

TESTO E COMPUTER: ELEMENTI DI LINGUISTICA COMPUTAZIONALE

Alessandro Lenci, Simonetta Montemagni, Vito Pirelli

1. I DATI DELLA LINGUA

Competenze del linguista computazionale:

1. Capacità di selezionare e raccogliere i dati linguistici più appropriati per i propri scopi;
2. Conoscenza di *metodi formali* (logico-algebrici, statistici, matematici e computazionali) per l'analisi di questi dati;
3. La padronanza di *tecniche informatiche* con cui condurre automaticamente le proprie analisi.

1.1 Le fonti dei dati linguistici

Dato linguistico → i prodotti del linguaggio che sono oggetto di un processo di analisi. Cosa viene scelto come dato dipende dalla tipologia di fenomeni che vogliamo indagare. I dati linguistici possono essere: *parole, frasi, enunciati*.

Problema preliminare: come individuare le fonti → due fonti principali:

1. I *testi dei parlanti* di una lingua, intesi come qualsiasi prodotto dell'attività linguistica dei parlanti elaborato come sequenza di caratteri;
2. I *parlanti*.

Il linguista procede preparando dei test ad hoc per lo studio di un particolare fenomeno; questi vengono somministrati a un gruppo selezionato: le risposte sono i dati oggetto di indagine.

1.1.1 Dati 'ecologici' e dati controllati

Una variabile fondamentale è la *naturalità* del contesto in cui i dati sono raccolti. Possiamo parlare dei dati testuali come *dati linguistici 'ecologici'*, a fronte dei *dati linguistici controllati* ottenuti somministrando ai parlanti i testi elaborati dal linguista.

Lo scienziato può osservare il soggetto x nel suo habitat naturale, limitandosi a registrare il comportamento; successivamente elaborerà le sue teorie e i suoi modelli analizzando il risultato delle osservazioni. In alternativa può seguire osservazioni in *laboratorio*. In questo caso l'osservazione è controllata: prepara un contesto sperimentale *ad hoc*.

I dati linguistici ricavati dai testi assomigliano molto da vicino a osservazioni naturalistiche. *Il testo rappresenta l'habitat naturale dei dati linguistici*.

I dati estratti da testi sono 'ecologici', in quanto osservati e raccolti nel proprio ambiente di cui dunque conservano tutta la naturalità. I dati linguistici ricavati attraverso la somministrazione di test ai parlanti sono controllati.

È possibile così ottenere dati altrimenti difficili da isolare nei testi reali prodotti dai parlanti.

La LC ha evidenziato la necessità di combinare in modo equilibrato i dati linguistici naturali con il ricorso ai dati controllati. I modelli e gli strumenti computazionali devono essere ben adatti all'ambiente linguistico per il quale sono progettati.

1.2 I corpora

Un *corpus* è una collezione di testi selezionati e organizzati in maniera tale da soddisfare specifici criteri che li rendono funzionali per le analisi linguistiche. I *corpora testuali* rappresentano la principale fonte di dati in LC.

Il fattore propulsivo fondamentale che ha promosso la creazione e l'uso dei corpora è stato lo sviluppo della tecnologia informatica. La raccolta di testi costituiva la prassi comune nello studio sul linguaggio prima della nascita della grammatica generativa chomskiana.

Il computer permette:

- a. Immagazzinare quantità sempre crescenti di testi;
- b. Ottimizzare la loro esplorazione e la ricerca di dati linguistici interessati;
- c. Sviluppare modelli computazionali della lingua.

Il ruolo dei computer nell'uso dei corpora è così cruciale che ormai il termine stesso di *corpus* è diventato di fatto sinonimo di *corpus elettronico*.

1.2.1 Tipi di corpora

Ogni corpus è il risultato di un'opera di selezione. Scegliere un corpus come fonte di dati linguistici per l'analisi computazionale richiede di valutare se il suo contenuto e organizzazione sono appropriati per i nostri scopi.

Parametri dei corpora:

1. *Generalità* → il grado di generalità di corpus dipende dalla misura in cui i suoi testi sono stati selezionati in maniera *trasversale* rispetto a varietà diverse di una lingua. I *corpora specialistici* o *verticali* hanno il grado minimo di generalità. In questo caso l'ampiezza del linguaggio che il corpus cerca di scrivere è ristretta: si tratta in genere di linguaggi settoriali.

All'estremo opposto si collocano i *corpora generali*, i cui testi appartengono alle diverse varietà- i corpora generali sono *plurifunzionali*, sono spesso progettati come *risorse trasversali di riferimento* per lo studio di una lingua → nome '*corpora di riferimento*', spesso articolato in *sottocorpora*.

2. *Modalità* → la diversità che caratterizza a tutti i livelli la lingua scritta e quella parlata rende la modalità di produzione dei testi un parametro rilevante per definire la fisionomia del corpus. Qui si possono distinguere:
 - a. *Corpora di lingua scritta* → solo testi originariamente prodotti in forma scritta;
 - b. *Corpora di lingua parlata* → prodotti in modalità orale e successivamente trascritti; devono essere distinti dai *corpora audio*.
 - c. *Corpora misti* → proporzioni variabili sia testi prodotti in modalità scritta sia trascrizioni di lingua parlata.

Un corpus ancora più di frontiera è il *corpus multimodale* o *audio-visivo*.

3. *Cronologia* → i corpora differiscono anche per il modo in cui i testi sono selezionati rispetto all'asse temporale. Un *corpus sincronico*: testi che appartengono a una stessa finestra temporale. *Corpus diacronico*: testi appartenenti a periodi diversi.
4. *Lingua* → i *corpus monolingue* / *bilingue* possono essere ulteriormente distinti in *corpora paralleli* e *comparabili*. Se le unità linguistiche dei testi L1 sono esplicitamente collegate alle unità di L2: *corpus parallelo allineato*.
Un corpus *bilingue* contiene testi originali in lingue diverse; è comparabile nella misura in cui i criteri di selezione dei testi sono gli stessi nelle varie lingue.
5. *Integrità dei testi* → un corpus può contenere *testi interi* o *porzioni*.
6. *Codifica digitale dei testi* → nei *corpora codificati ad alto livello* i testi sono arricchiti con *etichette* che ne rendono espliciti vari tipi di informazione.

Queste diverse dimensioni contribuiscono a comporre la fisionomia di un corpus (*estensione* → *numero di parole o token*).

Concentrandosi sui corpora generali, la loro evoluzione è contraddistinta dall'ampliamento costante dell'estensione.

La grandezza tipica dei *corpora* negli anni 60-70 è circa 1m di parole → modello riferimento: *Brown Corpus*: primo corpus elettronico progettato come riferimento per lo studio di una lingua.

Negli anni 80 la grandezza dei corpora è aumentata e va a crescere.

I corpora anche di grandi dimensioni sono comunque realtà 'chiuse': restituiscono una sorta di 'fotografia' di una lingua attraverso i testi selezionati, ma non sono adatti a seguire i mutamenti.

Un *corpus di monitoraggio* è una collezione 'aperta' di testi che muta nel tempo, conservando la fisionomia originale; viene usato in contesti lessicografici come fonte di dati per mantenere aggiornati i dizionari.

Linee di tendenza nell'evoluzione dei corpora:

1. I prodotti di prima generazione non misti. I corpora generali più recenti ospitano proporzioni variabili di parlato trascritto;
2. Numero crescente di corpora audio e multilingui;
3. Numero crescente di lingue per le quali esistono i corpora;
4. Includere in un corpus testi interi (ad eccezione del *BNC*);
5. Testi codificati in linguaggi di marcatura e schemi di codifica standardizzati;
6. Corpora annotati con la categoria grammaticale a cui si aggiungono informazioni sintattiche, semantiche ecc;
7. Strumenti informatici sofisticati che permettono la consultazione ed esplorazione dei grandi corpora di riferimento.

1.3 Il corpus come campione

Il grado di adeguatezza di un corpus come fonte di dati per una particolare analisi linguistica è determinato dalla dimensione *quantitativa* e *qualitativa*:

- *Quantitativa* → le dimensioni della finestra sono importanti per il tipo di osservazioni che vogliamo compiere e per avere maggiori probabilità di osservare quei fenomeni o strutture che sono rilevanti per i nostri scopi.
- *Qualitativa* → un corpus è il risultato di una scelta di testi che sono giudicati significativi per esplorare una lingua; è anche importante dove e come decidiamo di costruire il nostro punto di osservazione.

Nella *linguistica dei corpora*, la nozione di corpus ha trovato una definizione e trattazione scientifica rigorosa. Questa disciplina si è fatta portavoce della necessità di potenziare i corpora proprio per quanto riguarda il *controllo* delle modalità di selezione dei testi, accanto agli aspetti meramente quantitativi.

Un corpus si configura così come un *campione* di una lingua o una sua varietà.

Porre l'attenzione sull'uso dei corpora come campioni di testi ha conseguenze rilevanti sul modo di progettare e costruire i corpora.

1.3.1 *Rappresentatività e variabilità*

La possibilità di estendere le informazioni trattate da un campione alla sua popolazione dipende da quanto il campione è *rappresentativo* della popolazione stessa.

La rappresentatività agisce come vincolo qualitativo e quantitativo sulla capacità del corpus di fornirci un modello in scala delle proprietà di una lingua. Se questa condizione non è soddisfatta, non possiamo essere sicuri che l'evidenza del corpus corrisponda effettivamente a proprietà reali della lingua.

I corpora vengono generalmente distinti da collezioni di testi, archivi o biblioteche digitali. Il modo in cui i testi di una biblioteca sono raccolti è 'opportunistico' rispetto alle esigenze o ai gusti letterari dei suoi utenti oppure rispetto a una specifica politica culturale.

Per essere rappresentativo un corpus deve tenere traccia dell'intero ambito di variabilità dei tratti e proprietà di una lingua.

Diversi tipi testuali si legano a differenti situazione e scopi comunicativi.

Nel costruire un corpus rappresentativo di una lingua o di una sua varietà dobbiamo selezionare un campione di testi che "ci fornisca un'immagine il più accurata possibile delle tendenze della varietà in questione, comprese le loro proporzioni".

La complessità dell'operazione di selezione dipende dalla trasversalità o generalità della lingua che il corpus deve rappresentare. Il caso più semplice è costituito dai corpora specialistici.

In questi casi i parametri di variabilità interna della lingua sono limitati e comunque più facilmente controllabili.

1.3.2 *Corpora bilanciati*

Un ordine di complessità maggiore è presentato dai corpora che vogliono essere rappresentativi di una lingua nel suo complesso. L'obiettivo dei generali corpora è essere risorse di riferimento trasversali di una lingua.

La necessità di tener traccia dell'ampio spettro di variabilità dei tratti linguistici si concretizza nel requisito di *bilanciamento*. Il bilanciamento è assunto come condizione essenziale per garantire la

rappresentatività di un corpus che voglia essere plurifunzionale e trasversale rispetto alle diverse varietà di una lingua.

Il bilanciamento presuppone una descrizione accurata della popolazione di riferimento: è necessario definire una mappa della lingua tracciando:

- a. I confini spaziali e temporali;
- b. La tipologia dei testi.

Un esempio di bilanciamento particolarmente sofisticato e articolato è quello del *BNC*, che rappresenta uno standard qualitativo *de facto* nei corpora di ultima generazione. Il corpus contiene 90m di parole di testi scritti e 10m di parole di parlato trascritto.

Nella parte di lingua scritta, i testi sono stati selezionati sulla base del dominio e del 'medium'. La componente del parlato trascritto è suddivisa in una parte 'demografica', contenente trascrizioni di conversazioni spontanee ('contestualizzate') e trascrizioni di parlato prodotto in contesti comunicativi particolari.

Gli aspetti qualitativi del bilanciamento non possono essere disgiunti dalla dimensione quantitativa. Un bilanciamento corretto richiede una quantità consistente di testi selezionati per le diverse tipologie individuate nella popolazione.

1.4 I corpora in LC: istruzioni per l'uso

Il linguista computazionale usa tipicamente un corpus come *fonte di evidenza per definire modelli linguistici e sviluppare strumenti informatici per l'elaborazione della lingua*.

- *Evidenza è qualitativa* → riguarda *quali* strutture devono entrare a far parte della competenza linguistica rappresentata nel modello o nel programma.
- *Evidenza quantitativa* → ovvero *quante volte* una certa espressione o struttura linguistica ricorre in un corpus.

La composizione e l'estensione del corpus giocano un ruolo fondamentale nel determinare il grado di affidabilità dell'evidenza.

1.4.1 I limiti della rappresentatività

I corpora sono raccolte di 'osservazioni' dell'uso linguistico di parlanti reali. Come tali hanno spesso suscitato forti obiezioni circa la loro effettiva possibilità di costituire fonti adeguate di evidenza linguistica.

Il limite intrinseco dei corpora è quello di essere insieme *finiti* di registrazioni di usi linguistici e di essere comunque parziali.

L'affidabilità di un corpus come fonte di dati linguistici dipende dalla sua capacità di fornirci un modello fedele del lessico e della grammatica di una lingua. Secondo Chomsky, il modello offerto da un corpus è destinato a essere sistematicamente fuori scala e distorto.

