Una base di conoscenza lessicale per la mappatura semantica dei testi latini.

Stefano Minozzi
Dipartimento di Linguistica Letteratura Scienze della Comunicazione,
Università degli Studi di Verona,
Verona,
Italy
stefano.minozzi@lettere.univr.it

L'edizione digitale dei testi letterari e le fonti documentarie.
Il problema della rappresentazione del testo
Convegno nazionale di studi
Verona, 15 - 16 dicembre 2005

Sommario

Si affronta il problema dell'*information retrieval* semantico sui *corpora* testuali di documenti in latino. A partire da una analisi dei metodi di orientamento nel testo e delle implicazioni semiotiche delle attività finalizzate al rinvenimento dei significati e alla creazione di strumenti per l'orientamento testuale, si propone un modello linguisticamente orientato per l'indicizzazione lessicale dei documenti e un modello multilinguale di mappatura semantica degli elementi lessicali.

1 Creare mappe: indicizzazione e orientamento nel testo

1.1 Testualità, lettura e scrittura

Lettura e sapere venatorio

Intendo iniziare questa comunicazione con la definizione di *sapere venatorio* data da Carlo Ginzburg (1979, pp. 66 e 67). Egli definisce la nostra attitudine a leggere e a decifrare come frutto di un sapere venatorio, cioè della

> [...] capacità di risalire da dati sperimentali apparentemente trascurabili ad una realtà complessa e non sperimentabile direttamente

per questo motivo

> [...] il cacciatore sarebbe stato il primo a «raccontare una storia» perché era il solo in grado di leggere, nelle tracce mute (se non impercettibili) lasciate dalla preda, una serie coerente di eventi.

Il cacciatore che decifra tracce di animali compie un processo del tutto simile a quello del lettore che, attraverso la materialità di simboli geometrici, ricostruisce la rete di significati espressa dal testo: il percorso va da un oggetto materiale ad un oggetto immateriale, il significato, che nulla ha a che vedere con la propria rappresentazione fisica.

Scrittura: risultato di transcodifiche

La scrittura è un atto grafico, l'atto finale di un processo che, a partire dal senso, codifica quest'ultimo attraverso la sintassi e il lessico di un linguaggio naturale e ne dà una convenzionale rappresentazione nell'ordinamento lineare di una sequenza di *glifi*.

Allo stesso modo il processo di localizzazione dell'informazione nel testo è ancora un processo di ricerca simile alla caccia: da un insieme di orme e rami spezzati all'individuazione della preda, dalle tracce del significante al rinvenimento del significato, del quale il testo codifica graficamente la realizzazione linguistica.

Testo: luogo da mappare

Mi sia permesso di introdurre un'ulteriore paragone: il testo è un luogo in cui ci si può orientare attraverso strumenti fisici e gli indici e le tabelle stanno allo spazio testuale così come le mappe e le bussole stanno all'orientamento spaziale.

L'ente unificatore che permette di stabilire l'equivalenza semantica di ciascun codice è il soggetto-lettore, interprete in grado di superare attraverso la bontà delle sue inferenze l'ambiguità delle codifiche linguistiche e grafiche.

1.2 Il testo elettronico

Venendo a discutere della versione elettronica del testo ci accorgiamo come quanto detto fino ad ora si riproponga: il testo elettronico presenta una sua materialità nella codifica binaria che permette di renderlo memorizzabile nell'elaboratore e, allo stesso tempo, viene rappresentato graficamente sullo schermo del computer perché continui ad essere leggibile al lettore umano, al quale la codifica binaria rimane nascosta.

Testo elettronico come medium tra sistemi di elaborazione

Così, il testo elettronico è *medium* tra due forme di elaborazione: come testo grafico presentato a video, si presta all'interpretazione del lettore con caratteristiche simili a quelle del testo scritto; il testo come *data type*, formato di dati in codifica binaria, è l'oggetto del trattamento automatico. Pur tuttavia questa duplice dimensione non porta ad una equivalenza delle possibilità di elaborazione: infatti, il testo rappresentato a video continua ad essere oggetto di inferenza e di interpretazione per il lettore umano, né più né meno come il testo cartaceo, ma il testo come insieme di dati è un oggetto elaborabile da parte del *computer* soltanto nella dimensione del significante. O meglio, l'elaboratore elettronico opera matematicamente (e prima ancora fisicamente) su una codifica numerica del significante. L'operazione di marcatura del testo (*markup*), di cui molto si è discusso stamani, non è altro che il tentativo di esplicitare, attraverso l'aggiunta di segni, alcune delle caratteristiche del testo che, pur essendo evidenti al lettore umano, non possono essere oggetto di elaborazione elettronica. Si tratta in altre parole di realizzare, all'interno dello spazio testuale, una sorta di segnaletica che permetta la creazione di una mappa per l'elaborazione di elementi e di dimensioni del testo che vanno oltre la reticente sequenza simbolica.

