Pubblicato in:
"Parlare è comunicare?", a cura di A. Corrado e G. Di Martino, ESI, Napoli, pp.199-212.

Il paradigma della scienza statistica per l'analisi del linguaggio [1]

Domenico Piccolo
Dipartimento di Scienze Statistiche
Università degli Studi di Napoli Federico II
Via L. Rodinò 22 - 80138 Napoli
E-mail: domenico.piccolo@unina.it

1. Introduzione

In questo contributo intendiamo riflettere sui rapporti fra comunicazione, informazione e linguaggio che, a nostro parere, reclamano una relazione più intensa con la Statistica intesa come scienza e metodo di investigazione dei fenomeni collettivi. Lo faremo, in maniera deduttiva e ragionata, in modo che le argomentazioni siano disponibili (ed eventualmente confutabili) anche per chi non è abituato ad utilizzare l'apparato formale proprio degli studi statistici.

Il lavoro è così organizzato: nel paragrafo seguente introdurremo il senso delle relazioni tra comunicazione, linguaggio ed informazione insistendo sulle caratteristiche che essi assumono nella società contemporanea. Quindi, espliciteremo i contenuti sostanziali della disciplina statistica nella misura in cui essa si configura come una teoria della conoscenza sostenuta dall'evidenza empirica ed inquadrata in modo razionale in schemi formali, soggetti a verifiche e miglioramenti progressivi. Inoltre, evidenzieremo alcune situazioni sperimentali nelle quali l'analisi statistica consente di formalizzare con cura l'indagine linguistica apportando utili informazioni aggiuntive rispetto alle acquisizioni tipiche derivabili dalle scienze del linguaggio. Alcune considerazioni finali concernenti l'utilità della metodologia statistica negli studi linguistici chiudono il lavoro.

2. Aspetti linguistici nelle relazioni umane

La società contemporanea può essere descritta in molti modi e secondo molte prospettive ma, sicuramente, un aspetto che la caratterizza è la quantità, diffusione e velocità delle relazioni umane che attraversano il pianeta con modalità crescenti e persino impensabili sino a qualche decennio fa. Le stesse dimensioni di spazio e di tempo si fondono nella misura in cui la relazione tende a divenire istantanea e dialogica, un carattere che costituisce elemento fondante dei rapporti umani e motivo di successo dei recenti strumenti di comunicazione interpersonali (cellulari, Internet, etc.).

Anche se tale accelerazione potrebbe costituire un intralcio nell'analisi del fenomeno sopra indicato, crediamo che sia possibile descrivere alcuni elementi essenziali derivanti da una

[1] Il lavoro ha beneficiato dei commenti critici e costruttivi di Amelia Bandini, Angela D'Elia, Francesca Di Iorio, Francesca Marone, Emilia Mauriello, Francesca Vaghi, Marco Venuti che hanno letto una versione preliminare del testo; resta ovviamente all'Autore la responsabilità di quanto affermato. Si ringrazia la Sezione Linguistica del Dipartimento di Scienze Statistiche, Università degli Studi di Napoli Federico II, per gli stimoli quotidiani ricevuti a favore di un approfondimento delle relazioni tra Linguistica e Statistica, nonchè le strutture di ricerca del CFEPSR, Portici, che hanno concorso in modo significativo al completamento del lavoro.

riflessione delle componenti e della struttura della vita di relazione. In altri termini, un elemento decisivo per tale riflessione è la constatazione che *ogni relazione umana interpersonale nasce da un bisogno, si manifesta mediante uno strumento e trasmette un contenuto.*

All'origine di qualsiasi relazione tra persone (e quanto diciamo può estendersi a relazioni tra gruppi, comunità strutturate, nazioni, etc.) vi è un bisogno primario connaturato all'essere umano e alla sua vita quotidiana, collegato alle necessità biologiche e psicologiche di apprendere e conoscere, crescere e riprodursi, difendersi ed espandersi. Questo bisogno genera la ***comunicazione***, intesa come esigenza primaria ed insopprimibile di porsi in relazione ad un altro, interagendo e ricevendo comunicazioni di analogo spessore.

Storicamente, tale bisogno necessita di strumenti che l'ambiente, la cultura e via via le innovazioni tecnologiche forniranno alla persona in modo da soddisfare adeguatamente l'esigenza della comunicazione: gesti, simboli, segnali, immagini, scritti, codici, testi sempre più elaborati, etc. costituiscono modalità strumentali della comunicazione interpersonale e sono dettati di volta in volta dalla specificità dei messaggi, dal contesto in cui si effettuano, dalla cultura e dalla relazione interpersonale che si è instaurata (o che si intende instaurare). In tal senso, lo strumento diviene manifestazione di avviso o di chiamata, di allarme o di affetto, di paura o di amore, di legislazione o di regolamento, e così via.

