Linguaggio Naturale e Ontologie.

La Semplificazione Testuale come metodo per l'interazione tra uomo e macchina

Francesca Bonin

.

L'idea che Lotaria legga i miei libri a questo modo
mi crea dei problemi. Adesso ogni parola che scrivo la vedo giá
centrifugata dal cervello elettronico, disposta nella graduatoria
delle frequenze, vicino ad altre parole che non so quali possano essere,
e mi domando quante volte l'ho usata,[...],provo a immaginarmi
quali conclusioni si possano trarre dal fatto che ho usato
una volta o cinquanta volte quella parola.
Forse sará meglio che la cancelli...

tratto da *Se una notte d'inverno un viaggiatore*,
Italo Calvino, 1979.

.

Indice

Elenco delle tabelle

Elenco delle figure

Introduzione

Pensiero e Linguaggio

Pensiero e linguaggio, concetti e parole. Due mondi distanti o vicini?

Il problema della rappresentazione della conoscenza rappresenta, al giorno d'oggi, una delle sfide più avvincenti della ricerca scientifica. Una sfida che coinvolge discipline e aree di studio apparentemente distanti fra loro: dalla filosofia alla linguistica, all'informatica. Ma ancora più affascinante è stato ed è tutt'oggi il passo successivo: come permettere all'uomo di interagire in modo naturale, senza sforzo e senza limiti, con sistemi di rappresentazione della conoscenza?

E' interessante, in un mondo in cui queste interazioni stanno diventando sempre più frequenti, soffermarsi a riflettere sui problemi che emergono quando l'utente umano, con il suo linguaggio e la sua "visione del mondo", si trova a dover comunicare con un sistema di dati strutturato. Come arrivare ad una lingua comune, come raggiungere quel linguaggio intermedio che consenta una comunicazione libera, che non sia costretta a ''pagare il costo di un interprete''?

Fra pensiero e linguaggio

E' proprio la stretta relazione tra pensiero e linguaggio che ha posto il "natural language processing" al centro del dibattito dell'intelligenza artificiale, e lo ha reso un elemento centrale per lo sviluppo delle nuove tecnologie.

Questo campo nasce sulla sottile linea di confine fra le scienze cognitive e l'intelligenza artificiale, sebbene i cognitivisti abbiano sempre concentrato l'attenzione su "come la mente umana memorizza l'informazione", mentre i computazionalisti su come gestire automaticamente l'informazione. Nell'ambito dell'intelligenza artificiale si è ben presto riconosciuto che, perché una macchina possa esibire un comportamento intelligente, che le permetta di interagire con l'uomo, ha bisogno di gestire grandi quantità di informazione. Quindi uno dei problemi fondamentali che è necessario affrontare è riuscire a rappresentare la conoscenza in modo che sia manipolabile automaticamente nel modo più agevole possibile e che possa essere usata per ragionare ed effettuare deduzioni efficientemente. Dal momento della nascita di questa consapevolezza, quindi, vari sono stati i formalismi studiati per la definizione e la manipolazione di basi di conoscenza, e l'ultimo approdo della ricerca ha riguardato il vasto mondo delle ontologie, rappresentazioni di informazione strutturata a partire da concetti condivisi[1]. Allo stato attuale l'ontologia è considerata da molti come il metodo di rappresentazione della conoscenza più efficace (Borgida *e altri* (1989)), proprio perché presenta informazione formalizzata in un linguaggio logico su cui è possibile inferire informazione implicita.

Ma se da un estremo vi è la conoscenza, e la sua rappresentazione logica in un'ontologia, dall'altra vi è l'uomo e la sua necessità di interagire con essa. Come avviene questa

[1] Si veda Capitolo 2

interazione, quale linguaggio comune devono parlare questi due "soggetti" per riuscire a comunicare?

Il lavoro proposto andrà ad approfondire il problema della ricerca di quel linguaggio, che può essere condiviso dall'uomo e dai sistemi di rappresentazione della conoscenza, senza intermediari.

Studiando da un lato il linguaggio logico e la sua espressività, e dall'altro il linguaggio naturale con le sue caratteristiche, si cercherà il giusto equilibrio fra questi mondi in un frammento di inglese che sia esprimibile dal linguaggio logico, e, allo stesso tempo, soddisfi le esigenze dell'utente.

Sarà possibile raggiungere questo compromesso?

E se sì, sarà possibile allora permettere all'utente di interfacciarsi con la macchina usando una sintassi libera, che non sia costretta dai vincoli di un linguaggio controllato? Il fine ultimo che la ricerca può prefiggersi è un'interazione uomo macchina assolutamente naturale, un'interazione in cui l'uomo non debba più compiere alcuno sforzo di semplificazione del proprio linguaggio. Ma questo è un traguardo da raggiungere gradualmente, in un campo dove ogni piccola conquista rappresenta un passo sulla strada dell'incontro tra l'uomo e la macchina. Un traguardo ambizioso, ma proprio per questo una problematica decisamente affascinante.

Struttura della tesi

In questo lavoro si vanno innanzi tutto ad introdurre i concetti che sono stati al centro dello studio condotto: dalle ontologie fino alle *Description Logic*, linguaggi di rappresentazione

della conoscenza di cui si é voluta studiare l'espressivitá in termini linguistici (Capitolo 2). Nel Capitolo 3 si affronta, invece, il problema dall'altro punto di vista, con uno sguardo ai linguaggi controllati, come strategia per far incontrare le esigenze dell'utente con quelle dell'ontologia riducendo la libertá espressiva del'utente stesso. In seguito si descrivono i risultati emersi da analisi di corpora effettuate per studiare la distanza tra l'espressivitá delle *description logics* utilizzate e il linguaggio naturale; tali analisi hanno riguardato due tipi di corpora: corpora di queries per studiare i limiti del linguaggio di querying, e corpora di asserzioni, per approfondire i limiti del linguaggio di descrizione dell'ontologia (Capitolo 4).

Infine si analizzano i problemi emersi, cercando di suggerire possibili soluzioni attraverso forme di semplificazione testuale sintattica o semantica, dove possibile, o attraverso proposte di modifica dell'espressività del linguaggio logico. In particolare nel Capitolo 5 si approfondiscono le tematiche legate al querying dell'ontologia, mentre nel Capitolo 6 si trattano le tematiche legate all'arricchimento della conoscenza ontologica. Nel Capitolo 7 sono invece raccolte alcune alcune riflessioni finali

Parte del materiale di questa tesi è stato pubblicato, in particolare il Capitolo 4 estende i risultati pubblicati in (Bernardi *e altri* (2007a)).

Dal punto di vista dei programmi per l'analisi linguistica la scelta è ricaduta sulla catena di analizzatori C&C tools, con il parser CCG, sviluppato da James Curran e Stephen Clarck (Bos *e altri* (2004)) . A questi è associato l'analizzatore semantico Boxer, creato da Johan Bos (Curran *e altri* (2007), pgg 2-5)).

Capitolo 1

La ricerca di un frammento condiviso

1.1 Dai sistemi di rappresentazione della conoscenza all'utente

Al giorno d'oggi è sempre più importante riuscire a gestire grandi quantità di informazione, e, per questo motivo, i sistemi di gestione di dati strutturati sono diventati una risorsa fondamentale in vari ambiti.

Il ruolo centrale giocato dai sistemi di gestione di basi di dati (*Data Base Management System*, DBMS) ha aperto interessanti dibattiti non solo su come organizzare la conoscenza, ma anche su come estrarre la conoscenza.

L'analisi che si propone in questo lavoro si concentra proprio su questo secondo aspetto, affrontando il problema del recupero dell'informazione, da sistemi di gestione di basi di

dati, attraverso il linguaggio naturale.

Se infatti, da un lato, una base di dati può permettere un'ottima organizzazione della conoscenza, dall'altro può creare qualche problema in fase di estrazione della conoscenza stessa, costringendo l'utente non esperto ad interagire attraverso linguaggi di interrogazione, molto lontani dal linguaggio naturale. L'accesso diretto all'informazione immagazzinata in una base di dati, pone l'utente di fronte a due problemi fondamentali: innanzi tutto la necessità di conoscere il linguaggio usato dal sistema per formulare interrogazioni, e in secondo luogo la necessità di conoscere l'esatta struttura in cui sono organizzati i dati, struttura che, spesso, è lontana dai modelli concettuali con cui l'essere umano memorizza l'informazione.

Nel colmare la distanza fra il modello concettuale in cui sono immagazzinate le informazioni in un sistema di gestione di dati strutturati, e il modello concettuale con cui l'uomo memorizza l'informazione, giocano un ruolo determinante le ontologie.

Si è infatti notato che le ontologie[1], modelli di organizzazione della conoscenza, possono rappresentare un ponte fra l'uomo e la base dati, un ponte che consente di ridurre la distanza fra questi due diversi modi di strutturare l'informazione, offrendo una concettualizzazione più ad alto livello dei dati e proponendo un modello concettuale più vicino all'uomo (Calvanese *e altri* (2006)).

Ma anche le ontologie, ad uno sguardo attento, non permettono un accesso naturale all'informazione. I concetti e le relazioni che descrivono un dominio a livello ontologico sono espressi attraverso linguaggi formali, e possono essere interrogati solo con formalismi particolari con i quali l'utente finale può non avere dimestichezza. Quindi, sebbene l'ontologia rappresenti un passo avanti sulla strada che divide il dato puro dall'utente fi-

[1]Per dettagli si rimanda al Capitolo 2.

nale, esiste ancora una forte separazione fra il linguaggio naturale e il linguaggio "parlato dall'ontologia".

Il focus di questo lavoro é proprio l'analisi del *gap* che esiste fra il linguaggio naturale dell'uomo e il frammento di linguaggio naturale rappresentato dall'ontologia. L'obiettivo è costruire un ponte fra l'uomo e l' ontologia, un ponte che si realizza in un frammento di linguaggio che sia condiviso da entrambi e che rappresenti il miglior compromesso fra le esigenze delle parti. Come si vedrá, infatti, da un lato, si vorrebbe permettere all'utente un'interazione libera, in linguaggio naturale senza restrizioni, dall'altra però é necessario fare i conti con un linguaggio logico che, dovendo interfacciarsi anche con la base di dati, deve ridurre la propria espressività, per evitare un innalzamento eccessivo della complessità computazionale.

1.2 Descrizione delle componenti del contesto

Per avere una visione d'insieme del contesto in cui si situa il lavoro, si può immaginare il processo di interrogazione di una base di dati attraverso un'ontologia come un percorso che parte da un utente a cui è data la possibilità di esprimersi in linguaggio naturale, e arriva all'informazione frammentata nei dati di una base di dati. Fra questi due estremi si trovano l'ontologia, che specifica un modello concettuale del dominio rappresentato e la base di dati stessa che memorizza i dati in relazione ai concetti dell'ontologia.

Analizzando ogni componente, si ha:

1-Sistema di gestione di basi di dati.

Un programma in grado di gestire grandi collezioni di dati strutturati, organizzandoli in modo condiviso e persistente, che garantisce efficienza in termini di spazio e tempo, ed efficacia nelle prestazioni (Atzeni e Ceri (2002), pp 1-81). In questo contesto verranno presi in considerazione sistemi di basi di dati strutturati secondo modelli relazionali, dove per modelli si intende l'insieme dei concetti usati per strutturare i dati in modo che siano comprensibili ad un calcolatore, e per "relazionale" il modello più utilizzato che sfrutta il concetto di relazione per organizzare i dati in insiemi di tuple (visualizzate attraverso tabelle).

Nel modello relazionale, i dati sono memorizzati in tabelle (o relazioni) in cui ogni riga rappresenta una tupla, e ogni colonna un campo (o attributo) della tupla stessa. Ad esempio una tabella *IMPIEGATI* potrà essere visualizzata come in tabella 1.1.

Matricola	Nome	Età	Stipendio
111	Rossi	30	2000
222	Bianchi	25	1000

Tabella 1.1: Impiegati

Un linguaggio tradizionalmente usato sia per la descrizione (come DDL, *data description language*), manipolazione (DML, *data manipulation language*) e interrogazione di una base di dati è SQL, acronimo di *Structured Query Language* (Atzeni e Ceri (2002),pag 89-100) un linguaggio dichiarativo che consente la formulazione dell'interrogazione ad alto livello . La query in SQL viene poi passata ad un *ottimizzatore* (una componente del sistema di gestione di basi dati) che la trascrive in

una query equivalente in un linguaggio procedurale interno al sistema. Le operazioni tipicamente usate in fase di interrogazione a basi dati sfruttano gli operatori di selezione, proiezione e join, e permettono di restituire le tuple che soddisfano le condizioni specificate.[2]

2-Ontologia.[3]

Sistema di rappresentazione della conoscenza che, come tale, permette di inferire nuovi fatti da quelli memorizzati, a differenza di una base di dati che, al contrario, non ha capacità inferenziale. Un'ontologia può essere definita attraverso formalismi come le Logiche Descrittive, linguaggi formali usati per descrivere domini in modo strutturato.

In questo studio si andranno a trattare in particolare due linguaggi:

(a) i. DL-Lite: famiglia di formalismi che catturano gli operatori base delle logiche descrittive, mantenendo però bassa la complessità computazionale. Per questo presentano un'espressività ridotta rispetto ad una logica descrittiva completa. Si farà riferimento in particolare a *DL-Lite*$_{core}$, versione che comprende un nucleo di operatori condiviso da tutte le altre versioni di DL-Lite (Bernardi *e altri* (2007b)).

[2]Selezione: operatore che seleziona tuple che soddisfano certe condizioni.

Proiezione : operatore che seleziona i campi di una tabella.

Join: operatore che correla dati in tabelle diverse, sulla base di valori uguali in campi con lo stesso nome.

[3] Per dettagli si rimanda al Capitolo 2.

ii. Queries congiuntive: (*conjunctive queries,* di qui in avanti *CQs*) linguaggio di interrogazione dell'ontologia, che esprime gli operatori di selezione, join e proiezione delle queries SQL (Calvanese *e altri* (2006)).

3-Utente finale.

Mira a recuperare informazione proponendo un input in linguaggio naturale, che deve interfacciarsi con l'ontologia, e con il linguaggio dell'ontologia. Per studiare la distanza fra l'input generalmente usato dall'utente e l'ontologia, è necessario condurre un'analisi di corpora che metta in luce le caratteristiche del linguaggio naturale. Nel lavoro, che qua si propone, tale analisi è stata condotta sull'inglese, e in inglese sono quindi i corpora analizzati per studiare le caratteristiche dell'input in linguaggio naturale.

Riassumendo, questo lavoro si pone le seguenti domande: "*quali sono le strutture più comuni e quali gli operatori logici più frequenti in questo tipo di corpora?*", e ancora "*quanto è distante questo tipo di linguaggio dal frammento di inglese espresso e compreso dall'ontologia?*".

1.3 Obiettivi della tesi

La domanda dell'utente dovrà essere prima di tutto essere trascritta in modo "comprensibile" per l'ontologia (CQs), quindi nuovamente riscritta in un linguaggio di interrogazione a database (come SQL). La situazione ideale sarebbe quella di poter mantenere, passaggio dopo passaggio, ogni informazione, ogni operatore logico presente nella domanda iniziale, ma ciò sarebbe possibile a livello d'ontologia, solo se questa fosse definita da un

formalismo logico completo che coprisse tutti gli operatori logici esprimibili dal linguaggio naturale.

La necessità di mappare l'ontologia sulla base di dati, però, costringe a fare i conti con la complessità computazionale. E' per mantenere bassa tale complessità che si è scelto di definire le ontologie attraverso formalismi ridotti di logica descrittiva, come DL-Lite, guadagnando in efficienza, ma perdendo in espressività (Calvanese *e altri* (2005)).

Quello che l'ontologia riesce a riconoscere è quindi solo un frammento di linguaggio naturale, un frammento che esclude alcuni operatori (diversi a seconda della versione di logica utilizzata). Quanto costa questa semplificazione logica in termini di naturalezza? Quanto allontana il frammento di inglese gestito dall'ontologia dal linguaggio naturale? E, in ogni caso come trovare un compromesso fra i due?

Fra le strade percorse per colmare questo *trade off*, alcune procedono sul piano del linguaggio naturale, suggerendo di limitare l'utente al momento l'interrogazione (Dongilli *e altri* (2004)), (Schwitter e Tilbrook (2006a)). In altre parole l'utente viene costretto a formulare la domanda in un linguaggio controllato (si veda Capitolo 3), con una sintassi limitata che non accetti strutture non riconosciute dal frammento gestito dal linguaggio logico.

Ma se da un lato, l'uso di un linguaggio controllato può agevolare il riconoscimento della domanda e la sua gestione da parte dell'ontologia, dall'altro esso rappresenta un forte limite per l'utente e un passo indietro sulla strada della "naturalezza dell'interazione".

La riflessione che emerge è se sia veramente necessaria una limitazione dell'utente, un controllo a priori dell'input, o se, al contrario, l'input dell'uomo, lasciato libero di usare qualsiasi struttura, non sia in realtà già vicino al frammento di inglese rappresentato da

un linguaggio logico ridotto.

Con un'analisi dei corpora si è cercato di rispondere a questa domanda, analizzando la frequenza, in domande naturali, di quei costrutti che fuoriescono dal frammento di inglese espresso dal linguaggio logico, e notando come, in fondo, gli utenti semplifichino spontaneamente il proprio input, soprattutto in fase di interrogazione a database, se consapevoli di interagire con una macchina. In parallelo però ci si è soffermati anche sull'altro lato dell'interazione: la costruzione dell'ontologia; l'analisi di quei costrutti che non sono previsti dal linguaggio di descrizione dell'ontologia può aiutare a capire in che misura è ipotizzabile non solo interrogare la base di dati in linguaggio naturale, ma anche inserire informazioni in una base di dati attraverso asserzioni in linguaggio naturale.

