Carmela Cinquesanti

Corso di laurea magistrale in Informatica Umanistica, curriculum di Tecnologie del linguaggio

Matricola: 438526

Sviluppo di risorse linguistiche aperte e condivise per la lingua italiana

Relazione basata sul seminario "Verso una treebank estesa e condivisa per la lingua italiana", M. Simi, S. Montemagni.

L'argomento del seminario scelto per la presente relazione riguarda l'importanza che le risorse linguistiche aperte e condivise ricoprono nel campo delle tecnologie del linguaggio. In particolare, il discorso si concentrerà dapprima sui problemi relativi alla disponibilità di risorse linguistiche in lingua italiana, esponendo il contributo da parte di linguisti e informatici nella creazione di una treebank per l'italiano; in seguito, la relazione esporrà il mio apporto personale allo sviluppo di due risorse linguistiche, avvenuto tramite la realizzazione di due progetti.

Le tecnologie del linguaggio si inseriscono all'interno di un quadro metodologico più ampio conosciuto con il nome di Natural Language Processing (NLP), o Elaborazione del linguaggio naturale, ovvero il campo disciplinare al confine tra la linguistica, l'informatica e l'intelligenza artificiale che si occupa dell'interazione tra i calcolatori elettronici e le informazioni fornite in linguaggio naturale. Le tecnologie del linguaggio acquistano rilevanza nel momento in cui la comprensione del linguaggio naturale da parte della macchina può scaturire in una molteplicità di applicazioni pratiche, tra cui:

- l'estrazione di conoscenza strutturata da testi, mirata alla costruzione di ontologie[1], all'estrazione di relazione tra concetti, al *knowledge graph*[2];
- i sistemi di *Question answering*[3];
- l'analisi automatica di sentimenti, opinioni e tendenze;
- il monitoraggio di feed informativi in tempo reale, p. es. in ambito finanziario o per scopi di marketing;
- l'interazione vocale;
- la traduzione automatica.

[1] Un'ontologia è una rappresentazione schematica di un insieme di concetti relativi a un dominio.
[2] Il Knowledge Graph è una funzione di ricerca su base semantica introdotta da Google il 16 maggio 2012.
[3] Nel Question answering il calcolatore risponde automaticamente a domande poste in linguaggio naturale;

Tutti i sistemi di NLP adoperano un processo di analisi linguistica, ovvero prendono un testo in input e lo restituiscono arricchito di annotazioni a diversi livelli. E' possibile descrivere questo processo tramite il concetto di *pipeline*, procedimento raffigurabile come una catena i cui componenti descrivono i diversi livelli di analisi linguistica e dove l'output di un livello di analisi diventa l'input per il livello di analisi successivo. Il primo passo di una pipeline linguistica consiste nella suddivisione del testo in frasi, cui segue la suddivisione delle frasi in token; successivamente ogni token viene annotato con la propria categoria grammaticale (*PoS tagging*), si procede poi all'analisi sintattica delle frasi (*parsing*) e infine il testo può eventualmente essere arricchito con annotazioni di tipo semantico. Il livello più importante di una pipeline linguistica è proprio il parsing, che determina la struttura grammaticale di un testo utilizzando uno strumento chiamato *parser*, in grado di costruire l'albero sintattico della frase. Esistono due tipi di approcci per la rappresentazione degli alberi di parsing:

1. la **rappresentazione a costituenti**, in cui i nodi dell'albero sono etichettati con le categorie grammaticali e le foglie denotano le parole;
2. la **rappresentazione a dipendenze**, in cui ogni nodo dell'albero è una parola e le parole sono collegate tra loro da relazioni binarie di dipendenza,ciascuna annotata con un tipo.

Tra i due approcci, la rappresentazione a dipendenze viene solitamente preferita per il parsing dell'italiano, trattandosi di una lingua caratterizzata da un ordine libero dei costituenti e dalla frequente omissione del soggetto.

Tuttavia, indipendentemente da quale approccio si scelga, il parsing può essere basato su due sistemi di apprendimento differenti: può fondarsi su un insieme di regole grammaticali predefinite dal linguista, oppure può incentrarsi su un procedimento di apprendimento automatico, per cui la macchina impara ad analizzare le frasi a partire da un set di alberi correttamente costruiti. Durante l'apprendimento automatico, quindi, il calcolatore è in grado di costruire un modello statistico degli alberi sintattici a partire da un ampio e variegato numero di esempi, riuscendo a raggiungere livelli di accuratezza del 90%.

Per gli scopi dell'analisi linguistica, risulta fondamentale il concetto di corpus annotato, ovvero una collezione di testi arricchita con informazione relativa alla struttura linguistica. L'informazione può riferirsi a vari livelli di descrizione linguistica, sia essa morfologica, morfo-sintattica, semantica o pragmatica. L'aspetto sostanziale dell'annotazione è la sua capacità di costituire un ponte tra il testo e i contenuti, consentendo al calcolatore di accedere ed estrarre dal testo un tipo di informazione altrimenti invisibile. Nel caso dell'annotazione a dipendenze, l'informazione estraibile è di due tipi:

categoriale, se si osservano le categorie con cui vengono annotate le parole; relazionale, se ci si concentra sulla definizione dei tipi di relazione tra le unità linguistiche.