Tuttavia, perché la comunicazione sia percepita e produca effetti, risposte o comportamenti (e quindi divenga quella relazione interpersonale che ha generato la comunicazione medesima), gli interlocutori devono convenire su regole che rendano lo strumento efficace ed efficiente, cioè in altri termini devono convenire su (o definire) un ***linguaggio***. Per sua natura, tale regolamentazione si diffonde, si modifica e si arricchisce in un popolo organizzato e strutturato sulla base di vincoli spaziali, tribali, religiosi, economici, politici, culturali, sociali, etc. ed allora il linguaggio diventa *lingua*, intesa come *corpus* autonomo entro il quale la comunicazione si struttura in forma orale e scritta, si trasmette tra generazioni e viene assunta come modalità principale (e spesso esaustiva) della comunicazione.

Infine, la comunicazione veicola un suo proprio contenuto (informale o organizzato, semplice o complesso) che costituisce la ragione stessa della sua esistenza e questo contenuto è ***informazione.*** Tale informazione può essere parola, testo, grafico, immagine, libro, enciclopedia, ma nella sua semplicità costituisce una unità elementare che si trasmette da una persona all'altra e che consente il dialogo, il confronto, l'arricchimento, e così via.

Più avanti, evidenzieremo come questi aspetti abbiano costituito momenti fondanti dell'evoluzione umana e, soprattutto, della specificità dell'*Homo sapiens* rispetto a tutte le altre creature viventi. In ogni caso, la prospettiva qui evidenziata, peraltro abbastanza consolidata nell'ambito degli studi sul linguaggio e sulla comunicazione, intende sottolineare come la vita di relazione abbisogni sempre di un linguaggio strutturato entro il quale si pone la specifica informazione che si intende trasmettere.

3. Contenuti disciplinari della scienza statistica

Per avvalorare lo stretto legame tra disciplina statistica ed analisi linguistica, ripercorriamo brevemente gli elementi di una interpretazione della Statistica che giustificano tali rapporti; ciò è necessario soprattutto perché -a differenza di discipline ben più consolidate, come la Fisica o la Biologia- il contenuto proprio della Statistica è tuttora confuso e sovrapposto con aspetti marginali della sua investigazione, sottacendo troppo spesso il suo intrinseco carattere interdisciplinare.

La Statistica, etimologicamente nata come "scienza dello Stato" e sviluppatosi come "metodo di indagine dei collettivi" (reali e virtuali) è oggi divenuta, secondo una concezione che si va consolidando a livello internazionale "*la scienza delle decisioni in condizioni di incertezza*". Pur essendo presente sin dagli albori delle comunità primitive per la misurazione delle risorse umane (a fini militari, religiosi, economici, sanitari, etc.) e strettamente collegata agli organismi di governo (prima tribali e nazionali, poi comunitari e planetari), la Statistica assume un contenuto proprio di investigazione scientifica della realtà, quando assume al suo interno (per necessità ontologica) la teoria della probabilità. Infatti, a partire dall'inizio del '900 essa deve –sempre più spesso- rispondere alla permanente questione sperimentale della validità delle ipotesi: "*quanto è casuale ciò che osservo sperimentalmente? quanta forza posso derivare dall'analisi della realtà? che validità generale posso attribuire (sempre e ovunque) a questo risultato sperimentale (ottenuto ora e qui)?*"

Queste domande sono state poste originariamente dalla Genetica, dalla Biologia, dalla Psicologia, ma si sono diffuse in Fisica e Chimica nella misura in cui il principio dell'indeterminismo è divenuto il paradigma fondamentale della conoscenza sperimentale. Progressivamente, Medicina e Sociologia, Economia e Marketing assunsero come elemento centrale lo studio sperimentale effettuato con metodi statistici e, negli anni 1930-40, la Statistica ha sviluppato una serie di risultati inferenziali che esaltano il suo ruolo decisionale, in tutti i campi della conoscenza dedotta dalle evidenze sperimentali.

Purtroppo, mentre questa visibilità scientifica della Statistica è ben evidente per gli studiosi sperimentali, l'insegnamento (superiore ed universitario) della Statistica per troppo tempo è stato circoscritto alla gestione dei dati con mere finalità descrittive, producendo diffidenza e sospetto verso le analisi statistiche, essendo esse avulse da quella dimensione di incertezza intrinseca che invece le caratterizza. Segnale e conseguenza di tale atteggiamento sono le correnti utilizzazioni dei sondaggi intesi come risolutori di problemi invece che come strumenti conoscitivi la cui lettura va *sempre* accompagnata da una misura di qualità dell'indagine che ne espliciti i margini di dubbio.

In effetti, noi riteniamo che la Statistica abbia senso ed assuma valore (oggi, molto più che nel passato), perché riesce a pervenire dalla conoscenza sperimentale alla verifica di schemi concettuali e meccanismi logici, mediante i quali l'Uomo razionale ottimizza le sue regole decisionali: ciò avviene all'interno di una procedura probabilistica, e non matematica, ove l'incertezza assume un peso rilevante e consente di procedere in modo efficiente.

