Gian Babbeo e le banche dati full-text

Appunti sul metodo della ricerca on line

di Brunella Longo

Trent'anni fa la Lockeed, la NASA e poche altre grandi strutture al mondo iniziavano a sperimentare la ricerca in batch su banche dati di tipo bibliografico. I bibliotecari che negli anni Settanta hanno iniziato a compiere ricerche on line si sono confrontati con forme di immagazzinamento delle conoscenze di tipo referenziale, affini alle procedure utilizzate nella catalogazione delle raccolte librarie. Oggi, stando all'ultima edizione del *Cuadra Directory of Databases*, soltanto il 25 per cento delle banche dati esistenti al mondo è di tipo bibliografico "puro", contro un 40 per cento di banche dati full-text (d'ora di in poi BDF), cresciute vorticosamente con l'abbassamento dei costi di immagazzinamento delle informazioni.[1] Il rimanente 35 per cento comprende forme svariate di conoscenza (directory, statistiche, numeri, immagini, software) il cui genere è tuttavia molto più vicino al concetto di "testo" che non a quello di rappresentazione formalizzata di informazioni.

Gli host, i calcolatori centrali dei distributori di banche dati (d'ora in poi BD), contengono dunque milioni e milioni di unità elementari di conoscenza, organizzate prevalentemente secondo logiche testuali, la cui quantità raddoppia all'incirca ogni tre anni e mezzo. La tecnologia ci consente, costruendo combinazioni di parole e frasi del linguaggio naturale, senza alcun riguardo al modo con cui l'informazione è stata categorizzata dagli autori dei testi o dai produttori delle BD, di estrarre da questi giganteschi pozzi testuali la risposta ad un bisogno di conoscenza nel giro di una manciata di minuti.

L'information retrieval ha cambiato o sta cambiando la natura del lavoro di molti bibliotecari e documentalisti il cui ruolo, all'interno di organizzazioni diverse, è sempre più spesso di integrazione di strumenti e fonti informative vecchie e nuove e di traduzione di conoscenze specialistiche in linguaggi diversi per i diversi utenti.

"Voi potete cercare qualcosa soltanto dove c'è un metodo per cercarlo", diceva Wittgenstein. Questo articolo esamina le aree nelle quali occorre acquisire e sviluppare un *metodo* per condurre ricerche su BDF: crediamo che nel meccanismo di connessione e di scoperta di relazioni esistenti tra gli aspetti della tecnologia, del linguaggio e del nostro comportamento, risieda la chiave del metodo per la ricerca on line.

LA TIPOLOGIA DELLE BANCHE DATI FULL-TEXT

La Tenopir[2] ha individuato nove tipi di BDF:

1) legislative o normative (un esempio europeo è la banca dati della Comunità europea CELEX; uno italiano, su CD-ROM, è dato dalla raccolta delle Leggi d'Italia De Agostini);
2) documenti governativi, brevetti ed altre pubblicazioni ufficiali;
3) quotidiani;
4) agenzie di stampa nazionali ed internazionali (tra le quali esiste un elevatissimo grado di sovrapposizione);
5) newsletter, il cui genere testuale è spesso alfanumerico contenendo dati su singoli settori industriali;
6) opere di reference (enciclopedie, glossari o repertori);
7) periodici specializzati e tecnici;
8) periodici di consumo;
9) directory (come Who's Who, le Pagine gialle elettroniche, i repertori per gli affari come Dun & Bradstreet).

[1] Le banche dati full-text rappresentavano solo il 4 per cento del totale delle banche dati esistenti nel 1980, il 22 per cento nel 1986 ed il 33 per cento nel 1989. Fonte: Cuadra Directory Database.

[2] C. TENOPIR, *Users and Uses of Full-text Database*, in *12th International Online Information Meeting Proceedings*, Learned Information, 1988, p. 263-270. Della stessa autrice, tra i moltissimi contributi, si può segnalare anche: *Full-text Databases*, "Annual Review of Information Science and Technology", 1984, 19, p. 215-246.

Ci sembra opportuno aggiungere altre due categorie:
10) banche dati "ibride", nate come bibliografiche ma che comprendono ormai una notevole quantità di record full-text (come PTS Promt, ABI/Inform);
11) banche dati di tipo numerico-testuale del settore economico (come Standard & Poor's News, Extel, Mintel, ICC Business Research, Business International, ecc.).
Dal punto di vista della distribuzione disciplinare, la maggior parte delle BDF appartiene ormai al settore degli affari (34 per cento) e del giornalismo (22 per cento) mentre, come è noto, le origini dell BD bibliografiche appartengono ai settori della chimica, della scienza e della tecnica. Il fatto che il maggior utilizzo di banche dati full-text abbia sede, attualmente, nel mondo degli affari, della finanza, del marketing e dei mezzi di comunicazione in genere non è privo di conseguenze per la ricerca: come si darà modo di vedere più avanti, in queste aree prevalgono linguaggi settoriali poco distanti dalle lingue comuni che, insieme alla scarsità di strumenti per il controllo della terminologia, rendono la ricerca on line piena di trabocchetti.
Eppure è stato giustamente notato da diversi autori[3] come l'uso delle risorse on line favorisca il superamento dei confini disciplinari e della organizzazione delle informazioni in formati altamente strutturati. Pertanto ci sembra più utile alla riflessione sulle tecniche ed i metodi di retrieval non sopravvalutare la categorizzazione delle BD in base alla loro morfologia o al settore disciplinare a cui appartengono e preferire un approccio in cui prevalga il punto di vista dell'utente.
Mettere l'utente al centro del nostro tentativo di ragionare sulle BDF significa privilegiare i contenuti delle BD ed il loro valore d'uso piuttosto che le forme di immagazzinamento della conoscenza, significa porre più attenzione al contesto comunicativo nel quale si svolgono le ricerche piuttosto che isolare la discussione sui soli vantaggi e limiti delle tecniche di information retrieval.
L'utente che cerca riferimenti bibliografici sullo studio delle relazioni industriali nei servizi potrà oggi usare indifferentemente banche dati bibliografiche appartenenti tanto al settore specifico delle relazioni industriali (per esempio, Employee Benefits Inforsource) quanto al più generale campo del management (come Management & Marketing Abstracts) e banche dati a testo completo generaliste (come BIG/Il sole 24 ore) o specialistiche (come Harvard Business Review). Allo stesso modo, se fossimo interessati a sapere chi ha già applicato una nuova tecnologia nel settore farmaceutico, potremmo consultare decine di BD tanto del settore business che del settore chimico-farmaceutico, tanto bibliografiche quanto full-text. La determinazione delle fonti da preferire dipenderà soprattutto da quanto noi già sappiamo dell'argomento su cui verte la nostra ricerca e dal livello di completezza e di specificità che ci attendiamo di raggiungere attraverso la ricerca stessa.
Tenopir[4] riferisce delle prime ricerche compiute dall'American Chemical Society sull'utilizzo di BDF tra il 1979 ed il 1982, dalle quali risultava che in campo chimico le BDF venivano usate già all'inizio della loro comparsa sul mercato per una vasta gamma di obiettivi, tra cui:
— recuperare informazioni fattuali riportate nei testi degli articoli;
— recuperare informazioni marginali o periferiche rispetto all'oggetto principale dell'articolo;
— ottenere un aiuto in campo bibliografico.
Le indagini di Pagell[5] hanno confermato come, nel 1987, gli utenti della BDF Magazine ASAP (periodici di consumo statunitensi) utilizzassero questa fonte per diverse categorie di necessità:
— il reperimento di documenti (alla stessa stregua che nelle BD bibliografiche);
— recupero di notizie fattuali;
— recupero di informazioni marginali o periferiche rispetto al "fuoco" dell'articolo;
— ricerca di citazioni;
— recupero dei testi di articoli noti.
L'utilizzo di BDF in alternativa alla collezione delle versioni cartacee dei periodici della biblioteca veniva messo in forse da questi studi in quanto, si diceva, l'utente preferisce comunque consultare la copia cartacea di articoli dei periodici, provvista tra l'altro del suo corredo grafico originale.
Queste opinioni ci sembrano ormai datate, in quanto negli ultimi tre anni si sono maggiormente diffusi strumenti di office automation e desktop-publishing che consentono un ottimo repackaging degli articoli ottenuti on line.[6] Un progetto di sostituzione delle collezioni cartacee e dei prestiti interbibliotecari attraverso l'on line è stato messo a punto con soddisfazione degli utenti presso l'Università di Washington già nel 1989.[7] Presso il Centro di documentazione di Fininvest Comunicazioni abbiamo deciso di sostituire la sottoscrizione di abbonamenti a newsletter con la consultazione di BDF dopo aver verificato che gli utenti rimangono indifferenti alla forma nella quale presentiamo loro le informazioni tratte da queste fonti.
La più recente rassegna di studi sull'argomento compiuta da M.A. Siddiqui[8] aggiunge che le BDF utilizzate per la ricerca di riferimenti bibliografici consentono di poter verificare la pertinenza delle citazioni recuperate grazie allo scorrimento on line del testo completo. Il browsing, cioè lo scorri- ➤

