Aio

68

Isabella Chiari

Informatica e lingue naturali

Teorie e applicazioni computazionali per la ricerca sulle lingue

Copyright © MMV
ARACNE editrice S.r.l.

www.aracneeditrice.it
info@aracneeditrice.it

via Raffaele Garofalo, 133 A/B
00173 Roma
06 93781065

ISBN 88–7999–985–

I diritti di traduzione, di memorizzazione elettronica, di riproduzione e di adattamento anche parziale, con qualsiasi mezzo, sono riservati per tutti i Paesi.

Non sono assolutamente consentite le fotocopie senza il permesso scritto dell'Editore.

I edizione: febbraio 2005

Indice

Introduzione

Questo volume è stato pensato primariamente come manuale introduttivo allo studio della linguistica computazionale e in generale ai temi legati alle interazioni tra lingue storico-naturali e informatica.[1] È indirizzato dunque soprattutto a studenti che siano inseriti in corsi di laurea che richiedano lo svolgimento di questa disciplina (soprattutto lettere, scienze della comunicazione, mediazione linguistica, lingue e letterature straniere, informatica umanistica). Sono dunque date per scontate alcune nozioni minimali di linguistica generale, richiamate comunque sinteticamente nel testo. Altrettanto valga per le nozioni informatiche implicate nel testo.

Il secondo tipo di destinatario del volume è colui che sia già impegnato in ricerche di tipo linguistico e che voglia essere iniziato agli strumenti tecnologici e applicativi che si stanno rapidamente sviluppando nel campo della linguistica computazionale.

Per entrambi i destinatari si è inclusa una sezione di rassegna di alcuni strumenti sviluppati per la linguistica computazionale, in particolare i grandi corpora linguistici progettati per le principali lingue europee moderne, i software applicativi per l'analisi e l'elaborazione statistica dei testi, e un piccolo censimento dei centri più noti e delle associazioni che trattano i temi legati a questa disciplina. In tutto il volume si è cercato di dare spazio ai temi connessi con la multiculturalità e la comparazione linguistica, soprattutto in relazione alla traduzione interlinguistica.

Nel primo capitolo saranno forniti alcuni spunti storici e teorici che stanno alla base delle elaborazioni proposte all'interno della linguistica computazionale. Si mostreranno i principali approcci linguistici e i requisiti teorico applicativi che hanno dato vita ai principali filoni di ricerca legati a questa disciplina, che come si vedrà è costitutivamente disomogenea. In particolare si vedranno i contributi della teoria generativa, della teoria dell'informazione, della statistica linguistica e dell'intelligenza artificiale.

Il secondo e il terzo capitolo sono dedicati alla linguistica dei corpora, ossia a quella branca degli studi linguistici che si occupa di osservare i testi e analizzarli con gli strumenti informatici. Il secondo capitolo espone i temi di base per la costruzione e interrogazione dei corpora, mentre il terzo tratta specificazione delle annotazioni linguistiche e degli standard di codifica.

---

[1] La presente edizione è una versione di lavoro provvisoria da considerarsi come *pre-print*, pensata per gli studenti di *Linguistica generale e computazionale* della Scuola Superiore per Mediatori Linguistici "Carlo Bo" di Roma, e per gli studenti del corso di *Informatica umanistica* della Facoltà di Scienze Umanistiche dell'Università "La Sapienza" di Roma, per l'anno accademico 2003/2004.

Il quarto capitolo affronta uno dei temi centrali della linguistica computazionale ossia il *Natural Language Processing* (da alcuni identificato 'tout court' con la linguistica computazionale), o trattamento automatico del linguaggio.

Nel quinto capitolo è fornita una breve panoramica sulla traduzione automatica, uno dei primi e tuttora centrali obiettivi della linguistica computazionale. In particolare sono presentati i principali modelli di traduzione in uso per le applicazioni informatiche attuali.

Il sesto capitolo affronta le applicazioni tecnologiche legate alla lingua parlata. Si tratta di un settore tra i più recenti della linguistica computazionale che, soprattutto in vista dei possibili risvolti economici, ha avuto un certo sviluppo nelle applicazioni per la sintesi del parlato e per il riconoscimento vocale. Si fornisce inoltre anche una panoramica di uno dei settori maggiormente sviluppati e più antichi della linguistica computazionale, ossia la lessicografia: l'elaborazione e compilazione di dizionari elettronici.

L'ultimo capitolo vuole essere invece uno strumento di consultazione e di lavoro. Sono presentati i maggiori progetti di corpora delle principali lingue europee moderne (con particolare attenzione all'italiano) e i software più noti per condurre analisi di tipo linguistico su testi e corpora. È stata inoltre aggiunta una tabella delle principali espressioni regolari usate nelle ricerche complesse su corpora e dizionari e un elenco di alcuni centri e associazioni in Italia e nel mondo nel campo della linguistica computazionale.

Per quanto riguarda i riferimenti bibliografici, dato il carattere didattico del testo, si è preferito ridurli al minimo nel testo demandando alcuni riferimenti più significativi alle sezioni bibliografiche "Altre letture" che accompagnano ogni capitolo. Tali sezioni sono pensate come guida minimale all'approfondimento dei temi principali trattati nel capitolo, non mirano dunque ad essere esaustive. Per lo studente sono pensate anche le domande di verifica e le esercitazioni.