A tal fine, nel nostro insegnamento, introduciamo la Statistica come una disciplina che deriva da un percorso concettuale tipico della conoscenza umana e della sua razionalizzazione in funzione della fase decisionale. Esso, sequenzialmente, si snoda mediante le tappe della percezione, della comprensione e dell'azione.

In primo luogo, riceviamo sensazioni e percezioni che provengono dalla natura sensoriale della nostra esistenza e, rispetto alla totalità delle altre specie viventi, siamo in grado di interiorizzare questa percezione utilizzandola per migliorare l'esistenza. Tale fase la indichiamo come *vedere* perché consente all'Uomo di osservare, catalogare, organizzare e discriminare, producendo alla fine una sintesi di ciò che egli desidera in funzione degli obiettivi (specifici e generali) che persegue. Tale sintesi -che costituisce il patrimonio degli indicatori statistici più diffusi nelle applicazioni- assume utilità quando è chiaramente finalizzata e circoscritta dagli obiettivi. Peraltro, se la descrizione e l'analisi esplorativa dei dati sono il primo importante momento di ogni indagine sperimentale esse devono essere sempre filtrate dalla visione stessa (dall'agente, dalla prospettiva, dallo strumento, etc.) che le conferiscono la natura di esame parziale della realtà (perché circoscritto nel tempo e nello spazio).

In secondo luogo, la visione della realtà reclama spiegazioni soprattutto per *comprendere* l'origine della variabilità sperimentale, i legami con altri fenomeni, la persistenza rispetto al futuro, la possibilità ragionevole del controllo e della pianificazione. Peraltro, una concezione matura della conoscenza non accetta leggi deterministiche ma è consapevole che ogni induzione può solo essere probabilistica perché inevitabilmente soggetta ad incertezza. Quindi, probabilità ed inferenza costituiscono l'aspetto diretto ed indiretto, rispettivamente, della medesima concezione del reale. E' necessario assumere che il sapere è incerto, che la comprensione del mondo ha senso e valore scientifico solo se avviene entro un quadro probabilistico e che la valutazione esplicita dell'incertezza derivata dalla conoscenza fenomenica non costituisce nè un difetto né una limitazione ma un pregio peculiare del contributo epistemologico che l'analisi statistica moderna apporta alla conoscenza della realtà.

In terzo luogo, perché la percezione sensoriale e la comprensione del mondo si traducano in decisioni operative, occorre *agire* cioè porre in essere meccanismi incisivi che modificano la realtà in funzione degli obiettivi predefiniti. Tali azioni possono concretizzarsi solo se si possiedono schemi concettuali, razionalmente dedotti dalla percezione e dalla comprensione, che emulano la realtà e ne consentono di comprendere e ripetere i meccanismi: questi schemi vengono universalmente definiti "modelli statistici".

In un certo senso, la fase modellistica è l'aspetto finale della ricerca statistica ma, per il suo aspetto dialettico, essa merita una riflessione accurata perché è strettamente collegata alle successive considerazioni che esporremo sulle analisi del linguaggio.

4. Aspetti modellistici della Statistica

II modello è uno strumento semplificato della realtà, che richiama per analogia, e la cui origine deriva dagli scopi di investigazione della realtà. Di per sé, in quanto costrutto mentale di un essere razionale che, a posteriori, esamina e riflette sul mondo dei fenomeni, il modello non esiste né ha motivo di permanere e/o di essere utilizzato al di fuori dell'obiettivo per il quale è stato costruito e perfezionato.

Da ciò deriva che *tutti i modelli sono intrinsecamente sbagliati* perché, riassumendo in modo strumentale una realtà molto complessa, ne esaminano solo alcuni aspetti specifici; essi però sono utili, necessari e perfettibili. In tal senso, la fase modellistica della Statistica evidenzia l'aspetto dialettico della conoscenza che si sviluppa secondo il seguente schema:

Dati ω Modello ω Teoria

I *dati* apportano informazioni essenziali perché il metodo statistico pervenga alla specificazione e alla costruzione di un *modello,* che dovrà poi confrontarsi con la *teoria* acquisita. D'altra parte, la teoria va confrontata con i dati per accrescere la conoscenza e modificarla quando necessaria. Ora, il divario tra i dati proposti dal modello e quelli osservati costituisce la misura della difformità tra la nostra spiegazione della realtà e la realtà stessa: da ciò la necessità di pervenire ad una nuova spiegazione più verosimile.

La costruzione di un modello avviene sempre mediante la sinergia tra lo scienziato esperto del settore disciplinare in cui si opera e lo statistico, anche perché solo l'esperto potrà immettere e coinvolgere nella fase modellistica il patrimonio di conoscenze pregresse che migliorano la specificazione del modello e lo valicano nel momento critico di confronto con la realtà. Resta compito dello statistico quello di evidenziare -al di là di ogni ragionevole

incertezza- la presenza di una struttura intelligibile nei dati, pervenendo ad una teoria congiuntamente razionale e convincente per quello che l'evidenza sperimentale ha prodotto, in modo da consentire allo scienziato di costruire, aggiornare, modificare e finanche rigettare la sua precedente teoria.