[3] C. TENOPIR, *op. cit.*; R.K. SUMMIT, *Knowledge Online: Current Implications and Future Trends*, in *1st East-West Online Information Meeting Proceedings*, Learned Information, 1990, p. 19-28.

[4] C. TENOPIR, *op. cit.*; *Searching Full-text Databases*, "Library Journal", 1988, May, p. 60-61.

[5] R.A. PAGELL, *Primary Full-text Information: Databases for the End User*, in *12th International Online Information Meeting Proceedings*, Learned Information, 1988, p. 260; *Ruth Searching Full-text Periodicals: How Full is Full?*, "Database", 1987, October, p. 33-36.

[6] B. LONGO, *Il repackaging dell'informazione online*, "Biblioteche oggi", 9 (1991), 3, p. 313-328.

[7] C.S. KAAG, *Online Full-text Article Retrieval*, "Technical Services Quarterly", 1989, 2, p. 29-40.

[8] M.A. SIDDIQUI, *Full-text Databases*, "Online Review", 1991, 6, p. 367-371.

mento di interi record o di loro porzioni, è del resto diventato un modo di "fare" ricerca e non più solo di "vederne" i risultati, in quanto consente di colmare con una selezione soggettiva quella mancanza di raffinamento delle ricerche tipicamente causata dai limiti dei sistemi di information retrieval.
Siddiqui inoltre sottolinea un utilizzo delle BDF ancora poco studiato, benché sia già stato considerato come il rivoluzionario vantaggio della ricerca on line:[9] la possibilità di compiere ricerche sui concetti con infinite combinazioni di termini che li possono esprimere. Nei prossimi anni potremo perciò scoprire nuovi obiettivi per cui utilizzare la ricerca on line, grazie soprattutto alle banche dati giornalistiche.[10]
La banca dati VU/Text (contiene testi completi di oltre settanta quotidiani statunitensi), per esempio, è stata utilizzata per riaprire una inchiesta di polizia su undici omicidi commessi tra il 1985 ed il 1990 in sei diversi stati americani. Un reporter, confrontando le descrizioni delle vittime e delle circostanze dei delitti così come erano state riferite dai diversi giornali regionali,[11] aveva infatti dedotto che i crimini potessero essere imputati allo stesso maniaco. Ma, se è vero che "la varietà delle banche dati full-text e la eterogeneità della letteratura in esse disponibile ne permettono molti usi differenti" è anche vero che esse "pongono l'esigenza di nuovi passi avanti nello sviluppo delle strategie di ricerca".[12]
Stiamo diventando tutti di nuovo analfabeti di fronte alla "scrittura" delle nuove tecnologie? Si chiedeva su queste pagine il vicepresidente del Consiglio nazionale delle biblioteche francesi.[13] Riproponiamoci questo interrogativo, constatando che, finora, all'esplosione delle BDF ha corrisposto, in prevalenza, la nostra strenua concentrazione sui linguaggi controllati, forse gli unici che sappiamo leggere e che ripassiamo continuamente... alla stregua dei fratelli di Gian Babbeo.

IL RECUPERO DELL'INFORMAZIONE DALLE BANCHE DATI FULL-TEXT

La maggior parte degli host consente la ricerca on line, tanto su basi bibliografiche quanto su basi a testo completo, attraverso sistemi di information retrieval che utilizzano la logica booleana. Gli operatori logici "AND", "OR", "NOT", gli operatori di adiacenza e prossimità (come "ADJ", "SAME", "WITH"), il troncamento dei termini ed il mascheramento dei caratteri all'interno delle parole, costituiscono, per così dire, i *ferri del mestiere* della ricerca on line, insieme ai thesauri, ai qualificatori di campo, ai comandi di limitazione, ordinamento e scorrimento dei risultati di una ricerca.
Le BDF non hanno ancora modificato le tecniche di ricerca on line a cui siamo stati abituati dalle BD bibliografiche ed il meccanismo booleano ci è indispensabile per fare ricerche su entrambi i tipi di banche. Di fatto, le uniche innovazioni introdotte dagli host a seguito della grande crescita dell'offerta full-text possono essere ricondotte a due sole caratteristiche dei sistemi di information retrieval più noti:
— la possibilità di prendere visione di porzioni del testo in cui ricorrono i termini ricercati, tramite i formati KWIC (Dialog) o HITS (Datastar) o CONTEXT (FT Profile);
— le funzioni di evidenziazione (hilighting) dei termini durante lo scorrimento dei testi (browsing).
Anche i sistemi di information retrieval nati per l'interrogazione di archivi full-text, come quello di FT Profile, non hanno aggiunto granché alla tradizionale logica booleana. L'unica innovazione di FT Profile è quella della logica top-down secondo cui sono consentite solo ricerche nelle quali ogni passo agisca sull'insieme creato dal passo precedente. Si tratta di un modo approssimativo di restringere il campo partendo da richieste generiche che vanno raffinate via via (con il comando "PICK"), particolarmente apprezzato dagli utenti finali in quanto rispecchia l'approccio più naturale all'informazione. Se la "facilità" di questo approccio è indubbia, è anche vero che essa origina almeno due inconvenienti: l'utilizzo di termini — o di stringhe di termini — generici all'inizio della ricerca (al primo step) comporta un alto rischio di rumore. Inoltre non è possibile combinare a piacere i diversi passi di ricerca: questi sono legati dalla relazione in "AND" stabilita dal sistema seguendo il percorso gerarchico, come fossero scatole cinesi.
Textline, una banca dati di tipo ibrido che comprende sia abstract che testi completi, ha ripiegato molto più tradizionalmente sull'indicizzazione manuale di ogni record. Altri esempi di BDF indicizzate manualmente sono BIG/Il sole 24 ore e quelle basi nate come bibliografiche e diventate "ibride" come Promt e ABI/Inform. L'esperienza e l'abitudine a compiere ricerche su banche dati bibliografiche, dove si limita solitamente la ricerca di un termine del linguaggio naturale al dominio di un certo campo (l'autore, il titolo, l'editore, la data di pubblicazione), si rivelano preziose durante la ricerca su BDF che pure hanno un grado di strutturazione del testo molto inferiore. I produttori delle banche dati sembrano in genere attenti a consentire la ricerca dei termini del linguaggio naturale limitata alle porzioni di testo che si considerano più rappresentative del contenuto concettuale dei documenti (il titolo, il paragrafo iniziale del testo, il sommario o "table of contents" dei report).
All'utente dell'informazione on line il mercato offre quindi essenzialmente quattro modelli di ricerca, praticabili spesso all'interno di una stessa banca dati, per ottenere indifferentemente testi completi o riferimenti bibliografici:
1) ricerca con linguaggio controllato;
2) ricerca con terminologia libera;
3) ricerca con terminologia libera, raffinamento con linguaggio controllato;

[9] Si veda il bellissimo contributo di R. DAVIES, *The Creation of New Knowledge by Information Retrieval and Classification*, nella trad. it. di C. Revelli in "Biblioteche oggi nel mondo", suppl. a "Biblioteche oggi", 8 (1990), 6, p. 87-117.

