition. In the case of language, this assumption followed directly from Chomsky's (1972) approach to linguistics: that intuitions about the nature of language can be computed by rules written by experts who have intuitive knowledge of their (native) language.
>
> All this has now turned out to be false: no effective systems have ever been built on such principles, nor (outside machine translation, perhaps) are they ever likely to be. The revolution that has replaced those doctrines holds that such knowledge, world or linguistic, must be gained from data by defensible (i.e. non-intuitionistic) procedures like machine learning.

In realtà non molti decenni dopo i primi tentativi di costruzione di strumenti di linguistica computazionale (di traduzione automatica e NLP), comparvero le prime applicazioni di tipo probabilistico. Un'applicazione di *tagging* grammaticale (ossia di programma per l'etichettatura che associa a ciascuna parola di un testo la appropriata categoria grammaticale in un dato contesto, vedi Figura 2) se utilizza un sistema *rule-based* contiene una grammatica che definisce le regole di formazione dei diversi possibili sintagmi di una data lingua. I primi *taggers* furono di questo tipo: come TAGGIT, adoperato negli anni Settanta per etichettare il *Brown Corpus of Standard American English*, primo corpus linguistico elettronico dell'inglese americano, con un tasso di successo senza intervento manuale di circa 77% delle occorrenze.

Se invece si utilizza un *tagger* probabilistico il sistema che permette di disambiguare le occorrenza (ossia di determinare se la forma *porta* in un dato contesto linguistico è sostantivo o verbo) si basa su statistiche di frequenza delle parti del discorso e delle loro sequenze. Da dove si derivano le statistiche utili per applicare l'annotazione a nuovi materiali? Si usano i cosiddetti *training corpora*, «allenatori» del *tagger*

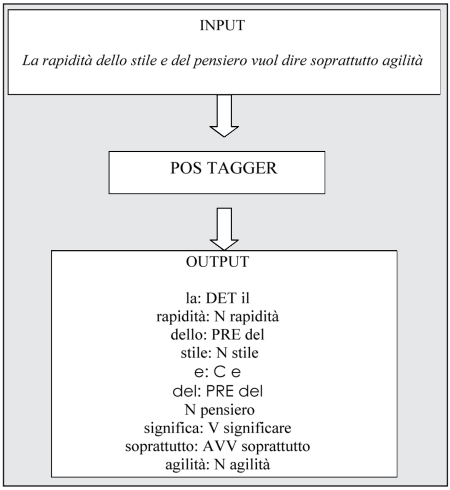

Fig. 2 Il POS tagging (cfr. Chiari 2007: 106)

nei quali le statistiche di occorrenza e di transizione sono derivate da corpora testuali già annotati in modo manuale. Uno tra i primi *taggers* probabilistici ad apparire, rimasto per questo famoso, è CLAWS (Constituent Likelihood Automatic Word-tagging System), sviluppato negli anni Ottanta all'Università di Lancaster, e usato nelle sue varie versioni per annotare il *British National Corpus* (corpus bilanciato dell'inglese britannico composto da circa 100 milioni di parole). Oggi un *tagger* probabilististico che analizza testi in lingua inglese raggiunge un grado di copertura corretta di etichette pari a circa 97-99% dei *token* (cfr. De Rose 1988).

Nel Natural Language Processing non solo *taggers* ma anche i *parsers* (analizzatori sintattici) oggi fanno uso di un modellamento statistico con training corpora e hanno dato vita al cosiddetto *Statistical Natural Language Processing* (cfr. Manning e H. Schütze, 1999). Anche la traduzione automatica ha sviluppato alla fine degli anni Ottanta un programma di ricerca di *Statistical Machine Translation* (SMT) che si fonda sull'accesso a corpora paralleli (in cui si trovano allineati testi originali e traduzioni) cui attingere per rilevare, sulla base delle porzioni da tradurre, strutture già tradotte utilizzabili per fare una sorta di «calchi» sugli esempi memorizzati nel corpus (*example-based* MT).