E' importante sottolineare come il modello sia una sovrastruttura mentale imposta al ragionamento per discutere *come se* fosse la realtà, e con la quale è in relazione perché esso è stato costruito a partire da quella medesima realtà con metodi statistici. Tuttavia, man mano che la realtà si evolve, che la mole dei dati si arricchisce, che gli strumenti di precisione si affinano, che nuove conoscenze invalidano precedenti acquisizioni, diventa necessario rigettare il modello acquisito e pervenire a nuovi modelli che, ovviamente, non solo dovranno fornire spiegazioni delle precedenti acquisizioni ma rendere anche ragione di quelle evidenze che il vecchio modello era incapace di spiegare.

Questo aspetto dialettico nella costruzione dei modelli ci conferma che la Statistica è *una teoria della conoscenza sostenuta dai fatti e verificata mediante strumenti analitici formalizzati rigorosamente* grazie alla Logica e alla Matematica. Tale conoscenza è intrinsecamente probabilistica, poiché nessuna conoscenza umana può divenire scientifica se viene ancorata a delle certezze, ma è anche intrinsecamente statistica, perché orientata alle decisioni operative, in modo da consentire di graduare le scelte, orientare i comportamenti, proporre e migliorare schemi di funzionamento, controllare l'immediato e prevedere il futuro.

5. Contenuti statistici del linguaggio e della lingua

Il ***linguaggio***, secondo il Dizionario Italiano De Mauro (2000, versione 1.0.3.5) è una "*capacità comune a tutti gli esseri umani di apprendere una o più lingue storico-naturali e di servirsene per ragionare, intendersi reciprocamente, comunicare sia oralmente sia, tra le popolazioni che conoscono la scrittura, graficamente, scrivendo e leggendo*". Tale definizione di "linguaggio" viene considerato un concetto di uso molto elevato.

In senso tecnico-scientifico, il medesimo dizionario definisce il ***linguaggio*** come una "*facoltà umana ricca di elementi innati, universali, presenti in ogni lingua; le sue manifestazioni maturano nel corso dei primi anni di vita fino a raggiungere la capacità d'uso di una o più lingue storico-naturali, per molti anche in forma scritta; nelle sue manifestazioni chiama a convergere un piano del contenuto, su cui si collocano le innumerevoli esperienze reali, possibili e, anche, naturalmente impossibili di cui è capace un essere umano, e un piano delle espressioni foniche o grafiche ecc., collegate ai contenuti semantici attraverso le infinite frasi generabili in ciascuna lingua*".

Ancora, il Dizionario Italiano De Mauro definisce la ***lingua*** come una "*parlata, idioma, ant. favella, loquela, talora linguaggio come facoltà umana; più spesso modo di parlare peculiare di una comunità umana, appreso dagli individui (in condizioni normali) fin dai primi mesi di vita, affiancato, per le popolazioni alfabetizzate, da modalità ortografiche e di stile connesse alla pratica dello scrivere e del leggere; nelle innumeri manifestazioni di tale modo di parlare e di scrivere si riconosce la presenza di un vocabolario comune alla generalità dei parlanti della comunità per le parole di più alta frequenza (vocabolario fondamentale o di base), integrato da parole e termini più rari in uso tra e per gruppi particolari (le classi più colte, categorie professionali, esercenti di particolari attività, varietà locali, ecc.): i significanti del vocabolario sono articolati e individuati con un sistema fonematico comune, affiancato da rappresentazioni grafiche (ideografiche o, più comunemente, alfabetiche) e il vocabolario noto viene usato per produrre e comprendere frasi e testi secondo un numero relativamente ristretto di norme grammaticali e sintattiche (sancite, di*

solito restrittivamente, nelle scuole) e secondo varie modalità stilistiche e pragmatiche legate alla varietà di situazioni comunicative sociali e alla diversa tipologia dei testi". In tale accezione, "lingua" costituisce una delle 2000 parole fondamentali dell'Italiano, cioè quei vocaboli di altissima frequenza, che da soli costituiscono circa il 90% delle parole che ricorrono nell'insieme di tutti i testi scritti o dei discorsi.

In senso tecnico-scientifico, il medesimo dizionario definisce la ***lingua*** come un "*insieme (cui spesso si attribuisce carattere di sistema) di morfi, il cui significante è costituito adoperando un insieme finito e poco numeroso di unità distintive asemantiche, dette fonemi; nei morfi in generale si riconoscono morfemi lessicali e morfemi grammaticali, che, combinati secondo regole sintagmatiche e regole di assegnazione di ruoli sintattici, consentono di generare (cioè descrivere in modo ordinato) un numero potenzialmente infinito di frasi (e, quindi, di discorsi o testi), ciascuna realizzabile in un numero indefinito di enunciazioni concrete (atti di parole, speech acts) consistenti in una espressione (fonica o grafica e simili) e di una significazione (o senso, riferimento), entrambe ricche di elementi (prosodici, sul versante dell'espressione fonica, pragmatici, sul versante della significazione) importanti nell'esecuzione, ma di più problematica attribuzione all'insieme in quanto sistema astratto*".