[10] R. BASCH, *Searching Newspapers Online: the times it is achangin'*, in *Online/CD ROM. Proceedings of the Conference, 5-7 November 1990.*

[11] *Reporter Uses VU/Text to Link Regional Murders*, "Worldwide Videotex Update", 1991, 6.

[12] C. TENOPIR, *Users and Uses*, cit., p. 269.

[13] M. MELOT, *Siamo tutti analfabeti, ovvero il futuro della lettura*, "Biblioteche oggi", 9 (1991), 4, p. 411-416.

4) ricerca con linguaggio controllato, ampliamento o raffinamento con terminologia libera.

Durante gli anni Ottanta la letteratura professionale anglosassone ha ospitato parecchi contributi volti a determinare se fosse preferibile la modalità 1 o la 2. Ci sembrano di particolare interesse le conclusioni di uno studio del 1985[14] che misurava la quantità di informazione recuperata da una banca dati (Harvard Business Review) provvista sia di testi completi che di abstract e thesaurus usando tre diverse metodologie: la ricerca sui testi raggiungeva un grado di richiamo del 73,9 per cento contro un grado del 28 per cento ottenuto con il vocabolario controllato e del 19,3 per cento raggiunto dalla ricerca sui termini degli abstract.

La precisione invece era del 18 per cento con il testo completo, del 34 per cento con il vocabolario controllato e del 35,6 per cento con gli abstract. Ogni metodo di ricerca comportava tuttavia il recupero di documenti rilevanti e non recuperati attraverso gli altri due metodi, per queste ragioni: il vocabolario controllato aveva permesso di recuperare anche documenti nei quali comparivano sinonimi dei termini scelti per la ricerca; la ricerca sugli abstract consentiva di legare concetti che nel testo potevano essere non espressi esplicitamente oppure espressi in differenti paragrafi; la ricerca in terminologia libera permetteva infine di recuperare documenti di argomento più vasto all'interno dei quali veniva comunque trattato, in modo pertinente alla richiesta, il tema desiderato.

Già da queste considerazioni possiamo dedurre che non è di grande utilità alla ricerca on line la contrapposizione tra linguaggio controllato e terminologia libera alla quale, a volte, ci si aggrappa per sostenere la necessità di uno sviluppo maggiore dell'indicizzazione nel campo delle BDF.[15] Lo sviluppo di liste terminologiche ed indici automatici consultabili on line è sicuramente da sollecitare ai produttori di banche dati full-text ma non dobbiamo dimenticare che il costo dell'indicizzazione manuale è improponibile, attualmente, viste le dimensioni del mercato, per la maggior parte dei produttori di BDF (queste nascono spessissimo come sottoprodotto dei processi di pubblicazione delle testate attraverso la fotocomposizione automatizzata). Una buona soluzione di compromesso sarebbe data, forse, dallo sviluppo di thesauri attivi sulle stringhe di ricerca in particolari settori disciplinari e per particolari categorie di utenti, sul modello recentemente testato presso un'università finlandese.[16]

Al momento, possiamo suggerire l'opportunità di verificare che la soluzione empiricamente più conveniente consista nella ricerca con terminologia libera associata o raffinata attraverso il linguaggio controllato. Per esempio, consideriamo che in BIG/Il sole 24 ore il maggior livello di precisione (2 documenti rilevanti su 2 recuperati) nella ricerca di "informazioni sul progetto di scioglimento dell'EFIM" si ottiene cercando:

(liquidazione OR scioglimento) ADJ2 EFIM

e raffinando il risultato di questo set con il ricorso al campo indicizzato "enti":

1 AND EFIM. ENTI.

Al contrario dell'esempio precedente, in una banca dati fornita di thesaurus ed indicizzata manualmente come PROMT la ricerca del "fatturato del Gruppo FIAT in Italia nell'ultimo anno" solo con l'uso del linguaggio controllato comporta un quasi assoluto silenzio (1 documento). Riusciamo ad accrescere il richiamo solo se utilizziamo tanto il thesaurus che la terminologia libera usata negli abstract. Il valore massimo di precisione si otterrebbe infine in questa ricerca abbreviando il termine "consolidato" come è d'uso nell'inglese economico in "cons" o "consol". Questi semplici esempi ci dimostrano la sterilità della contrapposizione tra linguaggio controllato e terminologia libera e la convenienza ad integrare i due sistemi dove è possibile.

Gli operatori booleani e gli operatori di adiacenza e prossimità consentono sì di stabilire relazioni tra i termini ma non c'è alcun modo di agire a livello testuale (del discorso). Ai fini della ricerca booleana ogni termine è considerato secondo la logica binaria: ha cioè peso uguale a 1 o 0 a seconda che venga o meno recuperato all'interno di un testo mentre, come sappiamo, ogni termine del nostro linguaggio può avere nel contesto del discorso per cui lo utilizziamo pesi molto diversi. Può assumere diversi significati a seconda del campo disciplinare in cui viene usato, può venire usato con lo stesso significato in frasi che esprimono concetti differenti attraverso differenti relazioni sintattiche oppure può essere sostituito da altri termini e/o espressioni che hanno lo stesso significato.

L'operatore "OR", nell'esempio della ricerca sull'EFIM, è stato impiegato supponendo siano stati usati nel linguaggio giornalistico indifferentemente i termini "liquidazione" e "scioglimento", anche se il termine più appropriato sarebbe "liquidazione". L'operatore "OR" consente in genere di risolvere problemi dovuti all'uso interscambiabile di sinonimi, di espressioni gergali e di abbreviazioni all'interno di un determinato contesto. Una precisazione riguardo all'"OR": più il contesto diventa generalista più si rischia di causare rumore. È opportuno perciò che l'uso dei vari termini venga verificato negli indici o nella letteratura prima di porre al sistema una ricerca in "OR".

Tenopir[17] riferisce le conclusioni di alcune indagini dalle quali emerge che il modo migliore di accrescere la precisione in una strategia di ricerca su basi full-text è dato dall'uso dell'operatore di prossimità "SAME" per combinare concetti presenti nello stesso paragrafo. "SAME" sarebbe quindi più efficace di altri operatori di prossimità e dell'"AND" per otte- ➤

[14] C. TENOPIR, *Contributions of Value Added Fields and Full-text Searching in Full-text Databases*, in *National Online Meeting Proceedings*, New York, Learned Information, 1985, p. 463-470.

[15] H. BECHTEL, *Problems Connected with Free-text Searching in CAS*, in *National Online Meeting Proceedings*, New York, Learned Information, 1990, p. 37-42.