Un corpus è dunque una fonte di 'curiosità' linguistiche, ma non una fonte di evidenza a partire dalla quale sviluppare modelli della conoscenza della lingua.

La linguistica dei corpora è andata affermando ripetutamente la natura relativa più che assoluta della rappresentatività. Per la linguistica dei corpora il controllo degli aspetti qualitativi di un corpus è lo strumento fondamentale per migliorarne la rappresentatività. Tutti i grandi corpora di riferimento disponibili oggi sono di fatto figli di questa strategia.

Ogni corpus è di fatto il risultato dell'applicazione di metodologie di campionamento rigorose miste a soluzioni pragmatiche e all'intuizione del progettista del corpus. La selezione dei tipi di testi è strettamente legata anche alla loro effettiva disponibilità. Spesso, una certa tipologia di testo compare in un corpus con un'alta percentuale semplicemente perché rappresenta tutto ciò che è recuperabile nei tempi e con le risorse di un progetto.

Alcuni corpora nascono anche con una vocazione esplicitamente "opportunistica", in quanto raccolgono materiale testuale selezionato semplicemente per la sua abbondante disponibilità in formato digitale.

Il concetto di corpus bilanciato presuppone una descrizione adeguata della popolazione di testi da campionare.

La composizione di un corpus, anche quando è realizzata seguendo rigorose tecniche statistiche di campionamento, è dunque *sempre* dipendente dalla particolare prospettiva con cui organizziamo e raggruppiamo i testi.

Per tali motivi, più che la nozione di bilanciamento "quello che conta è sapere che il proprio corpus è sbilanciato".

Nella LC esiste dunque un forte interesse a elaborare strumenti per controllare la variabilità dei corpora. In generale, in ogni tipo di analisi computazionale è necessario verificare accuratamente se e in che misura i risultati e le prestazioni ottenute siano stati influenzati dalla composizione del corpus da cui abbiamo estratto i dati linguistici. Se il corpus perfettamente rappresentativo non esiste, la LC è sempre più consapevole dell'importanza di metodi e strategie che permettano di controllare e limitare gli effetti dei possibili sbilanciamenti dei corpora.

1.4.2 Il corpus benchmark

Oltre che dal suo grado di rappresentatività, la scelta di un corpus come fonte di evidenza in LC può anche dipendere dalla misura in cui esso costituisce uno *standard di riferimento* per un particolare tipo di applicazione o analisi linguistica.

Diventa essenziale disporre di un corpus di dati testuali selezionati con cura allo scopo di 'mettere alla prova' i nostri programmi. Per tale motivo alcuni corpora hanno progressivamente assunto il ruolo di *standard de facto* per una certa comunità come dati di confronto nelle valutazioni.

La stessa tendenza però vale anche per corpora specialistici e per applicazioni ben definite.

Tra molti fattori che contribuiscono a rendere un corpus uno standard di riferimento vi sono chiaramente la qualità stessa del corpus, la facile disponibilità e il fatto che siano ben noti i suoi limiti e la sua composizione.

1.4.3 I corpora specialistici

I corpora specialistici sono estremamente utili per lo sviluppo di sistemi che siano fortemente adattati a un particolare tipo di linguaggio.

Al posto dei grandi corpora generali, per il linguista computazionale è dunque possibile optare per corpora molto focalizzati su un particolare dominio di interesse.

Un problema connesso all'uso dei corpora specialistici è che l'evidenza linguistica che essi forniscono è spesso generalizzabile solo in maniera limitata. È possibile sviluppare modelli e strumenti che operano su ambiti linguistici ristretti, ma che sono allo stesso tempo dotati della capacità di estendere e adattare rapidamente tale competenza a nuove varietà della lingua.

In alternativa, possiamo concepire la competenza generale di una lingua come il risultato di un processo di estensione progressiva a partire da competenze settoriali che vengono gradualmente ampliate e generalizzate grazie a un'incredibile capacità di adattamento a nuove varianti.

Invece di sviluppare sistemi dotati di conoscenze linguistiche generali, ricavate da fonti di dati generali, si preferisce spesso sviluppare sistemi e modelli "specializzati" su domini linguistici locali, e dunque su dati provenienti da corpora verticali, dotando però allo stesso tempo tali sistemi di capacità di adattamento linguistico sempre più sofisticate.

1.4.4 Corpora di addestramento

In quei settori della LC in cui maggiormente vengono usati metodi di analisi statistica, si è andata recentemente affermando una visione puramente quantitativa dei corpora come fonte di dati.

L'applicazione dei metodi statistici all'analisi computazionale del linguaggio si basa sulla possibilità di costruire modelli di un si basa sulla possibilità di costruire modelli di un fenomeno linguistico a partire dagli eventi osservati all'interno di un corpus.

Gli eventi osservati possono essere le parole del corpus, i loro significati o categorie sintattiche, ma anche espressioni e strutture più complesse come sequenze di parole, sintagmi, frasi, sequenze di frasi ecc.

In un corpus di addestramento è possibile raccogliere dati quantitativi sull'occorrenza di particolari eventi linguistici. I metodi statistici permettono quindi di trasformare le regolarità rilevate nei dati in modelli con cui effettuare previsioni su un dato fenomeno linguistico.

Questi modelli sono usati per l'analisi computazionale del linguaggio.

L'affidabilità e generalità dei modelli statistici dipendono da due fattori:

1. Quali espressioni sono attestate nel corpus;
2. Quante volte sono attestate.

Poiché il corpus è una porzione limitata di una popolazione linguistica, ci sono eventi linguistici che non siamo in grado di osservare. Anche relativamente agli eventi attestati, in un corpus vi è sempre una grande quantità di eventi linguistici *rari*.

Se osserviamo un evento una sola volta, non possiamo sapere se sia accaduto per caso, poiché i dati linguistici di un corpus sono rari è spesso estremamente difficile ricavare da essi modelli statistici affidabili.

Il bilanciamento di un corpus non può compensare la mancanza di dati sufficienti per ricavarne inferenze affidabili: ipotesi → la fonte di dati migliore è quella con l'estensione maggiore.

1.4.5 Usare il web come corpus

Il web è una miniera d'oro per gli studi linguistici: i suoi filoni sono in continua espansione e ancora largamente inesplorati.

È anche per sua natura una risorsa di informazione testuale multilingue, sebbene la prevalenza sia inglese (anche se la tendenza sta mutando).

Dal punto di vista di dati quantitativi il web è vincente rispetto a qualunque corpus esistente; resta da valutare in che modo possa costituire una fonte di dati linguistici.

I testi disponibili appartengono a un ampio spettro di generi, dai quali è possibile attingere per creare corpora generali e specialistici, mono-multilingui.

Il web stesso è il corpus su cui vengono effettuate le analisi computazionali.

Sebbene sia immenso, molte varietà di una lingua non sono rappresentate: es. parlato. Inoltre, il mezzo elettronico ha determinato esso stesso la nascita di un particolare tipo di varietà, la lingua del web.

Dall'altro lato è comunque una collezione di testi da utilizzare come fonte di dati per le analisi linguistiche.

I tipi di dati linguistici che possiamo estrarre dal web sono molteplici es. dati ortografici o varianti rare, difficilmente attestati nei corpora tradizionali.

I corpora, in particolare quelli generali, sono usati come risorse di riferimento per lo sviluppo di dizionari o lessici.

Il dinamismo del linguaggio nel web permette di ricavare informazioni interessanti su neologismi o nuovi sensi delle parole. A fronte di questi usi più semplici, è possibile impiegare il web anche per estrarre tipi di informazioni linguistiche più complesse.

Le dimensioni del web possono essere usate in particolare per alleviare il problema della rarità dei dati linguistici ma, non è ancora possibile affermare se e in che misura i nuovi ordini di grandezza dei dati disponibili sul web saranno in grado di portare un'innovazione negli studi computazionali sul linguaggio.

2. IL TESTO E LA SUA CODIFICA DIGITALE

Un testo è una struttura complessa, che contiene informazioni di tipo diverso, articolate su più livelli: la sequenza di caratteri che si combinano a formare parole, le quali entrano a far parte di una rete di relazioni e strutture linguistiche astratte e si raggruppano in unità testuali con funzioni specifiche, come titoli, capitoli e paragrafi. La rapidità e l'apparente assenza di sforzo con le quali siamo in grado di accedere ai molteplici livelli della struttura testuale per estrarne informazioni non diminuiscono la complessità di questa operazione.

Un computer non possiede queste conoscenze ed è in grado di vedere e manipolare solo sequenze di *codici binari*.

2.1 La codifica digitale del testo: il problema

I computer memorizzano ed elaborano dati sotto forma di sequenze di simboli 0 – 1 aggregati in sequenze di 8 cifre (*byte*). I testi per essere elaborati o trasmessi da un programma devono dunque avere una *rappresentazione binaria*. Ciascun carattere alfanumerico deve essere rappresentato nei termini di un codice binario composto da una sequenza di *bit*.

Accanto a questa dimensione lineare, ne esiste un'altra lungo la quale si sviluppano i livelli di organizzazione del testo e la sua struttura linguistica.

Con una codifica che si limiti ad associare a ogni carattere del testo una rappresentazione binaria, vi sarà una perdita di informazione.

Questa perdita si verifica perché una parte dell'informazione del testo non è convogliata dalla sequenza dei caratteri che lo compongono, ma è *implicitamente* veicolata attraverso la sua formattazione. È il caso dell'informazione degli aspetti macrostrutturali e coordinate metatestuali.

Per rendere esplicito questo tipo d'informazione è necessaria una codifica che si basi sull'identificazione di intere porzioni di testo e su indicazioni esplicite della loro funzione.

Esiste un ulteriore livello d'informazione testuale: la *struttura linguistica* del testo. È la chiave primaria di accesso al suo contenuto.

2.2 Livelli di codifica

La rappresentazione binaria di un testo è articolata su due livelli:

1. *Basso livello* → *codifica livello zero*: rappresentazione binaria della sequenza ordinata di caratteri del testo;
2. *Alto livello* → arricchisce il testo codificato al livello zero con informazione relativa alla struttura linguistico-testuale. A sua volta richiede:
 - a. Selezione degli aspetti funzionali e strutturali del testo che si considerano rilevanti e che si vogliono rendere accessibili al calcolatore con una rappresentazione esplicita;
 - b. Scelta di un linguaggio di rappresentazione.

È la codifica di alto livello che permette di colmare la lacuna rendendo esplicita l'organizzazione del testo o qualsiasi *interpretazione*.

2.3 la codifica di livello zero

La codifica dei caratteri consiste nell'associare a ciascun carattere del testo un *codice numerico*. Ai fini della codifica binaria, un carattere è un' *entità astratta*, distinta dalle sue possibili rappresentazioni grafiche.

Un set di caratteri è una tabella di associazioni biunivoche (1 a 1) tra gli elementi di un repertorio di caratteri e codici numerici (detti punti di codice). I codici sono tipicamente riportati in base decimale, ottale o esadecimale. Ciascun punto di codice è rappresentato in forma binaria come una sequenza di bit. La modalità di rappresentazione binaria dei punti di codice viene chiamata codifica di carattere. Il numero dei caratteri codificabili dipende dai punti di codice disponibili, questi dipendono dal numero di cifre binarie usate per la loro codifica.

I set di caratteri sono associazioni *convenzionali* tra carattere e codici numerici.

2.3.1 Il set di caratteri ASCII

Il più noto e diffuso set di caratteri è *ASCII*; ciascun carattere è codificato in *byte*, ma di questi sono usate per la rappresentazione del codice numerico le prime 7 cifre. Il set completo è formato da 128 caratteri.

L'insieme dei caratteri è limitato alle lettere dell'alfabeto anglosassone. Per ovviare a limitazioni, il codice è stato esteso a varie estensioni caratterizzate dall'uso di 8 *bit* → da 128 a 256.

A seconda del sistema operativo, talvolta i caratteri sono rappresentati in maniera diversa.

L'unica estensione standard di ASCII è ISO-Latin-1 → è uno dei membri della "famiglia" ISO-8859, che rappresenta il primo tentativo di estendere il processo di standardizzazione delle codifiche di caratteri al di là delle lingue dell'Europa occidentale.

Questo standard ha un limite: i set di caratteri della famiglia ISO-8859 sono tutti *mutuamente esclusivi*. Il programma interpreterà questo codice come l'uno o l'altro carattere in maniera esclusiva a seconda del set specificato.

2.3.2 Il set di caratteri Unicode

La soluzione ai limiti di ISO-8859 è fornita da Unicode. La differenza tra il primo e il secondo è che in ISO non esiste nessuna mutua esclusività tra caratteri di alfabeti diversi. Lo standard assegna a ogni carattere un punto di codice distinto.

Questo permette l'uso simultaneo nello stesso testo di caratteri appartenenti a sistemi grafici differenti.

Per l'assegnazione dei punti di codice, Unicode adotta un principio di *composizione dinamica dei caratteri*, grazie al quale caratteri complessi sono rappresentati come sequenze di caratteri elementari.