In lessicografia computazionale, sia nella costruzione di dizionari-macchina (per altre applicazioni computazionali) sia di dizionari informatizzati, l'approccio basato su corpora e sensi-

bile al dato statistico ha rivoluzionato il settore. Il primo prodotto di lessicografia *corpus-based* fu il *Collins Cobuild Advanced Learner's English Dictionary*, diretto dal linguista John Sinclair e basato sulla *Bank of English* (450 milioni di occorrenze), in cui il ricorso ai dati estratti dai corpora si fa sentire non solo nel prodotto stesso e nel suo disegno, ma anche nella totale revisione del lavoro che il lessicografo deve affrontare nella costruzione della risorsa. Si può con cognizioni di causa ad esempio indicare come prima accezione della parola inglese *gay* il significato di "omosessuale" e solo dopo indicare "gaio", basandosi sulle statistiche fornite dai corpora di riferimento.

Nelle *Speech Technologies* si è fatta strada la metodologia statistico-probabilistica spesso soppiantando completamente le forme concorrenti basate su regole. Nel riconoscimento del parlato si adoperano modelli probabilistici di tipo markoviano con buoni risultati, e si sono sviluppati modelli statistici che migliorano le prestazioni dell'applicazione mediante l'introduzione di fasi di addestramento. Alcuni sistemi attuali si servono di corpora di parlato spontaneo sotto forma di *training corpora* in modo da garantire un soddisfacente trattamento di diversi input linguistici. Si può parlare in questo caso di *corpus-based speech recognition*.

Un settore della linguistica computazionale che trasversalmente tocca tutti i precedenti è quello che si occupa di riprodurre il processo di «apprendimento» dall'esperienza mediante forme di inferenza statistica di nuove regole da materiale linguistico autentico e aggiungere potenza descrittiva e di trattamento alle singole risorse linguistiche. È il *machine learning*, programma che si suddivide in tre direzioni principali: apprendimento supervisionato (*supervised learning*, in cui il training è costituito da materiale già annotato e corretto manualmente); apprendimento non supervisionato (*unsupervised learning*, in cui il training è dato da materiale testuale grezzo e non trattato); apprendimento di rinforzo (*reinforcement learning*, in cui vi è una sorta di addestramento mediante premio o punizione secondo un modello comportamentistico). In generale la grande utilità dei

sistemi di apprendimento automatico è la possibilità, mediante algoritmi, di migliorare la performance dell'applicazione, ossia il suo comportamento di fronte a nuovi input. Il materiale da cui vengono estratte le regolarità statistiche è derivato da *training corpora*, che sollecitano il sistema a verificare la copertura di trattamento dell'applicazione su nuovo materiale.

La iniziale opposizione tra approccio basato su regole e approccio statistico si è andata gradualmente mitigando. Strumenti nati e sviluppati inizialmente come radicalmente *rule-based* come i *parsers* o i *taggers* sintattici e i sistemi di traduzione automatica per trasferimento sono oggi fruttuosamente integrati con moduli e sistemi di tipo probabilistico. I sistemi probabilistici, come si è visto, sono radicalmente legati alla utilizzazione dei dati empirici estratti da corpora.

Sottolineare il ruolo della statistica in linguistica computazionale finisce per coincidere con la assunzione della centralità dei corpora per questa disciplina. I corpora infatti non soltanto forniscono la base per l'estrazione delle regolarità di frequenza e co-occorrenza che servono per elaborare gli algoritmi di analisi degli applicativi probabilistici, ma permettono anche la soluzione di complessi problemi linguistici. Si pensi agli strumenti di traduzione automatica *example-based* con corpora paralleli e al ruolo giocato dalle *translation memory*. La maggioranza delle applicazioni computazionali di oggi non potrebbero funzionare ed essere realizzati senza l'accesso, l'elaborazione e il trattamento dei corpora, determinando una sorta di indiretta soggezione della linguistica computazionale dalla linguistica dei corpora.

Si determina così il circolo virtuoso della linguistica computazionale individuato dalla possibilità di creare applicazioni di NLP che servono per svolgere attività di trattamento e annotazione di corpora. Tali corpora vanno a loro volta ad affinare i modelli e gli algoritmi statistici degli strumenti computazionali e rendono possibile il loro training migliorando le prestazione degli stessi strumenti di NLP, che si applicheranno su nuovi corpora testuali.