Capitolo 1

La linguistica e i computer

I rapporti tra la linguistica e le scienze esatte e applicazioni tecniche, compresa l'informatica, sono sempre stati piuttosto turbolenti. A diversi livelli gli umanisti, letterati e linguisti, sono stati affascinati e sedotti dalle scienze in molti modi: usandole a modello metodologico, a modello metaforico, prendendo in prestito strumenti di analisi, rigore sperimentale e molto altro nel corso dei secoli, ma soprattutto nella seconda parte del Novecento.

In particolare l'informatica, essendo una tecnica di recentissimo sviluppo, si è effettivamente incontrata con la linguistica e le lingue in generale solo dagli anni Cinquanta in poi. Il cuore dell'informatica infatti non risiede nella progettazione della componente *hard* dei computer (microchip, circuiti, porte, motherboard, ecc.), ma della componente *soft* (ossia della gestione dei comandi, del linguaggio di programmazione e del trattamento delle conoscenze). L'informatica è definita in un noto dizionario come la "disciplina che si occupa della raccolta e del trattamento delle informazioni o, più specificamente, dell'elaborazione di dati per mezzo di calcolatori elettronici" (DM, 2000). Dato che la linguistica si occupa dell'analisi e dello studio dello strumento più potente che l'uomo possieda per esprimere informazioni era inevitabile il loro incontro. Si può dire metaforicamente che l'essere umano è un elaboratore di informazioni, se con informazione indichiamo qualunque forma di contenuto (vi sono contenuti emotivi, proposizionali e descrittivi, rappresentazioni simboliche). Le informazioni elaborate dall'uomo sono di moltissimi tipi, alcuni dei quali inaccessibili anche in linea di principio a una macchina, e una delle operazioni più complesse e determinanti per la vita e lo sviluppo di una società umana è appunto il linguaggio, l'espressione di tali contenuti per molteplici scopi diversi. Un calcolatore elettronico che deve trattare e usare informazioni dovrà dunque servirsi di qualche forma di linguaggio. È naturale che il primo luogo in cui osservare i meccanismi dell'elaborazione e dell'espressione sia dunque l'uomo stesso.

Negli ultimi 20-30 anni a moltissimi livelli vi è stata un'integrazione e progressiva assimilazione degli strumenti informatici negli studi linguistici e letterari e contemporaneamente anche in campo tecnologico si è cominciato a guardare alle lingue come possibili fonti e obiettivi di innumerevoli applicazioni informatiche. Così i contributi teorici della teorica dell'informazione, della statistica linguistica, dei linguaggi formali e dell'intelligenza artificiale si sono incontrati con le incredibili possibilità tecniche offerte dallo sviluppo dei primi processori e negli ultimi decenni dei microprocessori, della conservazione di grandissime quantità di dati multimediali, delle comunicazioni e trasmissioni di informazione in rete.

1.1 Che cos'è la linguistica computazionale?

Negli ultimi decenni del Novecento, insieme alla specializzazione e settorializzazione di numerose branche della linguistica, è nata anche la cosiddetta "linguistica computazionale". Rispetto ad altre branche, come la sociolinguistica, la psicolinguistica o la neurolinguistica, la linguistica computazionale si è configurata immediatamente come un settore frammentario, poco omogeneo e caratterizzato da programmi di ricerca fortemente distanti tra loro.

La linguistica computazionale può essere definita come il luogo d'incontro tra linguistica teorica (e applicata) e tecnologie informatiche. Forse l'unica caratteristica in comune ai diversi programmi che si raggruppano sotto questa branca di studio è proprio il fatto di congiungere problemi teorici e/o applicativi relativi al linguaggio con problemi teorici e/o applicativi relativi all'informatica e ai computer.

La parola *computazionale* ne richiama altre due: *computare*, che significa semplicemente "calcolare" in senso generale (dal latino COMPUTĀRE, composto di CON– "assieme, con" e PUTĀRE "calcolare") e *computazione* ossia "elaborazione elettronica di dati". La linguistica computazionale è una linguistica che fa calcoli ed elabora dati (linguistici evidentemente). Ma la linguistica computazionale è una branca della linguistica oppure dell'informatica? La domanda ha una certa rilevanza anche dal punto di vista teorico, e la difformità delle risposte rende conto della complessità e disomogeneità dei contributi a questa disciplina. Anche il tipo di competenze necessarie al linguista computazionale sono tecniche specialistiche ma richiedono anche una considerazione teorico-linguistica ben fondata e attenta alle caratteristiche meno formalizzabili delle lingue.

Per questo motivo la linguistica computazionale è essenzialmente **interdisciplinare** e multidisciplinare. Tutte le differenti branche della linguistica, dalla fonologia, alla morfologia, alla sintassi, possono essere utilmente studiate sotto un profilo computazionale. L'informatica, ma anche la statistica, la matematica, e l'intelligenza artificiale forniscono strumenti e metodi per le analisi linguistiche e per le loro applicazioni.

Le applicazioni computazionali possono avere dunque almeno due tipologie di obiettivi:

- lo sviluppo di strumenti informatici per lo studio e la ricerca specialistica delle lingue;
- lo sviluppo di applicazioni informatiche destinate al grande pubblico, ossia software che sfruttano le competenze linguistiche per produrre programmi di utilità generale (come traduttori, dizionari, sintesi del parlato, ecc.).