Uno schema di annotazione valido deve, inoltre, essere in grado di soddisfare una serie di esigenze, tra cui: la compatibilità con la teoria linguistica, l'usabilità per scopi diversi (siano essi applicativi o di ricerca), e infine la riproducibilità, data dal minimo grado di arbitrarietà possibile nelle scelte di codifica. Tuttavia, pur rispettando i suddetti princìpi, gli schemi di annotazione possono presentare alcune variazioni, come le regole stabilite per l'assegnazione della testa sintattica, il tagset (ovvero l'ontologia linguistica di riferimento) e i criteri di annotazione.

Per quanto riguarda i criteri scelti per l'assegnazione della testa sintattica, durente l'annotazione possono presentarsi casi alquanto controversi: p. es., mentre nei sintagmi preposizionali il ruolo di testa è sempre attribuito alla preposizione (nel sintagma PP "*a casa*" la testa è "*a*"), nei sintagmi nominali la testa può essere assegnata tanto all'articolo quanto al nome (nel sintagma SN "*la bambina*" la testa può essere sia "*la*" che "*bambina*"). La scelta dei criteri di annotazione della testa spetta pertanto al linguista, a patto che questa decisione venga esplicitamente chiarita e attestata nella documentazione di riferimento. Ritornando all'esempio sui sintagmi nominali, l'assegnazione della testa sintattica può incentrarsi tanto sul valore semantico delle unità linguistiche (attribuendo il ruolo di testa ai nomi pieni, portatori di significato) quanto sull’aspetto funzionale delle parole (e quindi la testa sarà assegnata all'articolo). Sebbene questi due criteri di annotazione siano molto diversi tra loro, in realtà la scelta dell'uno o dell'altro non ha un grande impatto sulla precisione dell'annotazione; piuttosto, l'aspetto che ne risente maggiormente è nella quantità e qualità di informazioni estraibili dal testo. Con un approccio funzionale, in cui la testa è sempre l'articolo, non è possibile estrarre informazione riguardo i partecipanti alle azioni. Considerando una frase come "*La bambina legge un libro*" e assegnando il ruolo di testa a "*la*", non si hanno informazioni relative a **chi** sta leggendo il libro. Pertanto, lo schema di codifica basato sul valore semantico delle parole risulta più conveniente per lo sviluppo di applicazioni basate su informazione semantica.

Un altro aspetto estremamente importante nella creazione di risorse linguistiche condivise risiede nella granularità del tagset, ovvero nell'insieme di attributi utilizzati per annotare il testo. Il tagset può variare a seconda delle scelte del linguista, spesso anche in base al dominio di riferimento del testo. Tuttavia, i requisiti fondamentali di un tagset sono la sua coerenza e la riproducibilità: in altre parole, non è detto che un tagset più ampio sia necessariamente migliore; l'importante è che soddisfi tutte le esigenze di codifica, coprendo in maniera completa il repertorio di dati da annotare.

L'urgenza della creazione di risorse aperte e condivise per la lingua italiana ha portato alla realizzazione di un progetto denominato Merged Italian Dependency Treebank (MIDT), che rappresenta il primo tentativo di unione e conversione di due treebank italiane già esistenti: la Turin University Treebank (TUT) e la ISST-TANL. La treebank TUT, sviluppata dall'Università di Torino, raccoglie sei corpora testuali di vario genere e presenta un insieme particolarmente ricco di relazioni grammaticali. La ISST-TANL è nata dal lavoro congiunto dell'Università di Pisa e dell'Istituto di Linguistica Computazionale del CNR, come revisione dell'Italian Syntactic-Semantic Treebank (ISST). Entrambe le treebank hanno partecipato in passato alle diverse edizioni di EVALITA, la campagna di valutazione degli strumenti di NLP per la lingua italiana.

Sebbene la MIDT sia nata a partire dalla conversione di TUT e ISST-TANL, essa presenta un numero molto ridotto di relazioni di dipendenza fra le parole: solo 20, a fronte dei 72 tipi di relazione della TUT; questo aspetto denota pertanto un livello di granularità minore nell'assegnazione delle relazioni di dipendenza durante il parsing, e rappresenta una particolarità che contraddistingue la MIDT dalle risorse linguistiche precedenti.

Nel 2012 il progetto ottiene i finanziamenti da parte di Google, che decide di convertire la MIDT in una nuova risorsa linguistica, l'Italian Stanford Dependency Treebank (ISDT): questa raccolta di testi si discosta dalla versione precedente poiché è caratterizzata da 64 tipi di dipendenza, in seguito alla reintegrazione di alcuni tipologie di dipendenza minimali.