Lo standard Unicode specifica varie modalità di codifica che utilizzano più di 1 byte per la rappresentazione dei caratteri. Tutte le codifiche condividono la stessa assegnazione di punti di codice ai caratteri, ciò che muta è il modo in cui tale codice è "tradotto" in sequenze binarie.

La codifica più comune è UTF-8; il vantaggio è la sua totale compatibilità con ASCII.

2.4 la codifica di alto livello: perché, cosa, come

Al termine della codifica di livello zero, il testo si presenta al computer nei termini di un flusso ininterrotto di codici binari.

Un testo così codificato è assimilabile a un manoscritto in *scriptio continua* → si presenta all'occhio umano come una schiera compatta di caratteri, all'interno della quale è difficile rintracciare un sentiero di lettura.

Si tratta di rendere esplicito ciò che è congetturale o implicito, con lo scopo di guidare il lettore nell'interpretazione del testo. Il lettore nel nostro caso è il computer e il compito della codifica di alto livello è quello di “dare forma” alla sequenza dei caratteri del testo rendendo esplicita quella parte di informazione che è veicolata attraverso le convenzioni tipografiche, testuali e linguistiche.

La codifica di alto livello rende esplicita l'informazione relativa ad aspetti specifici di un testo. Gli interrogativi che emergono in relazione a questo livello di codifica sono molteplici. Riguardano le *motivazioni* sottostanti a una codifica di questo tipo.

La codifica di alto livello gioca un ruolo cruciale nella trasformazione del dato testuale grezzo in *fonte di informazione linguistica*.

2.4.1 Perché codificare

La codifica di alto livello trasforma il dato testuale in fonte esplicita di informazione linguistica. I dati non hanno come tali un significato intrinseco. Un dato si carica di significato e diventa informazione nel momento in cui, inquadrandosi in uno schema di rapporti, viene legato a un contesto. L'informazione di un certo dominio è costituita da dati strutturati e organizzati in maniera esplicita.

Il valore di una base di dati si misura in rapporto alle informazioni che può fornire; nella costruzione di una base di dati un ruolo cruciale è giocato dalla *strutturazione* dei dati che la compongono.

Questa distinzione tra dato e informazione si applica in modo naturale anche alla rappresentazione digitale del testo. Il testo è un'entità altamente strutturata, dove i dati linguistici sono correlati secondo piani di organizzazione multipli. Questi piani comprendono:

1. *Struttura del testo* → con la sua articolazione in sezioni;
2. *Struttura del contesto* → in cui il testo è stato prodotto;
3. *Struttura linguistica*.

Quando i testi sono inseriti in un corpus, abbiamo un livello ulteriore di organizzazione dato dalla *struttura e composizione del corpus*.

Il testo, visto qui nella sua accezione estesa, diventa *fonte d'informazione linguistica*.

Il potenziale informativo di una codifica di alto livello sarà tanto più alto quanto più numerosi sono i livelli di organizzazione testuale e linguistica codificati esplicitamente. Il fatto che queste informazioni sono interconnesse attraverso livelli espliciti di rappresentazione rende immediato il loro reperimento con l'aiuto del calcolatore.

2.4.2 Cosa codificare

Nella fase di definizione di cosa codificare possono interagire vincoli di diverso tipo.

Il primo passo è individuare il livello di informazione che si intende codificare → dopo si può definire il repertorio di tratti giudicati rilevanti.

Ogni schema di codifica può essere descritto come comprendente:

- un repertorio di categorie per la codifica, corrispondenti alla tipologia dei tratti da rappresentare nel testo, generalmente espresso nella forma di attributi e dei loro possibili valori;

- la definizione delle regole di compatibilità tra categorie: ad esempio l'aggettivo non possiede un attributo inerente di persona, o un nome quello di tempo;
- la specifica accurata dei criteri di applicazione al testo delle categorie selezionate.

Dal momento che la codifica di alto livello non è mai fine a sé stessa, l'ultima parola spetta all'utilità dell'informazione codificata.

2.4.3 Come codificare

Esistono diversi formati digitali in cui può presentarsi un testo.

- Il *formato solo testo* che codifica il contenuto testuale come sequenza di caratteri: è il livello minimo di rappresentazione digitale di un testo. È un formato che al massimo della *portabilità* affianca una minima capacità espressiva.
- I formati come *doc* o *pdf* che strutturano un testo digitale in maniera fruibile per il lettore. Le sequenze di *bit* in un file *pdf* codificano i caratteri del testo secondo un certo set di caratteri ma il contenuto testuale è inframmezzato da istruzioni di formattazione. A una massima espressività del formato fa riscontro una minima portabilità.

Un'alternativa importante è la codifica con *linguaggi di mark-up* come *SGML* o *XML*. Dal punto di vista del formato, un testo codificato con un linguaggio di marcatura è ancora in *formato solo testo*.

Con i linguaggi di marcatura in linea di principio, non vi è limite alla tipologia di informazioni codificabili.

3. COSTRUIRE UN LINGUAGGIO DI MARCATURA

3.1 XML: principi di base

XML nasce alla fine degli anni '90 come evoluzione di SGML. Questo linguaggio rappresenta una versione semplificata di SGML che, pur mantenendo le potenzialità di quest'ultimo, si presta meglio a essere interpretato e manipolato da programmi automatici.

Tratti caratterizzanti di XML:

- Marcatura *dichiarativa* → indica la *funzione astratta*;
- Marcatura *strutturata* → per raggruppare porzioni di testo;
- Marcatura *gerarchica* → un'unità strutturale può a sua volta contenere strutture incassate.

XML non fornisce indicazioni riguardo alla *semantica* dei marcatori; si configura come un *metalinguaggio* di marcatura. Garantisce la completa *indipendenza dei dati codificati*: permette di concentrare la codifica dei dati testuali solo sugli aspetti strutturali o linguistici.

3.2 I componenti della marcatura XML

La tipologia di marcatori include: *elementi*, *attributi*, *riferimenti a entità*, *riferimenti a carattere*, *commenti*.

3.2.1 Elementi

Il termine tecnico usato per designare un'unità del testo, vista come componente della struttura linguistico-testuale è *elemento*.

Ogni elemento è identificato da un nome, selezionato per caratterizzare la funzione della partizione logica del testo che descrive; il nome associato è *identificatore generico*.

Tutto dentro lo stesso tag, scritto allo stesso modo, ciò che è contenuto dentro è l'elemento.

È con l'annidamento degli elementi che XML permette la rappresentazione di strutture gerarchiche. Un elemento deve essere contenuto dentro l'elemento *padre* → non è consentita la sovrapposizione tra elementi.

Ogni documento deve contenere *un solo elemento radice* → elemento 'orfano' e senza 'fratelli'.

Nella pratica della codifica XML di un testo, la dicotomia tra elementi che contengono solo dati di tipo carattere ed elementi che contengono solo elementi figli non è così rigida. Nella rappresentazione del testo è infatti molto frequente avere elementi con contenuto misto.

Esiste infine la possibilità di definire un elemento vuoto, a cui non è associato un elemento del testo.

3.2.2 Attributi

Gli *attributi* sono informazioni aggiuntive che si configurano come 'glosse'. I valori degli attributi devono essere sempre racchiusi dentro le virgolette.

es. <capoverso num="1">C'era una volta...</capoverso>

Gli elementi possono ricorrere più volte mentre un attributo può ricorrere al massimo una volta. Mentre nel caso degli elementi è possibile specificare l'ordine in cui devono apparire nella marcatura testo ciò non è possibile nel caso degli attributi.

3.2.3 Riferimenti a carattere e entità

Tra i marcatori: *riferimenti a carattere* e *riferimenti a entità*. Si ricorre ai primi per inserire nel testo caratteri non accessibili direttamente dai dispositivi di input disponibile.

Le entità sono sequenze arbitrarie di *byte* associate a nomi mnemonici; spesso vengono denominate *entità generali*. Solitamente si dividono in:

- a. *Interne* → valore dichiarato localmente;
- b. *Esterne* → valore rappresentato dal contenuto di una fonte esterna.

I riferimenti a entità hanno la forma `&nome_entità`; e la stringa di testo associata a `nome_entità`, a parte il caso delle entità predefinite, deve essere dichiarata nella DTD.

L'utilità delle entità consiste nel fatto che consentono di riutilizzare lo stesso frammento di testo in posizioni diverse.

Infine, ci sono le *entità predefinite* → non devono essere dichiarate nelle DTD.

3.2.4 Commenti

Il testo può contenere commenti segnati con:

es. `<!-- commento -->`

3.3 La definizione del tipo di documento (DTD)

La tipologia dei marcatori XML necessari a effettuare la codifica del testo e le regole della loro combinazione sono definite all'interno della cosiddetta Document Type Definition (o DTD). La DTD definisce la "grammatica" del linguaggio di marcatura associato a una specifica classe o tipo di documenti.

Nella DTD vengono dichiarati tutti gli oggetti necessari alla costruzione di un linguaggio di marcatura; ciascun oggetto deve essere dichiarato una volta sola. Una DTD è dunque una lista di dichiarazioni di tre tipi distinti:

1. gli elementi in cui si articola il testo;
2. gli attributi associati a ciascun elemento;
3. le entità richiamabili attraverso riferimenti all'interno del testo.

La lista non è ordinata.

3.3.1 La dichiarazione di un elemento

La dichiarazione si articola in due parti:

1. *Etichetta* o *tag* che lo identifica;
2. *Modello di contenuto*: descrizione del contenuto in termini strutturali.

es. `<!ELEMENT tag_elemento (modello di contenuto)>`

Nel caso in cui i sottoelementi siano più di uno, la loro co-occorrenza e il loro ordinamento reciproco sono specificati mediante connettori:

- la *virgola* specifica che i sottoelementi devono ricorrere all'interno dell'elemento in corso di definizione nell'ordine specificato;
- la *barra verticale* indica che i sottoelementi rappresentano scelte alternative.

3.3.2. La dichiarazione di un attributo

Una dichiarazione di attributo definisce l'insieme degli attributi pertinenti per la descrizione di un dato elemento e, per ciascun attributo, stabilisce vincoli sulla tipologia dei valori che può assumere, fornisce informazioni circa l'obbligatorietà della sua specificazione ed eventualmente sul suo *valore di default*.

Il tipo di valore associato a un attributo può essere dichiarato in due modi diversi: ricorrendo a parole chiave o specificando una lista di possibili valori. La parte finale della dichiarazione di attributo specifica la sua obbligatorietà/opzionalità e/o un eventuale valore di default.

La sintassi per specificare un valore di default è semplicemente il valore dell'attributo riportato tra virgolette, senza alcuna parola chiave.

I valori associati ad attributi di tipo ID hanno la funzione di identificare univocamente gli elementi a cui sono associati; lo stesso valore non deve apparire più di una volta e che nella DTD un solo attributo di tipo ID può essere associato allo stesso tipo di elemento.

3.3.3. La dichiarazione di un'entità

La dichiarazione di un'entità comincia con la parola chiave dedicata ENTITY, seguita dal nome dell'entità da dichiarare e dal valore a essa associato.

Le *entità parametriche*, il cui uso è ristretto alla DTD e la cui funzione è quella di rendere modulari alcune dichiarazioni. Ad esempio, può capitare di dover assegnare lo stesso gruppo di attributi a più elementi; le entità parametriche permettono di dichiarare tali attributi una sola volta e poi, nella dichiarazione dei singoli elementi, fare semplicemente riferimento alla corrispondente entità parametrica.

3.4 Struttura e validazione di un documento XML

Un documento XML si articola in due parti:

1. *Prologo* → dove sono contenute le informazioni che permettono di interpretare il documento come doc XML. Il prologo si divide a sua volta in due parti:
 - a. *Dichiarazione XML*;
 - b. *Dichiarazione del tipo di documento*.
2. La dichiarazione del tipo di documento specifica qual è l'elemento radice dell'istanza del documento che segue e contiene il riferimento alla DTD. Senza questa dichiarazione, i programmi non saprebbero quali marcatori o tag sono impiegati nel testo codificato e quali regole sintattiche in esso vigono.

Quando si sviluppa una DTD, può essere utile mantenerla con il testo codificato all'interno dello stesso file in modo da potere simultaneamente modificare i contenuti e verificarne la correttezza. È possibile avere una DTD suddivisa in due porzioni, una interna e una esterna al documento.

Esistono due livelli di correttezza, corrispondenti ai concetti di documento XML *ben formato* e documento XML *valido*.

1. Un documento XML è *ben formato* quando obbedisce a tutte le regole sintattiche di XML. Questo tipo di verifica viene effettuata solo sulla base delle regole sintattiche di XML.
2. Un documento XML è *valido* quando rispetta la tipologia di marcatori e le gerarchie di incassamento dichiarate nella DTD del documento. In questo caso il documento XML è un'istanza valida della classe di documenti definita dalla DTD.