Del primo tipo sono dunque le applicazioni informatiche destinate a specialisti del linguaggio, che mediante tali strumenti, tentano di portare alla luce

caratteristiche delle lingue altrimenti non rilevabili. Di tali strumenti si serve in genere, per esempio, la cosiddetta **linguistica dei corpora** (*corpus linguistics*), di cui parleremo più avanti, che esamina grandi quantità di produzioni linguistiche, scritte o parlate, osservandone le caratteristiche: il lessico, la sintassi, le cosiddette 'collocazioni', la catena fonica, le strutture morfologiche. La linguistica computazionale, per favorire tale studio, ha sviluppato strumenti informatici di analisi automatica o semi-automatica dei testi che evitano al linguista di analizzare e cercare per così dire 'manualmente' i dati linguistici. Un qualsiasi software di questo tipo, anche il più semplice, è capace per esempio di ordinare le parole di un testo in ordine alfabetico, di indicare quante volte una parola appare nel testo, ossia il numero delle sue occorrenze, la sua frequenza, con quali parole si accompagna più spesso. Contare, osservare i contesti, ordinare per frequenza o secondo qualunque altro criterio le parole di un testo (per non parlare dell'impegno necessario per un'analisi fonetica o fonologica) è un lavoro che, fatto senza l'ausilio di un computer e di un programma adeguato, risulta improponibile quando il testo preso in esame è molto lungo; e ovviamente, molte osservazioni linguistiche hanno senso solo se sono generalizzabili e osservabili in un numero significativamente ampio di testi. La gestione e l'analisi computazionale dei testi risulta dunque l'unico mezzo di osservazione di molte caratteristiche del linguaggio.

Il secondo approccio prevede invece l'uso di conoscenze linguistiche di diversi tipi per l'elaborazione di applicazioni informatiche di uso comune. Molti di noi oggi usano quotidianamente piccoli programmi che correggono l'ortografia dei testi che scriviamo al computer (o almeno propongono correzioni). Sono i cosiddetti correttori ortografici, o *spell-checkers*. Anche i dizionari che consultiamo su CD-ROM sono basati su analisi parzialmente automatiche del lessico, e ci forniscono informazioni, e soprattutto modi di accedere alle informazioni, totalmente diverse, flessibili e potenti, rispetto alle relative versioni cartacee. Dal CD-ROM del *Dizionario della lingua italiana* (Paravia), curato da Tullio De Mauro, per esempio, possiamo in pochi istanti sapere quanti e quali verbi dialettali sono oggi entrati a pieno titolo nella lingua italiana (e sono 13, *arrazzare, burlare, coccoveggiare, grifare, intorcinare, intorcinarsi, intruppare, sbiluciare, scalfare, sfacchinarsi, sprangare, svirgolare, zinnarsi*). Semplici applicazioni linguistiche d'uso comune sono anche i dizionari e i thesauri (dizionari dei sinonimi), spesso inglobati nei programmi di elaborazione elettronica dei testi. Programmi più sofisticati ci permettono invece di far leggere al nostro computer un testo scritto con una voce naturale, sono le applicazioni *text-to-speech* (dal testo al parlato), o all'inverso esistono programmi di riconoscimento vocale e software di dettatura, mediante i quali possiamo convertire il nostro parlato in un testo scritto, senza dover digitare sulla tastiera. Queste sono alcune delle più diffuse applicazioni per il mercato orizzontale della linguistica computazionale, ossia

software destinati a un pubblico non specializzato, elaborati per servire a scopi di interesse generale.

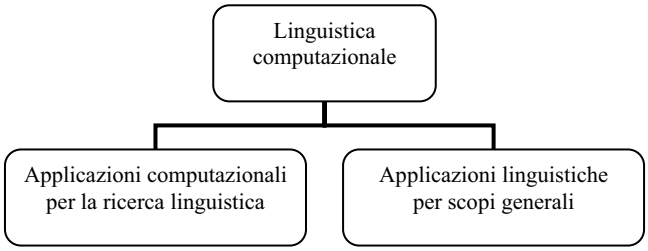

Figura 1 - Le applicazioni della linguistica computazionale

Il **filone teorico-linguistico** è legato a scopi prima di tutto connessi con una migliore comprensione del funzionamento dei vari livelli linguistici e solo in seconda battuta ha sviluppato obiettivi di carattere applicativo. Di questa tipologia fanno parte sia la già citata **linguistica dei corpora** sia una parte della **linguistica generativa**. Ed è singolare notare che questi due approcci che oggi maggiormente nutrono e sviluppano la linguistica computazionale siano dal punto di vista teorico per molti aspetti agli antipodi.

Il secondo **filone**, quello primariamente **applicativo**, nasce invece soprattutto in ambienti informatici con scopi commerciali (spesso sponsorizzati da grandi aziende) e concepisce l'oggetto linguistico come un qualsiasi contenuto (base di informazione). Spesso i progetti di questo tipo appaiono efficaci e comparativamente hanno fornito buoni risultati pratici sin dagli esordi, scontrandosi tuttavia in un secondo momento con una sorta di barriera oltre la quale non si ottenevano ulteriori miglioramenti nelle prestazioni.

I due filoni teorico-linguistico e applicativo non si sono incontrati che molto recentemente soprattutto in conseguenza del fatto che alcuni settori applicativi hanno raggiunto una condizione di stallo che ha avuto bisogno della linfa vitale offerta da approcci ben fondati teoricamente e soprattutto dall'integrazione con i dati reali forniti dai grandi corpora linguistici.

La linguistica computazionale non è dunque identificabile né come una specifica metodologia, né con uno specifico dominio di ricerca, né tanto meno con una teoria linguistica definita.