4. L'approccio statistico è non-linguistico?

La linguistica computazionale di oggi non può dunque fare a meno confrontarsi la gestione di grandi corpora testuali che non solo costituiscono il terreno di applicazione dei software sviluppati, ma istituiscono anche un modello e una base da cui partire per estendere caratteristiche linguistiche presenti nel corpus di training a nuovo materiale testuale. In secondo luogo la linguistica computazionale si serve oggi di modelli probabilistici per tecniche di induzione di regolarità di tipo statistico dai corpora per la estrazione di caratteristiche linguistiche che spesso non sono né osservabili né esplicitamente parte di ciò che in senso tradizionale chiamiamo grammatica di una lingua.

I sistemi di tipo probabilistico o statistico spesso vengono definiti come sistemi «non linguistici», o a volte «anti-linguistici», poiché fanno uso solo delle probabilità di co-occorrenza e delle frequenze delle parole, piuttosto che di regole di restrizione di tipo linguistico o grammaticale (cfr. Somers 2003, p. 513). Che cosa si intende dunque per linguistico e non-linguistico? La attribuzione di un etichetta di «non linguisticità» alle metodologie statistiche si basa sulla considerazione del fatto che la frequenza assoluta, le co-occorrenze, gli n-grammi, la probabilità di transizione, il clustering, la similarità formale sono caratteristiche che riguardano il comportamento statistico di qualunque tipo di oggetto, non specificatamente linguistico. Non è necessaria, si suppone, dunque alcuna conoscenza della lingua oggetto di analisi per estrarre informazione statistica da un corpus di testi.

Da un lato la sensibilità umana verso la frequenza con la quale si manifestano determinati fenomeni è chiaramente generalizzata, ossia non si applica specificatamente al dominio linguistico, ma vige in qualche modo in tutte le esperienze e in tutte le forme di apprendimento. Vi è tuttavia, come si è mostrato nei paragrafi precedenti, una forma di competenza di carattere specificatamente linguistico-quantitativo, che solo in parte si esercita a livello metalinguistico.

La posizione che assume la non- o anti-linguisticità dei metodi statistici sottovaluta tuttavia ciò che la psicolinguistica e la glottodidattica oramai da decenni hanno reso piuttosto esplicito, ossia il fatto che i fenomeni di frequenza giochino un ruolo significativo a ogni livello della produzione e della ricezione linguistica. Allo stesso livello sintattico, principale dominio di interesse chomskyano, fenomeni di apprendimento e riconoscimento statistico si sono provati fondanti anche per la elaborazione di gerarchie e restrizioni sintattiche (cfr. Saffran 2003).

Risulta difficile oggi rappresentarsi non solamente gli atti/processi di apprendimento, produzione e ricezione linguistica senza tener conto dell'influenza svolta dalla conoscenza delle probabilità di occorrenza delle unità e degli effetti di frequenza. Ancora più difficile pensare che la conoscenza su queste proprietà delle lingue non sia specificatamente parte di quelle competenze che permettono ai parlanti – anche – di formare giudizi di grammaticalità, ad esempio, usando tali informazioni stratificate nelle esperienze linguistiche. Rimangono numerosi punti da discutere e numerose questioni teoriche – e solo in seconda battuta applicative – aperte sul tema.

Ci si può domandare se parte della competenza linguistica sia effettivamente anche di natura probabilistica e se parte della grammatica di una lingua contenga informazioni sulle caratteristiche statistiche degli elementi linguistici. Esiste una competenza grammaticale probabilistica? Quanta di questa competenza è di natura metalinguistica? La statistica fa parte della grammatica di una lingua come patrimonio sociale? O invece in diverse misure è parte delle competenze del singolo locutore, manifestandosi nelle attività produttive e ricettive? Come si può rendere conto di questi aspetti in una teoria della produzione linguistica? In che relazione si trova la competenza statistica con le altre forme di competenza linguistica? In che modo, eventualmente, queste competenze si compenetrano e integrano a vicenda?

RIFERIMENTI BIBLIOGRAFICI

Alinei, M. 1965. *La lista di frequenza della Divina Commedia*, In: *Miscellanea Dantesca*. Utrecht & Antwerp, 138-270.

Boldrini, M., L. Faleschini, e A. Lonstergo (a cura di). 1948. *Statistiche letterarie e altri saggi*, Milano: Laboratorio di statistica, 6.

Chiari, I. 2002. *La procedura cloze, la ridondanza e la valutazione della competenza della lingua italiana*, «Italica, Journal of the American Association of teachers of Italian», 79, n. 4, 2002, pp. 466-481.

