Indice

La società contemporanea ci ha abituato ad un nuovo tipo di testo: il *testo elettronico*. In pratica, si tratta di un 'insieme di parole' che ha assunto una forma leggibile dalla macchina (con un'espressione ed una sigla inglesi *Machine Readable Form* – MRF); ciò lo rende insieme simile a e diverso da un normale testo scritto su carta. È simile perché è formato, sostanzialmente, dagli stessi elementi (parole, segni di punteggiatura e così via), ma anche diverso, perché ha in sé caratteristiche che il testo cartaceo non potrà mai possedere.

Ciò che rende differente un testo elettronico è la sua *flessibilità*. Molte operazioni compiute su un testo cartaceo corrono il rischio di distruggerlo; ad esempio, 'tagliare' una pagina da un libro per 'incollarla', poniamo, su un quaderno, compromette irrimediabilmente l'integrità del libro stesso, e così una sottolineatura troppo decisa o la sovra-scrittura di note e commenti possono rendere impossibile la lettura di parole e frasi.

Invece, una volta che il testo abbia assunto una forma leggibile dalla macchina, è possibile intervenire su quel testo in moltissime maniere. L'utente può copiare l'intero testo o una sua parte in altri *file* (quindi in altri testi), può spostare brani da una parte all'altra, tenere copie di riserva delle diverse stesure che sta realizzando, può variare forma e colore dei caratteri e così via.

Molti di questi interventi sono possibili con un comune programma di videoscrittura. Nel volume *Il testo elettronico* abbiamo cercato di mostrare come l'utente può arricchire il formato MRF fino a trasformare la videoscrittura in *videotestualità*. Qui ricapitoleremo brevemente queste conoscenze per mostrare il modo in cui il testo diventa un vero e proprio *database* e, soprattutto, vedremo cosa si può fare avendo a disposizione una base di dati testuali.

1. Esplorare un testo

Oggi, tutti noi abbiamo esperienza di testi e banche dati[1] e ci accorgiamo della differenza che passa tra questi due tipi di oggetti. Un testo appare (o, meglio, *è* per l'utente normale) come una sequenza di caratteri scritti, contenente segni di punteggiatura o anche immagini; un libro scolastico, un giornale, un cartellone pubblicitario sono tutti esempi di testi che si offrono all'utente per un'operazione di *lettura*.

Al contrario, una banca dati appare come un deposito da cui siamo in grado di estrarre informazioni o altro ancora; un terminale per conoscere un collegamento ferroviario, uno sportello bancomat da cui fare un prelievo o avere notizie sul proprio conto corrente, sono esempi di basi di dati che si offrono all'utente per un'operazione di *consultazione*.

La lettura è un processo complesso, attraverso il quale il parlante ricostruisce un contenuto concettuale partendo da caratteri che, in sé, non hanno un valore particolare. Ad esempio, il segno

(1)

potrebbe essere un qualunque cerchio tracciato su un foglio; esso acquista il valore di 'carattere alfabetico O' quando il parlante conosce un certo sistema convenzionale, socialmente condiviso, per costruire equivalenti

[1] L'espressione *base di dati* corrisponde all'inglese *database* e, in italiano, viene spesso usata come sinonimo di *banca dati*. In questo volume utilizzeremo i tre termini in modo interscambiabile, senza addentrarci in complesse questioni di terminologia tecnica. Per un approfondimento su questi temi rinviamo ad Ausiello *et al.* (1991). Per l'informatica umanistica e linguistica Busa (1987), Orlandi (1990), Gigliozzi (1997, 2003), Ciotti-Roncaglia (2000), Numerico-Vespignani (2002), Lenci *et al.* (2005).

grafici dei suoni di cui si compongono parole e frasi. In sostanza, quando sulla base delle sue capacità metalinguistiche, coglie la relazione sistematica tra *oralità* e *scrittura*.

La lettura è, dunque, un processo di *decodifica*, attraverso il quale un parlante riconosce prima delle espressioni materiali (i caratteri grafici), che poi collega alle sue conoscenze linguistiche arrivando infine a comprendere ed interpretare il messaggio verbale prodotto da un altro parlante (*codifica*). In esso entrano in gioco moltissimi aspetti (basti pensare, ad esempio, alla differenza di conoscenze e di intenzioni tra chi produce un testo e chi lo legge), ma la cosa fondamentale è che si tratta di un processo attivo, creativo, realizzato da soggetti umani intelligenti.