1.4 Contributi collaterali

Il primo impulso per la nascita della linguistica computazionale è venuto dal miraggio di poter costruire un programma che implementato su una macchina permettesse di tradurre automaticamente i testi da una lingua a un'altra. Si tratta di quel settore della linguistica computazionale che oggi chiamiamo **traduzione automatica**. Alla fine degli anni Quaranta, dopo la Seconda

guerra mondiale, W. Weaver suggerì, spinto dai risultati e dagli spunti offerti dalla teoria dell'informazione, che fosse possibile predisporre un programma che eseguisse, senza l'intervento umano, delle traduzioni da una lingua all'altra.

Dalle prime applicazioni in questo settore, che per l'eventuale ricaduta economica avevano richiamato numerosi finanziamenti, si passò a progetti di interesse più teorico e generale e all'idea che anche altri aspetti del linguaggio potevano essere utilmente trattati con strumenti informatici.

I principali paradigmi teorici che hanno arricchito lo sfondo delle ricerche computazionali sono la teoria generativa chomskiana e i linguaggi formali, la teoria dell'informazione di Shannon, la statistica linguistica e l'intelligenza artificiale. Una buona definizione dell'intelligenza artificiale è quella data da Kurzweil (1990) "The art of creating machines that perform functions that require intelligence when performed by people". All'interno dell'area di ricerca che va sotto questa etichetta vanno moltissime linee teoriche e applicative diverse. Una macchina che possa anche solo lontanamente essere paragonata per intelligenza a un essere umano, anche se non supera il noto e famigerato test di Turing, dovrà comunque operare qualche tipo di ragionamento e inferenza, possedere delle conoscenze e saperle attivare, essere capace di apprendere e soprattutto essere in grado di comprendere un linguaggio e di usarlo.

1.4.1 Il paradigma di Chomsky, modelli e linguaggi formali

Un impulso di poco successivo ai suggerimenti di Weaver si ebbe con i lavori del linguista americano Noam Chomsky, che con *Le strutture della sintassi* (1957), propose una nuova considerazione del linguaggio che prevedeva la possibilità di una completa formalizzazione delle lingue in regole, dalle quali sarebbe stato possibile dedurre l'insieme delle frasi ben formate della lingua stessa. Chomsky tuttavia non vedeva l'utilità di una implementazione di tali regole su una macchina, non vedeva dunque la sua applicabilità computazionale (né tanto meno vedeva l'utilità di integrare tale approccio con una considerazione statistica, che come si vedrà, procede spesso di pari passo con le indagini computazionali). Altri al posto suo hanno però perseguito tale obiettivo dando vita a un intero settore della linguistica computazionale chiamato ***Natural Language Processing*** (ossia trattamento – automatico – del linguaggio naturale). Questo ambito di ricerca si propone due obiettivi: l'implementazione di regole generali date le quali sia possibile far produrre al programma frasi ben formate della lingua (il programma potrà produrre la frase *il cane riposa sotto la scrivania*, ma non **scrivania la sotto il cane riposa*); e data una serie di frasi ben formate di una lingua, fornire un'analisi dal punto di vista morfologico, sintattico, ecc. Dunque data la frase sopra citata, ci potrà dire per esempio che *il cane* è un sintagma nominale e *sotto la*

scrivania è un sintagma preposizionale che dipende dal sintagma verbale retto dal verbo *riposare*.

L'approccio ispirato ai lavori di Chomsky pone al centro dell'attenzione da una parte la **competenza linguistica**, intesa come capacità interiorizzata anche se non esplicita del parlante nativo di discriminare tra frasi grammaticali e non grammaticali; dall'altra sottolinea come tale competenza sia formalizzabile in una serie di **regole** astratte che definiscono la grammatica di una lingua.

L'approccio che si ispira a queste considerazioni è detto **approccio modellistico** (cfr. Ferrari, 2000: 16). In questa prospettiva sono riferimenti teorici centrali non solo la linguistica generativa, ma anche la teoria degli automi, l'intelligenza artificiale e la teoria dei linguaggi formali. Uno degli sviluppi più significativi in questo campo è il *parsing* (di cui parleremo meglio più avanti, cap. 4) che consiste in un dispositivo (soft, astratto) che a partire da una serie di simboli in sequenza (la frase in *input*), opera una segmentazione (il parsing appunto) e restituisce un'analisi della sintassi della frase (albero in *output*).

Un ulteriore polo di ricerca della linguistica computazionale oggi molto perseguito è costituito proprio dall'integrazione del programma precedente con i metodi statistici, che aumentano notevolmente le capacità e la potenza dei modelli: si tratta del cosiddetto ***Statistical Natural Language Processing***. L'obiettivo di questo programma di ricerca, che descriveremo sinteticamente in seguito, è appunto quello di migliorare le capacità di produzione o analisi di frasi, servendosi della statistica. Quando parliamo e comprendiamo gli altri facciamo spesso, senza rendercene conto, ricorso a considerazioni statistiche sulle frequenze (per esempio sulla frequenza con la quale troviamo due parole o due tipi di parole una accanto all'altra in sequenza), appellandoci di conseguenza a considerazioni basate sulla nostra conoscenza del mondo e sulle situazioni comunicative. Di fronte all'affermazione *il gatto attacca la ragazza con le unghie affilate*, la nostra prima interpretazione semantica e di conseguenza la nostra prima analisi morfo-sintattica ci porta a considerare le unghie affilate come appartenenti al gatto e non alla ragazza, e dunque a considerare il sintagma preposizionale *con le unghie affilate* dipendente dal verbo direttamente e non dal sintagma nominale subordinato *la ragazza*. Questo perché statisticamente ci capita più di frequente di parlare di unghie affilate di felini, piuttosto che di umani (benché alcuni umani le possiedano). Lo *Statistical Language Processing* mira a introdurre considerazioni di questo tipo per migliorare la prestazione dell'applicazione computazionale nei due tipi di compito: produrre e analizzare frasi.