Il testo elettronico, quindi, per poter essere in qualche modo *letto* dal calcolatore deve essere mappato o come nel caso del *markup* di tipo *strongly embedded* deve

contenere segni che ne conducano il tracciamento di una mappa.
Mi sia ora permessa una breve citazione letteraria, da Borges, *Storia Universale dell'infamia*:

> ... In quell'Impero, l'Arte della Cartografia giunse a una tal Perfezione che la Mappa di una sola Provincia occupava tutta una Città, e la Mappa dell'Impero tutta una Provincia. Col tempo, queste Mappe smisurate non bastarono più. I Collegi dei Cartografi fecero una Mappa dell'Impero che aveva l'Immensità dell'Impero e coincideva perfettamente con esso. Ma le Generazioni Seguenti, meno portate allo Studio della Cartografia, pensarono che questa Mappa enorme era inutile e non senza Empietà la abbandonarono alle Inclemenze del Sole e degli Inverni.

Necessità di una scelta di scala

L'opera di chi si appresta ad editare il testo per renderlo elaborabile dalla macchina è accostabile a quella paradossale dei cartografi imperiali [ma speriamo che non sia simile anche negli esiti finali]: alla base della rappresentazione digitale del testo deve esserci una scelta, per così dire, di scala che ne selezioni le caratteristiche che si desidera rendere elaborabili, dato che la marcatura del testo, come operazione ermeneutica di esplicitazione di strutture di significato, può facilmente produrre una mappa più grande dell'impero.

La scala scelta

La proposta di un sistema di mappatura semantica dei testi latini, che qui si formula, opera nell'ambito di una *scala* che va dalla ricognizione della dimensione lessicale del testo alla sua categorizzazione semantica, nel tentativo di fornirne una sistemazione computazionale a partire dal riconoscimento dell'insufficienza degli strumenti di orientamento basati sull'ordinamento alfabetico.
La consapevolezza dell'utilità degli indici alfabetici per il reperimento delle informazioni nel testo è stata bene illustrata in Papia, che, nella prefazione del suo *Elementarium doctrinae rudimentum*, si prefiggeva di fissare *regulas certas*, perché il lettore potesse risalire velocemente ai contenuti, e sceglieva di disporre i lemmi in rigorosa successione alfabetica. Allo stesso tempo, però, Papia si rendeva conto della difficoltà che una tale disposizione comportava a causa della varietà delle grafie: egli ricorda come

> Hyaena a quibusdam per i, ab aliis per y vel per aspirationem cum diphthongo in penultima scribitur

La difformità colpiva anche la pronuncia dei vocaboli e lo stesso Papia segnalava che

> quam verbenam quidam, alii berbenam vocant herbam

Folia librorum querere

Un altro esempio di reperimento dell'informazione del testo attraverso i suoi aspetti materiali è la mnemotecnica che Ugo di San Vittore proponeva ai suoi allievi nel *De tribus maximis circumstantiis gestorum*:

> Multum ergo valet ad memoriam confirmandam ut, cum libros legimus, non solum numerum et ordinem versuum vel sententiarum, sed etiam ipsum colorem et formam simul et situm positionemque litterarum per

> imaginationem memoriae imprimere studeamus, ubi illud et ubi illud scriptum vidimus, qua parte, quo loco (suppremo, medio, vel imo) constitutum aspeximus, quo colore tractum litterae vel faciem membranae ornatem intuiti sumus.

Egli consigliava ai propri allievi di mandare a mente le caratteristiche grafiche della pagina, segnando nella memoria visiva la posizione delle *sententiae* nel testo. Si trattava di una forma di orientamento senza indice che si serviva della memoria posizionale e degli aspetti materiali del testo come elemento di collegamento tra il senso e la sua realizzazione in una posizione della pagina.