3.5 La codifica del testo in formato XML: esempio

```
<!-- prologo del documento -->
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE libro (View Source for full doctype...)>
<!-- qui comincia l'istanza del documento -->
- <libro>
- <titolo>
  Le avventure di Pinocchio
  <sottotitolo>Storia di un burattino</sottotitolo>
</titolo>
<autore>Carlo Collodi</autore>
- <parte p_id="1">
  <titolo>Parte prima</titolo>
  - <capitolo c_id="1">
    <titolo>Come andò che maestro Ciliegia, falegname, trovò un pezzo di legno, che piangeva e
    rideva come un bambino.</titolo>
    <capoverso num="p1c1c1">C'era una volta...</capoverso>
    <capoverso num="p1c1c2">- Un re! - diranno subito i miei piccoli lettori.</capoverso>
    <capoverso num="p1c1c3">No, ragazzi, avete sbagliato. C'era una volta un pezzo di
    legno.</capoverso>
    <capoverso num="p1c1c4">Non era un legno di lusso, ma un semplice pezzo da catasta, di quelli
    che d'inverno si mettono nelle stufe e nei caminetti per accendere il fuoco e per riscaldare le
    stanze.</capoverso>
    <!-- altri capoversi del capitolo qui -->
  </capitolo>
  <!-- altri capitoli della parte qui -->
</parte>
<!-- altre parti qui -->
</libro>
```

3.6 Standard e codifica del testo

La comunità scientifica ha da tempo avvertito la necessità di poter condividere e scambiare testi digitali codificati.

La codifica testuale effettuata sulla base di un *metalinguaggio di marcatura standard* come XML fa sì che il testo sia elaborabile da una ricca varietà di strumenti software. L'uso dello standard XML per la codifica digitale del testo rende indubbiamente più facile e immediata la condivisione di testi digitali.

XML non fornisce alcuna prescrizione relativamente alla tipologia, alla quantità e al nome dei marcatori ma si occupa di fornire un insieme di regole formali per la definizione di un linguaggio di marcatura. La conformità di un testo allo standard XML riguarda pertanto solo la *forma* della sua rappresentazione digitale.

Una reale condivisione del *contenuto* della codifica deve necessariamente passare attraverso la conoscenza del repertorio delle categorie usate. Ciò implica che un'effettiva interscambiabilità del testo digitale può essere garantita soltanto da una codifica basata su uno *schema di codifica* e *annotazione* il più possibile condiviso, ovvero su uno *schema standard*.

Un'opera di codifica che utilizzi un *metalinguaggio di marcatura standard* e uno *schema standard di codifica e annotazione* trasforma il testo digitale in una *risorsa di informazione* pronta all'uso.

3.6.1 Iniziative internazionali di standardizzazione della codifica del testo

Dal 1987, la Association for Computers and the Humanities (ACH), la Association for Computational Linguistics (ACL) e la Association for Literary and Linguistic Computing (ALLC) hanno avviato un progetto internazionale per lo sviluppo di un modello di codifica dell'informazione testuale in formato digitale che intende proporsi come punto di riferimento a livello internazionale. Questo progetto *Text Encoding Initiative (TEI)*.

Oggi la conformità alle specifiche TEI rappresenta un requisito per garantire l'effettiva condivisibilità dei testi codificati.

Le raccomandazioni TEI sono costituite a un insieme modulare di DTD che comprende circa 400 marcatori diversi. La flessibilità e generalità di queste raccomandazioni costituiscono al tempo stesso il punto di forza e di debolezza dello schema di codifica TEI.

Lo standard CES è nato come parte delle raccomandazioni del progetto dell'Unione europea EAGLES per quanto riguarda la codifica di corpora testuali. Con l'avvento di XML, lo standard CES si è evoluto in XCES. Lo standard CES/XCES è stato adottato per la codifica di numerosi corpora.

3.6.2 Lo schema di codifica XCES

Lo schema di codifica XCES distingue 3 categorie d'informazione che rivestono un ruolo importante nella codifica del testo finalizzata a elaborazioni automatiche:

1. *Documentazione* → include informazione globale sul testo;
2. *Dato linguistico primario* → vengono distinti la: *macrostruttura del testo*, che include elementi testuali fino a livello di capoverso es. volume, capitolo; *microstruttura del testo*, che include elementi che appaiono sotto il livello di capoverso es. periodi, citazioni, abbreviazioni;
3. *Annotazione linguistica* → arricchisce il testo associando al dato linguistico primario vari livelli di interpretazione linguistica.

Ogni testo codificato conformemente alle specifiche XCES è articolato in due parti, individuate rispettivamente dagli elementi `<cesHeader>` e `<text>`, entrambi obbligatori.

L'elemento `<cesHeader>`, corrispondente alla categoria "documentazione".

L'elemento `<text>`, contiene la codifica digitale del testo vero e proprio.

L'elemento `<cesHeader>` si articola in quattro sotto-elementi che forniscono la descrizione bibliografica del testo.

`<body>` è l'elemento che contiene il corpo del testo. L'elemento `<div>` è usato per la marcatura delle divisioni strutturali del testo il cui tipo è specificato dall'attributo *type*.

L'elemento `<head>` contiene il titolo di suddivisioni strutturali, mentre `<p>` è il tag che identifica l'elemento capoverso.

`<s>` *sentence*.

4. DAL BIT ALLA PAROLA

4.1 La “tokenizzazione” del testo

In LC le unità di base del testo digitale sono i “token”. Sebbene possano essere anche delle entità complesse, sono tutti accomunati dalla caratteristica di rappresentare le unità di base per i successivi livelli di elaborazione.

Il processo di segmentazione del testo in token è detto “tokenizzazione”. Non si basa su espliciti criteri morfologici. È generalmente considerata un compito semplice per le lingue che adottano la convenzione ortografica di delimitare le parole con spazi. Nelle lingue a ortografia continua la tokenizzazione del testo assomiglia al problema che si pone per il riconoscimento e l’interpretazione del linguaggio parlato.

4.1.1 I criteri per la tokenizzazione

Ciò che rende la tokenizzazione un’operazione non banale è che le convenzioni ortografiche e la struttura del lessico fanno sì che non esiste un rapporto biunivoco tra sequenze di caratteri delimitati da spazi e i token come unità linguistiche di base.

La *variabilità* linguistica incide enormemente sulla complessità della segmentazione del testo, così come su tutte le dimensioni di organizzazione linguistica. Una stessa parola può avere più varianti ortografiche.

La corretta tokenizzazione del testo richiede dunque tre passi fondamentali:

1. stabilire con precisione quali siano le unità linguistiche atomiche di interesse;
2. individuare la metodologia e i criteri più opportuni per l’identificazione dei token e le necessarie operazioni di elaborazione del testo.
3. esprimere i criteri e le trasformazioni necessarie in un linguaggio formale che sia traducibile in un programma eseguibile dal computer.

Le modalità con le quali va effettuata la tokenizzazione dipendono fortemente dal tipo di linguaggio e dal genere del testo.

È dunque fondamentale definire i criteri e i processi di tokenizzazione sulla base di un’analisi preliminare di un corpus sufficientemente rappresentativo del tipo di linguaggio e di struttura testuale su cui poi il tokenizzatore dovrà operare.

Punteggiatura e altri segni ortografici

La punteggiatura deve essere considerata come token indipendente. Il problema è che è *ambigua*, lo stesso segno d’interpunzione ha usi diversi, che richiedono trattamenti diversi in fase di tokenizzazione → *disambiguare*.

Maiuscole e minuscole

I computer sono *case-sensitive* → codificano i caratteri maiuscoli diversamente dai minuscoli. È necessario operare un processo di normalizzazione “intelligente” delle maiuscole nel testo, individuare criteri affidabili per decidere quando un token maiuscolo debba essere considerato equivalente al suo corrispettivo minuscolo e quando invece questa equivalenza non sussista. Un semplice criterio euristico può essere quello di convertire in minuscolo. Non esistono soluzioni univoche.

Acronimi e abbreviazioni

In molti tipi di testi l'uso di abbreviazioni e acronimi è diffuso. La loro corretta tokenizzazione è dunque necessaria non solo per un'appropriata preparazione del testo, ma anche per garantire che abbreviazioni e acronimi siano riconosciuti in maniera adeguata. La difficoltà maggiore posta da questi tipi di espressioni è la *produttività*.

La strategia più diffusa ed efficace adottata dai tokenizzatori per gestire le abbreviazioni e gli acronimi è quella di combinare la consultazione di elenchi e glossari contenenti le espressioni più comuni.

Token graficamente complessi

Esistono sequenze di caratteri che, sebbene contengano spazi, formano dal punto di vista linguistico delle entità unitarie. È opportuno considerare come singolo token:

1. *nomi propri* → es. *Los Angeles*
2. *espressioni multilessicali* → es. *per davvero*
3. *strutture alfanumeriche* → es. espressioni monetarie

In alcuni casi anche le sequenze nome-cognome possono essere considerate come token unitari.

4.2 Le espressioni regolari

Questo paragrafo introduce le espressioni regolari (ER) come uno strumento essenziale per l'elaborazione automatica del testo. Le ER permettono di individuare sequenze di caratteri che soddisfano una determinata struttura, come ad esempio numeri, date o parole con certe caratteristiche grafiche.

Nel contesto della linguistica computazionale, le ER sono particolarmente utili per compiti come la tokenizzazione, ossia la segmentazione del testo in unità significative (token). Le ER consentono di identificare automaticamente e con precisione pattern linguistici, evitando regole manuali complesse. Possono essere usate per distinguere, ad esempio, tra un punto che indica la fine di una frase e un punto usato in un'abbreviazione.

4.2.1 La sintassi delle espressioni regolari

Questa sezione approfondisce la sintassi delle espressioni regolari. Le ER sono costruite attraverso una combinazione di caratteri normali (come lettere e numeri) e caratteri speciali (come . * + ? [] () \ ecc.).

Alcuni simboli chiave:

- . (punto): corrisponde a qualsiasi carattere singolo.
- *: zero o più occorrenze dell'elemento precedente.
- +: una o più occorrenze.
- ?: zero o una occorrenza.
- [abc]: un singolo carattere tra quelli indicati.
- \d: cifra numerica.
- \s: spazio.
- \b: confine di parola.

Le parentesi tonde () servono per raggruppare porzioni della regex o per catturare sottosequenze, mentre l'operatore | permette di definire alternative (es. rosso|verde|blu).

È inoltre possibile utilizzare moltiplicatori per specificare il numero esatto o minimo/massimo di ripetizioni di un pattern (es. a{3} oppure a{2,4}).

4.3 Analizzare il linguaggio con le espressioni regolari

Questa parte mostra come le ER possano essere usate concretamente per analizzare il linguaggio, partendo da regole euristiche per gestire la punteggiatura. Ad esempio, una regola per identificare un punto come fine di frase (anziché parte di un'abbreviazione) può essere:

```
\b[a-z]+\.\s+[A-Z]
```

Questa ER dice: “una parola minuscola, seguita da un punto, da uno spazio e da una lettera maiuscola”.

Tuttavia, una singola ER difficilmente è perfetta. Serve raffinarla progressivamente, anche con verifiche empiriche su corpora reali.

Le ER sono inoltre utili per:

- riconoscere acronimi,
- individuare date (es. `\d\d?[-V]\d\d?[-V]\d\d(\d\d)?`),
- rilevare espressioni complesse (come indirizzi o sigle),
- e gestire fenomeni linguistici produttivi (es. nuove abbreviazioni o token complessi).

L'approccio tipico nei tokenizzatori è combinare glossari con ER, per bilanciare copertura e precisione. Alcune sequenze, pur contenendo spazi (es. “Los Angeles”, “al di là”), possono essere trattate come token unici se rilevanti per l'analisi linguistica.

5. PAROLE E NUMERI

Ogni testo è una miniera di dati quantitativi che possono essere oggetto di elaborazioni statistiche.

5.1 Popolazione e testo

L'analisi quantitativa di un testo inizia trattandolo come una popolazione linguistica composta da elementi osservabili: le parole. Questo approccio consente di applicare strumenti della statistica descrittiva. Si parte dall'individuare le "unità linguistiche" osservabili, che nel nostro caso sono le parole, con l'obiettivo di misurarne vari aspetti (frequenza, distribuzione, lunghezza, categoria grammaticale ecc.).

5.2 Parole unità e parole tipo

Si introduce la distinzione tra:

- Parole unità (token): ogni singola occorrenza di parola in un testo.
- Parole tipo (type): ciascuna forma diversa.

Ad esempio, in "un semplice pezzo di legno" ci sono 5 parole unità ma solo 5 tipi se tutte diverse; se "un" appare due volte, allora le unità saranno 6 e i tipi 5. Il confronto tra tipo e unità è utile per misurare la variabilità lessicale.

5.3 Frequenze e distribuzioni

Si analizza la distribuzione delle parole rispetto a diversi attributi, come lunghezza, categoria grammaticale, lemma ecc. Questo è analogo a come si distribuiscono le altezze in una popolazione scolastica. La frequenza è il numero di volte in cui un certo valore dell'attributo si presenta. Vengono introdotti strumenti per raggruppare e leggere i dati attraverso:

- istogrammi,
- curve di distribuzione,
- frequenze assolute e relative.