Queste analisi hanno portato a due conclusioni:

1- il linguaggio naturale, tipicamente usato dagli utenti, sembra non essere troppo distante dal frammento di inglese abbracciato dalla logica utilizzata.

2- nonostante *1),* esistono costrutti logici che non sono ammessi dalla logica, ma che non possono essere tralasciati in fase di riscrittura della query dal linguaggio naturale al linguaggio di interrogazione a ontologia.

Per gestire questi costrutti problematici si propone un approccio che cerca di muoversi in parallelo sia sul lato del linguaggio naturale che su quello dell'ontologia.

Il proposito è quello di analizzare la frequenza di tali strutture in modo da gestire quelle poco frequenti sul piano del linguaggio naturale (con semplificazione testuale) e pensare per quelle più frequenti ad un ampliamento del frammento di logica, per fare in modo che l'espressività del frammento logico sia plasmata sulle esigenze effettive dell'utente.

Tenendo presente che l'input in linguaggio naturale è sostanzialmente compatibile con il frammento logico scelto e cercando per questo di evitare un controllo a priori, si propone una semplificazione testuale solo per quei costrutti non accettati che, pur non essendo tanto frequenti da giustificare la modifica del frammento, devono comunque essere gestiti.

In questo modo si è cercato di fare qualche passo sulla strada verso il miglior compromesso modificando ora l'input dell'utente, ora l'espressività del linguaggio logico, e cercando di non perdere di vista né le esigenze di naturalezza, né il problema dei costi computazionali.

Capitolo 2

Ontologie e Logiche descrittive

2.1 Ontologie

2.1.1 Ontologie: definizione

Un'ontologia è un' esplicita specifica di una ***concettualizzazione****.*

A sua volta una concettualizzazione è l'insieme di oggetti, concetti ed altre entità che si può assumere esistere in una certa area di interesse e delle relazioni che esistono tra essi"

Gruber (1993)

Le ontologie possono essere descritte informalmente come rappresentazioni semantiche di un dominio attraverso concetti comuni che agevolano la comunicazione sia tra agenti software, sia tra agenti software ed essere umani.

La natura filosofica del termine riporta a quel ramo della metafisica che descrive i

vari tipi e modi di esistenza, trattando temi come classi, entità, proprietà intrinseche ed estrinseche.

Può essere definita come la disciplina filosofica che studia l'ordine e la struttura dell'essere in generale, e trova la sua prima espressione nelle dieci categorie aristoteliche per classificare ogni cosa che può essere detta o predicata.

Un' ontologia quindi è una rappresentazione del mondo, di un mondo o di parte di esso, delle entità che lo popolano e delle relazioni fra queste entità.

Con gli studi di Guarino (1998), le ontologie sono state comunemente definite come specificazioni di concettualizzazioni condivise.

Intuitivamente una concettualizzazione è una conoscenza informale che può essere estratta da esperienza, osservazione ed introspezione, mentre la specificazione è la codifica di tale conoscenza che avviene attraverso un particolare linguaggio di rappresentazione di concetti.

In altre parole dal concetto "tavolo" (a livello di concettualizzazione) si puó passare ad una prima forma di rappresentazione che puó essere chiamare arbitrariamente "concetto_tavolo" (a livello di ontologia), ed in seguito ad una lessicalizzazione di tale concetto con il termine tratto da uno specifico vocabolario di parole (livello del lessico), ottenendo, per esempio, "mesa" in spagnolo, "tavolo" in italiano, "table" in inglese etc.

Per approfondire la sottile differenza fra concettualizzazione ed ontologia, si puó affermare che la concettualizzazione è un modello astratto del mondo, che si concretizza con l'ontologia in una rappresentazione formale (non ancora lessicale). In seguito, i concetti dell'ontologia possono, a loro volta, essere realizzati dal lessico di una specifica lingua (come mostrato in tabella 2.1).

Concettualizzazione	Concetto astratto di un tavolo
Ontologia	<concetto_table>
Lessico	"tavolo", sostantivo maschile singolare.

Tabella 2.1: Concettualizzazione

Secondo questa definizione, quindi, il concetto di ontologia è indipendente dal linguaggio, perché la lessicalizzazione coinvolge la fase successiva, e non rientra in quella sfera che viene definita il livello dell'ontologia.[1]

Ontologia e lessico possono apparire come concetti simili, tanto che il lessico è stato visto come una sorta di ontologia che struttura logicamente una certa visione del mondo. Ma questa similitudine in realtà resiste solo a livello superficiale. Per descrivere, con un esempio, la sostanziale differenza fra i due concetti basti pensare alla relazione di sinonimia, che è molto importante a livello lessicale, ma non trova posto nelle ontologie formali dove due termini sinonimi vengono raggruppati sotto lo stesso concetto; non è infatti permesso che due termini siano in relazione con lo stesso significato.

E' per questo importante distinguere fra le ontologie formali e le ontologie linguistiche, dove l'ontologia formale rappresenta strutture espresse in logica formale e ben formate, mentre l'ontologia linguistica è una struttura concettuale realizzata e convenzionalizzata linguisticamente.

[1] Altri approcci (come Guarino (1998)) portano a comprendere nel termine ontologia, anche la lessicalizzazione della stessa e definiscono quindi l'ontologia come language dependent.

2.1.2 Concettualizzazione

La concettualizzazione è il processo che comprende l'estrazione e l'astrazione di informazione rilevante da un dominio di esperienza.

Con questo termine, quindi, si intende un momento non solo indipendente da una specifica situazione, ma indipendente anche da uno specifico linguaggio di rappresentazione, per il semplice fatto che non è ancora una rappresentazione, bensì un prodotto mentale che mostra il punto di vista del mondo adottato da un agente. Quando più agenti (umani o artificiali) condividono questo stesso prodotto mentale, allora è necessaria una formalizzazione attraverso un'ontologia che rappresenti la concettualizzazione con un determinato linguaggio di rappresentazione.

L'unica strada percorribile per arrivare ai concetti, per descrivere i concetti, è infatti il linguaggio. Fra agenti umani il linguaggio usato per "parlare" di concetti sarà il linguaggio naturale, ma se il rapporto di interazione è un rapporto ibrido fra agenti umani e artificiali, allora l'unico linguaggio utilizzabile sarà proprio il linguaggio formale con cui sono realizzate le ontologie.

L'ontologia diventa, quindi, un ponte irrinunciabile per la comunicazione e la "condivisione di conoscenza" fra uomo e macchina e fra macchina e macchina.

2.1.3 Specificazione

La specificazione è definita come la concettualizzazione espressa in un linguaggio di rappresentazione e la natura di tale linguaggio è un importante criterio di distinzione fra le ontologie formali, informali o semi-formali. Le ontologie formali, infatti, sono, come anticipato, espresse in linguaggi formali, mentre le informali vengono descritte da linguaggi

come quello naturale che possono presentare anche situazioni di sinonimia e ambiguità;

Diverso, inoltre, può essere il "mondo" coperto dall'ontologia, come dimostra il fatto che, sia nella ricerca che nella pratica, si distinguono tre diversi tipi di ontologie:

1. *top level, (upper level)*
2. *core level*
3. *domain level*

Le prime (top level) costituiscono il livello più alto, che descrive categorie generali del mondo, indipendentemente da domini specifici, mentre le ultime sono proprio le strutture logiche dei domini specifici. La distanza fra questi due livelli viene ridotta dalla presenza di una struttura intermedia, che rappresenta un ponte tra esse, rappresentando quelle relazioni che possono essere condivise da domini diversi.

L'introduzione delle ontologie *top-level* consente di affrontare uno dei principali problemi di questo tipo di sistemi di rappresentazione della conoscenza, l'integrazione fra domini diversi, permettendo di organizzare i concetti in una tassonomia generale che viene poi specificata in modo indipendente in ogni dominio (N.Calzolari *e altri* (2008)).

Lo scopo di tale struttura gerarchica sarà quello di rendere più ontologie accessibili attraverso la stessa ontologia top-level, come in DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering), ontologia formale top level, realizzata al Laboratory for Applied Ontology (LOA) (Gangemi *e altri* (2002a)) .

Un'ontologia che mira a organizzare concetti generici che esprimono categoria generali in cui è organizzato il mondo è, per esempio, Cyc (enCYClopedia), un progetto che nasce nel 1984, e ancora in corso, che include oltre un milione di concetti[2]

[2]Dati tratti dal sito: http://www.opencyc.org

Un'ontologia di dominio invece, pur non essendo ancora una lessicalizzazione, formalizza concetti di un dominio specifico come nell'ontologia di un sistema bibliotecario mostrato in figura 2.1.[3].

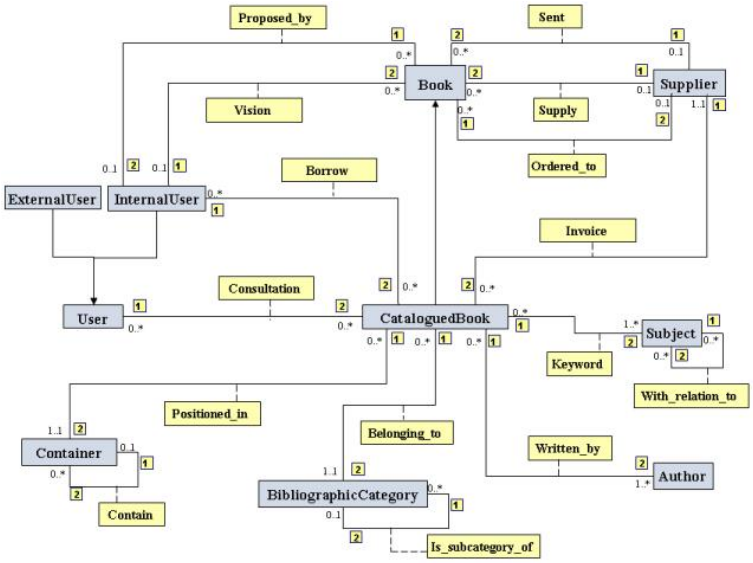

Figura 2.1: Biblioteca

2.1.4 Componenti dell'ontologia

Elementi fondamentali di un'ontologia sono concetti (nodi o classi) e le relazioni che intercorrono tra questi.

I concetti sono definiti come classi di individui, riprendendo l'esempio in figura 2.1, è possibile distinguere i concetti di BOOK, CATALOGUEDBOOK, AUTHOR etc. Le proprietà, invece, rappresentano le relazioni esistenti fra le varie classi, e, facendo ancora

[3]Dati tratti dal sito http://www.dis.uniroma1.it/~quonto/

una volta riferimento all'ontologia del dominio della biblioteca in figura 2.1, emergono fra le relazioni: *written_ by, beloging_ to, etc.*.

Gli individui, o istanze di una classe, consistono nei singoli oggetti che si riferiscono ad una classe, (ad esempio: l'autore *Verga*, il libro "i malavoglia", il etc.).

In una prospettiva formale è importante che ogni relazione sia ben definita, specifichi le classi a cui si riferisce e le proprie caratteristiche.

Si può quindi affermare che: "a CATALOGUEDBOOK is written_by an AUTHOR", e che la relazione "written_by" ha una cardinalità molti a molti (un libro può essere scritto da più autori e un autore può aver scritto più libri).

Infine, un'ontologia è definita anche da un insieme di assiomi: proposizioni sempre vere, usate per verificare correttezza di informazione e dedurre nuove informazioni. Un esempio di un assioma può essere l'asserzione "*Every external user is a user*".

Riassumendo gli elementi che compongono un'ontologia sono concetti, relazioni, assiomi e istanze: ma come sono rappresentati questi componenti?

2.2 Logiche descrittive

2.2.1 Linguaggi di rappresentazione di ontologie: le *description logics*

Fra i linguaggi che meglio si prestano alla definizione delle ontologie vi sono le *description logics* (DLs): una famiglia di formalismi di rappresentazione della conoscenza che possono essere usate per descrivere un dominio in modo strutturato e formale.

Con i termini *description* e *logic* si sottolinea quello che è da una parte il potere de-

scrittivo del linguaggio, che modella un dominio a partire dalla definizione di concetti, dall'altra il fatto che, a differenza di altri formalismi come frame e Reti Semantiche, le DLs usano una semantica basata sulla logica di primo ordine.

Uno dei primi approcci alla rappresentazione strutturata della conoscenza sono state le Reti Semantiche, generalmente attribuite a Quillian (Quillian (1967)). Tali reti propongono un insieme di nodi e archi dove i nodi possono rappresentare sia concetti, sia oggetti concreti, così come gli archi possono rappresentare sia relazioni fra concetti che relazioni fra oggetti del mondo. Questa imprecisione nella strutturazione semantica rende complesso un processo di ragionamento. Negli anni '70 sono stati introdotti i Frame, proposti da Minsky nel 1975: un frame rappresenta una classe (o concetto) a cui possono essere associati attributi che rappresentano le proprietà delle istanze. Anche in questo caso però il frame può rappresentare sia un concetto che un'istanza e questo fa emergere problemi simili alle Reti Semantiche quando si vuole affrontare un processo di ragionamento.

Come evidenzia Baader (Baader *e altri* (2003)), il processo di ragionamento, cioè il processo che permette di inferire informazione, è molto importante, tanto che un linguaggio, che ha l'ambizione di modellare un'ontologia, deve saper trovare il giusto equilibrio tra il potere espressivo e la complessità di tale meccanismo inferenziale (o *ragionamento automatico*). Il grande vantaggio delle ontologie è infatti quello di permettere non solo di rappresentare l'informazione, ma anche di dedurre informazione, grazie a processi inferenziali che sfruttano l'insieme di assiomi, cioè enunciati considerati veri, che che si é visto essere una componente fondamentale dell'ontologia. Data un'ontologia O, con un insieme di regole di inferenza e un insieme di assiomi, il processo di ragionamento permetterà di inferire se un fatto α , sia vero o meno.

Il linguaggio logico che descrive l'ontologia deve quindi agevolare tale processo inferenziale, e trovare il giusto compromesso fra espressività e complessità: più espressivo sarà il linguaggio più difficile sarà avere un algoritmo di ragionamento efficiente.

Baader, (Baader *e altri* (2003)), sottolinea l'importanza del *ragionamento*, come un processo che ritorna in molte fasi dello sviluppo di un'ontologia: in fase di modellazione, per esempio, può essere usato per verificare se i concetti sono in contraddizione tra loro, e successivamente per integrare ontologie diverse. Ma "ragionare" significa anche valutare che gli assiomi non siano in contraddizione, classificare automaticamente istanze in classi attraverso principi di sussuzione, o anche cercare di rispondere a queries complesse.

Le logiche descrittive sembrano rispondere bene a queste esigenze, (Calvanese *e altri* (2005)) perché permettono di mantenere i costi del processo di ragionamento a livelli accettabili, tanto che da quando sono nate per la prima volta negli anni ottanta, hanno concentrato su di sé gli sforzi di molti gruppi di ricerca.

Sintassi

Dal punto di vista sintattico le *description logics* presentano insiemi di concetti elementari, detti concetti atomici o concetti nome, e ruoli atomici o ruoli nome. I concetti atomici sono predicati unari, mentre i ruoli atomici sono relazioni binarie : in pratica ogni DL concepisce i concetti come concetti atomici che denotano un insieme di individui che appartengono ad esso, e le relazioni come relazioni atomiche, che denotano legami di dipendenza fra concetti.

La sintassi *DL* è riportata nella **Tabella 2.2**.

Semantica

Dal punto di vista semantico, un concetto atomico può essere rappresentato come un insieme di individui, e i ruoli diventano relazioni binarie sul dominio, quindi data la re-

Sintassi		Esempio
C,D		
A	Atomic concept	HUMAN
$C \cap D$	Congiunzione	$HUMAN \cap MALE$
$C \cup D$	Disgiunzione	$NICE \cup MALE$
$\neg C$	Negazione	$\neg MEAT$
$\exists R.C$	Quantificazione esistenziale	$\exists HasChild.Blond$
$\exists R$	Quantificatore esistenziale non qualificato	$\exists HasChild$
$\forall R.C$	Quantificazione universale	$\forall HaChild.Human$

Tabella 2.2: Sintassi DL-Lite

lazione R si ottiene l'insieme degli individui che soddisfano R in quel dominio. Così come dal punto di vista sintattico, anche semanticamente i concetti non atomici vengono definiti sulla base di regole ricorsive. Ogni linguaggio è caratterizzato da un insieme di costruttori che permettono di formare espressioni di concetti e ruoli complessi, e tali costruttori includono operatori che corrispondono ai connettivi proposizionali. Per esempio, la congiunzione viene interpretata come intersezione insiemistica, e la negazione come complemento del dominio di interpretazione. Inoltre è presente la *quantificazione universale* su un ruolo, data dall'espressione $\forall R.C$, dove R è un ruolo e C un concetto, che permette di specificare le proprietà che devono valere per tutti gli oggetti connessi da un ruolo,

La quantificazione esistenziale, $\exists R$, corrisponde in FOL alla formula con la variabile libera x, $\exists R(x, y)$, vale a dire, indica tutte le x tali che esiste un y e "x R y". La quantificazione esistenziale qualificata, $\exists R.C$ corrisponde alla formula in FOL , $\exists R(x, y) \wedge$

$C(y)$, cioè la formula di prima ma con una qualifica della variabile y. Linguisticamente questo significa che, nel primo caso, qualificazione esistenziale non qualificata, non viene qualificato il secondo argomento del verbo che esprime la relazione, come in (1a), mentre nel secondo caso tale argomento è qualificato, (es. (1b)): es:

(1a) Maria mangia (qualcosa). (∃MANGIA)

(1b) Maria mangia la mela. (∃MANGIA.mela)

Ne deriva che, data un'interpretazione I , si avrà:

1. $\neg C$ interpretato come l'insieme di tutti gli individui che non appartengono all'interpretazione di C Δ^I / C^I .