La situazione è ben diversa quando si consulta una base di dati. In questo caso, il computer offre all'utente delle informazioni che (i) sono state accuratamente preparate e (ii) sono recuperate ed elaborate partendo da una precisa richiesta (in inglese *query*) formulata dall'utente stesso. In questo caso, la macchina agisce come una sorta di filtro, di memoria e di elaborazione.

Un tipico esempio è rappresentato dalla consultazione dell'orario elettronico dei treni[2]. Utilizzando la procedura di consultazione della banca dati, l'utente indica una località di partenza ed una di arrivo e, inoltre, un orario (giorno e ora della partenza); come risultato ottiene un elenco dei treni che può prendere nella località di partenza e che lo portano alla località d'arrivo, all'interno di una fascia oraria che ha inizio con l'orario impostato.

La consultazione non prevede, ad esempio, la possibilità di indicare la sola località di partenza o la sola località d'arrivo; quindi non posso sapere quali siano «tutti i treni che partono da Perugia» o «tutti i treni che arrivano a Perugia». In più, non c'è la possibilità di impostare un orario *d'arrivo*, e quindi non posso sapere quali siano «tutti i treni che arrivano a Perugia entro le ore 20».

[2] Come nel sito delle Ferrovie dello Stato *www.trenitalia.com*.

Queste limitazioni sono tipiche di ogni banca dati e di ogni procedura di consultazione; nella prima si delimita un universo di informazioni e, con la seconda, si definiscono i modi in cui consultare tale universo.

La domanda che dobbiamo porci, a questo punto, è la seguente: quali sono le informazioni e le conoscenze che vogliamo ottenere (o, con un termine tecnico, *estrarre*) da un testo?

La prima risposta che possiamo dare è, insieme, la più immediata e la più ricca di implicazioni scientifiche: da un testo vogliamo estrarre gli elementi che lo costituiscono, per sapere *quali* siano e, anche, *come* siano combinati. Questa risposta è insita in ogni operazione di consultazione. Se sfoglio un vocabolario per cercare una parola di cui non conosco il significato, voglio sapere se la parola è presente nel vocabolario stesso e come è collegata alla definizione; se cerco informazioni nel *World Wide Web*, voglio trovare un sito che le contenga.

In tutti e due i casi, la consultazione è e diventa una *scelta*: cerco una parola tra le tante presenti nel dizionario, o un sito tra i milioni (forse miliardi) esistenti. I dati che la macchina elabora hanno un valore informativo per il soggetto umano ed i risultati della ricerca acquistano un preciso significato.

Partendo da questa idea, vediamo come operare. Per estrarre informazioni da un testo, il ricercatore può compiere due fondamentali operazioni. La prima consiste nello sfruttare la codificazione normalmente presente in un testo a stampa; i caratteri alfabetici, i segni di punteggiatura, le convenzioni ortografiche e tipografiche che si ritrovano in un testo scritto costituiscono un patrimonio considerevole, che può essere opportunamente sfruttato per guidare una procedura elettronica. Questa operazione viene detta tecnicamente *indicizzazione*, nel senso che il programma considera ogni elemento testuale come un'unità che può essere individuata e trattata. Così ogni carattere, ogni riga o pagina di testo diventa un'unità informativa che può essere contata, copiata, cancellata, in poche parole elaborata secondo gli scopi dell'utente.

Ad esempio, la ricerca delle *parole* può essere guidata dalla convenzione che prevede l'inserimento di uno spazio bianco all'inizio e alla fine di

ognuna di esse o, eventualmente, di uno spazio bianco all'inizio e di un segno di punteggiatura alla fine. Così, nelle sequenze:

(2)

```
a. la casa bianca
b. la casa bianca,
c. la casa bianca;
d. la casa bianca.
```

il programma potrà sempre individuare le tre parole *la, casa* e *bianca*.

Non mancano, naturalmente, difficoltà, ben note a chi si occupa dei sistemi di scrittura storici delle lingue naturali. A prima vista, esse possono sembrare apparentemente banali; se, ad esempio, troviamo una parola scritta con spezzatura a fine riga come deve comportarsi la procedura? Di fronte alla sequenza *ca-sa*, disposta su due righe contigue, il parlante italiano sa che si tratta della parola *casa*, ma adottando il criterio esposto prima, una procedura elettronica si troverebbe a ricostruire due parole distinte: *ca* e *sa*, di cui la prima delimitata da uno spazio bianco e da un segno di punteggiatura, la seconda da un inizio-riga e da uno spazio bianco.