5.3.1 La media aritmetica

La media fornisce una misura sintetica di una distribuzione. Viene calcolata come la somma dei valori osservati divisa per il numero di osservazioni. Tuttavia, non sempre la media è rappresentativa, soprattutto in presenza di dati molto dispersi.

5.3.2 La deviazione standard

La deviazione standard misura la dispersione dei dati intorno alla media. Una deviazione standard bassa indica che i dati sono concentrati, una alta indica grande variabilità. Questo è utile per capire se un testo è omogeneo (es. uso stabile di parole di una certa lunghezza) oppure no.

5.3.3 Leggere le distribuzioni

Le distribuzioni reali spesso non sono simmetriche. Nel caso linguistico, molte parole sono rare, poche molto frequenti. Si discute l'importanza di saper interpretare la forma della distribuzione per cogliere le caratteristiche stilistiche e strutturali del testo.

5.4 Il vocabolario di un testo

Qui si introduce il concetto di vocabolario di un testo come l'insieme delle parole tipo. Due sono gli indicatori principali:

- la diversità lessicale (quanti tipi diversi sono usati),
- la densità lessicale (rapporto tipo/unità).

5.4.2 Parole grammaticali e parole piene

Le parole si dividono in:

- grammaticali (articoli, preposizioni, pronomi...),
- piene (nomi, verbi, aggettivi...).

Le parole grammaticali sono poche ma molto frequenti; le piene più numerose ma meno frequenti. L'analisi di queste due classi permette di distinguere testi funzionali da quelli informativi o narrativi.

5.5 La legge di Zipf

La legge di Zipf stabilisce che la frequenza di una parola è inversamente proporzionale al suo rango: la 2^a parola più frequente occorre la metà delle volte della prima, la 3^a un terzo e così via. Ciò produce una curva decrescente regolare, tipica del linguaggio naturale.

5.5.1 La famiglia Zipf

Oltre alla legge principale, esistono estensioni che descrivono altri aspetti del comportamento lessicale, come la legge di Heap (che descrive la crescita del vocabolario) e la legge di Mandelbrot (una versione modificata della Zipf).

5.6 La dinamica del vocabolario

Questa sezione si focalizza sulla crescita del vocabolario durante la lettura/scrittura di un testo. Le parole nuove si accumulano all'inizio molto rapidamente, poi sempre più lentamente. Questo andamento è utile per capire quanto osservare un testo per coglierne l'intera varietà lessicale.

5.6.1 *La crescita di V*

$|VT(i)|$ rappresenta il numero di parole tipo dopo i parole. All'inizio cresce velocemente, poi si stabilizza. Il grafico della crescita mostra che, anche se rallenta, il lessico non smette mai di espandersi completamente. Si può distinguere anche tra vocabolario con punteggiatura inclusa o esclusa.

5.6.2 *La frequenza media*

La frequenza media è il rapporto tra numero totale di parole e numero di parole tipo ($f(i) = i / |VT(i)|$). È utile per misurare la “densità” linguistica di un testo, ma va usata con cautela, perché dipende fortemente dalla lunghezza del testo.

5.7 *Media e inferenza statistica*

Chiude il capitolo un confronto tra statistiche descrittive e inferenziali. Le prime (medie, deviazioni...) servono a descrivere un testo specifico. Le seconde cercano di trarre conclusioni generali da campioni. L'analisi statistica del testo deve sempre tenere conto della lunghezza e del contesto, altrimenti rischia di produrre inferenze errate.

6. PROBABILITÀ ED ENTROPIA

6.1 Il concetto di probabilità

Un evento che accade in maniera imprevedibile, o prevedibile con un margine d'incertezza è detto *aleatorio*; un *processo* o *sistema aleatorio* è un processo che produce eventi aleatori come effetto del suo comportamento. Un esempio di questo evento è il lancio di un dado.

La *probabilità* di un evento aleatorio può essere definita come la misura del grado d'incertezza del verificarsi dell'evento. La probabilità è dunque uno strumento quantitativo che consente di ragionare in una situazione d'incertezza, facendo previsioni sul possibile verificarsi di un evento. I valori possono variare con una continuità tra 0 e 1.

6.1.1 Spazio campionario e distribuzione di probabilità

La definizione classica della probabilità passa attraverso la descrizione di un *esperimento* e dell'insieme di tutti i suoi *esiti mutualmente esclusivi*. Un insieme così definito è detto *spazio campionario omega* dell'esperimento. In questo spazio si distinguono due tipi di eventi:

1. *Eventi aleatori semplici*: definiti da un insieme che contiene un solo esito semplice dell'esperimento;
2. *Eventi aleatori complessi*: definiti da un insieme che contiene più di un esito semplice dell'esperimento.

L'insieme di tutti gli eventi definibili a partire da uno spazio campionario omega viene detto *spazio degli eventi di omega quadrato*. La *probabilità* è una funzione che assegna un numero compreso tra 0 – 1 a ogni evento definito su omega e misura il grado di incertezza del verificarsi dell'evento. L'insieme delle probabilità di tutti gli eventi di omega definisce una *distribuzione di probabilità*.

6.1.2 Eventi congiunti

Evento congiunto: formato dagli eventi singoli.

6.1.3 Probabilità e frequenza

La definizione *frequentista* di probabilità ci dice che possiamo *approssimare* la probabilità di un evento con la frequenza relativa del suo verificarsi in un certo numero di esperimenti.

“approssimazione”: la definizione frequentista di probabilità dice che la probabilità di un evento A può essere stimata come il valore a cui *tende* la frequenza relativa di A, al crescere del numero delle volte in cui abbiamo eseguito l'evento.

La frequenza relativa di un evento A ci fornisce dunque una *stima empirica* della probabilità di A, effettuata *osservando* il numero di volte in cui A si è verificato, stima la cui accuratezza dipende dal numero delle osservazioni fatte. Questa definizione di probabilità è anche detta *a posteriori*.

La definizione di frequentista è solo un modo diverso per assegnare una probabilità a un evento.

La definizione frequentista di probabilità ha un ruolo centrale in tutti i tipi di indagini scientifiche, proprio perché permette di collegare la nozione di probabilità di un evento all'osservazione della frequenza con cui questo si verifica in una serie di esperimenti.

Il linguaggio si comporta un po' come un "dado truccato": il linguista sa che c'è un "trucco", perché le probabilità degli eventi linguistici non sono uniformi, e il suo obiettivo è proprio cercare di ricostruire dove il trucco risieda.

6.2 Lingua e probabilità

6.2.1 Modelli stocastici

Una metodologia comune in molti ambiti delle scienze per comprendere e spiegare la dinamica di un certo fenomeno o comportamento consiste nell'assumere che esso sia prodotto da un processo o sistema aleatorio. In questo modo è possibile usare la nozione di probabilità per costruire un modello.

Un modello probabilistico è un modello che assegna una certa probabilità agli eventi prodotti da un sistema probabilistico. Così facendo si ottiene un modello del sistema che regola la scelta che gli consente di fare delle previsioni.

È possibile spiegare un fenomeno assumendo che esso sia il prodotto di un sistema aleatorio, di cui costruiamo un modello probabilistico partendo da un corpus di osservazioni del comportamento del sistema. Questo corpus di osservazioni prende generalmente il nome di corpus di addestramento, poiché i dati da esso estratti sono usati per addestrare il modello.

La bontà del modello viene poi valutata in base alla sua capacità di prevedere in maniera corretta il comportamento futuro del sistema.

Un elemento chiave di questa metodologia è chiaramente il grado di generalità del modello: ciò a sua volta dipende dalla rappresentatività del corpus di addestramento rispetto alla tipologia del fenomeno da spiegare.

6.2.2 Modelli linguistici stocastici

Un testo può essere visto come una *sequenza* es. e_1, e_2 ecc. di *eventi aleatori semplici*, ciascuno dei quali rappresenta l'occorrenza di una parola specifica.

Il parlante/autore di un testo, con la sua grammatica e conoscenze delle regole e convenzioni della lingua, può dunque essere visto come il sistema probabilistico che ha prodotto la sequenza di singoli eventi aleatori che costituiscono il testo.

Una lingua è definibile come una sequenza potenzialmente infinita di eventi *generati* da un sistema aleatorio. Una descrizione della lingua come sistema probabilistico consiste così nella creazione di un *modello linguistico stocastico* del sistema che *genera* le sequenze di parole della lingua.

Dato un testo specifico T , un *modello linguistico stocastico* è in grado di associare a T una probabilità $p(T)$ compresa tra 0 - 1.

Le sequenze di parole che troviamo nei testi non sono prodotte a caso e non hanno tutte la stessa probabilità di essere prodotte.

Esiste una relazione tra il grado di correttezza grammaticale di una frase e la probabilità che la stessa frase *sia effettivamente* enunciata. Un campione rappresentativo di testi di una lingua può essere usato come corpus per *addestrare* un MLS del sistema che ha generato il campione. Il modello sarà tanto più accurato quanto più sarà capace di fornirci delle previsioni corrette su quali sono le sequenze di parole grammaticali della lingua dei testi del corpus.

Ricostruire il vero MLS responsabile della generazione di un testo o di un repertorio di testi è spesso un obiettivo difficile perché il numero di informazioni necessarie per definire il modello può essere così grande che i dati osservabili non sono sufficienti a offrirle tutte. Il modello migliore non è necessariamente quello che assegna la *massima* probabilità ai dati osservati.

L'adeguatezza di un MLS deve essere valutata sulla base della sua capacità di generare un insieme *molto grande* di testi; dobbiamo sempre considerare l'insieme di testi su cui impostiamo un MLS come un *campione* di una popolazione potenzialmente infinita di testi.

6.3 Modelli markoviani

Com'è possibile assegnare una probabilità a una sequenza *potenzialmente infinita* di parole? Possiamo ipotizzare che la probabilità di generare un testo sia uguale al *prodotto* della probabilità di generare ognuna delle sue frasi.

La probabilità di generare una singola frase può essere espressa come una *funzione* della probabilità delle sue parole.

Modelli markoviani → stimano la probabilità di una parola a partire da un certo numero di parole che la precedono direttamente nel testo. L'*ordine* del modello specifica in numero esatto di parole che il modello prende in considerazione.

L'ordine definisce anche il grado di complessità del modello markoviano.

Adottando la definizione frequentista di probabilità, useremo la *frequenza relativa* con cui le parole si distribuiscono nel testo campione per *stimare le probabilità* che il modello assegna agli eventi semplici. Questo equivale a rendere *massimamente probabili* le sequenze effettivamente osservate nel campione.

Il *metodo di massima verosomiglianza* equivale ad *addestrare* un MLS su un testo x .

6.3.1 Il modello base: l'urna lessicale

L'esempio in assoluto più semplice di modello markoviano si basa sull'ipotesi che la probabilità di una frase sia uguale al *prodotto* della probabilità di generare ciascuna delle sue parole in modo indipendente.

Siccome questa ipotesi equivale a rappresentare l'inserimento di una parola nel testo come il risultato dell'estrazione di una "pallina lessicale" da un'urna, chiameremo l'MLS corrispondente *modello a urna*.

Immaginiamo di descrivere un testo come il risultato dell'estrazione ripetuta di parole da un'urna lessicale. A ogni istante, una singola parola è estratta dall'urna. Il risultato dell'estrazione è una parola specifica che viene registrata su un foglio, accanto ai risultati delle estrazioni precedenti. La parola è poi reinserita nell'urna di partenza.

Un testo è dunque una sequenza di singole parole statisticamente indipendenti l'una dall'altra.

Un modello così fatto è anche noto come *catena markoviana di ordine zero*, l'evento che ricorre nell'istante i non ha memoria dell'evento incontrato nell'istante precedente.

Addestramento del modello

Le palline associate a un tipo lessicale assai frequente nel testo dovranno essere più numerose di quelle associate a un tipo lessicale più raro.

Questa intuizione si traduce nell'idea che la composizione dell'urna è un *parametro libero* del modello, che va *stimato* sulla base dell'evidenza distribuzionale delle parole nel corpus di addestramento. La

stima della composizione dell'urna rende massimamente probabili le distribuzioni osservate nel corpus.

È necessario che il testo sia effettivamente *rappresentativo* della sua popolazione e sufficientemente ampio da consentire stime affidabili.

6.3.2 Probabilità condizionate e catene markoviane del primo ordine

Una sequenza di parole *non* è il risultato di una serie di estrazioni casuali da un'urna lessicale. Le parole non si combinano tra loro con la stessa probabilità, indipendentemente dal contesto. La scelta di una parola modifica lo spazio di probabilità relativo alla scelta lessicale successiva.

6.3.3 Oltre le catene del primo ordine

Nel modello markoviano del primo ordine, si deve registrare soltanto l'ultimo evento in ordine di tempo, per poter ragionevolmente prevedere cosa succederà nell'istante successivo.

La scelta di un nome dipende solo dall'articolo che lo precede immediatamente, quella di un verbo dal nome alla sua sinistra e via dicendo. Ci aspettiamo pertanto che addestrare su questo testo una catena di Markov di ordine superiore al primo non porti alcun vantaggio apprezzabile.