2. il restrittore universale: $(\forall R.C)^I$: per ogni x, tale che x appartiene a Δ^I e soddisfa la relazione R, allora x è istanza della classe C^I

3. il restrittore esistenziale $(\exists R.C)^I$: esiste almeno un'istanza x di Δ^I , tale che x soddisfa la relazione R, e è istanza della classe C^I .

   Secondo quanto detto, se si ha P , insieme delle persone, e F insieme delle donne, allora l'insieme di tutte le persone che non sono donne può essere espresso dalla classe:

$$P \cap \neg F$$

Assumendo di voler definire il concetto:

A man that is married to a doctor and has at least five children, all of them are professors.[4]

può essere modellato dalla seguente descrizione di concetto:

$Human \cap \neg Female \cap \exists married.Doctor \cap (5 > hasChild) \cap \forall hasChild.Professor$

Un individuo che interpreta questa descrizione dovrà appartenere all'intersezione del set di Human, e quello di Non-Female, e contemporaneamente all'insieme interpretato dalla restrizione esistenziale definita da married.Doctor. Inoltre l'individuo dovrà appartenere all'insieme che soddisfa la restrizione numerica sul numero dei figli e quindi all'insieme di coloro i cui figli sono tutti professori.

Le DLs presentano, inoltre, due componenti definite **Tbox** , *terminological box* e **Abox**, *assertional box,* che rappresentano *come basi di conoscenza* terminologiche. In particolare, la A-Box, è la base di conoscenza terminologica a livello estensionale, e contiene asserzioni circa le istanze (o oggetti) del dominio, la loro appartenenza a classi, e i legami con altre istanze attraverso ruoli. La T-Box, invece, è la base di conoscenza terminologica a livello intensionale, che contiene le asserzioni di carattere generale sulle classi usate nella classe estensionale, ne stabilisce le proprietà e i mutui legami.[5]. Ne deriva che asserzioni come (2a) apparterranno alla Tbox, mentre espressioni come la (2b) andranno a costituire la A-Box

(2a) Every student is a person,

(2b) Mary is a student,

In realtà questi due tipi di asserzioni non sono trattate diversamente nella logica del primo ordine (che sottende buona parte delle DLs), ma la distinzione è interessante

[4] Baader *e altri* (2003), p. 4.

[5] (Calvanese (1996), p 4)

dal punto di vista formale. Prima di tutto è utile distinguere le espressioni del livello intensionale da quelle del livello estensionale in fase di progettazione, essendo importante mantenere la suddivisione logica fra gli assiomi sul mondo (la T-Box, livello intensionale) e la particolare manifestazione degli stessi (le asserzioni estensionali della A-Box). Ma tale distinzione viene conservata anche per agevolare il processo inferenziale, che dovrà affrontare problemi diversi nelle due componenti: i problemi di classificazione della T-Box non riguardano la A-Box, mentre problemi di "verifica dell'esistenza dell'istanza" della A-Box non riguardano la T-Box.

Come anticipato il processo inferenziale è ciò che permette di estrarre dalla conoscenza esplicita della base di dati, conoscenza implicita che viene appunto dedotta attraverso regole di inferenza. Fra queste regole alla base del ragionamento automatico si hanno la sussuzione, la classificazione, la verifica di consistenza o di equivalenza fra classi. Alcuni esempi di regole di inferenza che costituiscono il processo di ragionamento automatico (*reasoning*), sono descritte di seguito:

1. Verifica di coerenza.: se x è istanza della classe A e B, ma A e B sono disgiunte, allora l'algoritmo deve restituire un errore.

2. Classificazione: alcune coppie attributo/valore sono condizioni necessarie per essere istanza della classe A. Se l'istanza x soddisfa queste condizioni, possiamo concludere che x è istanza di A.

3. Appartenenza ad una classe per transitività: se x è istanza della classe A, e la classe A è una sottoclasse della classe B, allora x è istanza di B.

4. Equivalenza di classi: se la classe A è equivalente alla classe B e la classe B è equivalente alla classe C, allora A è equivalente a C.

Tutte queste regole di inferenza devono necessariamente mantenere una complessità bassa per fare in modo che il processo inferenziale non perda troppo in termini di efficienza, ma, contemporaneamente, ridurre la complessità computazionale significa ridurre lo spettro di operatori che il sistema è capace di gestire, quindi, ridurre l'espressività del linguaggio. Minore è l'espressività del linguaggio logico, più ristretto è il frammento di linguaggio naturale da esso rappresentato.

Ne deriva che ci sarà sempre uno scarto tra i due formalismi: l'espressività del linguaggio naturale, come libera espressione dell'utente, e l'espressività del frammento che corrisponde al linguaggio logico semplificato per motivi computazionali.

2.2.2 DL-Lite un frammento di *description logic*

Ridurre la complessità del linguaggio di rappresentazione della conoscenza consiste nel creare frammenti di tale linguaggio, capaci di coprire solo parte degli operatori logici gestiti dalla logica di primo ordine o da una *description logic* completa. Un interessante compromesso fra espressività e complessità è emerso con l'introduzione di DL-Lite (Calvanese *e altri* (2005)), un frammento di logica descrittiva studiato per ridurre i costi computazionali, nel tentativo di continuare a coprire i principali operatori logici, o meglio, quegli operatori logici che sono particolarmente utilizzati dagli utenti nell'interrogazione a basi di conoscenza.[6]

Precisamente DL-Lite rappresenta una famiglia di frammenti di logica descrittiva, più

[6]Nel Capitolo 4 si illustrerà un'analisi di corpora che mira proprio a valutare linguisticamente l'espressività del frammento di inglese espresso da DL-Lite

o meno espressivi, che trova il fulcro essenziale in *DL-Lite*$_{core}$ la versione che contiene il nucleo di espressività condiviso da tutte le altre. Lo scopo per cui è stato pensato DL-Lite, che nasce in un contesto di estrazione di conoscenza da basi di conoscenza, è quello di mantenere una complessità di ragionamento polinomiale rispetto al numero degli elementi della base di conoscenza (Calvanese *e altri* (2006)). Ragionare in questo contesto, infatti, non significa soltanto applicare regole di sussuzione, classificazione o verifica della consistenza, ma anche rispondere a queries su istanze immagazzinate in memorie secondarie.

Nel processo di interrogazione ad ontologia, come descritto nel primo Capitolo, è necessario che il linguaggio logico si relazioni con l'input dell'utente, ma allora vale la pena indagare fino a quanto è possibile semplificare la logica senza compromettere la comunicazione con l'utente, e quale sarà il punto di equilibrio che massimizza l'espressività pur mantenendo limitati i costi computazionali; e ancora, se il frammento espresso da DL-Lite sia sufficientemente ricco da permettere che utenti non esperti ricerchino e inseriscano informazione con interrogazioni libere. Per rispondere a queste domande è prima di tutto necessario descrivere qual è il frammento di inglese espresso da DL-Lite, per poi focalizzarsi sui costrutti che non rientrano nella sua sintassi, facendo però attenzione alla distinzione introdotta dalla terza domanda che ci si è posti. Sono infatti da tenere in considerazione due diversi momenti di interazione con l'ontologia: l'interrogazione e la costruzione, perché coinvolgono diversi frammenti di logica descrittiva :

1. da una parte si ha il linguaggio con cui è descritta l'ontologia, quello che è necessario “parlare” per inserire informazione nell'ontologia stessa. In questo caso è il linguaggio logico: DL-Lite.

2. Dall'altra si ha il linguaggio con cui interrogare l'ontologia, che subisce restrizioni ulteriori di espressività dovute alla necessità di dover tradurre quella che è la concettualizzazione ontologica in un database concreto che viene interrogato in SQL. Nel caso specifico questa "lingua" è rappresentata dalle cosiddette conjunctive queries, unione di selezione-proiezione e join SQL.

Questi due linguaggi comportano problemi diversi nel momento in cui li si vuole confrontare con il linguaggio naturale, e nei paragrafi che seguono si andrà ad esaminare la loro espressività nel dettaglio, dal linguaggio di descrizione a quello di interrogazione.

2.2.3 DL-Lite e la descrizione dell'ontologia

Essendo un frammento di logica descrittiva anche DL-Lite presenta due basi di conoscenza terminologiche: una T-Box che raccoglie le asserzioni universali, e una A-Box dove sono memorizzati gli enunciati che descrivono i fatti riguardanti le istanze. [7]

DL-Lite ha il vantaggio di riuscire a gestire grandi quantità di informazioni con costi relativamente contenuti: un costo polinomiale alla grandezza della T-box e logaritmico della grandezza della A-box, ma, allo stesso tempo, è necessario analizzare se il frammento di inglese, che questo linguaggio riesce a coprire, può rappresentare un limite per l'utente.

La base di conoscenza terminologica intensionale è costituita da un set di inclusioni del tipo:

$$Cl \subseteq Cr$$

dove Cl indica il concetto che compare nel contesto sinistro e Cr il concetto che compare in contesto destro. Tale relazione corrisponde alla formula FOL , *first order logic:*

[7] Si rimanda la paragrafo 2.2.1

$$\forall x.Cl(x) \rightarrow Cr(x)$$

notazione che mette in evidenza la natura universale di ogni assioma rappresentato. Il livello intensionale, infatti, può essere immaginato, informalmente come una sorta di "insieme di asserzioni che definiscono il dominio, e di norme che lo regolano".

Le strutture ammesse variano molto a seconda delle versioni di DL utilizzate, ma con riferimento alla versione base, *DL-Lite*$_{core}$ si possono così riassumere gli operatori accettati nei due contesti (Calvanese *e altri* (2006)):

Cl : A | ∃.R

Cr: A| ¬A | ∃.R | ¬ ∃.R

dove A indica un concetto atomico e R un ruolo atomico definito come *qualificatore esistenziale non qualificato,* cioè una relazione fra due concetti in cui l'argomento non è specificato.

Dal punto di vista linguistico i due contesti accettano solo un ristretto numero di costrutti, che possono andare a formare le seguenti strutture sintattiche:

a. **[EVERY NOUN]**VERB PHRASE

b. **[[EVERYONE [WHO VERB PHRASE]]** VERB PHRASE]

Il pronome "everyone" e l'aggettivo "every" sono parte integrante e fondamentale della sintassi della frase, perché, come si nota appunto dalla formula FOL, ogni asserzione esprime un concetto universale.

Il contesto sinistro (in grassetto in *a* e *b*) può contenere solo concetti atomici o relazioni con qualificatore esistenziale non qualificato. Questo, linguisticamente, si realizza con:

concetto atomico: *every + nome*

ruolo atomico: *everyone + who verb phrase.*

Se nel contesto sinistro è specificata una relazione, cioè la relativa che specifica l'universale "everyone", non è permesso inserire un sintagma nominale.

Ne deriva che in *DL-Lite*$_{core}$ sarà grammaticalmente accettato l'assioma in *(2a),* ma non quello in *(2b)*

(2a) everyone who eats left

(2b)every student who eats left.

Inoltre in contesto sinistro non è prevista la specificazione dell'oggetto del sintagma verbale perché il qualificatore esistenziale non deve mai essere specificato. In altri termini, si potrà avere

(2c) everyone who eats something left

ma non

*(2d) * everyone who eats an apple left.*

Infine, sempre nel contesto sinistro, non è accettata la negazione, né del concetto atomico, come in *(2e),* né della proposizione relativa, come in *(2f).*

*(2e) *Every no student is a professor*

*(2f)*Everyone who is not a student is a professor.*

Il contesto destro invece prevede un sintagma verbale, che realizzi un quantificatore esistenziale non qualificato, cioè un verbo intransitivo e la sua negazione o un verbo transitivo (privo di complemento oggetto) e la sua negazione.

Ciò che non è ammesso in contesto destro è l'operatore di disgiunzione, per cui situazioni come la *(3a)* risulteranno difficilmente esprimibili:

(3a) in order to borrow a book you should ask the the library staff or use the Selfcheck.

Esiste peró una versione ancora poco costosa computazionalmente, ma linguisticamente

più espressiva di *DL-Lite*$_{core}$, che è rappresentata da *DL-Lite*$_R$. Questo frammento di logica descrittiva accetta i seguenti operatori:

$$Cl \rightarrow Cl1 \cap Cl2$$

$$Cr \rightarrow \exists R.A$$

Il contesto sinistro quindi può accettare l'operatore di congiunzione che, sul piano linguistico, si concretizza nella possibilità di specificare il "NOUN" del sintagma nominale anche in presenza di una relativa (5a), e un aggettivo che determini N (5b).

(5a) Every student who studies left

(5b) Every nice student left.

Nel contesto destro, invece, è possibile, in *DL-Lite*$_R$ specificare il quantificatore esistenziale, cioè determinare l'argomento del un verbo transitivo, come in *(5c)*

(5c) Every student knows a girl.

In ogni caso però, sia che si utilizzi *DL-Lite*$_{core}$ che *DL-Lite*$_R$, permangono strutture che non possono, in alcun modo, essere espresse

1. negazione in contesto sinistro
2. disgiunzione in contesto destro
3. specificazione del complemento oggetto nel sintagma verbale del contesto sinistro.

2.2.4 Estrazione dati dall'ontologia: limiti ed espressività delle Conjunctive queries

Spostandosi sul problema dell'estrazione di dati dall'ontologia é necessario far riferimento al *query language*, cioè a quello specifico linguaggio di interrogazione in cui vengono

espresse le queries, e che dovrá interfacciarsi con il linguaggio di interrogazione della base di dati. Questo infatti ha delle limitazioni diverse rispetto a DL-lite, limitazioni che sono fondamentali per capire quali strutture separano la lingua parlata dall'utente e quella "capita" dall'ontologia.

Dato un database espresso in DL-Lite, le queries vengono espresse con *unioni di conjunctive queries* (UCQs), cioè espressioni del tipo:

$$q(x) = x|\exists y1\, conj(x, y1) \wedge .... \wedge \exists yn\, conj(x, yn)$$

dove x è la variabile da cercare (potenzialmente insieme vuoto), e ogni y è un insieme di variabili su cui applicare le relazioni. Se n=1 si parla di conjunctive query (CQ), altrimenti di unione di conjunctive queries (UCQs). Come in un Datalog (Atzeni et al. 2002, pp. 81-83), tutto ciò che segue il simbolo " | " viene definito "corpo" della query, mentre ciò che precede è definito "testa".

Ad esempio l'espressione "*which are the red books?*" corrisponderà ad una semplice CQ, mentre *"Which are the red books read by John?"* è definita da un'unione di CQs, una query complessa che si formalizza come:

$$\{x \mid book(x) \wedge red(x)\}\ \{x \mid book(x) \wedge read(john, x)\}$$

Un'interrogazione interessante, ma che provoca una situazione ancora più complessa, può essere:*"which are the students who attend a course which is taught by their father?"* , che può essere rappresentata come:

$$x|\exists y1 \exists y2(student(x) \wedge attend(x, y1) \wedge course(y1) \wedge teach(y2, y1) \wedge father(y2, x))$$

Da quanto emerge è interessante notare come le CQs, che esprimono in termini SQL selezione, join e proiezione, non prevedono alcuni operatori logici molto importati: *la negazione, la quantificazione universale e la disgiunzione.*

E' da queste limitazioni espressive delle queries che si è avviato lo studio dei corpora e l'analisi del linguaggio naturale al fine di stabilire in che misura questo scarto potesse essere un problema.

Capitolo 3

I linguaggi controllati

3.1 Modificare l'input in linguaggio naturale

Immaginando la ricerca del compromesso come la costruzione di un ponte che permetta all'uomo e all'ontologia di comunicare, si potrebbe affermare che, se fino ad ora si è lavorato dalla parte dell'ontologia (sulla sponda dell'ontologia), a questo punto si propone l'analisi del problema dalla prospettiva opposta, dal punto di vista del linguaggio umano. Se le modifiche del linguaggio logico consistono nel sottrarre ed aggiungere operatori per "bilanciare il frammento", la prospettiva che si scorge guardando il problema dal lato linguaggio naturale, ha suggerito approcci diversi.

Sforzi per ridurre il trade-off fra il linguaggio naturale e il frammento che corrisponde a forme logiche semplificate (come DL-Lite) sono stati, e sono, rappresentati dall'introduzione di *controlled natural languages* (linguaggi naturali controllati).

L'obiettivo dei sistemi che sfruttano linguaggi controllati è veicolare l'input dell'utente in modo che sia sempre compatibile con il linguaggio logico. In altre parole l'utente non

viene lasciato libero di scrivere ciò che vuole nella forma che vuole, ma viene costretto ad usare soltanto alcune forme sintattiche, alcune parole, alcuni costrutti logici.

La restrizione della libertà espressiva dell'utente può essere in effetti una soluzione per avvicinare il linguaggio dell'uomo e dell'ontologia, ma il prezzo da pagare in termini di naturalezza è piuttosto alto. Non essendo possibile, in situazioni di efficienza ottimali, coprire logicamente tutte le strutture del linguaggio naturale, si chiede all'utente di formulare la query in modo "comprensibile" per l'ontologia, un po' come se, in assenza del "ponte", si chiedesse all'utente di "guadare il fiume". Spesso però questo processo è supportato da interfacce intelligenti che sono in grado di supportare l'utente nella formulazione di una query precisa e adatta alle esigenze del sistema (Dongilli *e altri* (2004)).

I linguaggi controllati costituiscono, quindi, una limitazione per l'uomo: una limitazione che assicura la perfetta compatibilità fra l'input e il linguaggio logico dell'ontologia, a discapito della naturalezza dell'interazione. Allo stesso tempo, questo studio mira a capire se tale limitazione sia effettivamente necessaria, o se, al contrario, possa essere evitata qualora si riesca a creare un linguaggio logico che sia fondamentalmente compatibile con un frammento di linguaggio naturale. In altre parole, tale limitazione potrebbe forse essere evitata se il linguaggio logico fosse costruito con uno sguardo al linguaggio naturale.