L'avanzamento delle tecnologie linguistiche verso la creazione di nuove e più ricche risorse per la lingua italiana rappresenta un punto di partenza fondamentale per lo sviluppo di applicazioni efficienti nel campo dell'intelligenza artificiale. Inoltre, tale spinta innovativa costituisce una fase chiave nel superamento della scarsità di risorse per l'italiano, un aspetto critico che per lungo tempo ha caratterizzato lo stato dell'arte della ricerca linguistica nel nostro Paese. Questo argomento ha da sempre suscitato in me un enorme interesse, ragione per cui, durante il mio percorso accademico, mi sono accostata con grande impegno e dedizione allo studio della linguistica computazionale. In particolare, durante il periodo universitario, ho realizzato due progetti mirati proprio alla creazione e all'arricchimento di risorse linguistiche per l'italiano: il primo, svolto sotto forma di tirocinio presso l'Istituto di Linguistica Computazionale del CNR di Pisa, riguarda l'annotazione semantica di un corpus di frasi in lingua italiana, impiegate poi come dati di test da sottoporre ai sistemi di NLP nella campagna di valutazione EVALITA 2011; il secondo progetto, svolto durante il corso di *Tecnologie linguistiche per l'estrazione di informazione*, consiste in un'analisi della leggibilità di testi di dominio medico tramite il software DYLAN. Il motivo per cui ho scelto di esporre questi miei due progetti è che entrambi i lavori mi hanno fatto comprendere due aspetti diversi e rilevanti

della linguistica computazionale: il primo progetto ha rivelato l'importanza di creare un ambiente condiviso in cui linguisti e informatici possano collaborare in vista del miglioramento delle performance dei sistemi di NLP; il secondo progetto mi ha permesso di creare materialmente una risorsa aperta ed estensibile, di sottoporla a un software di analisi e di osservare i risultati ottenuti.

Primo progetto: preparazione del test set per EVALITA 2011

Il progetto di cui mi sono occupata durante il periodo di tirocinio ha previsto la revisione manuale di una collezione di frasi di dominio giuridico, pre-annotate automaticamente a livello semantico da Moses, uno strumento di annotazione sviluppato dall'Università di Roma, Tor Vergata[4] (Basili et al. 2009). L'annotazione delle frasi è avvenuta tramite l'assegnazione di categorie semantiche in stile FrameNet. I risultati dell'annotazione sono stati poi impiegati come dati di test da sottoporre ai sistemi di Frame Labeling nella campagna di valutazione EVALITA 2011.

Le campagne di valutazione, come EVALITA, sono delle iniziative che promuovono la valutazione dei sistemi di NLP. Una campagna di valutazione assume la forma di una competizione tra più squadre di sviluppatori di sistemi: dato un certo compito linguistico, ogni squadra partecipa all'esecuzione del compito con il proprio sistema. I risultati ottenuti dalle varie squadre vengono poi messi a confronto, per determinare quale sistema abbia portato a termine il compito in maniera più efficiente. Prima dell'esecuzione del compito, i sistemi vengono adeguatamente addestrati con dei dati di "esempio", chiamati *training data*. I dati su cui invece i sistemi si confrontano sono i dati di prova, o *test data*. I risultati finali vengono analizzati e discussi in una serie di workshop, e rappresentano il punto di partenza con cui valutare pregi e difetti dei sistemi partecipanti.

Lo scopo delle campagne di valutazione è di misurare le prestazioni di un sistema, per determinare se, e in quale misura, il sistema è in grado di soddisfare gli obiettivi per cui è stato creato. L intento è, quindi, di verificare che il prodotto risponda in maniera adeguata e coerente alle esigenze degli utenti. Gli sviluppatori di sistemi ricorrono spesso all'ausilio delle campagne di valutazione per due motivi principali: innanzitutto, perché la definizione di precisi criteri valutativi consente di specificare in modo altrettanto preciso i problemi che il sistema deve affrontare; in secondo luogo, la campagna di valutazione rappresenta un'occasione per confrontarsi su un determinato compito linguistico, comparando le soluzioni adottate e i risultati ottenuti. L'utilità di questi incontri risiede nella capacità di creare uno stretto rapporto di collaborazione tra linguisti e informatici, generando una sinergia fondamentale nella creazione di risorse comuni e nella condivisione dei criteri di valutazione.

[4] Moses è stato utilizzato per l'annotazione di una porzione di testo allineata dall'inglese all'italiano del corpus Europarl.

Nel 2007, in Italia, nasce il progetto EVALITA, mirato alla valutazione dei sistemi di NLP per il trattamento automatico dell'italiano. Ad oggi, EVALITA si è svolta nell'arco di tre edizioni, avvenute nel 2007, 2009 e 2011: ciascuna di esse presenta dei tratti distintivi, in quanto ogni edizione documenta i progressi raggiunti nella campagna precedente, sviluppandoli ed estendendoli a nuove sfide linguistiche.

La prima edizione è nata con la convinzione che la diffusione di metriche di valutazione condivise fosse un elemento cruciale per lo sviluppo del trattamento automatico della lingua. Questa campagna ha però coinvolto soltanto risorse di tipo scritto, tralasciando il trattamento automatico del parlato. La seconda edizione ha invece introdotto la valutazione dei sistemi per la lingua parlata, concentrando l'attenzione su alcuni aspetti della *speech technology* ispirati alle campagne di valutazione statunitensi. Tuttavia, lo scarso interesse dei linguisti italiani verso i sistemi di parlato e la situazione frammentaria, in questo campo, della comunità di ricercatori hanno comportato alcune difficoltà nel reperimento delle risorse in lingua italiana. La terza edizione ha infine introdotto alcuni compiti di annotazione semantica, tra cui il Frame Labeling over Italian Texts (FLaIT), nucleo del mio lavoro di tirocinio.