Precognizione dei risultati

Il reperimento dell'informazione da *corpora* testuali elettronici presenta elementi comuni al primo e al secondo metodo: è infatti quasi sempre un orientamento attraverso un indice (poco importa se questo sia un indice statico pre-generato o oppure venga creato al volo dalla scansione lineare dei documenti) e la consultazione di una base dati testuale prevede la formulazione di ipotesi su come i concetti siano stati lessicalizzati nel linguaggio (o nei linguaggi) dei documenti archiviati e allo stesso tempo deve ricostruirne la forma di rappresentazione grafica; in altre parole l'accesso al testo ideale è possibile soltanto formulando ipotesi sul testo materiale: riprendendo la metafora venatoria potremmo affermare che è come se il cacciatore per risalire alla preda dovesse ipotizzarne la forma delle tracce [e certamente, della caccia, questo tipo di ricerca condivide tutti gli aspetti aleatori]. È chiaro il contrasto con l'esempio medievale: laddove gli allievi di Ugo di San Vittore ritrovavano la posizione dei segni nella pagina, avendone memorizzato gli aspetti materiali dopo una accurata lettura, il ricercatore, che si serve dello strumento digitale, formula una ipotesi su un aspetto materiale del testo (la sua sequenza di glifi) per ritrovare un significato che non è stato, per così dire, visitato in precedenza.

2 Un modello di Information retrieval per i testi latini

IR su testi latini: un modello

Parlando di testi latini, la realizzazione di un modello specifico per l'*information retrieval* semantico si scontra con tre ordini di problemi:

- l'aspetto grafico-ortografico delle parole contenute nel testo;
- la flessione;
- l'imprevedibilità delle realizzazioni lessicali della dimensione semantica.

La proposta che qui si formula si basa sull'impiego di uno strumento di indicizzazione che permetta la compressione delle forme *allografe* o *alloglife*, la loro riconduzione ad un lemma e il collegamento dei lemmi ad una struttura, un *thesaurus semantico*, che permetta di organizzare i significati a partire dalla loro realizzazione lessicale, permettendo di simulare la competenza semantica nello strumento di ricerca.

I primi due momenti di indicizzazione operano una progressiva *reductio ad unum* degli elementi lessicali, che va dal catalogo delle forme ad un lemmario, organizzato

in modo da permettere l'individuazione della posizione dei lemmi nel testo. Il terzo momento collega i lemmi individuati ad una struttura più ampia (quindi si può considerare come un processo di espansione) che costituisce ed esplicita una rete di significazione.

2.1 Ricerche lessicali

2.1.1 Compressione delle allografie

Fuzzy problem

Il processo di compressione delle grafie alternative si confronta con un problema di tipo *fuzzy*, dato che decidere se una stringa di un indice *è allografa* di un'altra significa sindacare sulla loro similarità, valutazione che esula da un approccio di tipo *booleano*. Non necessita di dimostrazione l'affermazione che similarità e differenza tra stringhe si snodano in un continuo, tanto che ogni stringa di un testo è simile ad un'altra, purché esse abbiano almeno un carattere in comune. Inoltre, si dovrebbe parlare non soltanto di allografia ma anche di *alloglifia*, dato che l'ambiguità tipografica permette la rappresentazione degli stessi grafemi con glifi diversi [ae ; æ]. Pertanto, nell'implementazione del sistema, si è scelto di servirsi di una tabella di accelerazione elencante le allografie più consuete e di utilizzare una combinazione algoritmica per l'individuazione degli *alloglifi* inconsueti [di questo metodo, per brevità, si rimanda la descrizione formale ad una eventuale trattazione specifica in altra sede].

2.1.2 Lemmatizzazione

E pluribus unum

Il secondo momento di compressione dell'indice, vede la realizzazione di un meta-indice che raggruppi le forme per lemmi apponendo ad esse un codice di descrizione. Il tipo di implementazione realizzata nella presente ricerca si serve di un approccio semi-automatico e di una base di conoscenza lessicale[1] di circa quarantamila lemmi. Allo stadio attuale, la costruzione dell'indice dei lemmi avviene risolvendo le situazioni di omografia esolemmatica (tra forme di due o più lemmi)[2], con l'attribuzione dell'omografo a tutti i lemmi candidati, in quanto è ancora in fase di studio l'algoritmo di disambiguamento semantico sensibile al contesto della parola: questo tipo di organizzazione, comunque, permette ricerche lessicali molto più efficienti rispetto a quelle possibili sulle collezioni elettroniche che operano solo sulle forme flesse[3].