[16] J. KRISTENSEN - K. JARVELIN, *The Effectiveness of a Searching Thesaurus in Free-text Searching in a Full-text Database*, "International Classification", 1990, p. 77-84.

[17] C. TENOPIR, *Users and Uses of Full-text Databases*, cit.

nere il mix migliore tra richiamo e precisione.
Eppure, precisa la Tenopir, attraverso l'"AND" si ottiene non solo il maggior grado di richiamo, ma anche la massima precisione in tutti quei contesti nei quali due concetti sono espressi in un documento senza che il discorso li ponga in relazione all'interno di uno stesso paragrafo. L'esempio portato, molto suggestivo, è quello di una ricerca volta ad "accertare se le zanzare possano trasmettere l'AIDS". La ricerca in "AND" ha permesso di recuperare un articolo in cui si affermava che "non è possibile prendere la malattia dalle zanzare". Se avessimo utilizzato "SAME" in luogo di "ADJ2" nella ricerca sulla liquidazione dell'EFIM, raffinando comunque il risultato di questo passo con "1 AND EFIM. ENTI", avremmo ottenuto ben 81 documenti anziché due. Questi 81 documenti sarebbero rilevanti se il nostro fine fosse quello di comprendere non solo in che cosa consiste il progetto di liquidazione dell'EFIM ma anche come questo sia stato elaborato, quali reazioni ha incontrato presso operatori economici e politici e così via. Dunque siamo d'accordo con la Tenopir nel concludere che non dobbiamo farci ingannare dai risultati di ricerche sul grado di precisione ottenuto in questo o quel modo perché in casi come quelli esemplificati misure come richiamo e precisione diventano prive di senso.
Esse vanno attentamente considerate in relazione al tipo di banca dati cui si applicano e al contesto comunicativo in cui vengono utilizzate le informazioni recuperate (per cui diverse quantità di informazione, tanto 2 quanto 81 documenti, in risposta alla stessa richiesta possono costituire la quantità ideale di informazione recuperata). È possibile dunque spiegarsi come mai una indagine sulla banca dati di Harvard Business Review conclude che la ricerca in testo libero comporta il più alto grado di richiamo ed un più basso livello di precisione rispetto ad altre strategie di ricerca, mentre un'indagine[18] su una banca dati bibliografica come FROSTI (Food research on scientific and technical information) approda alle conclusioni opposte. Dai diversi usi linguistici dei testi contenuti nelle BD dipende anche la differente performance degli operatori di adiacenza. L'uso di questi operatori è particolarmente interessante in tutti quei casi in cui si vogliano estrarre dai testi informazioni fattuali come la ricerca del nome dell'amministratore delegato di una società o dati sull'andamento dei prezzi al consumo nell'ultimo periodo.
Usando termini combinati con "ADJ" (o "W" e "N", su Dialog) ci si aspetterebbe la massima precisione. Invece si ottiene frequentemente un basso numero di documenti rilevanti: ciò dipende dal fatto che lo stile linguistico, soprattutto nelle BD giornalistiche, frappone fra i due o più termini che esprimono il concetto ricercato incisi e frasi consecutive oppure, al contrario, pone i due termini vicini ma senza che tra essi esista la relazione concettuale che ci interessa.
In questi casi la Tenopir suggerisce due alternative per aumentare la precisione: l'uso dell'operatore "NOT" per escludere quei testi dove i termini potrebbero ricorrere solo casualmente o comunque in modo non rilevante ai fini della nostra ricerca (come editoriali, recensioni di libri, rubriche fisse, ecc.) oppure la costruzione di stringhe in modo "abile" dove i termini siano messi in adiacenza tra loro tenendo conto della struttura grammaticale "possibile" nei testi. Se l'orizzonte delle fonti a testo completo disponibili on line si è enormemente dilatato negli ultimi anni, i ferri del mestiere sono dunque rimasti sempre più o meno gli stessi, facendo emergere i limiti del meccanismo booleano, particolarmente evidenti laddove non esistano alternative alla ricerca con terminologia libera. Nessuna strategia di ricerca va considerata come la migliore per tutti i tipi di testo e per tutti i diversi tipi di utenti di una banca dati ed il ricercatore si trova, forse suo malgrado, costretto ad utilizzare, integrandole, più tecniche.

LA COMPETENZA DEL RICERCATORE

La strategia di ricerca "ideale" in un sistema booleano di retrieval consisterebbe dunque nel prevedere i termini esatti e le frasi con le quali possiamo trovare espresso ciò che stiamo cercando: uno studio sullo STAIRS, il sistema di information retrieval dell'IBM ha dimostrato che in media il sistema recupera un documento rilevante su 5 (ha cioè una capacità di richiamo assoluta non superiore al 20 per cento).[19] Ciò dipende proprio dalla difficoltà a prevedere i termini esatti che ricorrono nei documenti. L'attività di ricerca on line si colloca dunque in questa linea di giuntura tra ciò che è e ciò che potrebbe essere, si qualifica come una attività profondamente soggettiva nella quale il ricercatore deve "mettere in discussione di continuo il proprio dire con altri e con se stesso"[20] e deve saper, se non prevedere, quanto meno dominare una lingua.
È opinione piuttosto comune che il documentalista o il bibliotecario che svolge ricerche on line debba essere competente nel settore disciplinare nel cui ambito si svolgono le ricerche. A questo concetto di competenza disciplinare preferiamo opporre il concetto più ampio di competenza linguistica o, meglio, comunicativa.
La competenza linguistica è il sistema di regole grammaticali della lingua che abbiamo interiorizzato e che ci consente di generare frasi corrette e comprendere frasi ambigue. Secondo la più recente visione della lingua come strumento di comunicazione, essa va vista all'interno della competenza comunicativa o semiotica che consente di cogliere in un "testo" (nella più larga accezione di "messaggio" messa a punto dalla teoria della comunicazione) le relazioni tra aspetti grammaticali e semantici del messaggio e aspetti cognitivi dell'emittente e del destinatario della comunicazione.
In questa dimensione teorica la ricerca on line — ma saremmo portati a dire tutto il lavoro di intermediazione tra fonti

[18] R. BETTS - D. MARRABLE, *Free Text vs Controlled Vocabulary - Retrieval Precision and Recall over Large Databases*, in *15th International Online Information Meeting Proceedings*, Learned Information, 1990, p. 153-166.

[19] M.A. SIDDIQUI, *op. cit.*

[20] T. DE MAURO, *Ai margini del linguaggio*, Roma, Editori riuniti, 1984, p. 87.

dell'informazione e utenti — diventa una attività fortemente dinamica il cui punto fermo è da individuarsi proprio in quel reticolo di relazioni tra fattori linguistici ed extralinguistici che determinano la comunicazione, tra forme presenti nel testo e forme assenti dal testo dalla cui connessione emergono i significati che andiamo cercando. Il ricercatore on line deve misurarsi con la necessità di elaborare strategie di ricerca come se egli fosse un alchimista ed abbandonare la presunzione di poter controllare il linguaggio esclusivamente con gli strumenti terminologici, il cui potere non è mai né esaustivo né assoluto.