Vogliamo porci la domanda se, *in generale*, un modello markoviano del primo ordine sia un buon modello stocastico dell'italiano. La risposta è no.

Il modello non è in grado di generare una sequenza perfettamente legittima di categorie grammaticali come es. un verbo preceduto da avverbio.

Addestramento del modello

Le catene di Markov definiscono una famiglia di modelli linguistici di complessità crescente, in funzione della loro capacità di "memorizzare" le parole già viste del testo.

Ciascuno di questi modelli presuppone uno spazio di eventi semplici suo proprio: le singole parole per il modello di ordine 0; i bigrammi per il modello di ordine 1, i trigrammi per il modello di ordine 2 e via di questo passo. In teoria, non esistono limiti alla lunghezza degli eventi semplici.

In pratica, diventa difficile ottenere una stima accurata delle distribuzioni degli n-grammi al crescere di n. Quanto più lunghe sono le sequenze di parole, tanto più piccola è la probabilità di trovarle, perché i loro tipi tendono a moltiplicarsi. La stima di probabilità molto piccole rende dunque indispensabile il ricorso a testi molto grandi.

6.4 Linguaggio ed entropia

È possibile stimare l'accuratezza dell'approssimazione del testo da parte dell'MLS in modo più rigoroso? Gli studi di Claude Shannon sulla *teoria dell'informazione* e la nozione di *entropia* ci consentono di dare una risposta affermativa a questa domanda.

La teoria dell'informazione ci consente di collegare nozioni apparentemente distanti tra loro quali la capacità predittiva di un MLS, la teoria della codifica e il cosiddetto *rasoio di Occam*.

L'entropia nasce come misura del valore informativo di una classe di eventi esclusivi.

Gli eventi rari contengono maggiore informazione degli eventi probabili. Periodicità, ripetizioni o correlazioni di natura varia rendono gli eventi più prevedibili e quindi, in ultima analisi, meno informativi. Quanto più piccola è la probabilità, tanto più grande è l'informazione.

L'entropia è una *media* calcolata sulla distribuzione di *probabilità* di un insieme di eventi aleatori mutuamente esclusivi.

Si dimostra che l'entropia esprime la *lunghezza media minima* di una serie di passaggi in cui le parole più frequenti ricevono una codifica più corta e quelle più rare una codifica più lunga.

L'entropia di un vocabolario con distribuzione *non uniforme* è *inferiore* all'entropia di un vocabolario delle stesse dimensioni in cui le parole si distribuiscono in modo equiprobabile. A parità di numero di eventi l'entropia ha valore massimo quando gli eventi sono equiprobabili.

A parità di vocabolario, il testo in assoluto più imprevedibile è quelli in cui ogni parola può apparire con la stessa probabilità di un'altra senza alcun tipo di vincoli o restrizioni.

6.4.2 Testo ed entropia

Un testo è tipicamente formato da *sequenze* di parole. Dal punto di vista statistico, la nozione di sequenza di parole si traduce in quella di *evento congiunto*.

Passando da eventi singoli a congiunti, il nostro evento diventa multidimensionale.

L'entropia della distribuzione di singole parole nel testo tende a essere più piccola dell'entropia della distribuzione di coppie di parole. La ragione è → il numero di bigrammi tipo di un testo è generalmente maggiore del numero di parole tipo dello stesso testo.

Questo incremento può essere più o meno grande. Tutto dipende dal grado di *indipendenza statistica* tra le parole che formano i nostri eventi congiunti semplici.

L'entropia della loro distribuzione è massima quando la distribuzione è uniforme.

Maggiore sarà il numero di bigrammi, *maggiore* è l'entropia della loro distribuzione.

Ciascuna di queste classi di eventi di crescente complessità definisce un insieme di eventi aleatori mutuamente esclusivi. Da un certo punto in poi, ci aspettiamo che lo scarto tra l'entropia calcolata su sequenze di n parole e l'entropia calcolata su sequenze di $n + 1$ parole non diminuisce più. Questo vuol dire che allargare la nostra finestra di contesto non serve. La catena di condizionamenti che ci consente di predire l'occorrenza della parola corrente ha avuto termine.

Dal momento che il modello a bigrammi è anche quello che richiede di misurare il numero minimo di eventi mutuamente esclusivi, in base al principio del rasoio di Occam, secondo il quale *entia non sunt multiplicanda praeter necessitatem*, dobbiamo preferirlo ad altri modelli.

6.5 Considerazioni conclusive

I modelli linguistici stocastici sono meccanismi generativi che associano a ogni sequenza grammaticalmente corretta di parole, una probabilità maggiore di 0. La probabilità è stimata per approssimazione, *combinando* tra loro le probabilità di sequenze elementari più piccole.

Sequenze elementari corte possono catturare solo relazioni strutturali tra parole vicine tra loro. Maggiore la lunghezza: strutture più ampie.

L'entropia è un indice della quantità di struttura presente in un testo. Il valore di massima entropia si raggiunge quando le parole formano un flusso caotico, in cui ogni nuova parola ha probabilità costante, indipendentemente dal contesto. Il flusso è stimolato dall'MLS del modello a urna. Aumentando l'ordine nel modello markoviano, siamo in grado di catturare condizionamenti di ampiezza crescente tra parole.

La creatività del linguaggio consiste nella possibilità di generare frasi sempre nuove combinando in modo diverso un *numero finito* di schemi ricorrenti di *lunghezza finita*.

Questo excur.us può essere tradotto in *istruzioni all'uso*:

1. Il testo in questione deve essere *il più lungo possibile*;
2. Bisogna partire dal modello più semplice per complicarlo gradualmente;
3. Per ogni modello considerato, si deve calcolare l'entropia degli eventi semplici alla base del modello dopo aver stimato la loro distribuzione di probabilità sul testo;
4. Confrontare questa entropia con quella alla base del modello precedente in ordine di complessità e calcolare la differenza;
5. Se questa differenza smette di decrescere, si può accordare la preferenza al modello markoviano di ordine di complessità precedente a quello corrente.

Assunzioni di fondo: esiste un modello markoviano di ordine finito che sia in grado di rappresentare tutta la struttura linguistica che troviamo nel testo; avendo a disposizione un testo sufficientemente lungo, saremo in grado di osservare tutte le distribuzioni osservabili.

Alcuni modelli sono stati sviluppati a partire dalla distribuzione di *classi di tipi lessicali*: raggruppando parole e sequenze di parole rispettivamente in *classi astratte* e in *sequenze ricorrenti* di classi astratte, la cui distribuzione si presta a essere stimata a sua volta con accuratezza sempre maggiore. Questo ciclo incrementale, che si compone dei passi *stima* → *generalizzazione* → *nuova stima*, definisce una famiglia di algoritmi, *bootstrapping*, finalizzati a distillare dai testi informazione linguistica sempre più astratta.

7. ESPLORARE IL TESTO

La LC ha messo a punto una ricca serie di metodi che consentono di effettuare *esplorazioni avanzate del testo* allo scopo di individuare tratti e costruzioni rilevanti per una particolare indagine linguistica.

Un fattore fondamentale che condiziona i modi e i risultati delle nostre esplorazioni testuali è la possibilità di disporre di un *testo linguisticamente annotato*. Alla dimensione lineare di un testo, si accompagna una *dimensione verticale di analisi*, in stretto rapporto con i livelli di struttura linguistico-testuale.

Un testo annotato contiene alcuni *livelli* verticali di analisi, rappresentando il modo in cui il linguista *classifica* le forme lessicali concrete in categorie astratte, o *assegna* ai legami sintattici una struttura appropriata.

- A livello morfologico: possiamo definire un esponente lessicale, con i dovuti *caveat*, come la classe di tutte le forme flesse che condividono la stessa radice.
- A livello sintattico: queste stesse forme tendono ad accompagnarsi a contesti ricorrenti, selezionando anche classi di parole omogenee dal punto di vista semantico.

Da questo punto di vista, una classe linguistica astratta delinea il profilo comportamentale tipico dei suoi membri nel testo.

Le nozioni di *esplorazione* e *annotazione* del testo sono strettamente legate tra loro in quanto la dimensione distribuzionale e quella classificatoria delle unità linguistiche sono strettamente *interdipendenti*.

I livelli di analisi morfologica, sintattica e semantica consentono di scagliare la dimensione verticale di un testo *per gradi* di profondità crescente. Ciascun livello offre evidenza linguistica che può essere utilizzata un livello successivo per acquisire *ulteriori generalizzazioni*.

A un certo livello di astrazione, quest'analisi a cascata rappresenta una "simulazione" del processo attraverso cui il lettore umano giunge alla comprensione del testo.

7.1 Modi di esplorazione

I metodi di esplorazione del testo che si avvalgono dell'uso del calcolatore possono essere *qualitativi* o *quantitativi*.

- a. *Qualitativi* → rispondono all'obiettivo di individuare *esemplari linguistici* che illustrino un particolare uso o fenomeno della lingua;
- b. *Quantitativi* → cercano di definire la rilevanza di un tratto linguistico stimando la probabilità con cui esso ricorre in un corpus rappresentativo.

Entrambi questi metodi di esplorazione si basano sulla possibilità di recuperare dati linguistici all'interno del testo, isolandoli da tutto ciò che non è rilevante. Le espressioni regolari sono uno strumento fondamentale per l'esplorazione testuale.

7.2 Le parole e il loro habitat: le concordanze

Concordanze: sono una lista delle occorrenze di una parola tipo nel testo, ciascuna presentata nel suo contesto linguistico. Esse permettono di esplorare l'uso di una parola nei singoli 'habitat' linguistici in cui ricorre.

Il modo standard in cui vengono presentate le concordanze di una forma lessicale specifica (*keyword*) è il formato *KWIC*. Le concordanze *KWIC* contengono tante righe quante sono le occorrenze della parola chiave nel testo.

I programmi compilano di solito un *indice* del testo detto *inverted index*, una struttura dati che contiene, per ogni parola tipo, la sua frequenza e l'indicazione dei punti nel testo in cui la parola ricorre. Questo tipo di struttura dati consente di velocizzare il processo di compilazione delle concordanze.

Come tali le concordanze sono uno strumento fondamentale per il lavoro lessicografico, in quanto consentono di ancorare la descrizione del lessico di una lingua all'evidenza 'ecologica' dell'uso reale.

7.2.1 Tipi di concordanze

Esistono anche agenti, programmi simili ai motori di ricerca che usiamo quotidianamente per reperire informazioni sulle pagine web di tutto il mondo, che producono concordanze in formato *KWIC* di una parola chiave, usando come corpus enormi collezioni di pagine web. Tutti i maggiori corpora sono generalmente accompagnati da software per la loro esplorazione attraverso concordanze.

Il comando *grep* nei sistemi operativi *Unix/Linux* restituisce tutte le righe di testo in cui compare una stringa corrispondente a una particolare *espressione regolare*.

Tipicamente, i programmi di concordanze prendono in input un documento elettronico in formato solo testo, cui si applicano tutti i *caveat* relativamente al problema della codifica dei caratteri.

In genere, i programmi esistenti eseguono una rudimentale forma di segmentazione del testo, che usano come base per compilare l'*inverted index*. È preferibile, usare un testo già preventivamente segmentato in token.

Le funzioni principali dei programmi di concordanze sono quasi sempre le stesse, anche se differenze importanti esistono per quanto riguarda la velocità di compilazione delle concordanze.

È solitamente possibile definire la *lunghezza del contesto* che accompagna la parola chiave. Determinare la lunghezza appropriata del contesto dipende soprattutto dal tipo di analisi da effettuare.

Un secondo parametro modificabile riguarda l'*ordine di presentazione* delle concordanze. L'opzione predefinita è generalmente quella di presentare le concordanze in ordine di apparizione nel testo.

In molti casi, l'obiettivo dell'analisi delle concordanze è esplorare in quali contesti tipici tende a ricorrere una parola. A tale scopo, è utile ordinare le concordanze *alfabeticamente* rispetto alle parole del contesto sx o dx.

7.3 Funzioni di ricerca avanzate

Le concordanze di tutte le forme flesse di uno stesso lemma possono essere facilmente individuate utilizzando la possibilità offerta dalle ER di cercare classi di stringhe.

In generale, non è sempre facile formulare una ER in modo tale da trovare tutti e soltanto i dati rilevanti. Spesso succede che l'espressione ci restituisca erroneamente molto più contesti di quello che effettivamente cerchiamo, oppure che alcune delle strutture cui siamo interessati manchino all'appello.

7.3.1 Problemi e soluzioni

Definire un contesto con la restrizione che un'intera stringa di caratteri *non* vi ricorra è molto problematico se si usano le ER.

Mentre è relativamente poco dispendioso elencare tutte le forme dell'articolo, se vogliamo condizionare la nostra ER rispetto alle categorie grammaticali di classi lessicali aperte, ci troviamo costretti a ricorrere a una lista di radici aggettivali.

Il problema vero è che questa stessa ER diventa uno strumento *assai poco efficiente* per le nostre ricerche testuali. Ogni volta che una ER richiede la presenza di un aggettivo, il computer scorre il testo per verificare se la parola corrente è contenuta nell'elenco di forme che definiscono la classe dell'aggettivo nella ER.