Se quindi lo sforzo principale deve essere quello di un'attenta costruzione del linguaggio logico, affinché non sia necessario un controllo dell'input, dall'altra parte è anche utile per alcuni operatori e per alcuni costrutti particolari (vedremo il caso delle domande indirette), pensare ad una effettiva semplificazione dell'input, purché non sia una limitazione per l'utente. Il proposito è quello di una semplificazione a posteriori, cioè di un sistema che non limiti la libertà di espressione dell'utente in partenza, ma modifichi il suo input

in modo automatico.

In questo ambito rientrano le tre proposte di semplificazione testuale che verranno analizzate successivamente: una forma di indebolimento semantico per gestire il quantificatore universale in fase di interrogazione, e due forme di semplificazione sintattica nel caso degli atti indiretti e dell'inserimento automatico di informazione nel testo.

3.2 Linguaggi naturali controllati: definizione

Il linguaggio controllato è un sottolinguaggio che copre solo parte del linguaggio naturale e per questo si presta a giocare il ruolo di trait d'union fra i limiti di un sistema di rappresentazione di conoscenza e l'input dell'utente.

Punto di forza, al di là delle limitatezze espressive, è il fatto di basarsi su un set di regole ridotto e su un lessico ridotto, ma nello stesso tempo di essere perfettamente formalizzato e definito secondo costrutti adeguati allo scopo per cui viene creato. A prima vista un linguaggio controllato può richiamare alla memoria un sottocodice linguistico (Berruto (1998), pp. 154-157). Si definisce sottocodice, infatti, una varietà diafasica di una lingua che utilizza lessici speciali al fine di evitare ambiguità in particolari contesti (per esempio il sottocodice della tecnica meccanica automobilistica, dove "candela" è un particolare elemento del motore e non la "candela di cera"). Ma è necessario non confondere quello che è il lessico ridotto utilizzato da un linguaggio controllato, e il lessico speciale di un sottocodice, cioè un lessico che non necessariamente semplifica il vocabolario, ma al contrario, adotta significanti nuovi per significati nuovi, o assegna nuovi significati a termini già esistenti. Quindi sebbene in entrambi i casi si miri ad evitare ambiguità lessicale, solo nel sottocodice si ha una effettiva modifica del vocabolario utilizzato, mentre nel

linguaggio controllato si mantiene una generale corrispondenza con la lingua standard, andando solo a ridurre il vocabolario e non a modificarlo. Inoltre il linguaggio controllato prevede una forte semplificazione sul piano sintattico, accettando solo alcuni costrutti e periodi piuttosto semplici, mentre, i sottocodici non semplificano le strutture sintattiche. Ancora diverso è il caso della classe di variazione linguista che nasce spontaneamente in contesti sociali particolari e viene comunemente definita *gergo*. In questo caso la variazione linguistica non mira a raggiungere una formalizzazione o a ridurre la complessità, ma soltanto a veicolare un sentimento di appartenenza ad un gruppo sociale. Se quindi i sottocodici condividono con i linguaggi controllati l'obiettivo di una riduzione di ambiguità, le varianti gergali invece non hanno neanche questo obiettivo, ma nascono semplicemente come fenomeno prettamente sociolinguistico.

3.3 Le prime applicazioni

Le prime applicazioni che hanno spostato l'attenzione sui linguaggi controllati sono sistemi di *authoring,* cioè applicazioni in grado di controllare lo stile, la sintassi, il lessico di un testo scritto. L'obiettivo di questi sistemi consiste nell'agevolare il lavoro di un autore, e di veicolare un linguaggio che sia il più standard possibile, tale da favorire il recupero e la traduzione dell'informazione. Solitamente i sistemi di semplificazione testuale con questo scopo prendono in input il testo libero e lo ristrutturano proponendo in output una o più alternative.

Ma in realtà i linguaggi controllati possono oggi essere considerati un anello di congiunzione fra il linguaggio naturale e i linguaggi logici, perché esprimono, in linguaggio naturale, espressioni formali, strutturate e non ambigue. Ed è sotto quest'ottica che si

sono sviluppati gli studi più recenti.

3.4 Linguaggi controllati e ontologie: lo stato dell'arte

Negli ultimi anni molte ricerche sono partite dal presupposto che i linguaggi controllati siano una buona strada da percorrere verso l'interazione naturale con ontologie, un'interazione che, in questo modo, viene assicurata anche a utenti non esperti di linguaggi logici.

Perché questo avvenga, il linguaggio logico dell'ontologia deve essere traducibile in un frammento non ambiguo di inglese, come può essere un linguaggio controllato che per definizione è ben formato e non ambiguo. Il linguaggio naturale presenta, infatti varie forme di ambiguità che vanno dall'ambiguità lessicale a quella sintattica, o semantica e rendono il trattamento automatico del linguaggio più complesso. Prima di tutto molto frequenti sono i casi di ambiguità lessicale, ma anche situazioni di ambiguità sintattica o semantica possono causare problemi (in Tabella 3.1 si riportano esempi di diversi tipi di ambiguità, tratti da (Jurafsky e Martin (2000) pp.4-5)). L'ambiguità lessicale consiste, quindi, in situazioni in cui un termine può aver associati più significati (*her* pronome possessivo o dativo, *porta,* terza persona singolare del verbo portare o sostantivo singolare femminile). L'ambiguità sintattica invece si ha in situazioni in cui sono possibili più alberi sintagmatici grammaticale, sullo stesso enunciato, come in "l'uomo nel parco con il cannocchiale" ([l'uomo [nel parco] [con il cannocchiale]], [l'uomo [nel parco [con il cannocchiale]]]).

Sia l'ambiguità lessicale che quella sintattica, però possono essere gestite, almeno in parte, usando una grammatica controllata.

Ambiguità	Esempio: I made her duck	
Ambiguità lessicale	I cooked a duck for her I cooked a duck belonging to her	Her può essere pronome possessivo o pronome dativo
Ambiguità sintattica	I made [her duck] I [made her] [duck]	1)Interpretazione transitiva di to make: - Ho fatto la sua anatra(ho cucinato la sua). 2)Interpretazione dintransitiva di to make Ho fatto lei anatra (l'ho trasformata in anatra).
Ambiguità semantica	I cooked her duck I created a duck for her	To make: 1) fare, cucinare 2) creare, rendere

Tabella 3.1: Tipi di ambiguitá

Il problema dell'ambiguità lessicale, infatti, viene meno grazie all'uso di un vocabolario controllato che restringe il lessico ammesso, e i significati per ogni termine, rendendo il testo più chiaro e leggibile (Mitamura e Nyberg (2001)). D'altra parte le regole della grammatica mirano ad evitare la costruzione di periodi con strutture sintattiche equivoche, imponendo la formazione di frasi brevi, con sintagmi nominali non troppo complessi, e evitando costruzioni ellittiche del soggetto delle congiunzioni.

Sfruttando quindi agenti computazionali di riscrittura e semplificazione del testo è possibile ottenere un output che si presti ad interagire con il linguaggio logico.

Uno dei primi progetti in ordine di tempo che coinvolge i CNL (acronimo per la forma inglese Controlled Natural Languages), risale al 1999, e va sotto il nome di ACE (Attempto Controlled English).

ACE è un frammento di inglese controllato sviluppato, presso l'Università di Zurigo, nel progetto *Attempto* e modellato per agevolare l'interazione con sistemi di rappresentazione della conoscenza da Fuchs, Schwertel e Schwitter (Fuchs *e altri* (1999)).[1]

Anche se ad una rapida lettura dell'output può sembrare un linguaggio naturale, in realtà ACE è un linguaggio formale limitato ad uno stretto numero di costrutti, che esprimono concetti semplici senza ambiguità sintattica o lessicale. Si prevedono un vocabolario di parole "funzionali" e parole "contenuto", importate a seconda del dominio specificato, insieme ad una grammatica con cui si definiscono le regole di costruzione. Ad una panoramica sulle principali restrizioni previste da ACE si nota che non sono previsti verbi in persone diverse dalla terza singolare, i modali devono essere riscritti in parafrasi del tipo "it is possible that" o "it is necessary that", e la semplice struttura della frase ACE prevede: *NP - verbo - complemento* come in (1a) e (1b):

[1]Ricerche attuali di Kuhn, Kaljurand e Fuchs. Cfr. sito web: http://attempto.if.unizh/site

(1a) a customer wants the bill.

(1b) a customer give a card to a clerk.

E' possibile ampliare l'espressività con costrutti del tipo: *there are, there is*, o ancora coordinate, subordinate, o negazioni.[2]

Derivato di ACE è PENG acronimo di Processable ENGlish creato dallo stesso Schwitter e dal suo gruppo di ricerca al *Centre for Language Technology* della Macquarie University (Schwitter e Tilbrook (2006b)).

PENG è un linguaggio controllato dell'inglese standard, definito da una grammatica controllata che crea frasi con struttura sintattica non ambigua (deterministico). Anche in questo caso si avrà un vocabolario costituito da un set di parole predefinite (determinatori, coordinatori, pronomi, preposizioni) e un set di parole definite dal contesto. La struttura base di un'enunciato semplice è:

sentence: soggetto + predicato;

soggetto: determinatore [modificatore prenominale] testa nominale [modificatore post nominale]

predicato: [negazione] testa verbale complemento [altro modificatori].

Ne deriva che i seguenti esempi (2a,2b,2c) contengono enunciati grammaticali in linguaggio PENG:

(2a) The butler works

(2b) Every butler hates a person

(2c) Butlers are murderers

[2]E' interessante notare come la negazione venga indebolita attraverso la teoria del fallimento di probabilità. Per cui l'enunciato *John is not a costumer* verrà riscritto in *It is not provable that John is a costumer.*

3.5 Semplificare l'input o ampliare il frammenti di logica: un duplice approccio

Sistemi di semplificazione dell'input dell'utente, come quelli usati nell'ambito dei CNL, possono essere un nodo focale per l'interazione con le ontologie. Ma d'altra parte, l'idea che si vuole proporre è quella di limitare i processi di semplificazione solo a quei casi che non sono effettivamente gestibili dalla logica utilizzata, ma che non sembrano essere tanto ricorrenti da giustificare la modifica di tale logica. Il tentativo è quello di creare una logica che si adatti il più possibile al linguaggio umano, e usare la semplificazione dell'input, intesa come parafrasi dell'input, come rimedio estremo qualora vi siano discrepanze non colmabili. Nel prossimo capitolo si analizzerà in che misura i due linguaggi oggetto di studio (linguaggio di interrogazione e quello di descrizione) condividano le forme logiche effettivamente usate dagli utenti.

Per le strutture non condivise si proporranno due diversi approcci in base alla loro frequenza: i casi meno frequenti possono essere affrontati con il metodo della parafrasi testuale, mentre i casi piú frequenti suggerirebbero una modifica del linguaggio logico.

Capitolo 4

Analisi dei corpora

4.1 Introduzione all'analisi

Dopo aver introdotto sia il versante del linguaggio logico e delle sue restrizioni per motivi computazionali, sia il versante del linguaggio naturale e le sue possibilità di semplificazione, vale la pena adesso soffermarsi sulle analisi delle strutture linguistiche.

A tale scopo si studierà da un lato, il rapporto fra le domande tipicamente usate dall'utente nell'interrogare una base di conoscenza e l'espressività delle CQs, dall'altro il rapporto fra le asserzioni tipiche di testi in linguaggio naturale, in particolare testi normativi, e l'espressività di DL-Lite. Nel primo caso l'attenzione sarà rivolta all'interrogazione dell'ontologia, mentre, nel secondo caso, alla costruzione dell'ontologia attraverso la formalizzazione di una concettualizzazione.

L'analisi quindi si dividerà in due fasi: [4.2] analisi delle interrogative e [4.3] analisi delle affermative.

4.2 Analisi delle interrogative

Come pone le domande un utente seduto di fronte ad un computer? Come si rivolge ad un database? E in tutti questi casi: quali operatori logici tende ad utilizzare?

E' a questi interrogativi che l'analisi di corpora inglesi, lingua principale di studio, cerca di rispondere.

Escludendo le domande "how" e "why", le quali esulano dal campo di indagine riguardante le CQs, è interessante invece studiare le caratteristiche delle domande booleane e delle domande "wh"[1] costruite con "which", "who", "what", "when", "where". Tali interrogative infatti possono essere espresse da CQs, solo a condizione che non contengano operatori logici non definiti nel frammento di inglese rappresentato.

Affinché si ottenga una collezione che rappresenti uno spettro abbastanza ampio di strutture, le domande sono tratte non solo da domini diversi, ma anche da situazioni e contesti diversi. Si osserva infatti che il contesto in cui si situa la frase, così come la situazione psicologica in cui si trova il parlante influenza molto la scelta degli enunciati e delle strutture adottate; in altre parole, la consapevolezza di rivolgersi ad un database, o la consapevolezza di rivolgersi ad un altro essere umano, provocano sensibili variazioni nel registro linguistico dell'input.

Per questo motivo sono stati costruiti tre corpora:

1. Clinical questions: 435 domande, dalla struttura piuttosto lunga, contenenti richie-

[1] Si definiscono *wh,* o domande *x,* le domande introdotte da un sintagma interrogativo, comprendente un aggettivo, un pronome o un avverbio della serie interrogativa. Si definiscono alternative o, yes/no, o booleane, le domande che necessitano di una risposta alternativa (si-no).(L.Renzi *e altri* (2001), p. 676, vol. II).

ste di informazioni di medici a colleghi. Il vocabolario del corpus conta 3495 types, su un numero totale di tokens di 40489.

2. Answer.com: 444 domande tratte dal sito di Answer.com. Interrogazioni fatte da utenti internet su tematiche diverse. Il vocabolario del corpus conta 1639 types su un totale di 5971 tokens.

3. TREC: dati tratti dalle collezioni per la ricerca sul "Question Answering" di TREC 2004.

4.2.1 Struttura dei corpora

Nonostante fossero tutti oggetto della stessa analisi, i risultati dei tre corpora di riferimento sono stati mantenuti separati, allo scopo di poter confrontare i dati tenendo conto delle differenze intrinseche di ogni collezione, che, come anticipato, mettono in luce situazioni diverse in cui l'utente si trova ad interagire.

Descrivendo nel dettaglio ogni corpus si ha:

Clinical questions: presenta periodi complessi, con domande lunghe che introducono il contesto generale in cui la domanda si inserisce. La conversazione avviene fra medici e riguarda richieste di informazione o consigli su casi clinici. Tali frasi, sebbene molto difficili da gestire, hanno permesso di analizzare la varietà di inglese che emerge in una situazione di estrema libertà linguistica, quasi ai limiti del dialogo spontaneo. Si riscontrano fino a tre livelli di subordinazione e molte strutture anaforiche o riferimenti indiretti. Nonostante questo non sia esplicitamente il tipo di domande che un database si può trovare a dover gestire, si è rivelata un' interessante finestra sulle caratteristiche di una conversazione quasi naturale, per mettere in luce le peculiarità linguistiche di questo tipo di interazione.

1.

a. When treating someone with polymyalgia rheumatica, do you follow the erythrocyte sedimentation rate (ESR) or the symptoms?

If someone has a history of preterm labor or preterm delivery, do they need any special monitoring or surveillance during the present pregnancy?

Dati del corpus in Tabella 4.1

Domande yes/no	200
Domande wh	235

Tabella 4.1: Dati corpus Clinical

In Tabella 4.2 si riportano i dati in dettaglio della sezione wh-.

What	100
Where	7
Who	60
Which	24
Who	46

Tabella 4.2: Dettagli corpus clinical

Answer.com: domande che sono state selezionate dal sito web Answer.com, portale da cui è possibile trarre "frequently asked questions" di argomento vario poste da utenti internet. I temi spaziano dall'arte allo sport all'informatica, ma ciò che accomuna tutte le queries è il fatto di essere frutto consapevole di interrogazioni a basi di dati. La struttura

generale, che emerge, è estremamente più semplice rispetto alle domande precedenti, non vi sono periodi lunghi e si può notare al massimo un grado di subordinazione.

2.

a. What do you call someone from Connecticut?

b. What are some important parts of Japanese culture?

Un'analisi generale di questi due tipi di corpora dalle caratteristiche tanto diverse, fa immediatamente emergere una riflessione:

il linguaggio che l'utente usa per interrogare il database appare come una sorta di "linguaggio autocontrollato". I soggetti sembrano infatti propensi a ridurre spontaneamente la complessità del proprio input, soprattutto dal punto di vista sintattico, nel momento in cui si trovano, consapevolmente, ad interagire con una macchina. L'informazione viene frazionata in periodi distinti, evitando coordinazioni lunghe e riferimenti a lunga distanza.

Questa osservazione avvalora l'idea della possibilità di interrogazione di ontologie in linguaggio naturale, e stimola al proseguimento gli studi, perché il fatto che l'uomo tenda ad esprimersi spontaneamente in modo semplice, interfacciandosi con una macchina, fa sì che la distanza fra il suo input naturale e il linguaggio logico sia minore.

Dal punto di vista strutturale il corpus di "Answer.com" presenta 444 domande, di cui:

In Tabella 4.4 si riportano i dettagli sezione wh":

TREC: l'ultimo corpus su cui state sono effettuate le analisi è un estratto di dati di TREC (Text retrieval conference) 2004. TREC, un progetto nato nel 1992, supporta

Domande yes/no	203
Domande “wh”	241

Tabella 4.3: Dati corpus Answer

What	180
Which	10
Who	18
Where	17
When	16

Tabella 4.4: Dettagli corpus Answer

la ricerca nell'ambito del recupero dell'informazione e fra le varie aree si occupa anche di studi circa il Question answering[2]; i dati raccolti sono stati selezionati proprio dalla sezione riguardante questi studi. Di conseguenza i periodi di tale corpus non sono lunghi, né complessi, e sono domande rivolte a sistemi automatici, non domande estratte da dialoghi fra utenti.