"La ricerca su sistemi esperti ha dimostrato che sovente gli specialisti trovano difficoltà a ricordare e ad esprimere i concetti isolati del loro sapere se non quando li trovano nei relativi casi specifici".[21] La potenziale assenza di competenza linguistica nello specialista di una materia è la ragione di fondo per cui, ad esempio, una ricerca sulla "splenomegalia mieloide idiopatica" può essere condotta con maggiore successo — rispetto a quanto accade ad un medico — da qualcuno che, pur non essendo un esperto di medicina, abbia coscienza del fatto che questa malattia ha 12 sinonimi in inglese, 13 in tedesco e ben 31 in francese e che, nonostante esista una classificazione internazionale delle malattie pubblicata dall'Organizzazione mondiale della sanità "essa è ignorata dalla maggior parte dei medici e talvolta perfino rifiutata da certe scuole cliniche che pervicacemente seguono una loro minitradizione terminologica, in ossequio alle civetterie o ai capricci semiotici del loro Maestro".[22]

Il nostro linguaggio, riprendendo una celebre metafora di Wittgenstein, può essere considerato come una vecchia città ("un dedalo di stradine e di piazze, di case vecchie e nuove e di case con parti aggiunte in tempi diversi, e il tutto circondato da una rete di nuovi sobborghi con strade diritte e regolari e case uniformi") nella quale agli specialisti della materia, sprovvisti di competenza linguistica, viene preclusa la possibilità di movimento: essi sono in genere condannati a restare in periferia, in casette a schiera luminose ed accoglienti, ma tristemente isolati. La loro specializzazione li costringe a fare a meno di quella "fluidità" del linguaggio che "può collegare certi sensi a certi suoni, una fluidità che non pare avere limiti precostituiti",[23] nemmeno all'interno dei linguaggi specialistici, come vedremo fra poco.

Tuttavia, lo ribadiamo, la chiave della competenza comunicativa necessaria a chi compie ricerche on line è nella correlazione dei diversi aspetti del "testo", comprensivi dunque della competenza disciplinare, quell'humus che ci permette di leggere significati che nei testi non hanno significanti.

Su questo meccanismo della comunicazione portiamo un esempio cristallino fatto da Chomsky — che pure è il padre di una teoria della lingua come sistema acontestualizzato: "si consideri il fatto che oggi è giovedì. Supponiamo che io abbia un amico di cui so che tiene un corso il lunedì e gioca al tennis il giovedì. Supponiamo che egli mi dica 'oggi è stato un disastro'. Se è giovedì interpreto la sua frase come indicante che egli ha giocato una partita a tennis spaventosa, se è lunedì, che ha fatto una lezione terribile". Ci pare chiaro, dunque, il rischio al quale ci esponiamo quando isoliamo da una dimensione disciplinare la competenza linguistica esercitata durante l'attività di ricerca. Chomsky concludeva l'esempio dicendo che "questi fattori nell'interpretazione della frase sono facilmente separabili da quelli che determinano il significato letterale, intrinseco, della frase ed è sicuramente legittimo concludere che per fornire l'interpretazione interagiscono dei sistemi completamente indipendenti".[24] Ai fini della nostra riflessione noi possiamo affermare che l'interpretazione di ciò che stiamo ricercando on line e che ancora non abbiamo trovato influisce sul meccanismo che mettiamo in moto per cercarlo, da dove consegue che i sistemi sono assolutamente correlati ed interdipendenti e che siamo sempre noi — per caso o intenzionalmente — a scoprire attraverso la ricerca on line combinazioni concettuali per noi nuove quando impostiamo ricerche che non tengono conto di questa correlazione.

"La lingua riesce a essere un sistema perché è più che un sistema, perché ci permette di uscire da sé e dal suo ordine e cercare un ordine nuovo".[25] Per la ricerca on line è importante essere dotati di competenza linguistica arricchita da una dimensione comunicativa.

LE CARATTERISTICHE DEI LINGUAGGI SPECIALISTICI

Una ricerca di Tullio De Mauro sul *Lessico universale italiano* ha dimostrato che lo spazio occupato nel vocabolario italiano da parole che appartengono esclusivamente ad un linguaggio specialistico è pari ai due terzi del vocabolario stesso. "Un buon terzo del vocabolario non si lascia ricondurre a nessuna area particolare". Sul totale dei vocaboli riportabili ai linguaggi specialistici, De Mauro fornisce una indicazione del peso delle varie discipline, in valori percentuali:

biologia	16,0	scienza militare	2,9
medicina	10,9	storia	2,9
chimica	10,6	economia	2,6
geologia	9,6	geografia	2,4
diritto e Stato	4,4	fisica	2,4
artigianato	3,8	tecnologie	2,2
agricoltura	3,4	matematica	2,2
abbigliamento	3,2	cucina	2,1
linguistica	2,9	industria	2,1

I pesi in questione sono stati calcolati tenendo conto anche delle accezioni speciali di parole polisemiche che costituiscono comunque solo il 20,2 per cento delle 141.000 pa- ➤

[21] R. Davies, *op. cit.*

[22] T. De Mauro, *op. cit.*, p. 64.

[23] *Ivi*, p. 93.

[24] N. Chomsky, *Forma e interpretazione*, Milano, Il saggiatore, 1980, p. 108.

[25] T. De Mauro, *op. cit.*, p. 99.