Chiari, I. 2007. *Introduzione alla linguistica computazionale*. Laterza, Bari.

Chomsky, N. 1957. *Syntactic Structures.* Mouton de Gruyter, The Hague.

Coltheart, M. 1978. *Lexical Access in simple reading tasks*. In G. Underwood (ed.), *Strategies of information processing*, San Diego, Academic press, pp. 151-216.

Cortelazzo, M. 1984. *La dialettologia quantitativa in Italia*, In: H. Goebl (hrsg.), *Dialectology*, Bochum: Brockmeyer, 1-14.

De Mauro, T. 1966. *Alcuni aspetti quantitativi della lingua della Commedia*, In: *Dante e la Magna Curia*, Atti del convegno di studi Palermo-Catania-Messina, 7-11 novembre 1965, Centro di studi filologici e linguistici siciliani, Palermo, 1-15.

De Mauro, T. 1980 (1991[11]; 2003[12]): *Guida all'uso delle parole*. Roma, Editori Riuniti.

De Mauro, T. e I. Chiari (a cura di) 2005: *Parole e numeri. Analisi quantitative dei fatti di lingua*. Aracne, Roma.

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

De Rose, S.J., 1988: *Grammatical category disambiguation by statistical optimization*, in «Computational Linguistics», 14, 1, pp. 31-39.

Eagles. 1996. *Text Corpora Working Group Reading Guide. EAG-TCWG-FR-2*. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica computazionale.

Ellis, N. 2002a: *Frequency Effects in Language Processing. A review with Implications for Theories of Implicit and Explicit Language Acquisition,* in «Studies in Second Language Acquisition», 24, pp. 143-188.

Ellis, N. 2002b: *Reflections on Frequency Effects in Language Processing,* in «Studies in Second Language Acquisition», 24, pp. 297-339.

Ferrari, G. 2000: *Livelli di analisi del testo. Due approcci a confronto*, in *Linguistica e Informatica. Corpora, multimedialità e percorsi di apprendimento*, a cura di R. Rossini Favretti, Bulzoni, Roma, pp. 15-27.

Ferreri, S. 2005: *L'estensione delle conoscenze lessicali individuali*, in De Mauro, T. e I. Chiari (a cura di) 2005: *Parole e numeri. Analisi quantitative dei fatti di lingua*. Aracne, Roma, pp. 307-334.

Gladkij, A. V. 2002: *Mathematical linguistics*, in *Encyclopedia of Mathematics*, a cura di M. Hazewinkel. Springer-Verlag, Berlin Heidelberg New York.
(anche al sito http://eom.springer.de/M/m062650.htm).

Guiraud, P. 1954: *Les caractères statistiques du vocabulaire*, P.U.R., Paris.

Heilmann, L. 1961. *Statistica linguistica e critica del testo*, in *Studi e problemi di critica testuale,* 173-82.

Kirsner, K. (1994*). Implicit processes in second language learning*. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 283–312). San Diego, CA: Academic Press.

Köhler, R. 1995. *Bibliography of Quantitative Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kuiper, K. 1996. *Smooth talkers: The linguistic performance of auctioneers and sportscasters*. Mahwah, NJ: Erlbaum.

Luce, P. A. 1986. *A computational analysis of uniqueness points in auditory word recognition*. «Perception and Psychophysics», 39, pp. 155-158.

Manning e H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Mass.-London 1999.

Prosdocimi, A. L. 1977. *Sull'applicazione di metodi statistici e computazionali a corpora epigrafici*, In: A. Zampolli e N. Calzolai (a cura di), *Computational and mathematical linguistics*, Firenze: Olschki.

Rosiello, L. 1965. *Consistenza e distribuzione statistica del lessico poetico di Montale*, «Rendiconti», 11 : 397-421.

Saffran, J. R. 2003. *Statistical language learning: Mechanisms*, in «Directions in Psychological Science», 12**,** 110-114.

Somers, H., 2003. *Machine Translation: Latest Developments*, in *The Oxford Handbook of Computational Linguistics*, a cura di R. Mitkov, Oxford University Press, Oxford, pp. 512-528.

Wilks, Y. 2006. *Artificial Companions as a new kind of interface to the future internet*, in «Oxford Internet Institute, Research Report 13», October 2006, pp. 1-19.