L'annotazione semantica rende il significato del testo esplicito e accessibile al calcolatore, permettendo di superare l'ambiguità del linguaggio naturale: ciò avviene grazie all'annotazione delle parole tramite categorie semantiche. Rispetto agli strati inferiori di annotazione (morfologica e sintattica), l'annotazione semantica agisce a un livello più profondo: arricchisce i dati con informazioni legate al loro significato. Inoltre, addestrando il computer sul modo in cui le parole sono correlate e su come tali relazioni possono essere elaborate automaticamente, si è in grado di implementare operazioni di ricerca sempre più complesse. L'estrazione di conoscenza semantica dal testo costituisce uno strumento prezioso per la costruzione formale di ontologie, nonché per la creazione di applicazioni come i motori di ricerca semantici, l'estrazione di informazione e il recupero di documenti.

Il compito semantico di Frame Labeling, nucleo del mio primo progetto, consiste nella capacità di individuare, in una frase italiana, l'evento semantico (*frame*) evocato da un predicato (*target*) e i ruoli semantici che ruotano attorno al predicato (*frame elements*). I frames da individuare sono del tipo definito nel progetto FrameNet (Baker et al. 1998), un modello verso cui tendere per lo sviluppo di risorse in lingua italiana, proposito già in fase di crescita nel progetto iFrame[5]. FrameNet è nato da un'idea dell'International Computer Science Institute di Berkeley e si occupa dell'annotazione di testi tramite frames semantici, ispirandosi al precedente lavoro di Charles J.

[5] Sito del progetto iFrame: http://sag.art.uniroma2.it/iframe/doku.php.

Fillmore (Fillmore et al., 2003). Un frame costituisce il modello di un evento, di uno stato o di una situazione, evocato da un'unità lessicale detta target. Alla completa realizzazione del frame convergono una serie di ruoli semantici, o frame elements, che descrivono le relazioni semantiche tra il verbo e i suoi argomenti. Lo scopo di FrameNet è dunque di definire l'insieme di tutti i possibili frames e di annotare le frasi in modo da istanziare sintatticamente i frame elements attorno al target. A scopo esemplificativo, si consideri il frame *Commerce_pay*, evento in cui un compratore paga per un prodotto. I frame elements da individuare sono i seguenti:

- *Buyer* (il compratore);
- *Goods* (il prodotto ceduto al compratore);
- *Money* (il compenso per ottenere il prodotto);
- *Seller* (il venditore);
- *Rate* (il tipo di pagamento).

I frame elements elencati appartengono al novero degli elementi nucleari del verbo; tuttavia, esistono anche degli elementi non nucleari che si presentano sotto forma di sintagmi preposizionali o avverbi, e indicano fattori circostanziali come il luogo, il tempo, lo scopo o il modo in cui si svolge l'evento. Il database di FrameNet contiene più di mille frames, correlati tra loro mediante una rete di relazioni semantiche, e rende disponibile una molteplicità di situazioni semantiche utili per lo sviluppo di sistemi di NLP. Il ricorso ai frames, infatti, si rivela particolarmente proficuo nei sistemi di comprensione del testo o di disambiguazione, in cui bisogna associare alle frasi una o più strutture semantiche.

La mia partecipazione ad EVALITA ha previsto la preparazione dei dati di test da sottoporre ai sistemi di Frame Labeling durante la campagna di valutazione. Il lavoro si è svolto in due parti:

1. selezione, da un repertorio di 100 frasi, di un campione di dieci frasi per ciascuno dei 38 frames a disposizione;
2. revisione e correzione manuale delle frasi pre-annotate da Moses.

Durante la fase di selezione sono state rispettate tre regole:

- scegliere solo frasi con un target verbale (p.es. *donare*, invece di *donazione*);
- prediligere verbi quanto più diversi possibile, per fornire uno spettro ampio e articolato dei vari tipi di target.

- evitare target che si trovano in una subordinata relativa (p.es. *vediamo ancora animali (...) che muoiono*).[6]

Successivamente, si è svolta la fase di revisione e correzione degli errori. Gli eventuali dubbi relativi all'annotazione dei ruoli semantici sono stati risolti consultando il corpus di addestramento, impiegato poi per addestrare i sistemi in gara a EVALITA. Il corpus di addestramento è nato dalla fusione di due risorse già esistenti: una collezione di frasi sviluppata dalla Fondazione Bruno Kessler (Tonelli et al. 2008), e l'ISST-TANL Corpus, creato dall'Istituto di Linguistica Computazionale di Pisa come risultato della revisione di un sottoinsieme della ISST (Montemagni et al. 2003).

Lo svolgimento di questa campagna di valutazione ha messo in luce le grandi capacità dei sistemi di NLP nell'ambito dell'annotazione semantica: i punteggi migliori si sono rivelati abbastanza alti e simili tra loro, vantando una precisione di più del 70%. Questi risultati possono essere paragonati alle performance delle tecnologie per la lingua inglese, che tuttavia dispongono di una gamma di risorse certamente più ampia rispetto all'italiano. E' ragionevole, quindi, osservare i vantaggi che un approccio valutativo comporta per lo sviluppo di risorse aperte e condivise: innanzitutto, le sfide lanciate dal progresso tecnologico incoraggiano la definizione di compiti linguistici sempre più interessanti e sofisticati, e ciò aumenta l'interesse nei confronti delle prestazioni dei sistemi di NLP, spingendo alla creazione di prodotti adeguati alle esigenze degli utenti. Inoltre, non bisogna sottovalutare l'impatto sociale delle iniziative come EVALITA, che da un lato offrono un'opportunità di collaborazione tra linguisti e informatici, e dall'altro dischiudono uno spazio condiviso in cui i sistemi di NLP possono confrontarsi e migliorarsi costantemente.