Numero totale di domande: 405.

Domande yes/no	122
Domande “wh”	283

Tabella 4.5: Dati corpus TREC

[2] Sistemi che, prendendo in input domande (e non parole chiave), restituiscono in output risposte e non solo liste di documenti rilevanti. (Moldovan *e altri* (2002))

What	227
Which	7
Who	7
Where	1
When	41

Tabella 4.6: Dettagli corpus TREC

4.2.2 Risultati analisi

Oggetto di analisi sono state negazione, disgiunzione ed universale, le strutture non accettate da CQs. Al fine di estrarre dati circa la frequenza di questi costrutti logici nelle domande dell'utente, la ricerca si è focalizzata su termini che veicolassero tali operatori non gestibili. In Tabella 4.7 sono riportati i termini selezionati.

L'analisi effettuata non ha la pretesa di coprire tutte le possibili espressioni con cui il linguaggio naturale può esprimere tali operatori logici, ma, a nostro parere, è sufficientemente completa per tratteggiare la tendenza generale del linguaggio, analizzando quelle che sono le principali forme linguistiche utilizzate per esprimere i concetti logici oggetto di studio. Estraendo, in primo luogo, le ricorrenze di questi termini, rispetto all'intero corpus (al numero totale di tokens), e calcolando la frequenza relativa (Fr= (numero tokens/ totale Corpus) *100), si ottengono valori piuttosto bassi da cui si evince che le parole oggetto di studio non appartengono alle classi ad alta frequenza del corpus. Evidente è però la differenza fra i risultati del corpus di Clinical questions e i corpora di Answer.com e TREC. Ad eccezione della classe degli esistenziali, infatti, in tutti gli altri casi il corpus di domande cliniche, dove l'utente si rivolge ad altri esseri umani, presenta valori più

Operatore	**Termini linguistici**
Quantificatore universale	All, each, every, everybody, everyone, any (in contesto positivo), none, nothing.
Disgiunzione	Or
Negazione	Not (e sue abbreviazioni), without
Quantificatore esistenziale	Any, anything, anyone, anybody, some, somebody, something, someone, there is a, there are, there was a.

Tabella 4.7: Termini oggetto d'analisi

alti per quanto concerne la frequenza dei termini in esame, e questo avvalora l'idea che l'utente utilizzi un registro linguistico diverso a seconda della situazione in cui si trova ad agire. Tali dati vengono riassunti in Figura 4.1.

Più interessante però, può essere andare a vedere la distribuzione delle classi di termini, scelti come rappresentanti degli operatori logici non gestibili, sul numero totale delle ricorrenze degli stessi termini. In altre parole calcolare in che percentuale si distribuiscono la classe degli universali, degli esistenziali, della negazione e della disgiunzione rispetto al totale degli operatori non gestibili. Come emerge dai grafici in Figura 4.2, 4.3 e 4.4 , la negazione e la disgiunzione sembrano essere i due operatori che ricorrono con frequenza maggiore.

Da notare, inoltre che, oltre alle tre classi sopra indicate, i grafici riportano anche la

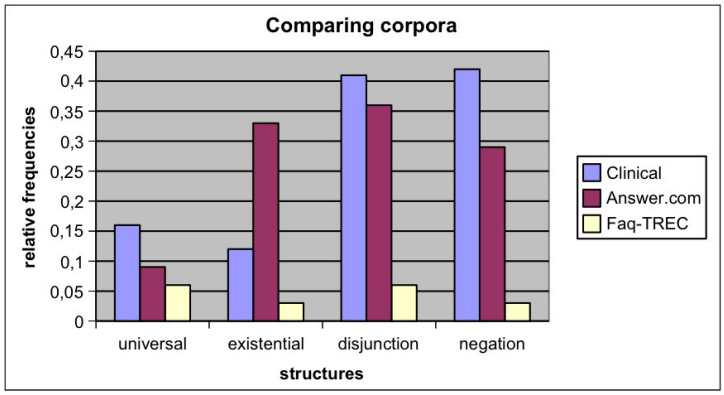

Figura 4.1: Confronto dei corpora

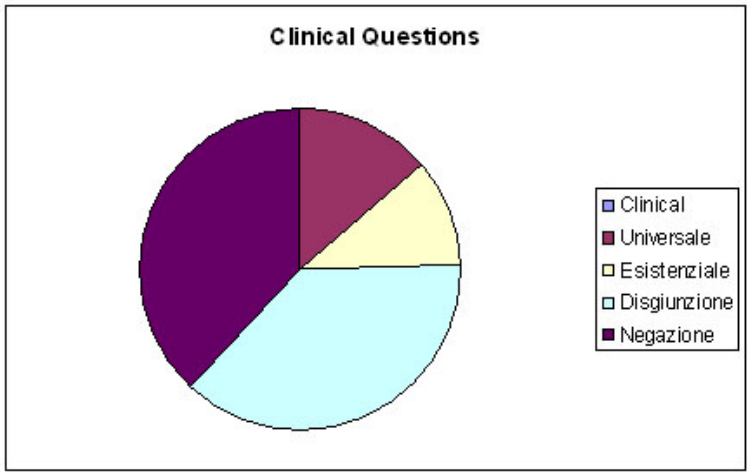

Figura 4.2: Operatori in Clinical question

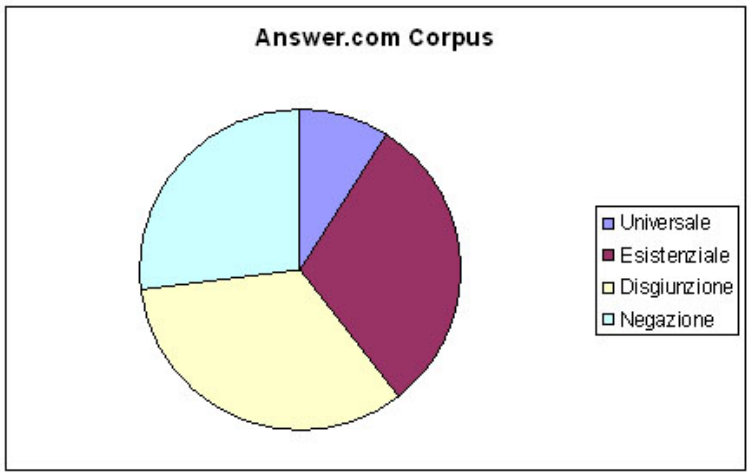

Figura 4.3: Operatori in Answer.com

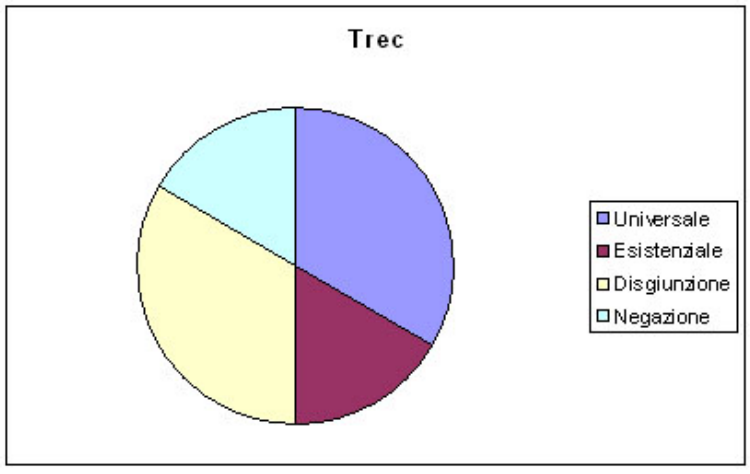

Figura 4.4: Operatori in Trec

frequenza dei termini legati al concetto logico di quantificatore esistenziale. Questo gruppo è stato analizzato sia per avere un metro di confronto con la classe degli universali, sia per una coerente gestione degli elementi a polarità negativa (quegli elementi con valore esistenziale in contesto negativo e interrogativo, ma valore universale in contesto positivo).[3]

Ma sicuramente più significativo è il dato riguardante la percentuale di domande non gestite dal frammento gestito da CQs, a causa della presenza degli operatori (e quindi dei termini) in questione. I risultati di tale analisi sono riassunti in Tabella 4.8.

Percentuale delle domande non gestibili			
	Universali	Disgiunzione	Negazione
Clinical Corpus	2.7%	13.00%	11.9%
Answer.com	1.3%	4.05%	2.9%
TREC	0.49%	0.5%	0.2%

Tabella 4.8: Percentuali delle domande non gestibili

Analizzando in dettaglio le varie classi, emerge quanto segue.

Quantificatore universale. Solo dodici domande del corpus Clinical questions (su un totale di 435 domande) presentano termini che esprimono l'operatore universale, cioè una

[3]Un esempio esplicativo di tale differenza puo' essere

-1. Mary doesn't want any present

-2 You can take any book!

In 1) *any* ha valore esistenziale (*non esiste neanche un regalo voluto da Mary*), mentre in 2), contesto affermativo, *any* assume valore universale (*puoi prendere qualsiasi libro*).

percentuale del 2,7. Nel corpus Answer.com tale percentuale scende al 1,3% (cioè sei domande su un totale di 444), e in quella della TREC si riduce ancora allo 0,49% (due domande sulle 405 totali). L' operatore universale, inoltre, sembra poter ricorrere in ogni posizione sintattica:

1

a. Should all pregnant women take a test for human immunodeficiency virus?

b. What is causing all the joint pain?

c. Can we use Energix for all five doses od DPT?

Negazione: gli elementi che introducono negazione logica all'interno delle interrogative risultano poco frequenti e non compaiono mai in posizioni tali da negare il verbo della proposizione principale (sono presenti solo in predicati della dipendente, sintagmi preposizionali introdotti da *without,* o come negazione di un aggettivo). La percentuale delle domande negative nel corpus Clinical questions si stabilisce sull' 11,9%, per un totale di 52 frasi negative; scende però al 2,9% in Answer.com (13 frasi), per ridursi ancora allo 0,24% nelle domande tratte dalla TREC (un solo caso negativo su 405 domande).

Esempi:

2.

a. Should I give you a full series of tetanus shots to an adult who does not know their immunization history?

b. What is the chance that aspirating a joint effusion that is not red or tender will yield anything diagnostically?

c. Is it possible for someone to get recurrent pelvic disease without a new exposure?

Disgiunzione: Analizzando la Tabella 4.8, si nota immediatamente come le percentuali

dell'operatore di disgiunzione siano sensibilmente più alte delle altre. Tale risultato però non deve trarre in inganno perché dal totale delle domande che presentano una struttura disgiuntiva è necessario sottrarre quelle in cui tale operatore coordina due sintagmi nominali dai casi in cui coordina due proposizioni. Per tali risultati si rimanda alla Tabella 4.9, dove sono riportate le percentuali circa la funzione dell'operatore di disgiunzione. Ogni valore è calcolato sul numero totale di domande del corpus: ad esempio nel corpus di Answer.com le domande con disgiuntive rappresentano il 4.5% del numero totale di domande (in numero di 18 su 444 domande totali), e, nel dettaglio: il 2.7% delle domande del medesimo corpus presenta un operatore di disgiunzione che coordina due sintagmi nominali, mentre l'1.3% presenta una disgiunzione fra proposizioni. Ne emerge che, sebbene la disgiunzione presenti valori più alti, compare soprattutto nella coordinazione di nomi, rarissimamente nella coordinazione di proposizioni, posizione in cui risulta più difficile da gestire.

3.

a. What is the word for the fear if viewing sports or playing sports?

b. Is it degenerative arthritis or might it be a stress fracture?

c. Is Liberia considered a rural or a urban country?

d. What are opinions on the best free on line photo editor or photo shop ?

A conclusione dell'analisi, emerge che le strutture non accettate da CQs sembrano non comparire ad alte frequenze nelle domande tipicamente usate dall'utente nell'interrogare una base di dati. D'altra parte, come anticipato precedentemente, emerge anche una

Corpus	**Numero domande**	**Disgiunzione fra SN**	**Disgiunzione fra proposizioni**	**Totale domande con disgiunzione**
Clinical	56	11.5%	1.4%	13,00%
Answer.com	18	2.7%	1.3%	4.05%
TREC	2	0.5%	0,00%	0.5%

Tabella 4.9: Percentuali delle domande con disgiunzioni

certa differenza, in termini di frequenza, fra il corpus di domande non consapevolmente indirizzate ad una base di dati e il corpus di domande rivolte ad una base di dati.

Seppur rari, i casi in cui tali operatori compaiono devono essere gestiti. Nel prossimo capitolo si descriverà una forma di "indebolimento semantico" che permetterà di gestire parzialmente l'universale, senza ricorrere ad una modifica del frammento di logica: un indebolimento che può essere definito come una semplificazione semantica dell'input proposto.

4.2.3 Ulteriori riflessioni emerse dall'osservazione dei corpora

Oltre alle analisi specifiche sugli operatori logici, l'osservazione dei corpora di queries fa emergere anche un altro fenomeno che può essere di ostacolo per una corretta interpretazione semantica dell'analisi automatica. In dettaglio, il fenomeno riscontrato riguarda il problema degli atti indiretti, enunciati che, pur mantenendo una struttura interrogativa, veicolano indirettamente la richiesta di informazione. Sebbene la questione esuli dal problema del frammento di inglese rappresentato dalle CQs, esso presenta comunque un tema da non sottovalutare, poiché l'utente dimostra di applicare all' interazione con la macchina quei processi inferenziali che sono tanto naturali nelle conversazioni reali, quanto di difficile interpretazione in ambito artificiale.

Nel corpus delle Clinical questions si trovano, infatti, domande come:

1

a. Can you please tell me what to do with this patient ?

b. I wonder if a rapid turnover of the bones could do this.

c. I don't know what is causing fever.

Questi tre esempi mostrano diversi livelli di "indirezionalità", e quindi di complessità, ma in tutti e tre emerge come il significato letterale dell'enunciato non veicoli la funzione comunicativa di "richiesta di informazione". Per approfondimenti circa la teoria degli atti diretti e indiretti si rimanda al Capitolo 5, sezione 5.4.

Sebbene tale fenomeno riccorra soltanto nel corpus di interrogazioni più complesse (domande cliniche) è sicuramente un campo interessante per analizzare come la semplificazione testuale, in questo caso semplificazione sintattica, può permettere di affrontare e superare questi ostacoli tipici del linguaggio naturale in esame.

4.2.4 Analisi comparativa di un corpus italiano

Lo stesso tipo di analisi è stato effettuato anche su un corpus di queries italiane, al fine di indagare se le tendenze generali riscontrate nei corpora dell'inglese, venissero sostanzialmente rispettate. E' interessante notare come da questo confronto emergano delle regolarità indipendenti dal linguaggio nell'inglese e nell'italiano tipicamente usati in interrogazioni a basi di dati.

Il corpus italiano comprende trecento domande poste da parlanti nativi ad una base di dati contenente informazione geografica sul territorio di Bologna[4], e la struttura generale delle queries si avvicina a quella delle domande del corpus inglese tratto da TREC: estrema semplicità sintattica e predilezione per la coordinazione piuttosto che la subordinazione.

Si riportano in 1 alcuni esempi tratti da questo corpus di domande italiane.

1.

[4]Dati forniti dall'Istituto di Linguistica Computazionale , CNR-ILC Pisa, utilizzati nel progetto FuLL (Fuzzy Logic and Language) un progetto di ricerca avente come obiettivo lo sviluppo di una tecnologia software che innova i sistemi di interrogazione dei database tramite l'utilizzo di una interfaccia in linguaggio naturale.

a. Qual è il corso d'acqua principale di Monzuno?

b. Qual è il nome della provincia di Bologna con il minor numero di residenti?

c. Esiste una pista ciclabile che costeggia il fiume Reno?

Il corpus, costituito da un totale di 2734 tokens, è stato analizzato per quanto riguarda le strutture logiche non gestite da CQs. Dal punto di vista linguistico, come precedentemente, la frequenza di questi operatori logici è stata calcolata sulla base dell'occorrenza dei termini in Tabella 4.10.

Operatori logici	**Termini linguistici**
Universale	Ogni, ognuno, tutto, tutti, tutte, tutta, nessuno, nessuna, alcuno, alcuni, alcuna, alcune
Negazione	Non, nessuno, nessuna, niente, nulla, senza.
Disgiunzione	O, sia

Tabella 4.10: Termini italiani oggetti d'analisi

I risultati emersi sono riassunti nella Tabella 4.11.

Operatore Logico	**Numero domande**	**Valore percentuale**
Quantificatore universale	0/300	-
Negazione	3/300	1%
Disgiunzione	2/300	0.6%

Tabella 4.11: Risultati sul corpus italiano

In dettaglio:

Quantificatore universale: non compare in alcuna delle domande della collezione.

Negazione: 1% delle domande presenta una struttura negativa.

2

a. Ci sono delle località non servite da farmacie?

b. Quali sono comuni che non sono attraversati da strade ?

c. A breve distanza dal Centro Leonardo (Imola) esistono rifornitori di benzina senza piombo che accettano carta di credito?

Disgiunzione: lo 0.6% delle domande presenta una forma disgiuntiva:

3

a. Il numero di telefono o la mail degli agriturismi nel Comune di Porretta.

b. Qual è il distributore ad Ozzano che permette sia di utilizzare la carta di credito sia di rifornirsi di metano?

4.3 Analisi delle affermative

In questa seconda sezione si esamina il frammento di inglese accettato dal linguaggio logico di descrizione dell'ontologia e usato per rappresentare gli assiomi della base terminologica intensionale, al fine di analizzare la distanza fra tale frammento e la varietà di linguaggio naturale ad esso più vicino.[5]

Nel contesto specifico, come descritto nel secondo capitolo, il linguaggio in cui viene descritta l'ontologia è *DL-Lite*$_{core}$, un frammento della famiglia delle DLs che permette

[5]Come si vedrà in seguito si è ritenuto interessante focalizzare l'analisi su corpora di regolamenti e norme che sono sembrati la varietà linguistica più vicina agli assiomi di una ontologia.

di gestire grandi quantità di dati, a discapito dell'espressività completa, ma che consente di esprimere i concetti che permettono al sistema di inferire conoscenza. Si riportano alcuni esempi esplicativi della struttura degli assiomi in una T-Box (base di conoscenza intensionale) in DL-Lite:

Everyone who studies left

Everyone who knows something left.