7.4 Collocazioni

Le parole hanno la capacità di combinarsi in espressioni sintatticamente complesse. Alcune potenzialità combinatorie sono determinate da tratti morfo-sintattici e semantici generali delle parole stesse.

Esistono però nella lingua altri tipi di combinazione lessicale che si basano su legami non riconducibili a classi linguistiche generali.

Il forte legame di associazione reciproca si manifesta in alcune proprietà che le collocazioni in generale condividono, sebbene in misura diversa:

1. *Elevata convenzionalità*: le collocazioni sono tendenzialmente espressione di usi convenzionali, tipici di particolari varietà linguistiche;
2. *Ridotta composizionalità semantica*: il significato di una collocazione è molto spesso non immediatamente ricavabile dalla composizione del significato delle parole che la formano;
3. *Forte rigidità strutturale*: spesso le collocazioni sono resistenti a modificazioni aggettivali o avverbiali.

Sinclair considera le collocazioni l'espressione tipica di un principio di combinazione linguistica non riconducibile a vincoli di grammaticalità generali. Questa strategia di combinazione è chiamata dallo stesso Sinclair *indiom principle*, in opposizione all'*open choice principle* che regola la costruzione di strutture complesse in base a vincoli combinatori generali validi per classi aperte di parole.

Un merito indiscusso della linguistica dei corpora è stato quello di aver restituito alle collocazioni pieno diritto di cittadinanza nell'analisi della lingua.

7.4.1 Alla ricerca di collocazioni

I corpora testuali sono miniere di collocazioni. La loro esplorazione ci permette di acquisire dati importanti sulla portata di questo fenomeno linguistico pervasivo. Le concordanze sono sicuramente uno strumento utile, ma non sufficiente, a garantirci procedure di ricerca soddisfacenti.

Un problema che si pone è come rendere più precisa la nozione di *associazione tra parole*; la LC può venire in aiuto, fornendo metodi di analisi automatica del testo che consentono di trasformare la nozione intuitiva di associazione lessicale in un indice quantitativo e misurabile. La LC ha messo a punto una serie di *misure di associazione* che servono a quantificare la forza del legame tra due o più parole nel testo; se due o più parole formano una collocazione in una certa varietà linguistica, è molto probabile che nei testi rappresentativi di quella varietà esse ricorrano insieme in maniera *statisticamente significativa*.

Misure di associazione

Lo studio di misure di associazione lessicale per l'identificazione di collocazioni occupa un posto importante nella ricerca linguisticocomputazionale recente. L'obiettivo è individuare indici quantitativi affidabili in grado di assegnare un valore alto di associazione a buoni esemplari di collocazioni.

È possibile ipotizzare che due parole siano tanto più fortemente associate quanto più spesso si presentano insieme *rispetto alle volte in cui ricorrono l'una indipendentemente dall'altra*.

Tra le misure di associazione, la *mutua informazione* è tra le più note.

Dopo aver valutato la congruenza tra la MI e le nostre intuizioni sullo status di collocazione di bigrammi specifici, proviamo a usare la MI per scoprire le collocazioni di un testo.

Sembra naturale assumere che se due parole ricorrono in un corpus sempre una accanto all'altra, il loro grado di associazione è direttamente proporzionale alla loro frequenza. La MI non è molto indicativa quando calcolata su bigrammi rari. Questo limite è aggravato dal ben noto problema della rarità dei dati linguistici.

L'applicazione di metodi quantitativi per la ricerca di associazioni lessicali deve dunque sempre tener presente il tipo di corpus che viene esplorato e il suo grado di rappresentatività rispetto alla varietà linguistica oggetto di indagine. Stabilire una soglia di frequenza per i bigrammi ha lo svantaggio di ridurre drasticamente la quantità di candidati individuati. Una soluzione alternativa è quella di cercare di ridurre la "rarità" dei dati ampliando il corpus da cui sono estratti i bigrammi.

Le dimensioni del web permettono di registrare valori di frequenza significativi per un numero molto più alto di bigrammi, tali da consentire una misura affidabile della loro forza di associazione. Infine, i limiti della MI possono essere risolti semplicemente ricorrendo ad altre misure di associazione, come la *log-likelihood*.

Bigrammi astratti

Un'altra estensione rilevante per il calcolo di associazione lessicale riguarda il modo in cui sono definiti i bigrammi. L'assunzione più comune è di considerare come bigrammi solo coppie di parole adiacenti nel testo.

Per estendere la tipologia di collocazioni possiamo dunque ampliare la nozione di bigramma, selezionando, tutte le coppie di parole che ricorrono all'interno di una stessa *finestra di contesto*, formata da un numero predefinito di parole. La ridefinizione della nozione di finestra di contesto: non più come sequenza di parole, ma come porzione di struttura sintattica.

8. L'ANNOTAZIONE LINGUISTICA DEL TESTO

L'annotazione linguistica di un testo consiste nella codifica di informazione linguistica associata al dato testuale.

In LC l'annotazione ha acquistato un ruolo centrale, in quanto permette di rendere esplicita, interpretabile ed esplorabile dal computer la struttura linguistica implicita nel testo. L'annotazione del testo presenta numerose analogie con la codifica della struttura testuale in titoli, capitoli, capoversi ecc.

Tuttavia, il loro ordine di complessità è diverso. I dati testuali si organizzano su più livelli, caratterizzati da gerarchie multiple di tratti linguistici, talvolta non perfettamente allineate, in alcuni casi sono parzialmente definite.

La codifica è integrante del lavoro di specifica dello schema di annotazione e interagisce su quest'ultimo su almeno quattro livelli fondamentali:

1. Il grado di *copertura* dello schema
2. La *riproducibilità* dell'informazione
3. L'*interazione* con altri livelli di descrizione
4. Il grado di *espressività* dell'annotazione → importante perché un testo può essere descritto e caratterizzato da molteplici punti di vista. Tipicamente l'annotazione viene in relazione ai tradizionali *livelli di descrizione linguistica*, ciascuno dei quali pone problemi specifici di rappresentazione dell'informazione sul testo.

8.1 Livelli di annotazione

8.1.1 Annotazione morfo-sintattica

L'*annotazione morfo-sintattica* rappresenta la forma basilare e più comune di annotazione linguistica, presupposta dalla maggior parte degli altri livelli di annotazione. Lo scopo dell'annotazione morfo-sintattica è l'assegnazione, a ogni parola dell'informazione relativa alla *categoria grammaticale* che la parola ha nel contesto specifico.

È a questo livello che vengono risolte omografie riguardanti le parti del discorso.

L'annotazione morfo-sintattica è spesso combinata con l'*annotazione per lemma* che consiste nel ricondurre ogni parola del testo al relativo esponente lessicale o *lemma*.

La lemmatizzazione permette invece di effettuare ricerche che astraggono da variazioni di tipo morfologico, riguardanti sia la morfologia flessionale sia quella derivazionale.

8.1.2 Annotazione sintattica

Annotazione sintattica → codifica di informazione relativa all'analisi sintattica delle frasi di un testo. Risente dei diversi approcci teorici alla sintassi; due principali approcci:

1. *Rappresentazioni a costituenti*: basate sull'identificazione di *costituenti sintattici* e delle loro relazioni di incassamento gerarchico;
2. *Rappresentazioni a dipendenze o funzionali*: che forniscono una descrizione della frase in termini di relazioni binarie di dipendenza tra parole che indicano relazioni grammaticali come soggetto, oggetto diretto ecc.

Un testo annotato sintatticamente diventa suscettibile di molte forme di esplorazione e analisi avanzata.

8.1.3 Annotazione semantica

L'annotazione semantica riguarda la codifica esplicita del *significato* o contenuto semantico delle espressioni linguistiche di un testo. Le forme che può assumere l'annotazione semantica sono molteplici proprio in relazione alla complessità **inerente alla nozione di significato**.

Un primo tipo di annotazione riguarda la classificazione delle parole lessicalmente piene di un testo rispetto a categorie semantico-concettuali predefinite.

Un'altra modalità la marcatura nel testo dei *ruoli semantici* che descrivono la funzione semantica svolta da un certo costituente nell'evento espresso dal predicato di cui è argomento.

L'annotazione semantica di un testo consente di aumentare la precisione delle ricerche testuali, evitando il "rumore" causato da eventuali ambiguità semantiche.

8.1.4 Annotazione pragmatica

L'annotazione pragmatica comprende sotto di sé fenomeni che riguardano la funzione comunicativa di una particolare unità linguistica, oppure relazioni che coinvolgono strutture linguistiche spesso al di sopra della frase.

L'identificazione della *funzione illocutoria* di un particolare segmento testuale. Questo tipo di annotazione è applicata essenzialmente alle codifiche di trascrizioni di parlato.

Altro tipo di annotazione è la marcatura delle *relazioni* tra un aggettivo o un pronome e il suo antecedente: *relazione anaforica*. Importante per lo studio dei meccanismi di mantenimento e gestione della coerenza testuale.

L'annotazione delle frasi di un testo relativamente alla loro *funzione retorica*.

8.2 Corpora annotati

I *corpora annotati* sono collezioni di dati testuali arricchiti con uno o più livelli di annotazione linguistica. L'annotazione morfo-sintattica rappresenta una sorta di annotazione di livello base, per molti aspetti quasi elementare.

Il *Brown Corpus* è stato il primo esempio di corpus annotato automaticamente a livello morfo-sintattico.

Per l'italiano, vale la pena ricordare che il corpus *La Repubblica* è anch'esso annotato a livello morfo-sintattico.

Le prime *treebank*, sono nate in Inghilterra nella metà degli anni Ottanta, ma hanno acquistato rapidamente fortuna in tutta la comunità della LC.

Per l'italiano il più grande *treebank* attualmente disponibile è *TRESSI*.

Ancora più nuova è l'idea dei corpora annotati a livello semantico. L'apripista in questo caso è stato *SemCor*, ovvero una porzione del *Brown Corpus* in cui nomi, verbi e aggettivi sono stati annotati con il loro senso, usando come risorsa di riferimento *WordNet*.

Per l'italiano, *TRESSI* ha anche un livello di annotazione semantica, con sensi desunti da *ItalWordNet*.

8.3 “Anatomia” di uno schema di annotazione

Per ciascun livello, i caratteri peculiari di uno schema di annotazione sono determinati da fattori diversi quali:

1. Gli scopi della ricerca e/o applicazione per la quale il corpus annotato è progettato;
2. La teoria linguistica di riferimento per la rappresentazione di un certo tipo d'informazione linguistica;
3. La modalità con cui l'annotazione viene effettuata e le risorse umane e temporali disponibili per l'attività di annotazione;
4. La “granularità” della descrizione linguistica che intendiamo codificare nel testo;
5. Le caratteristiche stesse della lingua dei testi da annotare ecc.

La conseguenza immediata di questi “gradi di libertà” è l'esistenza di *molteplici* schemi di annotazione linguistica per ciascuno dei livelli.

Al di là dei tratti di variabilità individuale, ciascuno schema di annotazione può essere in realtà visto come la risultante della combinazione di un ristretto insieme di *tipi di informazione linguistica di base*.

Tipi d'informazione linguistica:

- *Informazione categoriale*: l'assegnazione di *categorie* alle unità e relazioni linguistiche identificate in un testo
- *Informazione strutturale*: l'identificazione nel testo di strutture che possono o essere interne a un particolare token o raggruppate
- *Informazione relazionale*: la definizione di *relazioni* tra le unità linguistiche identificate.

8.3.1 Annotazione e informazione categoriale

L'informazione categoriale è tipicamente espressa nella forma di etichette che associano categorie o tratti linguistici alle unità identificate nel testo, così come le loro relazioni.

Al tipo di informazione categoriale può essere anche ricondotto il caso dell'annotazione semantica, intesa come classificazione delle parole in base al loro significato in un contesto specifico.

L'informazione di tipo categoriale entra in gioco anche per altri livelli di annotazione linguistica: si ricorre a essa per la categorizzazione dei costituenti sintattici, così come per la classificazione delle relazioni di dipendenza sintattica in soggetto, oggetto diretto ecc.

8.3.2 Annotazione e informazione strutturale

Lo stesso tipo di informazione strutturale può essere usata per rappresentare token che inglobano al loro interno due o più unità morfo-lessicali.

Un primo esempio di raggruppamento di più token in unità strutturalmente complesse riguarda l'annotazione delle *espressioni multi-lessicali* che non siano già state trattate come unità complesse al momento della tokenizzazione del testo. Le espressioni multilessicali sono costituite da sequenze di più parole che formano un'unità di annotazione o a livello morfo-sintattico o a livello semantico.

Nell'annotazione a costituenti le sequenze di token in una frase sono raggruppate in strutture progressivamente più ampie o *costituenti sintattici*. Una *rappresentazione sintattica a costituenti* si basa sull'identificazione di *costituenti* e sulle *loro relazioni di incassamento* gerarchico, dove le strutture identificate sono racchiuse tra parentesi quadre e le relazioni di incassamento gerarchico sono rappresentate attraverso l'inclusione di un segmento più piccolo all'interno di un segmento più grande.