1.4.2 La teoria dell'informazione

Un'altra fonte costante per alcuni approcci della linguistica computazionale è costituita dalla **teoria dell'informazione**, nata – per così dire – nel 1949 con la pubblicazione del saggio *La teoria matematica della comunicazione* di Claude Shannon e Wendell Weaver. La teoria dell'informazione non è una teoria della comunicazione in un senso semiotico-linguistico, come invece sembra suggerire il titolo del volume che ne ha definito la nascita. Claude Shannon (1916-2001), cui si deve l'elaborazione matematica e teorica vera e propria, non aveva un obiettivo tanto ambizioso. Shannon era stato crittografo durante la Seconda Guerra Mondiale e aveva studiato con particolare attenzione i codici e le strategie per trasmettere messaggi segreti in modo che fossero difficilmente decrittabili dalle forze nemiche, ed elaborando una teoria dei codici segreti (cfr. Shannon, 1949).

Il problema da cui parte l'elaborazione della teoria dell'informazione è un problema essenzialmente pratico e applicativo ossia la trasmissione materiale di informazioni su un canale. In particolare Shannon vuole determinare e misurare la capacità di un canale, inteso come mezzo fisico nel quale si trasmette un segnale, e in relazione a tale capacità stabilire quale sia il miglior sistema di codifica dei messaggi per garantire la ricezione dall'altra parte del canale.

Il termine **informazione** usato in questo contesto è molto diverso dall'accezione quotidiana, legata a contenuti o significati. Si può dire che la nozione di informazione venga usata da Shannon a significare un sistema di unità qualunque entro il quale colui che produce un messaggio sceglie una specifica unità (informazione) da trasmettere. La trasmissione è qualcosa di pratico e materiale che consiste nel fare in modo che una qualunque variazione di uno stato fisico (onde sonore, porzioni grafiche, o tattili, ma anche in linea di principio olfattive o gustative) passi su un canale e venga ricevuta correttamente da un apparecchio di ricezione di tali variazioni.

A Shannon non interessava dunque capire quale tipo di contenuti passassero sul canale, ma semplicemente avere la garanzia che le variazioni corrispondenti a qualunque ipotetico contenuto venissero ricevute senza errori. Non era dunque il sistema di conoscenze e contenuti semantici codificati a essere indagato, ma il sistema astratto che ne esprime le probabilità di scelta. Una informazione è rappresentata da una scelta tra due possibilità equiprobabili e viene misurata in *bits* (*binary digits*) che indicano quante scelte binarie è necessario compiere in un insieme dato di alternative equiprobabili. L'informazione è inoltre connessa con l'incertezza. È facile intuire che la portata di un'informazione molto prevedibile (per esempio dire "Buon giorno" quando si entra in una stanza) è molto minore di un'informazione totalmente imprevedibile (come esordire con "Arrivederci"). Dunque in un certo senso una informazione è proporzionale al livello di incertezza e sorpresa che si produce nell'uditore o ricevente. Dopo aver atteso il verificarsi di un

evento (tra un numero finito di eventi possibili), al prodursi dell'evento otteniamo un'informazione, e scopriamo dunque quale caso si verifica effettivamente eliminando così l'incertezza. L'informazione consiste proprio nella rimozione dell'incertezza della situazione iniziale. Maggiore è l'incertezza, maggiore sarà l'informazione.

Il modello proposto dalla teoria dell'informazione è abbastanza noto e prevede una serie di elementi nel processo di trasmissione (cfr. Figura 2):

- la **sorgente** o **produttore** o **emittente**, ossia il dispositivo che produce una variazione di stato fisico;
- il **ricevente** o **destinatario**, ossia il dispositivo che rileva le variazioni di stato fisico;
- il **messaggio**, ossia l'insieme delle variazioni di stato fisico emesse dal produttore (collegabili a una serie di contenuti anche complessi che però non entrano nel processo di trasmissione);
- il **canale**, ossia il mezzo fisico su cui viene trasmesso;
- la **codifica**, ossia il processo attraverso il quale sono formati i messaggi e la **decodifica** ossia il processo di ricostruzione del messaggio da parte del ricevente;
- il **rumore**, ossia il possibile disturbo che può intervenire sul canale;

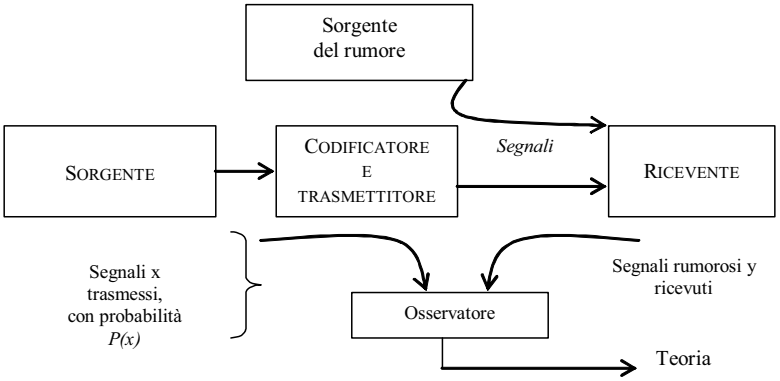

Figura 2 - Il modello di Shannon (*fonte*: Cherry, 1957: 199)

Il segnale ricevuto viene osservato, analizzato e, mediante un processo decisionale, avviene un'identificazione di ciò che potrebbe essere stato inviato dalla sorgente. Dunque nella teoria dell'informazione vi è una generale preminenza teorica affidata al momento della ricezione. In funzione dei problemi che si possono incontrare in questa fase va elaborata una adeguata codificazione per ottenere successo nella comunicazione.