Queste lunghe definizioni (che abbiamo integralmente riportato da un Autore che possiede notevoli esperienze e competenze linguistiche associate ad un'attenzione significativa per l'analisi statistica) esplicitano -dal punto di vista del nostro lavoro- un elemento comune sia al linguaggio (inteso come facoltà umana) che alla lingua (intesa come espressione storicamente e geograficamente determinata di quel linguaggio), e cioè la complessità. In effetti, entrambi sono il risultato di una lunghissima opera di selezione naturale, avvenuta nell'arco di millenni, durante i quali gli elementi biologici e psicologici, quelli personali e quelli collettivi hanno interagito a tutti i livelli e in modi spesso difficile da esplicitare. Peraltro, la diversità delle modalità del linguaggio (da un lato) e delle lingue (dall'altro) testimoniano con chiarezza come il bisogno di comunicazione, di cui si diceva in precedenza, abbia trovato canali differenti, sia in rapporto agli strumenti da utilizzare che ai contenuti da esprimere.

Negli studi sulla evoluzione della specie umana, la nascita del linguaggio -così come oggi viene interpretato- si collega sempre più al salto biologico che ha prodotto l'Uomo (qualche centinaio di migliaia di anni fa) secondo una strutturazione mentale che è sostanzialmente rimasta stabile sino alla data odierna. In termini semplici, molti scienziati concordano oggi nel sostenere che il passaggio da un *proto-linguaggio* (nel quale la comunicazione avviene mediante un cenno reale tangibile ed immediato ad una situazione-oggetto verso un altro essere per un obiettivo specifico: difesa, offesa, paura, uso, affetto, fame, etc.) ad un *linguaggio* avviene non solo e non tanto perché la comunicazione sia stata un bisogno degli ominidi, ma soprattutto perché essa rappresenta una necessità ontologica della evoluzione umana che ha consentito lo sviluppo e la creazione di cultura e conoscenza superiori rispetto a quelle degli altri viventi.

Non siamo qui interessati al come e al quando questo passaggio sia avvenuto nella storia della evoluzione; tuttavia, sembra abbastanza acquisito che il salto di qualità nel linguaggio (connesso alla posizione eretta che consente maggiori possibilità di sopravvivenza e produce suoni più articolati e quindi capacità combinatorie più elevate nella produzione linguistica) conduca ad una migliore capacità decisionale degli esseri in rapporto al territorio, all'ambiente di vita, alle altre specie. Quindi, chi ha sviluppato il linguaggio (inteso come modalità di produzione autonoma ed originale dell'espressione sull'esistente, ma anche come capacità di articolare e comunicare il pensiero) ha sviluppato anche il potere superiore di dominare sul contesto ambientale e sulle altre specie. A ben vedere, a differenza di quanto apparentemente accade negli animali anche sufficientemente vicini agli ominidi, dal punto di vista cerebrale ed anatomico, con la nascita del linguaggio si articola la possibilità di un pensiero non costretto dall'esistente ma capace di progetto,

di ipotesi, di simulazione del reale: questa facoltà costituisce un valore aggiunto eccezionale nella evoluzione naturale perché consente di anticipare, di adattare, di prevedere, di giocare di anticipo sulla preda e/o sull'avversario e, più avanti, finanche sulla natura organizzando e pianificando raccolto, produzione, concimazione, etc.

Tali asserzioni non negano che tutte le specie viventi abbiano necessariamente sviluppato forme sofisticate (e talvolta emozionanti ed affascinanti) per comunicare tra loro ed elaborare informazioni, né che esistano in molti contesti animali linguaggi degni di nota, che includano persino la capacità di conteggi elementari. Tuttavia, la sostanziale differenza del linguaggio umano è che esso, sin dall'inizio, contiene una capacità differenziale di elaborazioni complesse su se stesso che conducono la specie umana alla formalizzazione e all'astrazione, in altri termini alla logica del *pensiero simbolico*. Il linguaggio umano, infatti, non è solo comunicazione di ciò che esiste qui ed ora, non risponde solo alle esigenze dell'immediato presente e non elabora solo strategie per l'azione contemporanea: esso è umano perché rielabora ogni esperienza in categorie mentali, ne produce altre non necessariamente reali, sfrutta la concettualizzazione per procurarsi cibo e compagnia, riparo e cultura, conoscenza e potere. La produzione linguistica segue regole affini alla produzione formale del ragionamento matematico perché entrambe queste facoltà -squisitamente ed esclusivamente umane- derivano dal salto evolutivo che ha reso l'*Homo sapiens* una specie vivente strutturalmente diversa dai suoi stessi antenati.