Contestualizzando in un dominio bibliotecario

2

a. Every student can access the library.

b. Every student has a Student Card.

c. Every professor has a Campus Card.

L'analisi si è concentrata su alcuni operatori logici che per motivi di costi computazionali rimangono al di fuori dell'espressività DL-Lite: disgiunzione e negazione. In realtà tali operatori non sono gestibili solo in alcune parti dell'assioma, ma un'analisi più specifica sarà oggetto nel Capitolo 6, dove si descriverà lo studio condotto, a mano, su un terzo dell'intero corpus di affermative, al fine di estrarre le caratteristiche non gestibili con l'analisi automatica.[6]

4.3.1 Struttura del corpus

I corpora analizzati sono stati raccolti da collezioni di testi normativi e di regolamenti, considerati il genere più vicino alle strutture sintattiche tipiche degli assiomi di un'ontologia

[6] Tale studio necessita di informazioni sull'espressività e le peculiarità delle varie versioni di DL-Lite, che verranno fornite nel sempre nel Capitolo 6. Questa necessità ha reso impossibile anticipare qui i risultati di tale analisi. Per completezza però si è scelto di anticipare in questa sezione i dettagli sulla costruzione del corpus e i risultati dell'analisi automatica.

che possono essere descritti come le "regole" del mondo rappresentato dall'ontologia. Vista la varietà degli stili e dei generi all'interno di quest'area d'indagine, la collezione è stata strutturata scegliendo diversi tipi di testi: testi governativi dal registro molto formale, e regolamenti in un registro più informale, come linee guida o risposte a FAQ.

Quello che si è ottenuto è un corpus di 33200 parole, formato da:

- regolamento sul trasporto aereo canadese - 17000 tokens (circa): legge del emanata dal Governo canadese;

- linee guida per il cittadino canadese, raccolte dal sito di *e-governement* (circa 8000 tokens);

- regole e servizi per un passeggero della compagnia aerea British Airways (circa 7000 tokens).[7]

Circa la metà del corpus è costituita da un testo di legge tradizionale e presenta articoli come:

1

a. Every applicant for a licence or for an amendment to or renewal of a licence, and every license, shall file with the Agency, in respect of the service to be provided or being provided, as the case may be, a valid certificate of insurance in the form set out in Schedule I.

L'altra metà del corpus invece raccoglie testi in registri meno formali, ma molto frequenti sul web:

1. linee guida su un sito di e-government dove sono descritte dettagliatamente le pro-

[7]Dati tratti da: http://canadabusiness.gc.ca/gol/cbec/site.nsf/en/index.html e http://www.britishairways.com/travel/home/public/it_it.

cedure che il cittadino deve eseguire per ottenere documenti, o VISA, o passaporto o servizi di vario genere.

b. To apply for a Social Insurance Number, you must complete an application form.

c. You can obtain an application from your local office or download one;

2. risposte estratte dalle *frequently asked questions* del sito British Airways, volte ad informare il passeggero sulle tariffe, i servizi e le misure di sicurezza.

c. Liquids must be carried in individual containers, not exceeding 100 ml (even if they are not full).

d. All customers may purchase any non-liquid item before or after security.

Il principio comune dei testi raccolti è l'essere portatori non di fatti, ma di asserzioni e di definizioni universali, che possono costituire la base di principi universali di un dominio.

4.3.2 Risultati

I termini su cui l'analisi si è focalizzata sono raccolti in Tabella 4.12.

Operatore logico	**Termine**
Negazione	Not, without
Disgiunzione	or

Tabella 4.12: Termini oggetto d'analisi

I risultati emersi sono riassunti in Tabella 4.13, dove i valori sono normalizzati rispetto alla grandezza totale del corpus.

Termine	**Frequenza relativa**
Not, without	0.5
or	2.46

Tabella 4.13: Risultati analisi sulle affermative

Oltre a notare che i costrutti negativi non sono eccessivamente frequenti, è interessante vedere che, sul totale dei costrutti non gestibili, la percentuale maggiore è legata a enunciati che presentano la disgiunzione. Tale percentuale infatti raggiunge l'83% del totale delle situazione problematiche.

La frequenza delle disgiunzioni è significativa, perché, dal punto di vista quantitativo risulta a valori più alti rispetto agli altri operatori, e perché presenta strutture molto difficili da gestire. Per un'analisi dettagliata delle difficoltà di tale operatore si rimanda al Capitolo 6. E' da notare comunque che, in questo corpus, l'operatore di disgiunzione ricorre con una frequenza maggiore rispetto all'operatore di coordinazione copulativa *and*. Sul numero totale di congiunzioni coordinanti (avversative, disgiuntive e congiuntive), la disgiunzione copre il 50%, la coordinazione copulativa con *and* il 48% e l'avversativa con *but* il 2%. Sia *and* che *or* appartengono alla casse di parole ad alta frequenza del corpus.

La frequenza dell'operatore *or* non è quindi da sottovalutare perché nei testi normativi o di regolamenti emerge la generale tendenza all'uso di complesse strutture disgiuntive. In particolare tale tendenza è da rimandare alla presenza di elenchi di possibilità alternative come in 2a.

2.

a. You can book your the Skyflyer Solo service by contacting your local British Airways office or your travel agent or by mail.

b. All customers may purchase any non-liquid item before or after security.

c. You can obtain an application from your local office or download one.

d. It is important that documents are originals and that they are written in English or French.

4.3.3 Conclusioni

A conclusione di tali analisi, si può osservare che le domande spontaneamente poste dall'utente in fase di interrogazione non presentano alte frequenze di strutture problematiche. Una situazione leggermente diversa si ha nel corpus di affermative, dove la disgiunzione compare con una frequenza relativa superiore al 2%, addirittura superiore alla coordinazione con *and*.

Rispetto al frammento gestito da CQs, negazione e universale hanno frequenza piuttosto basse, quindi per consentire una gestione di tali strutture verrà descritta una forma di indebolimento semantico, applicata per ora al caso del quantificatore universale. D'altra parte le frequenze degli operatori problematici per Dl-Lite risultano più alte, soprattutto nel caso della disgiunzione, e questo, porta a riflettere sugli effettivi costi dell'operatore di disgiunzione: se da un lato è un operatore fortemente gravoso in termini di complessità, dall'altro è comunque un operatore presente in un input naturale, la cui rinuncia crea un vuoto importante fra l'utente e l'ontologia. [8]

[8] D'altro lato,come si vedrá nel Capitolo 6, prima della disgiunzione vi sono altri costrutti che limitano la copertura del frammento di logica sul linguaggio delle affermative; la gestione dei quali permeterebbe di avere una copertura abbastanza soddisfacente, anche senza trattare il problema disgiunzione.

Capitolo 5

Dal linguaggio naturale alle CQs

5.1 Verso CQs: la semplificazione del testo

Nella sezione precedente si è analizzata la frequenza degli operatori logici non gestibili da CQs concludendo che, anche se non ricorrono ad alte frequenze, dato il ruolo fondamentale che giocano, la loro presenza deve essere ugualmente gestita. D'altra parte sono anche emerse situazioni complesse, come atti indiretti, che possono essere un ostacolo nel passaggio dal linguaggio naturale alle CQs.

In questo capitolo verranno quindi descritti due approcci che permettano di gestire le situazioni problematiche emerse dall'analisi. Due approcci che, pur partendo entrambi dall'idea della semplificazione testuale, si delineano l'uno sulla strada della semplificazione semantica e l'altro sulla strada della semplificazione sintattica.

In [5.3] verrà descritto l'approccio, presente in letteratura e rielaborato in (Carbotta e Calvanese (2007)), di indebolimento semantico, per la gestione dell'universale attraverso la teoria del fallimento della negazione.

In [5.4] verrà invece descritta una proposta personale di semplificazione sintattica, volta alla gestione del problema degli atti indiretti, attraverso parafrasi automatica.

Sebbene da un lato l'attenzione sia sull'aspetto semantico e dall'altro su quello sintattico, in entrambe le situazioni si richiamano approcci basati sulla semplificazione testuale che permetta di parafrasare l'input in una frase proiettabile sul frammento di logica.

Tale semplificazione ricorda i processi che sfruttano i linguaggi controllati (Capitolo 3), ma se in un approccio basato esclusivamente sul linguaggio controllato la strategia è quella limitare l'utente a scrivere in un linguaggio ridotto, in questo caso l'utente viene lasciato assolutamente libero di esprimersi. Solo in un secondo momento, e in modo completamente invisibile all'utente stesso, l'input viene automaticamente semplificato in un enunciato compatibile con il linguaggio logico.

Siddharthan (Siddharthan (2003)) definisce la semplificazione del testo come *il processo, basato su un insieme di regole scritte a mano, di riduzione di complessità di un input allo scopo di facilitare la comprensione per l'uomo e per la macchina.*

Lanciando uno sguardo ai lavori precedenti, si nota come gli studi sulla semplificazione testuale sono nati con due diversi obiettivi che richiamano la definizione di Siddharthan: in primo luogo la riduzione della complessità per gruppi di soggetti con disturbi linguistici o della comunicazione (come pazienti afasici o soggetti affetti da disturbi uditivi); in secondo luogo la riduzione della complessità come trattamento preliminare del testo prima di processi o analisi automatiche.

Come evidente quindi, gli approcci qua descritti, sia quello basato sull'indebolimento semantico (Carbotta e Calvanese (2007)), sia la proposta personale di un modulo di riscrittura sintattica per la semplificazione degli atti indiretti, si discostano dai progetti che hanno segnato la nascita della semplificazione testuale.

5.2 Precedenti lavori sulla semplificazione testuale

Dalla metà degli anni '90 alcuni gruppi di ricerca hanno concentrato i propri sforzi sul processo definito come semplificazione testuale.

Da una parte il gruppo dell'Università della Pennsylvania (Chandrasekar *e altri* (1996)) si è occupato della semplificazione come processo preliminare ad un sistema di parsing automatico. Dall'altra il gruppo PSET (Carroll *e altri* (1998)) si è interessato della semplificazione di articoli di giornale per la lettura degli afasici.

Lavoro di Chandrasekar. (Chandrasekar *e altri* (1996)). Al fine di ottenere un testo riformulato, facile da analizzare sintatticamente, propone un processo basato su due fasi successive: analisi e trasformazione. La fase di trasformazione (nel primo studio del 1996) sfrutta una grammatica di regole di semplificazione scritte a mano che permette di riscrivere frasi relative come la (1a) in due proposizioni come in (1b)

1.

a. John, who was the CEO of a company, played golf.

b. John played golf. John was in the CEO of a company.

In un secondo momento invece Chandrasekar (Chandrasekar e Srinivas (1997)) propone un approccio *machine learning* (di apprendimento automatico) per creare automaticamente la grammatica di riscrittura sulla base di un algoritmo di apprendimento supervisionato. Confrontando un corpus di frasi complesse con un corpus delle stesse frasi semplificate manualmente, estrae le regole di riscrittura che permettano di ripercorrere lo stesso processo di semplificazione automaticamente.

Progetto PSET: legato allo studio e progettazione di programmi di accessibilità per afasici (Carroll *e altri* (1998)). Il processo è simile a quello di Chandrasekar, basato su

una grammatica di regole di riscrittura e i costrutti analizzati sono: forme attive, passive e legami anaforici.

Nel 2003, la semplificazione sintattica viene ripresa da Advaith Siddharthan nella Tesi di Dottorato Syntactic simplification and text cohesion (Siddharthan (2003). Siddharthan analizza l'interazione fra i processi di semplificazione e le teorie del discorso, per ottenere un testo finale che, pur semplificato, mantenga la coesione e la fluidità dell'originale. Per questo divide il processo in tre fasi successive, *analisi, trasformazione e rigenerazione.* Le prime due corrispondono alle fasi di Chandrasekar:

1. analisi sintattica del testo,

2. trasformazione sulla base di regole scritte a mano.

Mentre, con la terza fase, cerca di ricostruire la coesione del discorso modificando l'ordine delle frasi, controllando i connettori e ricreando legami anaforici.

Le regole di Siddharthan formano una grammatica che gestisce le frasi relative, le frasi coordinate e le apposizioni, e presentano una struttura del tipo:

$VW_{NP}X[_{RL}RELPRY]Z$,

dove la relativa incassata viene estratta in due proposizioni indipendenti come in (1).

1. *Mary who likes honey eats.*

   1a. Mary eats

   1b. Mary likes honey

Quindi, se una frase presenta una relativa $RELPR\ Y$ legata ad un sintagma nominale W, queste due parti vengono scisse in due frasi distinte che possono essere coordinate.

5.3 Semplificazione semantica del testo

L'approccio di semplificazione "semantica" del testo è stato pensato per gestire la mancanza dell'operatore universale nell'espressività di CQs e può essere definito come un "indebolimento semantico".

Per introdurlo si ripropone, con una panoramica generale, il processo di interrogazione.

L'input naturale da cui ha inizio tale procedimento deve essere analizzato linguisticamente, secondo il tradizionale processo di analisi. In questo contesto, l'analisi linguistica si avvale dell'analizzatore sintattico CCG (Bos *e altri* (2004)), della catena di agenti C&C[1], e quindi dell'analizzatore semantico Boxer, compreso nella catena. La catena comprende una grammatica categoriale che è il nucleo dell'analizzatore sintattico, ma anche un analizzatore morfologico, un chunker e un riconoscitore di entità nome. L'output di questo processo (costituito da dipendenze CCG, annotazioni di entità nome e annotazioni morfologiche) viene sfruttato da un modulo Boxer (Curran *e altri* (2007)), per produrre strutture in rappresentazione semantica. Boxer infatti riesce a generare strutture "a scatole" (conosciute come strutture di rappresentazione del discorso, DRSs Discourse Representation Structures)[2]. Ogni DRSs comprende un insieme di referenti del discorso (x0, x1...) e un insieme di condizioni che possono essere semplici come relazioni unarie o binarie fra due referenti, o complesse come negazione, disgiunzione, implicazione.

Boxer, quindi, prendendo in input un enunciato analizzato sintatticamente con la specifica delle componenti da CCG, vi assocerà una rappresentazione DRSs dove nomi, aggettivi e verbi diventeranno relazioni unarie, il cui significato è dato dal lemma corrispondente

[1] Cfr. sito:http://svn.ask.it.usyd.edu.au/trac/candc/wiki/Documentation.

[2] Le DRSs sono rappresentazioni grafiche legate alle teorie di rappresentazione del discorso, DRT (*discourse representation theory* (Kamp e Reyle (1993)).

(*Bob(x1) , write(x2), event(x2)*) , mentre i ruoli semantici diventeranno relazioni binarie (*agent(x2,x1)*).

Es. Bob writes

referenti del discorso: *x0, x1.*

condizioni: *Bob(x0), write(x1), event(x1), agent(x1,x0).*

x0, x1
Bob(x0)
write(x1)
event(x1)
agent(x1,x0).

$\exists x0, \exists x1 (Bob(x0) \wedge write(x1) \wedge event(x1) agent(x1, x0))$

Prendendo in considerazione enunciati interrogativi, l'output di Boxer sarà espresso nella forma:

$$\{\ x | \exists y(\varphi(y) \wedge D(x) \wedge \exists z((\psi(x, y, z))))\ \}$$

dove x è il focus della domanda, D è il dominio in cui *si domanda*, φ è la conoscenza che l'utente esprime nella domanda, e ψ è il corpo della domanda stessa.

Il dominio D sarà vuoto se la domanda è booleana, mentre dipenderà dal pronome utilizzato se la domanda è wh: *modo, persona, luogo, unità di tempo, cosa e ragione* corrispondono rispettivamente a *how, who, where, when, what* e *why.*

Quindi in una domanda introdotta da *who*, x apparterrà al dominio delle persone (*per-*

son(x)), come in (1a), mentre in una domanda con *what,* x apparterrà al dominio delle cose (*thing(x)*).

Data la query Boxer assegnerà i ruoli tematici a seconda delle categorie sintattiche presenti nel file analizzato sintatticamente. Per esempio la query: "*Who may use the Interlibrary Loan service?*" verrà rappresentata come in (1a), dove si notano le relazioni unarie associate ai nomi e agli aggettivi della query (per esempio *loan(y1)*) e le relazioni binarie associate ai ruoli (con *agent(z,x)*).

(1a)

$$\{\mathrm{x}|\ \exists y^1 \exists y^2 (loan(y^1) \wedge Interlibrary(y^2) \wedge service(y^2) \wedge nn(y^1, y^2)) \wedge person(x) \wedge \exists z (use(z) \wedge event(z) \wedge agent(z,x) \wedge patient(z,y^2))\ \}^3$$

In altre parole, dalle categorie sintattiche espresse da CCG (ad esempio *who_ soggetto),* Boxer estrae i ruoli tematici (ad esempio: *who_ agente*) .

L'output di Boxer non corrisponde, però, al formalismo UCGs (unione di queries congiuntive), e questo costringe ad una prima traduzione che trascriva le formule in logica di primo ordine di Boxer in CQs. E' in questo momento del processo che, passando da una logica più espressiva ad una meno espressiva, si è costretti a rinunciare a qualche operatore logico (come l'universale).

[3]La formula in (1a) potrebbe essere informalmente tradotta : dato che esiste un *servizio*, esiste un *prestito*, esiste un *servizio di prestito* (*nn-* identifica il sintagma nominale complesso), e dato che x è una persona (dominio che dipende dal wh- *who*) e esiste un evento z che è *usare*, restituire un soggetto x che sia agente di z quando y è paziente di z.