Secondo progetto: analisi della leggibilità di testi medici

Un altro ambito di ricerca molto fertile nel campo delle tecnologie del linguaggio riguarda l'analisi della leggibilità di testi scritti, al fine di misurarne la facilità di comprensione. L'analisi della leggibilità ha occupato, durante gli ultimi 80 anni, uno spazio centrale nel campo della linguistica computazionale, sia a vantaggio della società dell'informazione (come nel caso dell'informazione amministrativa, della comunicazione medico-paziente o dell'accessibilità nel Web), sia per

[6] I sistemi infatti tenderebbero a vedere il *che* come parte integrante del target, quando esso, invece, istanzia i ruoli semantici a cui si riferisce, espressi nella proposizione principale (*animali*);

circostanze relative a persone con scarse competenze linguistiche (per esempio, soggetti con un basso livello di alfabetizzazione, parlanti che apprendono una seconda lingua o pazienti con disabilità intellettive). In tutti questi casi, infatti, è utile monitorare la facilità di comprensione di un testo, in modo che il suo contenuto venga correttamente appreso dal lettore.

Il processo di comprensione di un testo è dato dalla costruzione di una rappresentazione coerente dei contenuti nella mente del lettore, che può avvenire soltanto se si instaura una relazione tra il testo e l'utente. Questo processo di rappresentazione cognitiva segue tre fasi: il lettore identifica le parole e le frasi del testo, le immagazzina in memoria e infine costruisce una rete di relazioni inferenziali tra i concetti appresi durante la lettura. L'elaborazione del testo presenta dunque uno svolgimento sequenziale, e attiva una funzione cognitiva chiamata *Working Memory*, facoltà responsabile della memorizzazione temporanea e della manipolazione simultanea di informazioni. Il compito della working memory è quello di accrescere le informazioni rilevanti per la comprensione del testo, sopprimendo gli elementi irrilevanti. Lo studio approfondito del funzionamento della working memory ha portato due teoreti, Daneman e Carpenter, a individuare una misura chiamata *Working memory span* (Daneman e Carpenter, 1980), definita come il numero di frasi che un soggetto può elaborare e di cui può richiamare l'ultima parola. Appare chiaro, quindi, il ruolo fondamentale che la working memory ricopre nella comprensione testuale: più essa viene sollecitata, più la comprensione del testo risulta difficile. Inoltre, Miller e Kintsch hanno formulato un modello computazionale per l'elaborazione di testo in prosa (Miller e Kintsch, 1980), dimostrando che un lettore umano incontra difficoltà di comprensione negli stessi punti in cui il modello fatica a individuare relazioni coerenti, soprattutto in presenza delle inferenze. Questo avviene perché solo una porzione del testo già letto può essere mantenuta nella working memory e, se si legge un segmento di testo non collegato al contenuto corrente immagazzinato, sarà compito della memoria a lungo termine ricercare informazioni rilevanti nel testo già processato. La working memory occupa un ruolo centrale anche nell'ambito delle *Intellectual Disabilities (ID)*, disturbi cognitivi caratterizzati dall'indebolimento delle abilità di lettura. Le ID causano una limitazione delle funzioni cognitive e mnemoniche, difficoltà di apprendimento a livello fonologico e incapacità di attivare la conoscenza semantica veicolata dal testo. Dunque la possibilità di valutare la leggibilità di un testo e di semplificarlo è di massima importanza anche per i soggetti affetti da ID, che in questo modo hanno l'occasione di esercitarsi nella lettura.

La semplificazione di un testo si identifica innanzitutto come semplificazione lessicale e sintattica. Le **misure tradizionali** per la valutazione della leggibilità si basano su funzioni lineari che coinvolgono pochi fattori, di tipo lessicale e sintattico. I fattori lessicali riguardano la difficoltà delle

parole, p. es. la misura della loro lunghezza o frequenza; i fattori sintattici monitorano il grado di complessità delle frasi, p. es. contando il numero di parole in una frase. Il vantaggio principale delle misure tradizionali è certamente la loro semplicità di calcolo, tuttavia presentano due svantaggi: una scarsa affidabilità nei criteri di misurazione (in quanto non sempre le frasi più corte sono le più comprensibili) e l'assenza di fattori legati al *discourse processing*, ovvero la sfera del discorso in cui si collocano l'argomento del testo, la sua coesione e coerenza, le intenzioni comunicative dell'autore e le conoscenze possedute dal lettore.