role considerate del lessico (sul totale di 263.000 lemmi: la differenza è data da andronimi, toponimi, etnici, titoli).
"Sperare di dominare una lingua ignorando non solo e non tanto le terminologie tecniche speciali, ma le accezioni specifiche di parole di larga diffusione, è sperare nell'impossibile,[26] conclude De Mauro.
Sebbene possano presentare un grado di specializzazione lessicale molto basso, i testi delle BDF sono tuttavia da considerare espressione di "linguaggi specialistici": prodotte in determinati contesti disciplinari o di attività professionali vengono distribuite sul mercato in funzione della domanda di particolari gruppi di utenti. I linguaggi specialistici "non si differenziano dalla lingua comune per il possesso di regole linguistiche speciali e non comprese nella lingua comune quanto per un uso quantitativamente diverso di tali convenzioni"[27] e questo avviene anche all'interno delle BDF. L'orientamento degli studi linguistici attuali riguardo ai linguaggi specialistici o settoriali colloca questi ultimi nell'ambito delle varietà situazionali e contestuali, dove la lingua, in quanto strumento di comunicazione, prende forma in relazione all'argomento della comunicazione e al destinatario di questa. Da ciò dipende la caratteristica più importante, ai fini delle ricerche on line, dei linguaggi specialistici, la monoreferenzialità: in un determinato contesto semantico esiste per un dato termine un unico significato. Per esempio, nel linguaggio pubblicitario il termine "budget" significa "investimento pubblicitario stanziato da una azienda" e non "bilancio" o "preventivo" come nel linguaggio dell'economia. Il significato del termine è dunque molto preciso e il concetto non viene espresso con altri termini o con eufemismi. La monoreferenzialità dei linguaggi specialistici può essere considerata un grosso punto d'appoggio dal ricercatore on line anche quando egli ha poca familiarità con la materia della ricerca in quanto gli sarà sempre possibile scoprire qual è o quali sono i termini preferiti in un determinato contesto per esprimere un concetto (attraverso la letteratura, i glossari e dizionari, un esperto o alla peggio gli stessi testi delle BD, recuperati partendo da termini più specifici o più generali di quello che va cercando).
La monoreferenzialità produce inoltre un effetto di reiterazione del termine o dei termini "fuoco" all'interno di un discorso specialistico: questa particolarità può essere molto utile a valutare la pertinenza di un documento recuperato in quei sistemi di information retrieval che forniscono l'indicazione della ricorrenza dei termini all'interno dei testi (VU/Text). E tuttavia la quantità di ricorrenze di un termine è, come è noto, un'arma a doppio taglio in quanto, se il termine non denota un concetto "specialistico", possiamo ricadere nella trappola della ridondanza cioè nell'apparente paradosso che costituisce il principio di base della teoria dell'informazione: il contenuto di informazione è inversamente proporzionale alla probabilità di ricorrenza di un termine. Quanto più è probabile la ricorrenza di un termine tanto maggiore è la ridondanza, opposto simmetrico della nozione tecnica di informazione (quantitativa).
Si può non essere del tutto convinti di questa legge della teoria dell'informazione quando ci capita di scoprire nuova conoscenza partendo dalle sfumature di significato rinvenute nei diversi contesti in cui un termine ricorre, nelle pieghe che potremmo definire "qualitative" della ridondanza.
Un'altra caratteristica lessicale dei linguaggi specialistici è la tendenza alla sinteticità: in molti contesti linguistici i termini si riducono al loro interno o nella loro parte terminale. Nel linguaggio pubblicitario il termine inglese "ads" è di gran lunga preferito al termine "advertisements" (annunci pubblicitari). L'esigenza di sinteticità nel discorso specialistico causa il ricorso ad acronimi ed abbreviazioni, ragione di veri e proprii muri di "silenzio" documentale tanto che qualcuno, evidentemente esasperato, è arrivato a proporne l'abolizione.[28] In molti casi gli acronimi diventano talmente radicati nell'uso linguistico che viene riservata alla forma estesa una diversa sfumatura di significato: si pensi al caso, nel linguaggio economico, dell'acronimo PIL. Esso viene sempre preferito a "prodotto interno lordo" quando l'argomento viene affrontato dal punto di vista statistico o si parla degli indicatori della contabilità nazionale di un paese, mentre è più frequente l'espressione estesa nei discorsi di teoria economica.
La tendenza al tradizionalismo, ovvero l'inserimento nel discorso di termini obsoleti o, peggio, l'attribuzione di significati desueti a termini che nella lingua comune vogliono dire altre cose, è stata ravvisata in particolare nel linguaggio medico ed in quello legale e costituisce un'altra ragione di ambiguità. Quest'ultima viene a determinarsi in tutti quei casi in cui si incontrano eccezioni al principio della monoreferenzialità: presenza di sinonimi e quasi sinonimi, linguaggio metaforico o tono emotivo del discorso, imprecisione o "vaghezza" linguistica.
Il ricorso ai sinonimi è particolarmente frequente nel linguaggio medico, dove pare sia originato dall'uso di eponimi per ricordare lo scopritore di una certa malattia o rimedio ed è complicato dal fatto che spesso una stessa scoperta è rivendicata da più studiosi: il megacolon è il morbo di Hirschsprung per i danesi, il morbo di Ruysch per gli olandesi e corrisponde a due eponimi, il morbo di Battini e il morbo di Mya per gli italiani. A vantaggio dei ricercatori on line in campo medico e chimico-farmaceutico va comunque ricordata la straordinaria ricchezza di strumenti di controllo della terminologia che non ha pari in altri settori: cercando "megacolon" nel thesaurus on line della banca dati Medline si rintracciano i 10 sinonimi usati in luogo del termine preferito per l'indicizzazione, che è "Hirschsprung-disease". I sinonimi ricorrono più frequentemente man mano che ci si allontana dal linguaggio specialistico e ci si avvicina agli "umori" della lingua comune: il linguaggio del giornalismo economico, per esempio, presenta una ricchezza lessicale, con inevitabile abbondanza di sinonimi, in tutto corrispondente alla lingua comune.
Il concetto di "aiuti statali" può essere espresso attraverso

[26] *Ivi*, p. 79.

[27] M. GOTTI, *I linguaggi specialistici*, Firenze, La Nuova Italia, 1991, p. 7.

[28] H. BECHTEL, *op. cit.*

una lunga serie di altri termini e di perifrasi (agevolazioni, finanziamenti, incentivi fiscali, esenzioni fiscali, sussidi, ecc.). La ricerca di informazioni sulla "politica degli aiuti statali al Mezzogiorno intrapresa dal Governo Amato" (nella banca dati BIG/Il sole 24 ore) può essere condotta così attraverso almeno tre strade, con risultati ovviamente diversi:
1) Usando l'operatore "AND" per combinare i descrittori di tematica "Mezzogiorno", "spesa pubblica e sovvenzioni statali", "politica industriale e commerciale, sostegni pubblici a imprese" con l'espressione "Governo ADJ Amato". Questa strada ci permette di recuperare 1 solo documento (basso richiamo).
2) La seconda strada, che consente di recuperare 4 documenti dei quali 3 rilevanti, consiste nel porre al sistema la ricerca in OR di tutti i termini sinonimi che abbiamo elencato e di combinare il risultato di questo passo di ricerca con l'espressione "Mezzogiorno", tramite l'operatore di prossimità "SAME" ed infine di combinare questo secondo risultato, tramite l'"AND", con "Governo Amato".
3) La terza strada infine, che possiamo qualificare come quella ideale, ci permette di recuperare soltanto i tre documenti rilevanti ottenuti nel secondo modo, senza il quarto documento ritenuto non rilevante: questa strada consiste nel cercare in AND "Governo ADJ Amato" e "(Intervento ADJ straordinario) SAME Mezzogiorno". Può darsi che al ricercatore non venga in mente, se non ha molta familiarità con questa tematica, il fatto che quando si parla di aiuti statali al Mezzogiorno d'Italia nel linguaggio politico ed economico italiano si parla di "intervento straordinario", due parole del linguaggio comune che sono state investite dal linguaggio legale di un significato specialistico.
Una seconda ragione di ambiguità del linguaggio può risiedere nella predominanza di un tono emotivo ovvero di finalità del discorso più persuasive che informative: si tratta di contesti comunicativi in cui l'autore ricorre ad eufemismi, a perifrasi e ad espressioni metaforiche. Tipico esempio di questo genere di contesto comunicativo è dato dal linguaggio economico, dove sono frequentissime metafore ormai talmente radicate anche nella lingua comune da aver raggiunto lo stadio, come usano dire i linguisti, di "metafore morte" (si pensi ad espressioni come elasticità della domanda, equilibrio del mercato, concorrenza fra imprese, depressione economica). Un esempio magistrale di uso intenzionale della metafora per esprimere nuovi significati e porre nuovi concetti è stato indicato da Gotti nella *General Theory* di Keynes ed in particolare nel capitolo 12 sugli investimenti di lungo periodo, dove Keynes paragona l'investimento ad un gioco d'azzardo e la borsa valori ad un casinò in cui prevalgono il gusto del gioco e del rischio: termini come "player", "game", "gambling" sono usati proprio per suggerire i concetti di casualità, di imprevedibilità e di rischio associati a questo genere di attività umana. La ricerca di informazioni on line in quest'area, come del resto in quella delle acquisizioni e fusioni di imprese ed in molte altre degli affari, diventa praticabile solo utilizzando termini della lingua comune "reinventati" dalle metafore degli economisti. Quali società quotate hanno avuto la migliore performance finanziaria durante il crollo della borsa del 1987? Converrà utilizzare termini quali "sopravvissute" o "resistere" o "salve, salvate", poiché la metafora più probabile in questo contesto è quella della "resistenza" delle imprese alla "catastrofe" (il crollo della borsa, appunto).
L'ambiguità di espressioni mutuate dalla lingua comune ed utilizzate all'interno di metafore con funzioni di catacresi (cioè per dare a termini già esistenti nuovi significati) è ad ogni modo piuttosto vischiosa e trae in inganno anche gli specialisti della materia: uno studio citato da Gotti riferisce un caso di errata comprensione della metafora "parent company" che invece di essere interpretata nel suo significato di "azienda che controlla un'altra" è stata decodificata come "azienda che ha dato origine ad un'altra azienda". Naturalmente, esistono e con minori implicazioni concettuali anche metafore di tipo stilistico, il cui valore informativo è nullo: esse costituiscono un pericolo per il ricercatore on line in quanto causano rumore, ricorrendo nelle pieghe della lingua comune in tutt'altro senso rispetto a quello che ci interessa: si pensi alla ricerca di informazioni sul mercato delle trappole per topi. Usare una espressione come questa in una banca dati giornalistica senza limitarla al contesto in cui essa può assumere il significato di "categoria di prodotto" comporta il rischio elevatissimo di pescare una montagna di spazzatura! L'imprecisione, l'oscurità, la trascuratezza o l'instabilità semantica che si possono notare nella lingua comune sono fattori di disturbo né più né meno anche nei linguaggi specialistici. Un esempio di polisemia all'interno dello stesso contesto disciplinare, altamente specialistico, si trova nel linguaggio informatico laddove il termine "entry" ricorre sia nel senso di "una unità di informazione" che nel senso di "indirizzo della prima istruzione di un programma". Il linguaggio informatico è inoltre ricco di prestiti dalla lingua comune e di neologismi come ogni linguaggio giovane e nato in fretta.
Conviene cercare "alphameric" o "alphanumeric", "optronics" o "optoelectronics", "floppy drive" o "floppy disk drive"? Il troncamento e la mascheratura per quei termini che si sanno soggetti a grafie o a declinazioni differenti, possono permettere di prevenire il silenzio che è solitamente causato da questo genere di "disturbi". Certamente, ancora una volta, la presenza di un indice o di un thesaurus può offrire la chiave migliore per avviare una ricerca on line. Tutto sommato piuttosto rara all'interno di uno stesso linguaggio specialistico (contraddice infatti il principio della trasparenza e della monoreferenzialità), la polisemia è il maggior rischio di rumore della consultazione di BDF di tipo giornalistico.
La soluzione migliore nella ricerca su BDF di questo tipo può essere a volte quella di individuare modi di dire ed espressioni che ci portano dritti... "al dunque": se stessimo cercando commenti e supposizioni sul voltafaccia di Bush in tema di politica fiscale, cercare in adiacenza i termini "read" e "lips", sostiene una collega americana,[29] "may be quick and dirty, but it works", poiché, presumiamo che il cronista scriva espressioni come "ma gli si legge sulle labbra".
Anche nel caso degli aspetti sintattici del linguaggio, i linguaggi specialistici non si differenziano dalla lingua comune: piuttosto, in essi ricorrono regole sintattiche che ➤