Ebbene, alla base della nascita del linguaggio vi è la possibilità combinatoria delle parole e la possibilità parallela che esse stratifichino nel cervello scenari corrispondenti: quindi, il pensiero si articola con il linguaggio e l'astrazione cresce con la comunicazione (con tutto quello che ne conseguirà poi in termini di logica, formalismo, modellistica, etc.). Che tale possibilità sia genetica è indubbio: basta riflettere sugli sforzi sovrumani che occorrono per addestrare qualche scimpanzé all'articolazione di espressioni sonore circoscritte e limitate a ciò che esiste fisicamente attorno a lui. E, per converso, basta constatare miliardi di volte come in tutto il pianeta bambini sani di appena pochi anni gestiscano con naturalezza alcune migliaia di parole con regole grammaticali generalmente precise, producendo anche in modo creativo richieste, pressioni, aspettative, desideri che derivano da una capacità di astrazione fondamentale, la quale sicuramente non può essere il risultato del breve tempo trascorso nella vita familiare e sociale.

Ai fini del nostro ragionamento, questa breve (e sicuramente imprecisa) digressione sul legame tra evoluzione della specie e sviluppo del linguaggio mira a sottolineare un punto centrale per il prosieguo. Cioè, se il linguaggio nella sua forma superiore è la risultante di una capacità propriamente ed esclusivamente umana, allora il linguaggio contiene nella sua formulazione corrente e nelle sue modalità espressive l'effetto risultante di uno sforzo eccezionale e lunghissimo, che va studiato con quelle metodologie proprie dei fenomeni complessi, derivati da molteplici cause interagenti. La Statistica rivela in tale ambito notevoli potenzialità perché esamina e riassume, comprende ed interpreta ciò che è avvenuto al di là della variabilità individuale ed accidentale, costruisce modelli di comportamento e di funzionamento del sistema linguistico utili ed efficaci, funzionali e sempre perfettibili.

6. Studi linguistici e metodologia statistica

L'analisi linguistica consiste nello studio scientifico dei testi e dei discorsi che gli umani pongono in essere; essa cerca di esplicitare elementi comuni, modalità di produzione e diffusione, caratteristiche salienti, schemi di comportamento, e così via. Per questo, appare corretto introdurre la metodologia propria della Statistica nello studio linguistico: perché si tratta di osservare, comprendere, modellare un insieme di informazioni complesse -non generate in modo

deterministico- che costituiscono una modalità viva di trasmettere la comunicazione interpersonale e le cui evidenze empiriche costituiscono esse stesse delle modalità di comprensione linguistica. Se ogni lingua è uno strumento di relazione interpersonale, osservando come essa si articola e si evolve, anche in modo inconsapevole per coloro che la utilizzano, si accresce la conoscenza del meccanismo di produzione linguistica sia nel tempo che nello spazio.

Queste considerazioni generali possono essere confermate agevolmente mediante il riferimento a specifiche situazioni di studi statistici applicati ad analisi linguistiche, come ora dettaglieremo.

L'analisi della distribuzione di frequenza delle lettere dell'alfabeto, delle parole più diffuse nella lingua parlata e/o in quella scritta, l'uso (e la modifica) dei tempi verbali in funzione del contesto della comunicazione (burocratica, politica, giornalistica, televisiva), le modifiche generazionali della lingua e la presenza sempre più diffusa di neologismi e di espressioni standard (anche straniere), nonché i vari indici proposti per misurare la comprensibilità di un testo, etc., costituiscono esempi di studi linguistici che si sono sviluppati mediante le tecniche proprie dell'analisi esplorativa dei dati.

Le prime analisi linguistiche sviluppate con metodi statistici, per esempio, hanno confermato un principio generale di *economicità degli sforzi* che si ritrova in tutte le attività umane e che permea anche l'attività di comunicazione che si estrinseca nel linguaggio. Così, si assiste in tutte le lingue al fenomeno per cui la brevità di una parola è inversamente proporzionale alla sua frequenza d'uso (secondo la distribuzione di Zipf, dal nome dello studioso che la introdusse). Secondo tale legge statistica, lo sforzo di usare molto frequentemente una parola breve è compensato dalla sua lunghezza minima; il viceversa accade per una parola rara che è invece, spesso, di lunghezza notevole. In effetti, ciò trova conferma nel fatto che il vocabolario fondamentale di una lingua è sostanzialmente composto da parole brevi mentre parole di lunghezza superiore alla media (5-6 lettere) si ritrovano per circostanze specifiche e/o nel linguaggio specialistico.