Come reazione alla scarsa affidabilità delle misure tradizionali, nascono le **tecniche di NLP**, che adottano sistemi di semplificazione testuale più elaborati. La semplificazione lessicale avviene calcolando la frequenza delle parole e creando delle liste di parole "facili", al fine di verificare che percentuale di termini presenti nella lista venga riscontrata anche nel testo. La semplificazione sintattica si compie, invece, grazie alla costruzione di alberi sintattici e alle tecniche di riconoscimento dei pattern[7]: l'osservazione di questi due fattori permette di trasformare i periodi lunghi e complessi in frasi corte e più semplici. Il vantaggio degli approcci adottati dai sistemi di NLP risiede nella possibilità di rendere il processo di semplificazione automatico, quindi meno costoso e più affidabile. Tuttavia, anche le tecniche di NLP presentano alcuni svantaggi: come le misure tradizionali, ignorano gli aspetti legati al discourse processing. Inoltre, trasformando le frasi lunghe e complesse in frasi più brevi, aumentano la lunghezza del testo, incrementando quindi il numero di informazioni da elaborare.

Per queste ragioni, le tecnologie del linguaggio hanno migliorato le loro capacità nell'ambito della semplificazione testuale, a partire dal lavoro di Pitler e Nenkova, i quali hanno realizzato un sistema in grado di integrare i tre livelli linguistici (lessicale, sintattico e del discorso). Il modello di Pitler e Nenkova si è dimostrato particolarmente valido nell'individuazione della coesione lessicale del testo e nella costruzione delle relazioni inferenziali (Pitler e Nenkova, 2008), aspetti del tutto innovativi per un sistema di NLP. Le ricerche in questo campo sono proseguite con l'intenzione di mantenere le informazioni rilevanti a discapito di quelle irrilevanti, le quali vengono semplificate o totalmente eliminate dal testo. In proposito, sorgono due questioni principali:

1. Come identificare le porzioni di testo difficili per il lettore;
2. Scegliere, tra più alternative di semplificazione, la soluzione ottimale.

Il metodo più idoneo per risolvere tali questioni risiede nell'impiego di uno **strumento automatico** per la valutazione della leggibilità. Lo strumento automatico opera sul testo tramite fasi distinte e

[7] Le tecniche di riconoscimento dei pattern individuano delle espressioni linguistiche fisse, che tendono a ricorrere nel testo.

incrementali: dapprima, viene impiegato sul testo per individuare i passaggi più complessi; in seguito, se esiste più di una soluzione semplificativa, valuta la leggibilità di ognuna e sceglie l'alternativa ottimale; infine, valuta il procedimento complessivo misurando la leggibilità del testo prima e dopo l' attività di semplificazione.

READ-IT è il primo strumento di valutazione della leggibilità per la lingua italiana (Montemagni et al. 2011). Esso definisce una serie di *features*, o fattori di leggibilità, e crea un modello statistico usando le features estratte dal corpus di riferimento. Il modello, inoltre, è in grado operare su testi annotati a dipendenze e di effettuare una distinzione binaria tra testi "facili" e "difficili".

Oltre a READ-IT, è stato sviluppato un altro strumento per l'analisi dei testi italiani: si tratta di DYLAN, un software realizzato dal DylanLab del CNR di Pisa, un laboratorio di modelli computazionali del linguaggio. L'impiego di DYLAN si è rivelato fondamentale per lo svolgimento del mio progetto incentrato sulla leggibilità di testi di dominio medico. I testi da analizzare sono stati suddivisi in tre gruppi, in base al registro comunicativo: testi divulgativi, testi scientifici e testi farmaceutici (ovvero i foglietti illustrativi allegati ai farmaci). Lo scopo primario del progetto era quello di valutare l'efficacia della comunicazione medico-paziente, verificando se, e in che misura, i testi in ambito medico si presentano comprensibili al lettore. Come obiettivo secondario, il progetto si proponeva di semplificare alcuni testi complessi, al fine di ri-sottoporli a DYLAN per riscontrare i miglioramenti ottenuti con la semplificazione. In particolare, la scelta del materiale si è focalizzata su testi di ginecologia e ostetricia. Le risorse da cui sono stati estratti i testi sono il sito della SIEOG (Società Italiana di Ecografia Ostetrica e Ginecologica)[8], il sito di AOGOI (Associazione dei Ginecologi Italiani)[9] e il sito "Prontuario Farmaci"[10]. Le ragioni per cui ho deciso di concentrarmi su testi di questo genere vertono sulla necessità di migliorare la comprensibilità delle informazioni destinate a un certo tipo di pubblico, quello femminile. L'urgenza di semplificare si manifesta innanzitutto nell'informazione destinata alle ragazze più giovani: capire in modo chiaro la natura e l'uso dei metodi contraccettivi previene non solo lo sviluppo di gravidanze indesiderate, ma anche l'insorgenza di rischi per la salute. Inoltre, la semplificazione testuale giova anche all'informazione destinata alle donne adulte, relativamente all'utilità di determinati esami (p. es. ecografie, pap-test) e all'approfondimento di alcune patologie gravi da non sottovalutare. L'analisi dei foglietti illustrativi, invece, interessa l'informazione farmaceutica del paziente, in modo che quest'ultimo sia in grado di comprendere con facilità i benefici e le controindicazioni di un farmaco (p. es., se l'assunzione di alcune sostanze da parte di una donna incinta nuoce al feto).

[8] www.sieog.it
[9] www.aogoi.it
[10] www.prontuariofarmaci.com

Tutti i testi selezionati sono stati sottoposti all'analisi di DYLAN, sia da un punto di vista lessicale che sintattico.