2.2 Ricerche semantiche

Semantica e NLP

La possibilità di operare ricerche attraverso l'astrazione dei contenuti rappresenta il limite di frontiera degli attuali studi sulla elaborazione dei linguaggi naturali (NLP): di particolare interesse può risultare l'applicazione di queste tecniche alle lingue concluse, come il greco antico e il latino. Qui si propone di applicare alla

[1] Realizzata a partire da fonti di pubblico dominio

[2] Per una classificazione tassonomica dell'omografia latina si rimanda a: Passarotti e Ruffolo (2004)

[3] Unica eccezione la base dati di *PerseusProject* dove l'omografia esolemmatica è comunque trattata allo stesso modo

struttura della frase latina una indicizzazione semantica, rifacendosi al modello di Roussey *et al.* (1999). La base per l'implementazione del modello è la realizzazione di un *thesaurus* semantico che permetta di unire più lemmi all'interno di una definizione di dominio e che permetta di chiarire le relazioni tra i lemmi, definendo due livelli di conoscenza:

1. un livello concettuale che dia un modello del campo di studio formato dai concetti e dalle relazioni che intercorrono tra di essi;

2. un campo terminologico che rappresenti l'insieme delle manifestazioni linguistiche di un concetto nel testo.

2.2.1 *Thesaurus* semantico

WordNet

Tra i tipi di modellizzazione della conoscenza lessicale è parsa particolarmente interessante la rappresentazione dei rapporti semantici nei dizionari definita dagli studi di Miller *et al.* (1990) e Fellbaum (1998): a partire dal riconoscimento della natura del tutto accidentale dello *spelling* delle parole, nel modello di *WordNet* le parole sono organizzate per blocchi di significato, denominati *synset*, che raccolgono tutti i lemmi che lessicalizzano lo stesso concetto; i *synset* sono collegati tra loro per mezzo di relazioni che includono, assieme alla sinonimia, anche l'iponimia, la meronimia e l'antinomia. L'ipo/iperonimia mette in relazione significati subordinati e superordinati fornendo così una struttura gerarchica di concetti. La relazione meronimica induce una gerarchia delle parti sull'insieme dei significati. In questo modo il livello lessicale è chiaramente separato da quello concettuale e questa distinzione è rappresentata dal *medium* semantico-concettuale e dalla relazione semantica che uniscono rispettivamente *synset* e parole.
L'applicazione di tale modello ad un *thesaurus* semantico per la lingua latina rappresenta l'orientamento dei prossimi studi di chi vi parla: si ritiene utile riportare in questa sede una schematizzazione del metodo di costruzione di tale rete semantica.

Mapping

Risultando impraticabile e antieconomico per un singolo studioso costruire da zero l'insieme di relazioni di una rete di questo tipo, il punto di inizio del progetto di costruzione si basa sull'ipotesi che la rete dei significati, definita per la versione inglese, possa essere in gran parte portata verso altri linguaggi. Una ipotesi plausibile, se ci si limita alle principali lingue indoeuropee che presentano una vasta sovrapposizione culturale, ma che deve essere ancora verificata per una lingua conclusa come il latino.

2.2.2 Matrice lessicale multilingue

Multilingual lexical matrix

La matrice lessicale della *Wordnet* inglese è bidimensionale (si estende nelle due dimensioni dei lemmi e dei significati); sulla scorta di quanto effettuato in progetti analoghi[4] per le lingue moderne, aggiungendo una terza dimensione alla matrice (la dimensione delle lingue) diventa possibile considerare la lingua latina. Per realizzare la matrice multilingue, si rende necessario mappare i lemmi latini sui significati corrispondenti, andando a costruire l'insieme dei *synset* per il latino. Il risultato

[4]Alcuni tra i più significativi: Artale *et al.* (1997); Chen *et al.* (2002); Lee *et al.* (2004)

è una completa ridefinizione delle relazioni lessicali, mentre per la creazione della rete di relazioni semantiche vengono impiegate, per quanto possibile, quelle già definite per l'inglese. La dimensione dei significati, pertanto, viene considerata costante rispetto alle possibili lessicalizzazioni a livello linguistico. In un primo tempo, attraverso l'utilizzo di algoritmi di *matching* e di un dizionario bilingue, viene verificata la corrispondenza tra i lemmi delle lingue che devono essere aggiunte alla rete. Questa prima fase di automazione produce un insieme di possibili connessioni tra una parola latina e i significati nella rete *WordNet*. A questa fase segue l'intervento umano per validare le scelte proposte.