Casi più complessi sono rappresentati da sigle d'uso comune come 'N.B.' per 'Nota Bene'. Il parlante riconoscerebbe nell'*intera* sequenza una parola (punti fermi compresi), mentre per la procedura automatica avremo le parole 'N' e 'B' separate e non caratterizzate dai punti fermi.

In sintesi, l'operazione di indicizzazione è guidata da elementi già inseriti nel testo in MRF (normalmente ben noti al parlante che conosce il sistema di scrittura della sua lingua e che rappresentano già una base significativa per l'esecuzione della procedura elettronica), ma essi sono insufficienti per ottenere risultati ottimali. In primo luogo, i segni convenzionali, di ortografia e punteggiatura, veicolano soltanto informazioni interne al testo[3] e, in più, non tutte le informazioni. Consideriamo questo esempio:

[3] Su questo argomento cfr. la "pillola" dedicata a *Il testo elettronico* in questa collana.

(3)

```
La Casa Bianca
È il nome che tradizionalmente si dà alla residenza del presidente degli Stati Uniti d'America. La Casa Bianca si trova a Washington e deve il suo nome al caratteristico colore dell'edificio.
```

In (3) non ci sono differenze tra *La Casa Bianca* come 'titolo' della breve definizione e *La Casa Bianca* come parte della definizione stessa. In più, come parlante dell'italiano ho la precisa intuizione che *Casa Bianca* costituisca un'unità lessicale (una specie di 'nome proprio') e non una sequenza, occasionale, di due unità come in:

(4)

```
Guarda quella casa bianca là in fondo alla strada
```[4].

La seconda operazione che il ricercatore può compiere sul testo è quella della *marcatura*, di cui si è parlato estesamente nel volume *Il testo elettronico*. In breve, egli può inserire delle sequenze particolari di caratteri (i cosiddetti *tag* o *marcatori*) che il computer interpreta come 'segnali'. Ad esempio, in (3) è possibile marcare come <titolo> la prima sequenza *La Casa Bianca* e come <parola> la sequenza *Casa Bianca*. I *tag* possono avere vari formati (ad esempio, possono essere racchiusi tra parentesi angolari '<...>' o preceduti da caratteri quali '$', '%'), ma hanno un'unica funzione, quella di introdurre informazioni che il parlante ritiene estremamente importanti e significative, ma non sono recuperabili sfruttando il normale apparato delle convenzioni ortografiche[5]. Il ricercatore potrà in-

[4] La presenza del carattere maiuscolo non è, di per se stessa, un indicatore affidabile; basti pensare a casi come *Città del Capo*, *Castel del Piano* ecc.

[5] Essi sono elementi metacognitivi che agiscono come istruzioni (ad esempio: «tratta le due parole Casa e Bianca come qualcosa di unitario, vale a dire come se fosse un'unica parola»).

serire *tag* per marcare ogni tipo di informazione, 'aprendo' e 'chiudendo' gli specifici elementi testuali[6].

La marcatura può riguardare un gran numero di informazioni e conoscenze sul testo. In modo molto sintetico, possiamo dire che esistono due fondamentali insiemi: da un lato, le informazioni esterne al testo, dall'altro, quelle interne. Le *informazioni esterne* sono quelle conoscenze che fanno da sfondo al testo vero e proprio, che spesso lo precedono o lo seguono e, a volte, possono non comparire. Prima di tutto abbiamo l'autore (o gli autori), quindi il tempo e il luogo in cui il testo è stato prodotto[7]. Altro aspetto fondamentale è la modalità con cui si presenta il testo, che tipicamente può essere orale o scritta (con le relative sottocategorizzazioni).

Le informazioni *interne* al testo si possono, semplificando, suddividere in sintetiche, analitiche e ipertestuali. Le informazioni sintetiche sono quelle che l'autore fornisce come riepilogo del contenuto concettuale del testo; si tratta, in particolare, dei titoli (dal titolo generale del testo a titoli relativi a parti di esso) e dei delimitatori di unità coerenti: paragrafi, capitoli e simili. L'autore usa questi strumenti metalinguistici per indirizzare il lettore nell'opera di comprensione e interpretazione, e, inoltre, per dare un ordine interno alla sua produzione.