8.3.3 Annotazione e informazione relazionale

L'informazione relazionale: è con questo tipo d'informazione che possono essere annotate relazioni di dipendenza e ruoli semantici.

Si ricorre abitualmente a informazione di tipo relazionale anche per rappresentare unità linguistiche *discontinue*.

È basata su informazione relazionale anche la rappresentazione di tutti quei fenomeni e costruzioni che nella letteratura linguistica sono trattati attraverso il meccanismo della *coindicizzazione*. Due o più elementi della struttura linguistica *si dicono coindicizzati* quando si riferiscono alla stessa *entità*.

8.3.4 Tipi di informazione di base e schemi di annotazione

Al cuore dell'informazione sintattica a costituenti c'è l'informazione di tipo strutturale, che riguarda l'identificazione dei costituenti e delle loro relazioni di incassamento gerarchico.

Vi sono schemi di annotazione dove la categorizzazione dei costituenti può includere altra informazione. Questo è il caso della *Penn Treebank-II* dove viene fatto ricorso a informazione di tipo categoriale per la codifica di alcune funzioni grammaticali.

Uno schema di annotazione a costituenti può anche includere informazione relazionale.

Un'alternativa alla rappresentazione sintattica a costituenti è data da una *rappresentazione a dipendenze o funzionale*. Uno schema di annotazione a dipendenze "puro" si fonda su informazione relazionale tipicamente integrata da informazione categoriale che assegna un tipo alle relazioni identificate tra le parole del testo.

8.4 Tipi di informazione e rappresentazione XML

8.4.1 Rappresentazione XML di formattazione categoriale

Il modo più intuitivo offerto da XML per la rappresentazione di informazione categoriale è rappresentato dalla sua codifica attraverso *attributi* associati all'*elemento* che descrive l'unità linguistica che si vuole categorizzare.

8.4.2 Rappresentazione XML di informazione strutturale

es.

```
[8.14] <struct><cat>F</cat>
      <struct><cat>SN</cat>
        <struct><cat>SN</cat>
          <orto>Maestro</orto>
          <orto>Ciliegia</orto>
        </struct>
        <struct><cat>SN</cat>
          <orto> falegname</orto>
        </struct>
      </struct>
<struct><cat>SV</cat>
  <struct><cat>V</cat>
    <orto>trov&#x00F2;</orto>
  </struct>
  <struct><cat>SN</cat>
    <struct><cat>SN</cat>
      <orto>un</orto>
      <orto>pezzo</orto>
    </struct>
    <struct><cat>SP</cat>
      <orto>di</orto>
      <struct><cat>SN</cat>
        <orto>legno</orto>
      </struct>
    </struct>
  </struct>
</struct>
```

8.3.4 Rappresentazione XML di informazione relazionale

La rappresentazione XML di informazione relazionale presuppone che a ogni elemento della struttura XML sia associato un identificatore univoco, ovvero un attributo XML di tipo ID. La rappresentazione della relazione consisterà dunque nello stabilire un collegamento tra diverse unità linguistiche attraverso i loro identificatori.

La stessa tipologia di rappresentazioni XML può essere adottata per la codifica di espressioni multilessicali discontinue.

Può essere necessario anche stabilire relazioni tra strutture più complesse.

8.5 Annotazioni stand-off

Il modo più semplice di immaginare l'annotazione linguistica di un testo è come una "griglia" di etichette metatestuali intercalate al dato testuale primario.

Ci sono ottime ragioni per sostenere che i dati e le loro annotazioni siano mantenuti il più possibile indipendenti.

Nel momento in cui predisponiamo ad adottare un testo è bene tener a mente:

- Lo stesso testo può essere annotato a diversi livelli di descrizione linguistica;
- Per uno stesso livello di annotazione, più annotazioni alternative possono coesistere;
- Nello stesso testo devono potersi integrare diverse prospettive di analisi, offerte da diversi livelli di annotazione.

Tutti questi aspetti trovano una loro naturale collocazione all'interno dell'*annotazione stand-off* o *annotazione distribuita*.

In essa il dato testuale primario è fisicamente separato dall'annotazione. L'annotazione è associata al dato testuale primario mediante collegamenti ipertestuali e viene aggiunta virtualmente al dato testuale primario.

Questo approccio distribuito all'annotazione linguistica presenta vantaggi:

- Il dato testuale primario si mantiene leggibile;
- Il dato testuale primario rimane stabile e permanente;
- Si rendono possibili annotazioni su diversi livelli che strutturano gli stessi dati secondo gerarchie "disallineate" o "incompatibili";
- Si creano i presupposti per ricerche linguistiche;
- Sul versante pratico, si rende possibile lo sviluppo di annotazioni parallele e distribuite nello spazio e nel tempo in relazione allo stesso testo.

Grazie all'uso dell'annotazione *stand-off*, il testo e la sua annotazione si presentano come un ipertesto organizzato in una serie di moduli: *modulo base* e *modulo extra indipendente*.

8.6 Standard e annotazione linguistica

La comunità scientifica e quella industriale nella LC hanno a disposizione sistemi di codifica standard per la rappresentazione del dato testuale.

La LC e la linguistica dei corpora hanno infatti individuato tre ordini di esigenze cui ogni schema di annotazione dovrebbe idealmente ispirarsi:

1. La *compatibilità* con diverse teorie linguistiche di riferimento;
2. L'*usabilità* del testo annotato per scopi diversi sia di tipo applicativo sia di ricerca;
3. La *riproducibilità*, ovvero la minimizzazione dei margini di arbitrarietà nelle scelte di codifica.

Sono state promosse numerose iniziative internazionali, finalizzate a formulare raccomandazioni per una rappresentazione standard dell'informazione linguistica nel testo.

La varietà dei formati con i quali l'annotazione è rappresentata ostacola fortemente l'interscambio e il riuso di corpora testuali annotati. Lo standard XCES per l'annotazione linguistica costituisce il primo passo in questa direzione.

Alla base di questo approccio risiede la consapevolezza che la standardizzazione a livello del repertorio di nozioni o categorie linguistiche è un obiettivo estremamente arduo da raggiungere.

Una strada alternativa è quella di immaginare la standardizzazione dell'annotazione come la condivisione di una struttura comune “leggera”.

Seguendo questo tipo di filosofia, XCES propone uno standard per l'annotazione basato sulla combinazione di un *metamodello strutturale* con un insieme di *categorie di dati*.

Il metamodello fornisce una struttura comune condivisibile da più tipi di annotazioni, mentre le categorie di dati, che forniscono la “semantica” dell'annotazione, rimangono di stretta pertinenza di chi progetta un particolare corpus annotato; la compatibilità tra corpora annotati con i diversi formati è garantita dall'esistenza di un *repertorio di categorie di dati*.

Questo approccio alla standardizzazione, se da un lato permette variazioni al livello dei contenuti dello schema di annotazione, crea al contempo le basi per il confronto, la valutazione, e l'integrazione di annotazioni diverse.

9. VERSO IL TRATTAMENTO AUTOMATICO DELLA LINGUA

9.1 Insegnare la lingua al computer

Le parole presentano al loro interno strutture ricorrenti di costituenti più piccoli, chiamati solitamente *morfemi*. A loro volta le parole sono inserite in sistemi di relazioni di ordine lineare, di incassamento gerarchico tra sintagmi e di dipendenze a lungo raggio che nel loro insieme vengono a comporre la struttura sintattica delle frasi. Queste ultime si articolano infine in complesse strutture retoriche che formano la tessitura del discorso. *La struttura linguistica è la porta di accesso al contenuto del testo*, ma essa rimane *implicita*, come *nascosta* al suo interno.

All'interno di questo processo di estrazione e classificazione di dati linguistici con l'aiuto del calcolatore, l'annotazione linguistica del testo svolge un ruolo di primaria importanza. È attraverso l'annotazione, infatti, che il linguista può riproiettare sul testo le generalizzazioni ottenute nella prima fase, ancorandole esplicitamente a concreti esempi d'uso e rendendole accessibili al computer.

Il ciclo dati > informazione > annotazione non cambia, come tale, il rapporto tra testo e computer. Quest'ultimo accede ai tratti linguistici del testo secondo le modalità standard di ricerca di stringhe di caratteri, con la sola differenza che nel testo annotato l'oggetto della ricerca sono proprio le etichette che codificano i tratti.

In una sorta di "rivoluzione copernicana" dell'elaborazione dell'informazione testuale, invece di arricchire il testo per avvicinarlo al computer, questa seconda strategia cerca di *arricchire il computer con le conoscenze linguistiche necessarie per accedere al testo comprendendone la struttura e il contenuto*.

9.2 Un esempio: la morfologia

In un formario, ciascuna forma flessa è accompagnata da una serie di informazioni linguistiche, tra le quali l'indicazione dell'esponente lessicale.

Per ogni parola unità del testo, sarà sufficiente individuarne il tipo corrispondente nel formario di riferimento e registrare il relativo esponente lessicale.

Mettere insieme il formario completo di una lingua può essere arduo.

Il problema di fondo posto dai formari risiede però nel loro carattere *statico*. Un computer che ha a disposizione soltanto un formario *non* è in grado di estendere in modo autonomo le proprie conoscenze lessicali.

Una soluzione relativamente semplice per evitare la trappola del numero chiuso di forme è fornita da un tipo di procedura automatica detta *stemmer* → programma che rimuove le terminazioni delle parole con l'obiettivo di ricondurre queste ultime alla loro *radice*. Ha a disposizione solo due tipi di informazioni morfologiche:

1. Le terminazioni di una lingua;
2. L'ordine nel quale le terminazioni possono disporre all'interno di una parola.

La forza dello *stemmer* risiede nella sua semplicità e generalità. La semplicità dell'algoritmo costituisce però anche il suo punto di maggior debolezza. Lo *stemmer* non contiene nessuna lista delle radici legittime, e questo provoca due tipi di errori comuni:

1. L'individuazione di una radice sbagliata;
2. Il mancato riconoscimento di una radice.

Un modo per risolvere questi problemi è quello di dotare il computer di conoscenze riguardanti l'organizzazione del lessico morfologico.

Uno strumento software che integri questi tipi di conoscenza linguistica per produrre automaticamente l'analisi morfologica di una parola del testo è chiamato *analizzatore morfologico* che ha a disposizione 4 tipi di informazione:

1. Un *lessico di radici* lessicali;
2. Un *lessico di affissi*;
3. *Regole di combinazione* tra radici e affissi;
4. Regole di normalizzazione.

Associa alla radice una forma normalizzata e converte le terminazioni in *tratti* morfosintattici.

Differenze tra formari, *stemmer*, e analizzatori morfologici:

- a. Il livello di astrazione delle rappresentazioni prodotte;
- b. Il grado di generalità della conoscenza linguistica che incorporano (sistemi chiusi vs. sistemi aperti);
- c. La complessità del modello di morfologia che definiscono.

L'aumento di accuratezza del modello è controbilanciato dalla sua maggiore complessità, dovuta alla necessità di strutturare e codificare un repertorio morfologico di radici, affissi e regole di combinazione. Questa dialettica tra accuratezza e complessità è un aspetto comune in molti settori di ricerca della LC.

9.3 Alcune conclusioni (in forma di introduzione)

Dotare il computer di conoscenze morfologiche per analizzare la struttura delle parole è chiaramente solo il primo dei molti passi necessari per metterlo in grado di comprendere il contenuto di un testo.

Un passo fondamentale per analizzare la struttura del testo è dunque assegnare a ciascuna parola la categoria grammaticale appropriata, che dipende a sua volta dal contesto linguistico in cui la parola appare. I programmi in grado di operare questo tipo di assegnazione sono detti *part-of-speech tagger*.

Il problema dell'*ambiguità* riguarda tutti i livelli di informazione linguistica, dalla fonologia alla pragmatica. A seconda del tipo di ambiguità, l'informazione contestuale necessaria per la sua disambiguazione può essere anche molto complessa.

La LC ha spesso adottato come sua mascotte *HAL 9000*, il computer parlante del film *2001: Odissea nello Spazio* di Stanley Kubrick. La padronanza linguistica di HAL 9000 rappresenta una sorta di esempio limite della capacità del computer di comprendere e usare il linguaggio naturale. L'avvicinamento a tale limite sta impegnando la LC in un percorso lento e complesso.

Dotare il computer di capacità sempre più avanzate di comprendere il contenuto dei testi apre la strada a prospettive applicative potenzialmente illimitate di grande impatto tecnologico, sociale, economico e culturale.

Una seconda ragione: non è possibile immaginare un autentico progresso nelle capacità dei computer di comunicare nella nostra lingua, senza che ciò si accompagni a una maggiore conoscenza dei misteri che tuttora avvolgono il linguaggio umano.

Integrare tutte queste *conoscenze* all'interno di un *modello computazionale* sofisticato che sia in grado di gestire ed estendere aspetti non banali dell'uso linguistico rappresenta non solo una grande sfida tecnologica per i prossimi anni, ma un modo per affrontare i problemi di sempre in una prospettiva nuova e originale.