Non solo IR

Il campo di applicazione di una rete così costituita va dal disambiguamento dei contesti, all'IR semantico.

Il collegamento tra lemmi presenti nel testo e rete semantica, costituisce una mappa della catena di significati realizzati, che prescinde della loro lessicalizzazione: l'insieme dei riferimenti ai *synset* contenuti nella struttura di una frase diventa, così, il descrittore del contenuto di quella frase e l'elemento su cui deve operare il processo di ricerca. L'accoppiamento che si intende ottenere tra la matrice lessicale inglese e quella latina dovrebbe permettere anche la possibilità di formulare richieste di rinvenimento di contesti che, operando sulla lessicalizzazione in inglese, restituiscano documenti in lingua latina.

3 Conclusione

Alla fine della descrizione di quest'ultima ipotesi di ricerca, intendo sottolineare che ci si rende conto che la capacità di rappresentazione e, di *produzione* di senso propria dell'oggetto testo non è limitata, e non si esaurisce, nella sua dimensione lessicale, ma si estende anche nella dimensione sintattica e nelle capacità attributive del soggetto, a tal proposito, rimando a quanto detto precedentemente sulla necessità di operare una *scelta di scala* nella costruzione del modello.

Il percorso fin qui descritto ha portato alla realizzazione di una serie di strumenti di analisi automatica del testo, come aspetto applicativo (o, se vogliamo, come sottoprodotto) della ricerca: il già citato sistema di identificazione delle allografie e il meccanismo di indicizzazione e lemmatizzazione. Si segnala inoltre che recentemente è stato rilasciato al pubblico dominio il primo dizionario latino compatibile con il *software* WinEdt per il sistema di editoria elettronica LaTeX: il dizionario composto da una lista di 1243950 forme latine, permette il controllo ortografico dei testi ed è utilizzabile, con lievi adattamenti, nelle maggiori applicazioni di riconoscimento dei caratteri (es. *Caere OmniPagePro*), per migliorare la qualità dei risultati di acquisizione. È in via di completamento un *thesaurus* di sinonimi bilingue (inglese/latino), che costituirà la base di partenza della costruzione della rete semantica sopra descritta.

Nel chiudere questa comunicazione, che ha voluto proporre un metodo per il rinvenimento automatico del significato, mi siano concessi un'ultima breve citazione e un augurio. Prima la citazione:

> So d'una regione barbarica i cui bibliotecari ripudiano la superstiziosa e vana abitudine di cercare un senso nei libri, e la paragonano a quella

> di cercare un senso nei sogni o nelle linee caotiche della mano... Ammettono che gli inventori della scrittura imitarono i venticinque simboli naturali, ma sostengono che questa applicazione è casuale, e che i libri non significano nulla *di per sé*.

L'augurio è che i libri possano sempre, e comunque, significare *per noi*.

Riferimenti bibliografici

Artale, A., Magnini, B. e Strapparava, C. Proceedings of ACL/EACL-97 Workshop Lexical discrimination with the Italian version of WordNet. In P. V. et al., cur., *Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications (Madrid, Spain, July 1997)*. Association for Computational Linguistics, ACL 1997

Chen, H.-H., Lin, C.-C. e Lin, W.-C. Building a Chinese-English wordnet for translingual applications. In *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 1(2) [2002]:pp. 103–122

Fellbaum, C., cur. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press 1998

Ginzburg, C. Spie. Radici di un paradigma indiziario. In A. Gargani, cur., *Crisi della ragione*. Einaudi 1979

Lee, C., Lee, G. G. e Seo, J. Multiple Heuristics and Their Combination for Automatic WordNet Mapping. In *Computers and the Humanities*, vol. 38(4) [2004]:pp. 437–455

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. e Miller, K. J. Introduction to WordNet: an on-line lexical database. In *International Journal of Lexicography*, vol. 3(4) [1990]:pp. 235 – 244

Passarotti, M. e Ruffolo, P. L'utilizzo del lemmatizzatore LEMLAT per una sistemazione dell'omografia in latino. In *Euphrosyne*, vol. 32 Nova Série [2004]:pp. 99–110

Roussey, C., Calabretto, S. e Pinon, J.-M. Etat de l'art en indexation et recherche d'information. In *Revue Document Numérique*, vol. 3(3-4) [1999]:pp. 121–150