Il contributo più proficuo e conosciuto della teoria di Shannon è dunque l'elaborazione del suo *teorema fondamentale per un canale discreto con disturbo*, secondo cui «è possibile inviare informazioni alla velocità *C* attraverso il canale con una frequenza di errori o una equivocazione piccola

quanto si vuole mediante una opportuna codifica» (1949: 77). Il teorema afferma dunque che una codifica che permetta la trasmissione senza il prezzo di una velocità troppo bassa è sempre possibile. Nelle comunicazioni, dice Shannon, «vi sono, comunque, dei modi di trasmettere le informazioni i quali sono ottimali nel combattere il disturbo» (ivi: 72). Si implica quindi che l'introduzione di sistemi appropriati di ridondanza possa garantire una buona trasmissione al di là di ogni possibile rumore. La ridondanza infatti è una codificazione del messaggio che ripetendo alcune porzioni oppure variando le probabilità di occorrenza delle unità permette di garantire la corretta ricezione del messaggio (cfr. Chiari, 2002, cap. I).

La teoria di Shannon rimane un passo fondamentale nella storia del Novecento poiché è alla base di tutte le applicazioni informatiche, ivi compresa la trasmissione via internet. Per diverse ragioni la teoria dell'informazione è diventata un punto di riferimento, anche se spesso negativo, per molte elaborazioni teoriche in campo linguistico. In particolare Shannon si rese immediatamente conto di come le variazioni nelle probabilità di occorrenza degli eventi influenzassero grandemente il processo di trasmissione e anche di ricezione delle informazioni. Dato che in genere ogni sistema di comunicazioni veicola messaggi e unità in modo non casuale, né regolare, ma con frequenze differenziate secondo alcuni principi, tenendo conto di tali variazioni probabilistiche e anche delle restrizioni che si impongono alle sequenze di unità, è possibile elaborare una migliore codificazione e trasmissione delle informazioni.

L'importanza di quelli che potremmo dire "scompensi" statistici delle unità linguistiche (o semiotiche) è dunque centrale nella comprensione dei fenomeni non solo di produzione, ma anche di trasmissione e ricezione linguistica. Da Shannon in poi l'uso delle conoscenze statistiche sui codici ha notevolmente migliorato, come si vedrà nei prossimi capitoli, le prestazioni della maggior parte dei modelli attuali di linguistica computazionale. Tra i meriti della teoria dell'informazione va sicuramente annoverata quindi la centralità e preminenza della dimensione statistico-probabilistica rispetto alla dimensione logico-formale (dominante invece nei modelli di stampo generativo).

1.4.3 La statistica linguistica

A gettare luce sulle caratteristiche quantitative dei codici linguistici ha provveduto anche la linguistica soprattutto novecentesca, con quella branca che va sotto il nome di **statistica linguistica** (ma anche a volte *linguistica quantitativa* o *linguistica matematica*). La linguistica da sempre ha evidenziato più o meno sistematicamente l'utilità degli strumenti matematici per lo studio del linguaggio. Già i greci e i romani avevano osservato una significativa differenza nella frequenza dei diversi tipi di parole, distinguendo le parole di alto uso da quelle rare fino agli *hapax legomena*. La vera svolta, dal punto di

vista storico, per l'elaborazione dei dizionari fondamentali e di frequenza fu la disponibilità dei calcolatori elettronici per automatizzare una buona parte del processo di spoglio e analisi.

Nel Novecento diverse teorie linguistiche hanno fatto uno a diverso titolo di strumenti matematici. Da una parte la linguistica matematica ha sviluppato modelli formali che si sono incontrati con le formalizzazioni di stampo logico (cfr. Marcus, Nicolas e Stati, 1970; Gladkij e Mel'chuk, 1983), dall'altra vi è stata una tensione verso le considerazioni quantitative delle occorrenze testuali. In quest'ultima ottica si collocano una serie di studi su tutti i livelli di analisi linguistica legati soprattutto ai nomi di George K. Zipf e Benoit Mandelbrot negli Stati Uniti, di Pierre Guiraud e Charles Muller in Francia, dei praghesi e di Gustav Herdan nell'Europa orientale.

Un tratto comune agli autori appena citati è una maggiore attenzione alle realtà testuali rispetto alle formalizzazioni astratte delle regole linguistiche. A interessare sono soprattutto le caratteristiche dei testi parlati o scritti concretamente prodotti e raccolti in corpora nella loro dimensione fonologica o grafemica, morfo-sintattica, testuale e stilistica, ma soprattutto lessicale.