5.3.1 Indebolimento semantico

Per parlare di indebolimento semantico, è necessario, in primo luogo, riflettere sulle basi di conoscenza; esse si basano sull'ipotesti di "mondo chiuso" (*closed world assumption*), ipotesi che permette di considerare completa la conoscenza di cui si dispone, derivandone il fatto che ciò che non é presente nella banca dati sia falso. Nell'approccio logico al querying di una banca dati ed in particolare nell'utilizzo di un'ontologia di supporto a tale compito, si assume, invece, una situazione di conoscenza incompleta, ipotesi di "mondo aperto", da cui deriva che, se un'informazione non é presente nella banca dati non é necessariamente falsa. Questo é dovuto al fatto che l'interrogazione della conoscenza (immagazzinata in banca dati e ontologia) viene gestita come un' implicazione logica di cui si verifica la validità, cioè la sua verità in ogni modello possibile. In ipotesi di mondo chiuso, al contrario, la verifica di validità viene effettuata solo nella specifica istanza rappresentata dalla banca dati che si sta interrogando.

Inserendo l'operatore di conoscenza K si può asserire $K\exists y.\varphi(c,y)$, dove si assume la conoscenza epistemica dell'implicazione che segue (cioé del fatto che esista un y per cui $\varphi(c,y)$). L'ontologia, che possiede informazione certa nella sua base di conoscenza, può estrarre tuple per cui vale la relazione $\varphi(c,y)$.

In questo contesto però, il quantificatore universale diventa problematico, perché non è possibile inferire conoscenza circa tutti gli elementi dell'insieme. Semplificando, l'ontologia sarà in grado di dire, per esempio, che *"gli studenti x, y, z.....n dell'università di Pisa possono usare il prestito interbibliotecario"* , ma non avrà mai la certezza che *"tutti gli studenti dell'Università di Pisa possono usare il prestito interbibliotecario",* a meno che non esista un assioma che espliciti questa informazione.

Come si è messo in evidenza, le CQs non contengono il quantificatore universale. Per poter riuscire a gestire domande in linguaggio naturale che presentano nella rappresentazione formale tale operatore, in (Bernardi *e altri* (2007b)) é stato proposto un procedimento di indebolimento semantico di tale rappresentazione, che si avvale di operatori epistemici. L'intuizione, che sta dietro a questo indebolimento, è l'idea che tutto ciò che il sistema può fare, sia trovare un controesempio ad una data implicazione; se fallisce allora si può assumere l'implicazione come vera. L'ontologia può solo "credere" che la condizione valga universalmente, sulla base della conoscenza che possiede, e per questo sostituire l'operatore di conoscenza K con B *(believe).* Lavorando in un insieme chiuso, si può sfruttare l'idea che, qualora non si trovino costanti che non "rispettino" la condizione, allora tale condizione può essere assunta come vera per tutte le costanti del dominio. Per cui si sfrutta l'equivalenza:

$$B \equiv \neg K \neg \varphi$$

"credo che tutti gli studenti possano prendere in prestito libro, perché non sono a conoscenza di alcuno che non può".

Facendo un passo indietro all'input in linguaggio naturale dell'utente, una query del tipo:

(2a) what is causing all the joint pain?

Potrà essere tradotta dal sistema come:

(2b) I am not aware of any joint pain that is not caused by the following diseases.

Si può quindi dire che l'interrogazione dell'utente è stata parafrasata in una forma più semplice, anche se non si è trattato di una parafrasi a livello di struttura sintattica, ma a livello semantico. E' per questo interessante proporre questo approccio di indebolimento

semantico come una forma di semplificazione testuale semantica (Carbotta e Calvanese (2007)). D'altra parte, è importante rendere esplicito questo indebolimento, informando l'utente che la risposta alla sua interrogazione è riferita ad una domanda semanticamente diversa da quella che lui aveva precedentemente inserito; è per questo motivo che la risposta restituita dal sistema va a completarsi con una premessa come*"I am not aware of...*, che in (2b) diventa " *I am not aware of any joint pain that is not caused by the following diseases"*[4]. In questo modo si esplicita linguisticamente sia l'impossibilità del sistema di rispondere alla query letterale, sia lo "sforzo" di restituire un' informazione che si ritiene essere comunque utile per l'utente, piuttosto di dichiarare un fallimento completo nel recupero.

5.3.2 Sviluppi futuri e limiti dell'indebolimento semantico

La semplificazione testuale che è stata presentata, si discosta dalla tradizionale letteratura a riguardo. Infatti, è a livello di forma logica che la frase viene ripensata e modificata in modo da essere gestibile, nonostante la presenza dell'universale. La teoria è nota in logica come "negazione del fallimento", ed, in questo ambito, si può affermare che tale operazione proponga una forma di semplificazione: *tutti gli studenti possono usare il prestito interbibliotecario?,* viene semplificata a livello di forma logica in *esiste uno studente che non può usare il prestito interbibliotecario?.* Se qualcosa di simile può essere applicato alla negazione, non è possibile invece riuscire a gestire la disgiunzione, che rimane al di fuori delle potenzialità di CQs.

La disgiunzione rappresenta quindi un problema aperto in un approccio che utilizza CQs,

[4] *Non sono al corrente di alcun dolore articolare che non sia causato dalle seguenti malattie.*

dal momento che, essendo un operatore estremamente costoso in termini computazionali non può essere inserita in un frammento di logica, se non a costo di un esponenziale incremento della complessità. D'altra parte è emerso che, nelle analisi condotte nel capitolo precedente, tale operatore caratterizza solo il 4% delle domande tipicamente usate dagli utenti in fase di interrogazione e di tale percentuale solo raramente si trova un operatore di disgiunzione fra proposizioni (solitamente infatti coordina due sintagmi nominali). In questi rari casi la domanda può essere gestita usando UCQs, (unioni di CQs)[5] dividendo la frase nei dei due elementi della disgiunzione.

Tra i progetti futuri che riguardano questo processo di semplificazione a livello semantico è interessante pensare ad un test di valutazione dell'accettazione da parte degli utenti. La risposta fornita dal sistema infatti si discosta leggermente da quella che l'utente si aspetta di ricevere. Per questo motivo vale la pena potare avanti un test di valutazione su un campione rappresentativo di utenti per calcolare il livello di soddisfazione e gradimento della risposta.

5.4 Gli atti indiretti

Esiste una sorta di rapporto preferenziale tra il tipo di frase (dichiarativa, interogativa, etc.) e lo scopo che il parlante si prefigge di raggiungere con frase utilizzata. Ma questo nn é sempre vero. Se solitamente un ordine viene veicolato da un frase imperativa ("chiudi la porta"), o una richiesta di informazione da una domanda ("dov'è la stazione?"), ci sono casi in cui il parlante può decidere di raggiungere il suo obiettivo comunicativo, (dare un

[5] Si veda 2.2.4

ordine, o chiedere un'informazione) in modo "indiretto". Sono, per esempio, situazioni in cui si tende a celare un ordine dietro ad una domanda di cortesia ("puoi chiudere la porta?"), o una richiesta di informazione dietro ad una affermazione ("non so dov'è la stazione").

Questo fenomeno, che caratterizza una notevole percentuale delle conversazioni umane, viene trattato in letteratura negli studi sulle teorie degli atti comunicativi, e in particolare degli atti indiretti (si veda di seguito): studi che sono stati al centro dell'interesse di quel settore della linguistica che va sotto il nome di pragmatica e che si occupa delle relazioni fra lingua e contesto.[6]

Tale fenomeno non emerge solo nelle conversazioni uomo-uomo, ma si riscontra anche nelle conversazioni uomo-macchina. Nelle queries analizzate, infatti, si notano situazioni in cui l'utente tende a modificare in qualche modo quel rapporto diretto fra tipo di frase e intenzione comunicativa proprio come farebbe in una situazione di conversazione reale. Come emerso in 4.2.3, infatti, nei corpora analizzati vi sono casi come:

(1a) Can you please tell me what to do with this patient ?

(1b) I wonder if a rapid turnover of the bones could do this.

(1c) I don't know what is causing fever.

Si noti come in (1a) il parlante non ha intenzione di sapere quali siano le potenzialità del sistema di "essere in grado di fare x", ma vuole in realtà avere informazioni circa x. Così come in (1b) e (1c) non vuole solo affermare la sua "non conoscenza di x", ma richiede *indirettamente* informazione circa x.

[6]Levinson (1983),(pp.17-30), mette in luce le difficoltà di una definizione esaustiva del termine, ma in questo contesto, si può definire la pragmatica come lo studio delle relazioni fra lingua e contesto.

Ma se in situazioni conversazionali reali, è possibile risalire, anche in questi casi, a quelle che sono le reali intenzioni del parlante, celate dietro ad un certo tipo di enunciato, ed è possibile grazie al mondo di conoscenze condivise tra ascoltatore e parlante, lo stesso non si può dire nel caso di conversazioni uomo/macchina, dove non ci si può appellare ad informazione di tipo contestuale.

5.4.1 Contesto teorico: la teoria degli atti

Come anticipato, il fenomeno sopra descritto si riallaccia a quello che, in letteratura, é l'ampio e complesso dibattito sugli atti comunicativi, introdotto per la prima volta da Austin nel 1962 (Austin (1962)); a lui infatti si deve la prima definizione di atto comunicativo: chi parla, secondo Austin, non si limita a "dire" qualcosa, ma "fa" qualcosa, in altre parole non si limita a pronunciare enunciati, ma compie azioni. Ogni enunciato infatti, oltre a significare qualcosa, esegue azioni (Levinson (1983), p.242), in virtù di quella che Austin stesso ama definire "forza specifica dell'enunciato":

Oltre alla questione, ampiamente studiata in passato, di cosa significhi un enunciato, ce n'è un'altra, di natura diversa, relativa a quale sia la forza, come la chiamiamo noi, di quell'enunciato. Può essere chiaro cosa significhi "chiudi la porta", ma può non esser chiaro se, pronunciata in un certo contesto, la frase sia un ordine, una supplica o altro. Oltre alla vecchia dottrina sugli enunciati ci occorre quindi una nuova dottrina sulla possibile forza degli enunciati.

(Austin (1962), p. 251)[7].

[7]Da notare che Austin non distingue fra concetto di frase (come struttura linguistica astratta) e enunciato (concretizzazione di una frase in un'istanza fisica)

Austin individua tre atti che avvengono simultaneamente nel momento in cui un parlante pronuncia un enunciato:

1. *atto locutorio:* il significato intrinseco dell'enunciato;
2. *atto illocutorio*: l'affermazione, l'offerta, la promessa, la richiesta... che viene veicolata dall'enunciato in virtù della forza che lega parole e azioni;
3. *atto perlocutorio*: la produzione di determinati effetti sull'ascoltatore.

La teoria degli atti linguistici si interessa soprattutto degli atti illocutori, tanto che, spesso, il termine atto linguistico, viene usato per riferirsi esclusivamente all'atto illocutorio.

Da Austin in poi, molti hanno elaborato diverse tassonomie di enunciati che veicolano diversi tipi di azioni, e qua si riporta la classificazione di Searle (1975); l'allievo di Austin distingue fra seguenti tipi di atti illocutori:

1. direttivi: tentativi del parlante di indurre l'ascoltatore a fare qualcosa (richiedere, domandare...);
2. rappresentativi: impegnano il parlante nei confronti della verità di una proposizione (asserire, concludere...);
3. commissivi: impegnano il parlante a fare qualcosa nel futuro (promettere, offrire...);
4. espressivi: esprimono uno stato psicologico (ringraziare, scusarsi);
5. dichiarativi: provocano cambiamenti immediati della realtà, spesso sono legati a complesse situazioni extralinguistiche (battezzare...);

Ognuno di questi tipi di enunciati é portatore di una forza illocutoria, convenzionalmente associatagli.

Come riassume bene (Levinson (1983), p. 243):

L'atto illocutorio è ciò che viene eseguito direttamente per tramite della forza convenzionale associata alla produzione di un certo tipo di enunciato secondo una procedura convenzionale: ne consegue che esso è (almeno in linea di principio) determinato.

Perché Lenvinson, però, si sente costretto ad attenuare la forza della definizione precedente con l'inciso "almeno in linea di principio"? Perché, sebbene si possa pensare, a questo punto, che ogni tipo di enunciato veicoli in modo determinato e non ambiguo uno e un solo atto illocutorio, in realtà questo non è sempre vero: vi sono situazioni in cui la forza illocutoria letterale di un enunciato non corrisponde alle reali intenzioni del parlante in quel contesto. Questo fenomeno, in letteratura, è definito *atto indiretto*, e si manifesta in enunciati come (1) discussi all'inizio di questa sezione.

Nonostante esista un rapporto preferenziale tra certi tipi di enunciati e certi atti illocutori (domanda-richiesta di informazione, imperativa-ordine, etc.), vi sono casi, gli atti indiretti appunto, in cui tale rapporto non viene rispettato, e il parlante decide di veicolare indirettamente le proprie intenzioni. In questi casi, l'enunciato, sembra avere, oltre alla forza letterale, una forza indiretta che deve essere inferita, così come deve essere inferita la reale intenzione comunicativa del parlante. Se infatti, negli atti indiretti ci si può avvalere del principio per cui ogni enunciato ha una forza illocutoria letterale che permette di inferire le intenzioni del parlante, gli atti indiretti costituiscono l'eccezione a tale principio. Venendo meno la corrispondenza enunciato-intenzione, si pone il problema di capire quale azione il parlante voglia perseguire con un certo enunciato.

Un esempio come (2) può essere significativo:

2 *Puoi chiudere la porta?*

Se (2) venisse espresso in una conversazione tra amici, probabilmente celerebbe, dietro alla struttura interrogativa di cortesia, un atto di "richiesta/ordine"; il parlante, infatti, non sarebbe, in questo caso, realmente interessato a sapere se l'ascoltatore sia potenzialmente capace di chiudere la porta, ma sarebbe interessato soltanto all'esito finale dell'atto (ottenere che qualcuno chiuda la porta). Se invece (2) venisse pronunciato da un medico che fosse interessato a valutare le capacità motorie di un paziente, allora si potrebbe pensare che l'atto illocutorio letteralmente veicolato da questa domanda corrisponda al reale obiettivo comunicativo del parlante.

Come si evince dal questo esempio, ciò che permette di disambiguare le intenzioni comunicative del parlante, in caso di atto indiretto, è il contesto in cui i due soggetti si trovano ad agire e l'insieme di conoscenze che condividono. Ma queste conoscenze sono disponibili solo nel caso di una conversazione reale uomo/uomo. Come è possibile invece gestire il problema degli atti indiretti nel caso di una conversazione uomo/macchina, che non può avvalersi dell'informazione pragmatica, contestuale, normalmente condivisa dai parlanti?

Il problema comincia ad emergere negli anni '70, quando gli studi sulla teoria degli atti comunicativi richiamano l'attenzione dei ricercatori che si occupano di IA(Jurafsky e Martin (2000), p. 728). Come identificare l'intenzione comunicativa di un enunciato in un contesto se non ci si può più basare sul principio di forza illocutoria letterale, indebolito dalla presenza di atti indiretti?

Per superare questo ostacolo, si sono formulate diverse ipotesi fra le quali due hanno dato origine a possibili approcci che si trovano agli estremi del concetto di idiomaticità: da un lato l'approccio idiomatico e dall'altro l'approccio inferenziale (Jurafsky e Martin (2000), pp.732-738).

Con il primo viene sfruttata la struttura idiomatica delle forme linguistiche usate per esprimere atti indiretti. Un esempio si trova nelle richieste formali come:

can you X...

could you X...

may you X...

Una grammatica inglese potrebbe ricreare un elenco di tutte le possibili varianti di queste forme e usarle come indicatori di una funzione di richiesta. Tale teoria, però, solleva molti problemi di ambiguità, perché non è detto che l'espressione usata in modo idiomatico (come "can you X"), non sia talvolta utilizzata in senso letterale (si veda l'esempio (2), nel caso di conversazione medico-paziente). Come afferma Levinson (Levinson (1983), p.274):

la teoria delle espressioni idiomatiche dovrebbe in realtà essere integrata da una potente teoria pragmatica che colmasse il vuoto tra ciò che si dice e ciò che si intende dire, che spiegasse cioè qual è l'interpretazione giusta da assegnare ad un determinato enunciato in un certo contesto.

L'altro estremo è invece caratterizzato da modello inferenziale che è stato proposto per la prima volta da Gordon e Lakoff (1971) e ripreso da Searle (1975). Esso prevede un processo di inferenze che l'ascoltatore (o la macchina) deve seguire per estrarre, dalla funzione illocutoria letterale dell'espressione, la funzione che effettivamente il parlante vorrebbe esprimere.

La catena di inferenze proposta da Searle prevede che l'ascoltatore:

1. riceva l'input e decodifichi la funzione illocutoria letterale,

2. realizzi che la funzione illocutoria letterale presenta delle incongruenze, rispetto al

contesto condiviso di conoscenze[8],

3. decida di cercare un'altra possibile funzione illocutoria che il parlante avrebbe potuto voler trasmettere,

4. sfrutti il contesto, la situazione, tutti gli strumenti extra linguistici che ha a disposizione, per capire quale sia l'intenzione comunicativa del parlante e, quindi, rispondere.

Questo approccio (modellato da Allen (1995)) è sicuramente molto potente e affascinante, ma nello stesso tempo estremamente fragile e difficile da realizzare poiché comporta la modellazione di un contesto di conoscenza condivisa (Jurafsky e Martin (2000), p.738). Dall'altra parte il sistema idiomatico, nonostante i suoi limiti, può essere una soluzione interessante, perché non comporta modellazione di un mondo di conoscenza condivisa. Ed è per questo motivo che il modulo di riscrittura qua proposto si rallaccia all' approccio idiomatico.