[29] R. BASCH, *op. cit.*

esistono già nella lingua comune ma con frequenza differente. Come si è già visto, la logica booleana ci consente di articolare una ricerca (stringa) con operatori di adiacenza e prossimità attraverso cui, di fatto, cerchiamo di ipotizzare la costruzione di una frase o la disposizione dei concetti all'interno di un periodo.
Nei linguaggi specialistici la sinteticità porta ad usare il paragrafo "come quella parte di testo che esprime un determinato concetto o che svolge una funzione pragmatica ben precisa".[30] Tuttavia, abbiamo già visto (nell'esempio dell'AIDS/zanzara) che "ben raramente esiste un rapporto di 1:1 tra paragrafo fisico e paragrafo concettuale".[31]
Perciò la costruzione dei periodi è fonte di grandi incertezze nell'uso degli operatori "AND" e "SAME" o "WITH".
La presenza di frasi subordinate, benché sembra che esse non costituiscano più del 25 per cento delle frasi di un testo specialistico,[32] può vanificare l'utilizzo dell'operatore di adiacenza, come abbiamo già detto.
La disposizione dei periodi all'interno del testo va in un certo senso immaginata perché anche nelle BDF meno strutturate, essa costituisce un elemento utile a determinare la strategia di interrogazione più efficace. Nelle BDF di tipo giornalistico, è conveniente ricercare i termini che ci interessano (che esprimono concetti "fuoco") nel "lead paragraph", cioè nel paragrafo iniziale degli articoli, perché l'apertura di un testo giornalistico contiene solitamente tutti gli elementi su cui si articolerà il discorso nel seguito del testo. La consapevolezza del genere testuale presente nella banca dati da interrogare ci sembra costituire quindi una garanzia per impostare bene la strategia di ricerca. Se, per esempio, la banca dati contiene anche abstract (è il caso delle BD già definite di tipo ibrido) può esser utile limitare la ricerca ai soli testi degli abstract per ottenere maggiore precisione. Nei testi completi di ricerche di mercato le informazioni vengono presentate, analizzate e commentate secondo uno schema predefinito ed il più delle volte accessibile attraverso il sommario o la "table of contents": è conveniente allora limitare la ricerca a queste porzioni iniziali del testo per poi visualizzare i paragrafi o le pagine in cui ricorrono i termini ricercati, scorrendone prima il testo attraverso formati come il KWIC, per sicurezza.
Un preciso genere testuale può inoltre costituire esso stesso la risposta ad una ricerca on line: è il caso di ricerche volte ad ottenere sintetici giudizi su situazioni economiche o brevi valutazioni finanziarie di una azienda. In questi casi le strategie di ricerca dovranno contenere elementi formali che hanno il potere di disambiguare o qualificare termini generici (a patto che tali elementi formali siano resi ricercabili dal produttore della banca dati, naturalmente). Per esempio nella banca dati Country Report Service della PRS, una tabella sintetica di indicatori economici sull'Italia si ottiene cercando l'espressione "fact sheet".
Similmente, si possono estrarre dati dalla banca dati Business International specificando che la caratteristica speciale dei testi da ricercare è "SF=Table". L'attenzione posta dai produttori delle BD nella strutturazione dei testi ai fini della ricerca compensa spesso la mancanza di un'indicizzazione manuale. Nelle BDF di giornali quotidiani è molto importante poter distinguere tra i vari tipi di articoli (editoriali, inchieste, interviste, ecc.) in quanto essi corrispondono a diversi generi testuali che hanno più o meno probabilità di contenere le informazioni che cerchiamo. "L'uso convenzionale di canoni stilistici appartenenti ai vari generi testuali crea nel pubblico dei destinatari precise aspettative"[33] e nel ricercatore può dunque originare precise strategie di ricerca.
"La standardizzazione testuale è costante in tutti i campi disciplinari e diviene massima nei casi in cui il testo non viene scritto appositamente ogni volta, bensì equivale alla rielaborazione di un testo precedente, in cui vengono inclusi i dati riferentesi alla nuova situazione".[34] È questo il caso dei bilanci e degli Annual Reports aziendali.
La sinteticità espressiva che caratterizza i linguaggi specialistici — causando fenomeni quali l'omissione di articoli e preposizioni, la sostituzione di frasi relative con aggettivi, la trasformazione del verbo di una frase relativa in participio presente (in specie nell'inglese) — non ci pare causare problemi nella ricerca on line. Anzi, in teoria, questi fenomeni potrebbero rappresentare un vantaggio poiché riducono la ridondanza di lemmi poco significativi. La ellitticità del linguaggio specialistico comporta tuttavia l'esigenza di prevedere un uso accorto degli operatori di adiacenza e prossimità per prevenire le ambiguità causate dalla aggettivazione nominale, cioè "dell'uso di un sostantivo che ne specifica un altro in funzione di aggettivo di esso",[35] particolarmente frequente nella lingua inglese che a differenza dell'italiana tende a "costruire la frase a sinistra" anziché a destra.
Profonde differenze di significato sono insite per esempio in forme quali "a small car factory". Questa espressione può voler dire tanto "a small factory for making cars" quanto "a factory for making small cars". In questi casi, non si riesce a trovare una soluzione all'interno del sintagma: per prevenire il problema l'unica strada consiste nell'aggiungere alla stringa di ricerca altri elementi, estranei al contesto di quel sintagma, che possono contraddistinguere il dominio di significato a cui vogliamo che il nostro testo appartenga, prendendoli dalla terminologia libera o da quella controllata, come abbiamo già visto ("EFIM. ENTI", per esempio).
Quanto alle forme verbali, sembra prevalente nei testi specialistici la forma presente dei verbi. Ciò è da tenere presente nel caso di inclusione di forme verbali in una stringa di ricerca: queste possono essere utili in funzione di "qualificatori" con grande potere di richiamo. Se volessimo recuperare per esempio tra i comunicati stampa di una azienda solo quelli che riportano notizie su investimenti e conti economi-