Linfa vitale per la statistica linguistica è fornita da ambiti periferici e applicativi. I primi studi statistici delle strutture linguistiche li dobbiamo principalmente agli **stenografi** o a ricercatori che intendevano fornire strumenti per facilitare l'elaborazione di metodi stenografici. Uno dei primi lavori di lessicografia statistica è stato il *Häufigkeitswörterbuch der deutschen Sprache* (1898) coordinato da F. W. Kaeding che nacque appunto a scopo stenografico, e che operava uno spoglio di circa 11.000.000 di parole (in modo ovviamente manuale e non elettronico), al fine di individuare le frequenze dei grafemi, delle sillabe e delle parole della lingua tedesca.

A segnare l'inizio di più accurate analisi linguistiche sempre a scopi stenografici fu invece *Gammes sténographiques* (1907) di J. B. Estoup nel quale si individuò la legge che lega la frequenza al rango di una parola e si definì per la prima volta la centralità della nozione di rango. Il rango è il posto occupato da una parola posta in una lista di frequenza decrescente: dunque il rango di una parola molto frequente è basso (la parola più frequente ha rango 1).

Come abbiamo detto, il dominio che ha polarizzato maggiormente l'interesse dei linguisti statistici è stato infatti il lessico. Da considerazioni soprattutto sulla composizione e le caratteristiche del lessico parte anche quello che potremmo definire il primo fondatore della statistica linguistica ossia **George Kingsley Zipf**. Il quadro entro cui si muovono le ricerche di Zipf parte da una visione del linguaggio che mira ad adottare una metodologia analoga alle scienze esatte, attraverso l'applicazione di principi statistici. Si tratta di una visione globale del linguaggio (da lui chiamata anche *filologia dinamica*) che per definizione non può fare a meno di tener conto dei contenuti del discorso, né delle caratteristiche personali, sociali e culturali degli utenti del linguaggio (Zipf, 1935).

Zipf osserva una serie di regolarità nella composizione delle parole che costituiscono il **lessico** di una lingua, in particolare osservando queste parole come occorrono nei testi effettivamente prodotti:

I. La lunghezza delle parole tende a mantenersi in un rapporto inverso al numero di occorrenze nei testi. Come conseguenza di queste osservazioni Zipf (1935) individua la legge seguente: la lunghezza di una parola tende a presentarsi legata da una relazione inversa – non necessariamente proporzionale, ma probabilmente secondo una funzione matematica non lineare – con la sua frequenza relativa. Questa legge per Zipf implica una tendenza del linguaggio a mantenere un equilibrio tra la lunghezza e la frequenza delle parole, e una soggiacente legge di economia;

II. Il numero delle parole differenti sembra diventare più grande in rapporto alla diminuzione della frequenza di occorrenza, la varietà e la brevità. Vi sarebbe dunque una tendenza a mantenere un equilibrio tra la frequenza e la varietà. Un testo è costruito per lo più da un ristretto numero di parole di alta frequenza e da numerose parole di bassa frequenza;

III. Il prodotto della frequenza di una parola per il suo rango è costante (detta *legge di Zipf-Estoup* o anche solo *legge di Zipf*);

IV. le parole più frequenti sono semanticamente più generiche.

Anche a **livello fonologico** esistono una serie di regolarità statistiche che Zipf mette in evidenza e cerca di spiegare:

I. Vi è un equilibrio tra il grado di complessità di un fonema e la sua frequenza relativa di occorrenza, nel senso che il grado di complessità è in rapporto inverso alla frequenza relativa;

II. Il sistema fonetico di ciascuna lingua cerca di mantenere costantemente questo equilibrio;

III. Il mantenimento di questo equilibrio è la probabile causa dei cambiamenti fonetici che producono le scissioni dialettali che terminano infine nella creazione di nuove lingue;

IV. Quanto più un fonema è frequente tanto meno tende a essere nettamente articolato (detta anche *legge Zipf-Martinet*);

V. Il numero di fonemi di una parola è direttamente proporzionale al suo rango, cioè decresce con l'aumento della frequenza (detta anche *legge Zipf-Guiraud*);

Queste e altre regolarità statistiche delle lingue dipendono non dal caso, bensì dalle caratteristiche di finitezza psico-biologica dell'essere umano. Ogni attività umana infatti per Zipf è governata dal *principio del minimo sforzo*. Questo principio governa la totalità del comportamento umano (non solo dunque quello linguistico) e gli conferisce una certa economia.

È difficile provare in modo definitivo la correttezza di tutte le ipotesi statistiche suggerite da Zipf. Come sottolinea George Miller (1965), Zipf aveva una visione scientifica del linguaggio secondo la quale erano considerati rilevanti gli aspetti biologici, psicologici, sociali e soprattutto statistici delle lingue, ma allo stesso tempo le sue idee e alcune delle sue applicazioni ma-

tematiche mostravano una certa ingenuità, giustificata però dall'approccio matematico-statistico che all'epoca in cui Zipf lavorava poteva essere considerato in tutto pionieristico. Il contributo di Zipf alle scienze del linguaggio è infatti da molti considerato controverso: basti citare l'osservazione di Benoît Mandelbrot (1974: 313), che definisce Zipf «autore di numerosi libri che combinano strettamente e in modo inconsueto verità e follia».