In particolare, ci pare utile accennare ad un elemento di notevole importanza nella preparazione di testi ad uso pubblico (sussidi didattici, testi legislativi, moduli, questionari, etc.): la *leggibilità del testo*. Essa è intesa come semplicità di formulazione dei concetti tramite parole e forme grammaticali corrette che agevolino per il massimo numero di utenti la immediatezza della sua comprensione e la correttezza della sua interpretazione. Sul piano tecnico un indice di leggibilità è una misura della difficoltà a leggere ed interpretare un testo scritto in una lingua nota sulla base di una scala di valori predefinita. Sono sorte numerose misure di tali aspetti, e tutte sono derivate da indicatori statistici di sintesi della complessità di un testo. Tali indicatori (entro certi limiti e con opportune convenzioni) sono agevolmente programmabili e quindi consentono una rapida ed oggettiva quantificazione tramite elaboratore di qualsiasi testo scritto.

Nella letteratura corrente, oltre all'adattamento che Vacca ha prodotto dell'*indice di Flesch* per la lingua italiana (che include nella valutazione sia la media delle parole per frase che il numero medio di sillabe per parole), sembra di un certo interesse l'*indice Gulpease* (proposto dal gruppo Universitario Linguistico-Pedagogico, Università di Roma "La Sapienza") che è stato direttamente costruito per la lingua italiana. In effetti, tale indice misura la facilità di lettura di un testo in funzione del numero di lettere per 100 parole e del numero di frasi per 100 parole. Oltre ad essere una misura basata su variabili testuali di agevole determinazione, tale indice merita di essere diffuso perché esso è tarato anche in funzione del livello di scolarizzazione del lettore. In tal modo, la corrispondenza tra leggibilità del messaggio e livello culturale del ricevente può essere sfruttata per agevolare e migliorare la comunicazione interpersonale, tra gruppi e quella istituzionale.

Altre analisi, di livello successivo, includono gli studi sul cosiddetto *vocabolario di un Autore*, cioè sull'ammontare di parole conosciute ed utilizzate da parte di una persona specifica. Il problema ha assunto rilevanza nei tempi più recenti perché si è notato come, a livello di generazioni successive, si assiste ad un impoverimento di tale vocabolario indotto dai mezzi di comunicazione di massa che, essendo sostanzialmente stereotipati e condizionati da una logica consumistica (di assuefazione al messaggio più che ad una sua elaborazione), riducono il corpo linguistico del parlante a pochi elementari fonemi. Ebbene, la valutazione di tale "dimensione" non può che avvenire con metodi statistici essendo di fatto impossibile enucleare per ciascuno e con assoluta certezza la lista delle parole che una persona utilizza; in pratica, essa avviene mediante procedure che hanno numerose connessioni con la dinamica dei sistemi biologici e deriva dal meccanismo probabilistico di estrazione da un'urna le cui palline sono numerate ma di cui non si conosce il valore massimo.

In numerosi altri contesti (soprattutto grazie alle scuole francese ed italiana), gli studi statistici concernenti le analisi multivariate sono state applicate a numerosi *Corpora* allo scopo di individuare caratteristiche proprie, elementi classificatori, aspetti discriminanti, elementi centrali della comunicazione. In tale ambito si ritrovano l'esame dettagliato dei discorsi, dei messaggi pubblicitari, lo studio delle risposte a domande aperte nei questionari, e così via. Il complesso di tali studi richiede, ovviamente, una formalizzazione più accurata e la verifica preliminare che le ipotesi -generalmente accolte in altri settori di ricerca- possano essere sostenute anche allorché si tratti di espressioni verbali.

7. Analisi statistica e impronta testuale

Una elaborazione statistica sulla quale intendiamo soffermarci con maggiore ampiezza è il *riconoscimento di un Autore*, cioè la possibilità di pervenire sulla base di un testo sufficientemente lungo (e spontaneo) alla caratterizzazione possibilmente significativa del suo estensore, mediante raffinati metodi statistici. Tale tipo di studi sottende un principio che, stante le precedenti argomentazioni su linguaggio e lingua, potrebbe essere accolto senza troppe riserve: in sostanza, ogni persona adulta possiede una sua specifica *impronta testuale (o letteraria)*, cioè una modalità propria e caratterizzante per esprimere il suo pensiero, i suoi desideri e le sue opinioni, perché queste sono la risultante della sua vita biologica, psicologica, affettiva, culturale e sociale. Tale impronta individua una persona allo stesso modo delle impronte vocali, digitali, etc. ma trattandosi di una caratterizzazione pienamente umana (e non meramente fisiologica, organica, anatomica, etc.) la sua individuazione non soggiace a leggi fisiche e chimiche di agevole determinazione ma a quelle statistiche, essendo intrinseco di tale impronta un elevato contenuto di incertezza derivante dalla varietà e dalla molteplicità della libertà umana.