Tra gli elementi sintetici vanno considerati anche gli indicatori di *genere*, ad esempio, *prosa*, *poesia*, *romanzo*, *racconto*, ecc. Essi possono essere inseriti direttamente dall'autore, ma anche derivare da interventi successivi (ad esempio, glosse e commenti).

La parte analitica del testo è quella che costituisce il vero e proprio 'corpo'; è l'insieme di parole che forma il 'contenuto' visibile del testo. In

---

[6] Ad esempio, racchiuderà tra marcatori specifici le parole straniere, le sigle, le date e così via.

[7] In un testo a stampa le informazioni precedenti saranno tipicamente presenti nella copertina o nelle pagine iniziali e formeranno la prima 'guida' alla lettura di un'opera. Vale la pena sottolineare l'importanza di queste conoscenze perchè il parlante, il tempo e il luogo dell'enunciazione rappresentano delle àncore importantissime sia per la produzione che per la comprensione e l'interpretazione del testo.

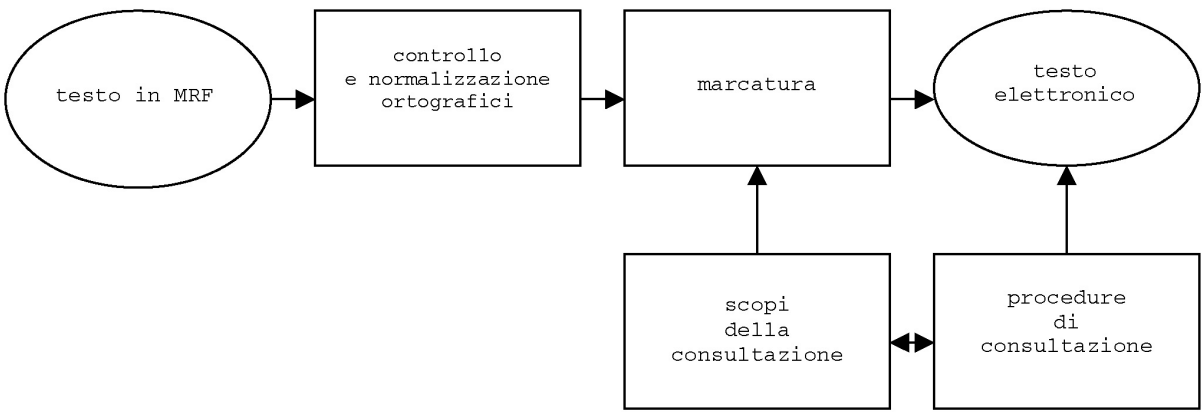
esso[8] possono essere mescolate espressioni appartenenti a più lingue, parole e immagini, sigle e abbreviazioni, cifre e formule e così via, in combinazioni che a volte mettono a dura prova le capacità di comprensione e interpretazione. Tra l'altro, possono trovarsi rinvii ipertestuali: citazioni di altri testi, note (a piè di pagina o in fondo ai capitoli in un libro stampato), indici o altro ancora.

Introducendo le marcature, il criterio generale che il ricercatore deve tener presente è quello del rapporto tra *dati* e *programmi*. I codici che egli inserisce nel testo sono funzionali ad ottenere determinati risultati, e quindi devono consentire e rendere agevole l'applicazione di un programma che porta a questi ultimi. In più, egli ha l'esigenza di registrare i dati in un formato che possa servire anche per altri programmi, e quindi non sia troppo legato ad una specifica procedura. Va da sé la considerazione che le operazioni manuali di codifica sono onerose, in termini di tempo e costi, e ciò porta a privilegiare la parte di elaborazione automatica.

Il lavoro che svolge chi prepara un testo elettronico in vista di una sua consultazione automatica (o, per usare un termine tecnico, di uno *spoglio*) si può schematizzare nel modo seguente:

[8] O anche nelle parti riconoscibili come sintetiche o ipertestuali.

(5)


testo in MRF
controllo
e normalizzazione
ortografici
marcatura
testo
elettronico
scopi
della
consultazione
procedure
di
consultazione