Per quanto riguarda i contributi teorici e applicativi europei, particolarmente attiva, soprattutto per gli studi di statistica lessicale, è la Francia con due personalità: Pierre Guiraud e Charles Muller. Uno dei contributi più significativi del francese Pierre Guiraud (in particolare in *Les charactères statistiques du vocabulaire*, 1954) fu l'aver esplicitato il modo con il quale le parole si distribuiscono statisticamente nei testi: poche, pochissime parole coprono una percentuale altissima della maggioranza dei testi (le prime 100 coprono circa il 60%, le prime 1.000 circa l'85%, e così via), mentre vi sono un grandissimo numero di parole registrate nei comuni vocabolari che sono rare, rarissime, quasi assenti. Mentre al cecoslovacco Gustav Herdan si deve invece soprattutto una visione dei caratteri della lingua come alternarsi tra scelta del parlante e casualità (*choice* e *chance*). Anche la scuola di Praga produce molti risultati di linguistica quantitativa, ereditati con uno sforzo teorico notevole anche dal funzionalista André Martinet che li applicò allo studio dei mutamenti fonetici con l'elaborazione tra le altre della nozione di rendimento funzionale (cfr. Martinet, 1955).

Un'altra fonte di particolare interesse per la statistica linguistica è la **didattica delle lingue**. Ci si rese infatti conto che l'insegnamento delle parole più frequenti migliora sensibilmente le produzioni linguistiche di un apprendente, soprattutto in considerazione del fatto che la maggior parte dei testi sono costituiti dalle stesse 6.000-7.000 parole. Nella prima metà del Novecento si incominciarono quindi a produrre i cosiddetti ***word books*** contenenti le parole del vocabolario di alto uso, partendo dall'inglese (cfr. Thorndike, 1932; Faucett e Maki, 1940; Fries e Traver, 1940; Thorndike e Lorge, 1944), e successivamente prodotte per il francese (cfr. Vander Beke, 1929; Henmon, 1924), per lo spagnolo (cfr. Buchanan, 1927; Keniston, 1933) per il tedesco (Pfeffer, 1964) e per altre lingue come svedese, russo, portoghese. Per la lingua italiana la prima lista di parole di alta frequenza fu curata da Knease (1931).

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 tutte le altre. Per l'inglese è molto noto il *Basic English* (1928) di Ogden e Richards. Per la lingua francese, si registra il caso più noto di dizionario fondamentale creato con forti basi statistiche (cfr. Gougenheim, 1958). In un secondo tempo compaiono dizionari fondamentali anche per il tedesco (Pfeffer, 1970) e le altre lingue europee. Per l'italiano compaiono un volume di Bruno Migliorini (1943), che tuttavia non ha una solida costruzione statistica, Sciarone (1995)

elaborato su liste di frequenza precedenti, e soprattutto i lavori di De Mauro che sono confluiti prima nella costituzione del Vocabolario di base (De Mauro, 1980) e successivamente in senso più propriamente lessicografico nella serie di dizionari per fasce d'età e scolarità pubblicati con l'editore Paravia.

Ancora di stampo statistico e lessicografico ricordiamo i **lessici di frequenza** di molte lingue, che hanno iniziato a farsi strada dagli anni Sessanta ad oggi. I lessici di frequenza coniugano diversi aspetti della ricerca computazionale (a seconda della ricchezza e complessità della progettazione): linguistica del corpus, analisi statistiche, lessicografia, analisi terminologica. Si sviluppano sulla scia di una tradizione molto antica che si occupa di individuare le frequenze e le concordanze delle parole presenti in testi letterari o religiosi. Ciò che veniva prima calcolato e ordinato manualmente, con l'avvento dei primi elaboratori permette spogli molto più ampi. Pionieristico in questo campo che coniuga le analisi linguistiche statistiche con esigenze computazionali è stato il lavoro di Roberto Busa sull'*Index Thomisticus* (1951) negli anni Sessanta. A Busa si deve inoltre anche il primo manuale di *informatica linguistica* (1987). I lessici di frequenza, cogliendo l'eredità degli studi statistici su concordanze, hanno per la prima volta un obiettivo più ampio: descrivere non la lingua di un autore o di un opera, ma la lingua viva, parlata e/o scritta di una intera comunità linguistica, la lingua così come manifestata in una moltitudine di generi testuali (lettere, diari, lezioni, conversazioni, articoli giornalistici, teatro e cinema, ecc.). Per l'inglese il primo esempio è il famoso Kučera e Francis (1967) *Computational analysis of present-day American English*, seguito molti altri lavori; per il francese e lo spagnolo rispettivamente Juilland, Brodin e Davidovitch (1971) *Frequency Dictionary of French Words*, e Juilland e Chang-Rodriguez (1964) *Frequency Dictionary od Spanish Words*. Per l'italiano ricordiamo invece il LIF (*Lessico di frequenza della lingua italiana contemporanea,* 1971) di Bortolini, Tagliavini e Zampolli e il più recente LIP (*Lessico di frequenza dell'italiano parlato*, 1993) di De Mauro, Mancini, Vedovelli e Voghera.[1]

L'utilità dello studio statistico del linguaggio combinata con le grandissime possibilità di trattamento, gestione e diffusione dei dati offerte dalle tecnologie informatiche hanno fortemente potenziato una branca della linguistica computazionale chiamata *linguistica dei corpora* (cap. 4).

Altre letture

Come introduzione ai problemi generali che riguardano il rapporto tra linguistica, linguaggi formali e aspetti computabili del linguaggio, ricco di riferimenti bibliografici, è il lavoro di Elisabetta Gola (2002) "Linguistica com-

[1] Dei lessici di frequenza si parlerà più in dettaglio nel cap. 3; mentre per i lessici collegati a corpus accessibili si veda § 7.1).