I testi divulgativi della SIEOG sono stati prelevati da una sezione del sito denominata "Canale Donna", in cui le tematiche in campo ginecologico vengono affrontate in modo più semplice e discorsivo rispetto ai trattati scientifici. L'analisi della leggibilità di questi testi ha rivelato:

- un livello di difficoltà lessicale medio (55%), dovuto principalmente a un'alta densità lessicale (ovvero l'uso di molti termini diversi);
- un livello di difficoltà sintattica estremamente elevato (98%), a causa della presenza di periodi lunghi e caratterizzati da molte subordinate.

I testi scientifici della SIEOG hanno mostrato:

- una difficoltà lessicale alta (98%);
- una difficoltà sintattica alta (90%), ma inaspettatamente inferiore rispetto ai testi divulgativi; questo risultato si deve all'uso ridotto delle subordinate, alle frasi corte e alla massiccia presenza di liste, piuttosto che di testo in prosa.

Il materiale divulgativo dell'AOGOI è stato ricavato da una sezione chiamata "Teen Aogoi"[11], in cui sono presenti testi rivolti alle ragazze molto giovani. La valutazione della leggibilità ha evidenziato:

- un lessico di facile comprensione (tra il 2% e il 34%, a seconda dei testi), grazie alla presenza di termini semplici e alla bassa densità lessicale;
- una complessità sintattica molto alta, che raggiunge addirittura il 100% di difficoltà a causa dell'abbondanza di subordinate e delle grandi distanze tra testa e dipendente.

Nei testi scientifici dell'AOGOI si è osservato che:

- il lessico è formale e ricercato (92,5%);
- la sintassi non è eccessivamente complessa (81,7%): sono infatti presenti molte proposizioni principali e poche subordinate.

Infine, per quanto riguarda l'analisi della leggibilità dei testi farmaceutici, sono stati impiegati una serie di foglietti illustrativi di farmaci utilizzabili durante la gravidanza. L'esame linguistico ha rivelato:

[11] www.teen.aogoi.it

- una grande variabilità nel tipo di lessico (tra il 5% e il 59% di complessità), caratterizzato per lo più dall'uso di termini comuni, frequentemente ripetuti nel testo;
- una difficoltà sintattica piuttosto elevata (tra l'88% e il 100%), dovuta principalmente alla scarsità di teste verbali, spesso sostituite dagli aggettivi (p.es., *indicata per*).

Un aspetto molto interessante riscontrato durante la valutazione dei testi di tipo farmaceutico risiede nel fatto che i farmaci generici (p. es., la Tachipirina) presentano un lessico estremamente semplice, al punto da raggiungere un livello di complessità pari allo 0%. Questa osservazione spinge a chiedersi se la genericità del farmaco possa influire in qualche modo sulla difficoltà lessicale del testo: la mia opinione è che i farmaci generici, dovendo trattare patologie altamente diffuse, tendono a utilizzare termini di uso comune (p. es., *tosse*, *mal di gola*, *influenza*), a differenza dei farmaci specifici, i quali devono necessariamente menzionare patologie di minore diffusione, e quindi estranee al vocabolario comune.

Il processo di valutazione della leggibilità dei testi di dominio medico ha evidenziato una serie di elementi alla base delle complessità riscontrate. In particolare, si è osservato come negli articoli divulgativi prevalga una difficoltà di tipo sintattico, dovuta essenzialmente alla lunghezza eccessiva delle frasi: questo avviene perché i testi, nel tentativo di fornire spiegazioni esaurienti alle tematiche trattate, caricano la prosa con periodi molto lunghi e articolati. La soluzione più adatta sarebbe quella di suddividere le porzioni testuali lunghe con passaggi più brevi, caratterizzati da parole semplici e spesso ripetute. Il materiale scientifico ha esibito una difficoltà lessicale alta, a causa della presenza di termini prettamente medici, e una sintassi mediamente complessa, ma tuttavia minore rispetto a quella dei testi divulgativi. Si tratta di un risultato inaspettato, in quanto solitamente si ritiene che i trattati medici usino una sintassi tortuosa e articolata; in realtà, l'andamento sintattico è attenuato dal frequente ricorso alle liste e agli elenchi puntati. Infine, nei testi farmaceutici, la complessità della sintassi si è rivelata nell'uso predominante degli aggettivi al posto dei verbi, condizione che comporta il collocamento di molti dipendenti per testa verbale; l'alternativa di semplificazione più idonea sarebbe di inserire nelle frasi un maggior numero di teste verbali.

Al termine dell'analisi, per verificare la validità dei risultati ottenuti, ho selezionato e riscritto in maniera più comprensibile due articoli divulgativi e un foglietto illustrativo. I documenti, scelti tra quelli che hanno presentato maggiori complessità a livello lessicale e sintattico, sono stati poi nuovamente sottoposti all'analisi del software DYLAN, al fine di determinare l'incidenza delle semplificazioni sulla leggibiltà. Gli articoli divulgativi sono stati modificati segmentando i periodi lunghi in frasi più corte, utilizzando sinonimi più semplici e ripetendo più volte le stesse parole. Il

foglietto illustrativo è stato trasformato solo a livello sintattico, introducendo un maggior numero di teste verbali. I risultati dell'analisi effettuata sui testi riscritti hanno evidenziato un soddisfacente miglioramento della leggibilità: nel caso degli articoli divulgativi, la complessità (sia sintattica che lessicale) è scesa dal 100% al 96,8%, mentre il testo farmaceutico ha visto una drastica diminuzione della difficoltà grammaticale, dal 94,8% al 73,1%.