In sostanza, se la comunicazione promana da un bisogno personale e se tale bisogno si estrinseca mediante lo strumento del linguaggio, ogni comunicazione contiene in sé le risultanti del lungo processo di sedimentazione biologica e psicologica che ha prodotto la specificità di colui o colei che trasmette informazioni. Allora, l'analisi di un testo dovrebbe essere in grado di caratterizzare in modo univoco il suo Autore nella misura in cui esso è genuino, sufficientemente lungo e comparabile con testi di riferimento, acquisiti ed incontrovertibili, sul medesimo Autore.

Ovviamente, il punto di riferimento dell'impronta testuale può essere un qualsiasi elemento linguistico e l'analisi empirica dovrà dimostrare la migliore efficacia di uno tra

essi oppure la convergenza necessaria di più elementi. Tra quelli più accreditati citiamo, per esempio: la lunghezza delle parole usate, il numero di parole per ciascun periodo, l'abbondanza relativa di parti del discorso (avverbi, congiunzione, aggettivi), l'uso di parole "rare", la frequenza relativa di specifici punti di interpunzione, e così via.

La struttura dell'impronta testuale può essere percepita, compresa e modellata secondo le linee dell'analisi statistica delineate all'inizio di questo paragrafo. In dettaglio, si può studiare la distribuzione di frequenza dell'elemento prescelto (per esempio, la lunghezza delle parole utilizzate da un Autore nel suo testo); si possono far derivare i dati osservati da uno schema probabilistico complesso mediante il quale si generano le parole in funzione dello scopo che assolvono nel discorso; si può verificare se esistono elementi per confermare o rifiutare l'ipotesi che il brano in esame sia stato prodotto dal medesimo Autore di cui si ha una estesa conoscenza, ovvero se due brani derivano da Autori differenti, e così via. Tuttavia, ciò che ci preme sottolineare in questo tipo di analisi è il contenuto proprio dell'inferenza statistica, e quindi il suo carattere di incertezza decisionale che può essere valutato correttamente solo in termini di verosimiglianza e plausibilità (cioè, in termini probabilistici). Difatti, ogni decisione è corretta solo se accompagnata da una valutazione della probabilità di commettere un errore: se tale probabilità è molto piccola, la nostra decisione possiederà elementi forti per essere sostenuta, anche se non assurgerà mai al ruolo di certezza definitiva.

In tali ambiti, è di notevole interesse esplorare il legame tra le caratteristiche dell'individuo e la struttura della sua impronta testuale, perché tale relazione consente di esplicitare (e finanche di quantificare) l'effetto delle condizioni culturali, sociali ed economiche sulla lingua e sulle sue modalità espressive. Generalmente, tali studi si limitano a confronti tabellari di tipo descrittivo, ma l'analisi statistica moderna consente oggi di esplicitare la funzione-legame tra la struttura probabilistica della variabile di interesse e alcune caratteristiche dell'Autore mediante i cosiddetti *modelli lineari generalizzati*. In tal modo, si può addirittura visualizzare come talune caratteristiche si modifichino con il genere, l'età, la cultura, il reddito, e così via.

Il complesso di tali indagini apporta agli studi linguistici una significativa valenza interpretativa che non va sottovalutata in tempi di veloci mutazioni del linguaggio e delle sue manifestazioni. In altri termini, assieme alle analisi storiche e socio-culturali di una lingua, diviene ora possibile -sulla base di evidenze empiriche- pervenire ad un'analisi sistemica dell'uso della lingua medesima e delle sue trasformazioni in funzioni delle caratteristiche del soggetto. Questo tipo di indagini non solo possiede un valore scientifico per l'analisi linguistica, ma coinvolge anche numerose implicazioni operative: per esempio, nella didattica delle lingue, nella teoria della comunicazione, nelle indagini per l'attribuzione dei testi, negli studi di similarità, e così via.

8. Considerazioni finali

In questo lavoro, si sono discussi alcuni dei principi che stabiliscono un rapporto essenziale fra comunicazione, informazione e linguaggio e che richiedono un impegno congiunto tra statistici e linguisti per acquisire maggiori e più estesi risultati, sia per finalità conoscitive che confermative.

La discussione ha evidenziato come linguaggio e lingua costituiscano collettivi di elementi la cui stratificazione progressiva è avvenuta, e tuttora avviene, in modo complesso e dinamico. Per questo, la disamina del sistema linguistico si arricchisce in misura significativa anche grazie alla

Statistica, nella misura in cui tale scienza elabora la formulazione di regole decisionali secondo criteri ottimali derivanti dalla conoscenza dei collettivi. Trattandosi di espressioni umane, ove storia e cultura, psicologia ed ambiente modellano senza soste e modificano ripetutamente le modalità linguistiche, ogni acquisizione ha significato scientifico *se e solo se* si inserisce in un contesto inferenziale di natura statistica.

L'approfondimento regolare e rigoroso di questi studi reclama formalizzazioni impegnative e sinergie sistematiche tra competenze diverse, ma la novità dei risultati e l'ampiezza degli interessi che tali approcci suscitano -sia in ambito linguistico che in quello statistico- confermano la validità culturale e metodologica di tali innovazioni.