[30] M. GOTTI, *op. cit.*, p. 117.
[31] *Ivi*, p. 118.
[32] *Ivi*, p. 83.
[33] *Ivi*, p. 116.
[34] *Ivi*, p. 121.
[35] *Ivi*, p. 72.

ci insieme a dichiarazioni di rappresentanti dell'azienda stessa potremmo usare "ha ADJ (detto or dichiarato)" per assicurarci un buon grado di precisione.
Questo esempio ci consente anche di puntualizzare che la scelta delle forme verbali è legata non tanto all'asse temporale quanto a fattori retorici legati al tipo di testo e all'argomento trattato, per cui se è vero che nelle BDF di quotidiani e agenzie di stampa il presente è certamente il tempo più usato, è anche vero che per estrarre dai testi dati o notizie economiche di tipo previsionale può essere utile l'inserimento nella stringa di ricerca di forme verbali al futuro.

IL COMPORTAMENTO DEL RICERCATORE

Una interessante indagine compiuta di recente in una biblioteca americana[36] sul comportamento di un gruppo di utenti finali durante la ricerca on line su BDF, conclude che il comportamento di una persona alle prese con una ricerca on line, afferisce a tre dominii:
— un dominio emotivo; il comportamento rispecchia gli obiettivi e gli scopi della ricerca, i bisogni e la motivazione dell'utente;
— un dominio cognitivo: il comportamento dipende dalla conoscenza del ricercatore e le sue decisioni possono essere viste come l'esecuzione di sequenze simboliche di piccoli passi per la soluzione di problemi a lui noti;
— un dominio sensomotorio, che concerne l'attività di scansione del testo, di lettura nonché di digitazione.
Ci sembra che i tre dominii siano strettamente correlati non solo per gli utenti finali ma anche per i professionisti dell'informazione. Questi ultimi possono avvantaggiarsi di una maggiore capacità di "governo" del dominio emotivo grazie all'esperienza professionale negli altri due. Eppure si pone loro, di continuo, nella ricerca di informazioni e conoscenze nuove che contraddistingue l'utilizzo delle BDF, l'esigenza di trovare una linea guida. Quale dovrebbe o potrebbe essere il comportamento ideale durante la ricerca on line su BDF?
La questione si pone in una prospettiva che esula dagli aspetti più tecnici e trova, secondo noi, risposta nella competenza linguistica, intesa come creatività, da un lato, e in una disposizione psicologica di stupore davanti alla conoscenza dall'altro, "quello stupore il cui significato consiste nel non presumere di sapere troppo".[37]
La competenza linguistica a cui abbiamo voluto fare riferimento in precedenza, intesa nell'accezione ampia di competenza comunicativa del ricercatore di informazioni on line, ha un corrispondente nel concetto di creatività di cui parlano i linguisti. Questa può essere intesa come "disponibilità alla variazione delle forme di un sistema o di un codice semiologico insita negli utenti del sistema o codice e riconoscibile come proprietà del sistema o codice stesso".[38] La creatività permette di "muoversi all'interno e, per dir così, all'esterno di sistemi e codici linguistici e non linguistici diversi, dandosi diverse tecniche e all'occorrenza mutandole".[39] De Mauro individua nel personaggio della novella di Andersen, Gian Babbeo, il prototipo degli individui creativi: uno che si rapporta alle cose ed alle situazioni in modo "strano", che non conosce a memoria tutto il vocabolario latino e le ultime tre annate del giornale locale, né ha studiato tutti i paragrafi dei regolamenti delle corporazioni d'arti e mestieri, come hanno fatto i suoi due fratelli. Gian Babbeo è uno che si avvia a corte cavalcando un caprone, perché il padre ed i fratelli gli hanno rifiutato un cavallo, e lungo la strada raccoglie strani oggetti che gli saranno utili per improvvisare davanti alla figlia del re una conversazione di cui nessun altro, nemmeno i suoi eruditi fratelli, sarà capace.
Assumere un atteggiamento di stupore significa permettere alla nostra mente di non restare imbrigliata nella ovvietà sterile di ciò che già sappiamo, o presumiamo di sapere: "i nostri concetti sono i cartelli indicatori degli interessi che dirigono i nostri problemi e le soluzioni che ne approntiamo".[40] E allora dovremo abbandonarli o cercare di farlo: cambiare punto di vista disciplinare, rifiutare i primi termini che ci vengono in mente per tradurre un concetto e cercarne altri, magari in un primo momento bizzarri come lo zoccolo rotto o la cornacchia morta o la manciatina di fango che Gian Babbeo raccoglie lungo la strada, sbeffeggiato dai fratelli: non importa. Lo stupore alimenta la creatività, ci premette di seguire l'intuizione che ci è data da un sinonimo a cui non avevamo pensato prima. Accettiamo di procedere un po' a tentoni, di usare nella ricerca anche termini ridondanti come le forme verbali in congiunzione all'aspetto a prima vista più marginale e meno significativo dell'argomento, di utilizzare il browsing su un campione casuale di record recuperati con una ricerca che si è rivelata troppo generica, alla caccia di altre idee. Accettiamo cioè l'evidenza del "testo": "il linguaggio non è mai solo; quando è solo aspira al possesso dei fatti, dei valori, anziché realizzare la condizione del loro significato e della loro possibilità [...]. Il processo della svolta del tempo nostro è costituito dalla rinuncia alla verità come tensione verso il cumulo e il possesso dei fatti, per realizzare invece l'analisi e la condizione del loro senso". Questo processo non si può compiere solo con lo strumento tecnologico, né solo con la competenza linguistica e men che meno solo con la competenza disciplinare ma piuttosto "ha origine nella disposizione etica, secondo la quale guardiamo aspetti delle cose".[41]
"E così Gian Babbeo divenne re, ebbe una sposa, una corona e un trono: l'abbiamo giusto letto nel giornale del vecchio decano — ma di quello non c'è da fidarsi", concludeva la fiaba di Andersen. ■

[36] C. TENOPIR, D. NAHL-JACOBOVITS, D. LEE HOWARD, *Full-text Search Strategies and Modifications: the Role of the Searcher and the Role of the System*, in *National Online Meeting Proceedings*, New York, Learned Information, 1990, p. 389-399.

[37] A. GARGANI, *Lo stupore e il caso*, 3ª ed., Roma-Bari, Laterza, 1992.

[38] T. DE MAURO, *Minisemantica*, Roma-Bari, Laterza, 1986, p. 53.

[39] *Ivi.*

[40] A. GARGANI, *op. cit.*, p. 67.

[41] A. GARGANI, *op. cit.*, p. 53 e 67.