L'esposizione di entrambi i progetti è avvenuta con lo scopo di dimostrare la validità delle tecnologie lingustiche nel recupero di informazione dai testi. Ogni livello di estrazione linguistica (sia esso morfologico, sintattico o semantico) costituisce uno strumento prezioso per lo sviluppo di sistemi per l'elaborazione automatica del linguaggio, e rappresenta il punto di partenza nella creazione di applicazioni sempre più sofisticate. Inoltre, il tragitto verso il progresso tecnologico segue un andamento sequenziale, per cui le informazioni linguistiche estratte a un dato livello rappresentano gli elementi necessari per l'estrazione di informazione al livello successivo. Tuttavia, la sola definizione dei metodi di codifica non è sufficiente allo svolgimento dei compiti linguistici: la parte teorica deve necessariamente accompagnarsi a un'adeguata disponibilità delle risorse. Questo è il punto su cui la ricerca linguistica, a mio avviso, dovrebbe insistere, promuovendo la diffusione libera e condivisa delle fonti testuali. Soltanto con l'impegno nella costruzione di risorse ampie e variegate si può aspirare allo sviluppo di sistemi automatici veramente innovativi, che siano in grado di simulare la naturalezza tipicamente umana del linguaggio. I progressi già in corso si devono indubbiamente alla realizzazione di iniziative come EVALITA, che promuovo lo sviluppo dei sistemi di NLP tramite la definizione di nuove sfide linguistiche. Il futuro delle tecnologie del linguaggio è certamente orientato all'approfondimento dell'annotazione semantica, una questione fortemente innovativa ma ancora in fase di crescita, nonché al miglioramento delle performance delle applicazioni, possibile solo grazie a una divulgazione estesa e condivisa delle risorse linguistiche.

Bibliografia

Baker, C.F., Fillmore, C.J., Lowe, J.B. 1998. *The Berkeley FrameNet Project*. In: Proceedings of the 36th ACL Meeting and 17th ICCL Conference, Morgan Kaufmann.

Basili, R., De Cao, D., Croce, D., Coppola, B., Moschitti, A. 2009. *Cross-language frame semantics transfer in bilingual corpora*. In: Proc. of 10th Int. Conf. On Intelligent Text Processing and Computational Linguistics (CICLing 2009), Mexico City, Mexico.

Basili, R., De Cao, D., Lenci, A., Moschitti, A., and Venturi, G.: Evalita 2011: *the Frame Labeling over Italian Texts Task*. In: Working Notes of EVALITA 2011, 23-24th January 2012, Rome, Italy, ISSN 2240-5186 (2012).

Daneman, M., Carpenter, P. A. 1980. *Individual differences in working memory and reading*. In: Journal of Verbal Learning and Verbal Behavior.

Fillmore, Charles J., Christopher R. Johnson e Miriam R.L. Petruck. 2003. *Background to FrameNet*. In: "International Journal of Lexicography", 3, pp. 235-250.

Lenci A., S. Montemagni, V. Pirrelli. 2009. *Annotazione sintattica di corpora: aspetti metodologici.* In: Corpora di italiano L2: tecnologie, metodi, spunti teorici, Guerra Edizioni, Perugia.

Miller, J. R., & Kintsch, W. 1980. *Readability and recall of short prose passages: A theoretical analysis*. In: Journal of Experimental Psychology: Human Learning and Memory.

Montemagni, S., F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M.T. Pazienza, Saracino, F. Zanzotto, N. Mana, F. Pianesi, R. Delmonte. 2003. *Building the Italian Syntactic-Semantic Treebank.* In: Anne Abeill'e (ed.), "Building and using syntactically annotated corpora", Kluwer, Dordrecht.

Montemagni S., F. Dell'Orletta, G. Venturi. 2011. *Assessing readability of Italian texts with a view to text simplification*. In: Proceedings of the 2nd workshop on speech and language processing for assistive technologies, pp. 73-83, Edinburgh, Scotland, UK.

Pitler E., M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, A. Joshi. 2008. *Easily identifiable discourse relations*. In Coling 2008: Companion volume: Postersand Demonstrations, Manchester, UK.

Sito web di EVALITA, pagine *Evalita 2007*, *Evalita 2009*, *Evalita 2011*

http://www.evalita.it

Sito web di FrameNet, pagina *About FrameNet*

https://framenet.icsi.berkeley.edu/fndrupal/about

Tonelli, S., E. Pianta. 2008. *Frame information transfer from english to italian*. In: Proc. of LREC Conference, Marrakech, Marocco.

Wikipedia, voce *Google Knowledge Graph* (pagina visitata il 28 agosto 2013);

Wikipedia, voce *Question Answering* (pagina visitata il 28 agosto 2013);

Working Notes of EVALITA 2011, 23-24th January 2012, Rome, Italy, ISSN 2240-5186 (2012).