5.4.2 La soluzione proposta al problema degli atti indiretti: modulo di riscrittura

Come accennato precedentemente, una soluzione al problema degli atti indiretti può passare attraverso sistemi di semplificazione testuale, quindi di parafrasi automatica. Al tale fine si descrive di seguito il progetto per un modulo di riscrittura con cui modificare l'input da atto indiretto ad atto diretto.

[8]Searle per esempio afferma che ogni atto di *DOMANDA* ha come prerequisito fondamentale la "non conoscenza della risposta da parte del parlante". Se questo prerequisito non è rispettato (come nel caso delle strutture *can you X*) allora l'atto è viziato e non può essere l'interpretazione corretta. (Lenci (1992))

Un modulo, cioè, che ristrutturi sintatticamente l'enunciato dell'utente, tenendo invariato il significato. Come selezionare però la funzione illocutoria che l'atto indiretto può volere veicolare? Lavorando in un contesto di interrogazione a base di dati, e prendendo in considerazione solo un input costituito da domande, si può assumere come costante il fatto che ogni interrogazione veicoli una funzione di richiesta di informazione, senza dover affrontare alcun problema di ambiguità. Non è necessario quindi, in questo caso, colmare quel vuoto tra "ciò che si dice", e "ciò che si intende dire" (Levinson (1983), p. 274), perché si può sfruttare la certezza che la forza illocutoria, sottintesa al processo in corso, sia una richiesta di informazione.

Stabilito quindi che l'intenzione comunicativa dell'utente sia sempre quella di "ottenere informazioni", il modulo di riscrittura deve gestire l'input in modo da restituire domande dirette, senza strutture di cortesia come "can you please...", o strutture indirette come "I wonder if...". Un modulo di riscrittura consisterà in una grammatica di regole che, data in input una domanda dalla forza comunicativa ambigua, restituisca una frase sintatticamente ristrutturata in modo tale che veicoli la propria funzione comunicativa senza ambiguità.

Sia dato ad esempio un input come: *I don't know if this patient can take antibiotic*, (enunciato che può veicolare sia una asserzione, sia una richiesta di informazione), la grammatica restituirà in output una domanda disambiguata dal punto di vista della funzione comunicativa : *can this patient take antibiotic?* (enunciato che esplicita una richiesta di informazione).

Indicatori lessicali

Come riconoscere però un input che necessita la parafrasi da un input che non necessita tale trattamento?

E' interessante notare come vi siano delle classi verbali e delle strutture ricorrenti che possono essere sfruttate come indicatori lessicali. Il modulo infatti può avvalersi di informazione morfologica e lessicale che viene annotata ed esplicitata durante l'analisi linguistica.

Le strutture linguistiche che sono state prese in considerazione per il riconoscimento di atti indiretti sono soprattutto lessicali; si nota, infatti, come gli enunciati interrogativi indiretti siano introdotti da un insieme chiuso di predicati nella proposizione principale:

-predicati di richiesta ("to ask, to wonder", ma anche nomi che appartengono allo stesso campo semantico come "question"),

-predicati o perifrasi dubitative ("not to be sure", "to doubt"),

-predicati che indicano un atteggiamento mentale di conoscenza già acquisita ("to remember"),

-predicati indicanti decisione ("to decide"),

-predicati di non conoscenza ("to ignore").

La presenza di queste classi lessicali in una frase principale, seguite da congiunzioni subordinative interrogative come "if", "what", etc. può essere considerato un segnale forte dell'esistenza di una struttura indiretta.

Es: *I don't know what...*

Viceversa, la presenza di un elemento *wh* all'inizio della domanda è un indicatore di una forma interrogativa diretta di tipo wh-.[9]

[9] In realtà è teoricamente possibile trovare costruzioni sintattiche marcate dove la subordinata interrogativa è anticipata rispetto alla principale (come in "what I can do, I don't know"), quindi dove anche

Contesto in cui inserire il modulo

Volendo sfruttare informazioni linguistiche di cui sopra, è necessario inserire il modulo di riscrittura all'interno della catena di strumenti computazionali che annotano il testo con tali informazioni; per questo motivo una soluzione interessante sembra essere quella di interrompere il processo dopo l'analisi morfologica e la segmentazione in chunks[10] del testo, e inserire in quell'ambito il modulo di riscrittura, che, in questo modo, si va a posizionare tra l'analizzatore morfologico e il parser. Tale scelta ovviamente, se da un lato permette di sfruttare l'importante informazione ricavabile dalla segmentazione e dalla struttura morfologica, dall'altro costringe il modulo a restituire un output che sia formalmente compatibile con il parser; la riscrittura della frase non può quindi limitarsi alla parafrasi del testo semplice, ma deve prevedere anche una modifica delle strutture sintattiche, in modo che l'output possa essere re-immesso direttamente nel processo di analisi linguistica. In (3) si riporta l'input del modulo di riscrittura, cioè la query annotata morfologicamente e segmentata, utilizzando la catena di analizzatori C&C [11].

3.

I|PRP|-NP wonder |VBP|I-VP if |IN|I-SBAR the|DT|I-NP rapid|JJ

|I-NP turnover|NN|I-NP of|IN|I-PP bone|NN|I-NP could|MD|I-VP do|VB

|I-VP this|DT|I-NP.

In (3) è possibile distinguere l'annotazione che esplicita la categoria grammaticale e i

in caso di struttura indiretta, l'elemento wh si trova all'inizio del periodo; tali strutture però non sono state riscontrate nei corpora analizzati, e si può assumere che non rientrino nel linguaggio tipicamente usato per interrogazione a base di dati.

[10] Si definisce chunk un'unità sintattica di parole adiacenti analizzate morfologicamente.

[11] Si veda capitolo 1.

chunks, presenti nel testo: un termine come *I* viene completato con la categoria morfologica (*PRP, pronome personale*) e quindi con il chunk di appartenenza (*I-NP*).

Una struttura di questo tipo fornisce informazioni molto utili per l'individuazione dei suddetti indicatori lessicali che veicolano la riscrittura, ed è proprio per sfruttare questa informazione che è conveniente modificare l'input in questa fase del processo, prima che venga rielaborato dall'analizzatore sintattico.

Dato un input come (3), quindi, il modulo dovrà restituire in output una struttura come (4), che, pur presentano una frase sintatticamente diversa da quella di partenza, mantiene una struttura ben formata e adatta ad essere rielaborata dal parser.

4.

Could|MD|I-VP the|DT|I-NP rapid|JJ|I-NP turnover|NN|I-NP of|IN|I-PP bone|NN|I-NP do|VB|I-VP this|DT|I-NP.

Le regole

Il modulo di riscrittura proposto comprende una grammatica di regole che si occupano sia della selezione che della riscrittura effettiva dell'input, intervenendo attraverso spostamenti e modifiche sui chunks della domanda analizzata. Tali regole sono costituite da una fase riconoscimento di sequenza di chunks, una di test, con cui si verifica l'esistenza di alcuni indicatori lessicali (presenza di wh- o di *if* preceduto da verbi di conoscenza o di domanda), e una fase di "azione" che consiste nella riscrittura effettiva.

Per quanto concerne la struttura interna della regola, essa si basa sul binomio analisi e riscrittura, due fasi strettamente legate.

1. In fase di analisi la regola attraversa la domanda per estrarre, se presente, la posi-

zione di un indicatore lessicale, ad esempio un verbo di conoscenza (*I don't know*), verbo di domanda (*to wonder, to ask*), o un pronome *wh-* non in prima posizione.

2. In fase di riscrittura, qualora l'analisi confermi la presenza di una struttura indiretta, la regola modifica la query e restituisce una nuova query al parser.

Astraendo da ogni caso specifico, le regole possono essere espresse nella forma:

5.

(Regola 1)[12]

1) Pattern

2) test(condizioni)

3) {azione}

La formalizzazione in (*5*) può essere spiegata in dettaglio come:

1. Pattern: identifica un'espressione regolare su una sequenza di chunks che si vuole recuperare; ad esempio la sequenza *I-VP I-NP* seleziona un sintagma verbale seguito da un sintagma nominale.

2. Test: fase in cui si verificano condizioni lessicali o grammaticali particolari. In altre parole questa fase impone delle restrizioni ulteriori sui sintagmi che vengono recuperati: ad esempio se il pattern richiede una sequenza di *I-VP I-NP*, in fase di test, è possibile specificare ulteriormente che tale sintagma nominale sia un elemento wh-, o che il lemma di quel sintagma verbale appartenga ad una certa classe lessicale. Esempi di test possono essere:

[12]La numerazione ad ogni linea è volta esclusivamente ad agevolare la lettura e la spiegazione, ma non appartiene al formalismo adottato.

I-VP.categoria= MD : in cui si stabilisce che il predicato del sintagma verbale recuperato deve essere un modale.

I-NP.categoria=WP in cui si stabilisce che la categoria del sintagma nominale recuperato, deve essere WP, cioè un elemento wh-.

3. Azione: è la fase di modifica vera e propria dell'input. Una volta selezionata la frase da modificare, cioè una volta trovata una domanda che soddisfa la sequenza di pattern descritta nella regola e le condizioni specificate nella fase di test, la regola di riscrittura deve provvedere alla parafrasi; tale processo passa prima attraverso un momento di cancellazione della frase principale, e quindi di ristrutturazione in forma interrogativa della subordinata. I due momenti che si distinguono sono:

   1. cancellazione della principale

      6 *a. I wonder if a rapid turnover of the bone can do this.*

      *b. * a rapid turnover of the bone can do this.*

   2. ristrutturazione della subordinata secondo una corretta forma interrogativa, attraverso la strategia di inversione soggetto-verbo, o di inserimento di ausiliare interrogativo ("do/does").[13]

      c. can a rapid turnover of the bone do this?

Per un esempio di formalizzazione delle regole si rimanda all'Appendice.

[13] In fase di ristrutturazione è necessario tener presente la strutture sintattica inglese, la quale prevede una costruzione interrogativa del tipo *ausiliare+soggetto +verbo*. Qualora l'ausiliare sia esplicitato nella frase, l'interrogativa si otterrà grazie allo spostamento dell'ausiliare stesso in posizione pre- soggetto. Nel caso in cui non vi sia un ausiliare espresso, la struttura interrogativa si otterrà grazie all'inserimento dell'ausiliare interrogativo “do/does”.

La struttura della grammatica

Se fino a questo momento si è analizzata la struttura interna di una regola, si vanno ora a descrivere le classi in cui è possibile suddividerle. Ogni regola, infatti, propone pattern e test specifici in modo da coprire diverse situazioni di input, legate a diverse soluzioni di riscrittura, che vengono riassunte nella Tabella 5.1.

Tipo	**Ristrutturazione**	***Indiretta***	***Diretta***
Wh-soggetto	nullo	*I don't know what is causing fever.*	*What is causing fever?*
Wh	Inversione	*I don't know what you should do.*	*What should you do?*
	Inserimento	*I wonder what you want.*	*What do you want?*
Booleana	Inversione	*I wonder if you can come*	*Can you come?*
	Inserimento	*I don't know if he comes.*	*Does he come?*

Tabella 5.1: Classi di regole

In dettaglio:

Regole wh– soggetto: classe di regole che gestisce i casi in cui l'elemento interrogativo è un pronome che ha funzione di soggetto della frase. Tali situazioni vengono individuate grazie all'assenza di un pronome personale o di un sintagma nominale fra l'elemento *wh-* e il verbo della subordinata, (7).

7 *a. I don't know what is causing fever.*

b. what is causing fever?

La regola deve provvedere alla fase di cancellazione della principale, ma non è necessaria nessuna modifica ulteriore della subordinata.

Regole wh-inversione: classe di regole che gestisce i casi in cui sia presente un'interrogativa *wh-* indiretta, con un ausiliare o un modale espresso nel sintagma verbale della subordinata. La regola deve provvedere alla cancellazione della subordinata e allo spostamento dell'ausiliare in seconda posizione. [14]

8 a.*I don't know what you should do.*

b.what should you do?

Regole wh- inserimento: classe di regole che gestisce i casi in cui sia presente un'interrogativa wh- indiretta, senza ausiliare espresso nel sintagma verbale della subordinata. Le regole devono provvedere alla cancellazione della principale e all'inserimento dell'ausiliare in seconda posizione. Inoltre vanno a distinguere la persona del sintagma verbale grazie all'informazione morfologica sul verbo, per rispettare i vincoli di accordo con l'ausiliare ("does" per la terza persona singolare e "do" in ogni altro caso).

9 *a. I wonder what you want.*

b. What do you want?

c. I wonder what he wants.

d. What does he want?

[14]La struttura interrogativa dell'inglese, in situazione di domande wh prevede che la prima posizione sia solitamente riservata all'elemento wh. (Andreolli *e altri* (1996)). Di conseguenza, in questo caso, l'ausiliare dovrà occupare la posizione immediatamente successiva all'elemento wh-.

Regole - booleane inversione: classe di regole che gestisce i casi in cui sia presente un'interrogativa booleana, introdotta dalla congiunzione interrogativa "if", con un ausiliare o un modale espressi nel sintagma verbale della subordinata. Le regole deveno provvedere alla cancellazione della subordinata e allo spostamento dell'ausiliare in posizione iniziale.

10 a.*I wonder if you can come.*

b. Can you come?

Regole - booleane inserimento: classe di regole che gestisce i casi in cui sia presente un'interrogativa booleana, introdotta dalla congiunzione interrogativa "if". Come sopra, le regole devono provvedere alla cancellazione della subordinata, ma in fase di riscrittura devono gestire l'inserimento dell'ausiliare in posizione iniziale. Inoltre vanno a distinguere la persona del sintagma verbale grazie all'informazione morfologica sul verbo, per rispettare i vincoli di accordo con l'ausiliare.

11 *a. I wonder if he comes.*

b. does he come?

E' infine possibile individuare un altro gruppo di regole, che gestisce casi particolari come le strutture di cortesia (12).

12 *a.Can you please tell me what I can give to this patient?*

b.What can I give to this patient?

Per esemplificare i casi riscontrati nel corpus, che possono essere gestiti dal modulo di riscrittura, si riportano in Tabella 5.2 alcune proposizioni tratte dai corpora stessi.

Frase originaria	***Frase modificata***
I don't know what is causing fever	*What is causing fever?*
I dont' know what she has at this point.	*What does she have at this point?*
I wonder if it could be something else.	*Could it be something else?*
I wonder if this patient could have a rotator cuff.	*Could this patient have a rotator cuff?*
I would like to know what the natural history is	*What is the natural history?*
I asked if she needs a tuberculin skin test.	*Does she need a tuberculin skin test?*
Can anyone tell me what I can do with this patient?	*What can I do with this patient?*
Can you tell me if it is a vascular problem?	*Is it a vascular problem?*
I don't know if I should buy a new plug.	*Should I buy a new plug?*

Tabella 5.2: Esempi di parafrasi

5.4.3 Valutazione del prototipo

Allo scopo di concludere con lo studio di fattibilità del modulo di riscrittura proposta, si presentano i risultati di un' analisi valutativa, effettuata su una test suite di 60 domande. Tale test suite è stata ottenuta dalla selezione di 19 domande dirette e 41 domande indirette: le prime hanno lo scopo di valutare la capacità del sistema di riconoscere i casi da modificare, rispetto ai casi di atti diretti da lasciare invariati, mentre le seconde mirano a valutare la correttezza delle parafrasi. La maggior parte delle domande dirette e delle domande indirette presenti nella test suite sono state estratte dal corpus di Clinical questions. Una piccola parte invece è stata creata appositamente per valutare casi

particolari.

Oltre alla formalizzazione quindi è stato implementato un prototipo, in linguaggio Java, che ha permesso la valutazione di questa selezione di queries. Pur consapevoli che tale prototipo non è che il primo passo verso un trattamento più completo e dettagliato del problema dal punto di vista implementativo, è comunque un approccio utile per avere una metrica di valutazione e per far emergere alcune situazioni problematiche, che verranno descritte in seguito.

Su un corpus quindi di 60 domande, di cui 19 dirette e 41 indirette, il prototipo è in grado di riconoscere 31 delle 41 domande indirette. Di queste 31 domande però si ha una parafrasi corretta per 25 domande, mentre nei restanti 6 casi emergono problemi di riscrittura.

E' interessante utilizzare questi dati per calcolare le percentuali di precisione e completezza del sistema, dove per precisione si intende la percentuale dei casi gestiti correttamente, sul totale dei casi gestiti, mentre per completezza la percentuale dei casi gestiti sul totale dei casi che il sistema avrebbe dovuto gestire.

Precision = riscritture corrette/riscritture effettuate

Completezza = riscritture effettuate/totale riscritture da effettuare

In questo contesto il prototipo presenta una precisione dello 0.8 ed un recall dello 0.75.

Si riporta un esempio di parafrasi corretta: in (13a) la domanda originale, e in (13b) la domanda riscritta.

13.a

```
'I' 'PRP'| 'I-NP'

'wonder' 'VBP'| 'I-VP'

'if' 'IN'| 'I-SBAR'
```

```
'this' 'DT'| 'I-NP'
'patient' 'NN'| 'I-NP'
'could' 'MD'| 'I-VP'
'have' 'VB'| 'I-VP'
'a' 'DT'| 'I-NP'
'rotator' 'NN'| 'I-NP'
'cuff' 'NN'| 'I-NP'
'thing' 'NN'| 'I-NP'
```

13.b

```
'could' 'MD'| 'I-VP'
'this' 'DT'| 'I-NP'
'patient' 'NN'| 'I-NP'
'have' 'VB'| 'I-VP'
'a' 'DT'| 'I-NP'
'rotator' 'NN'| 'I-NP'
'cuff' 'NN'| 'I-NP'
'thing' 'NN'| 'I-NP'
```

Completezza

Si contano dieci casi di proposizioni indirette che non vengono gestiti dal prototipo. Le cause sono principalmente due:

1. Presenza di forme contratte che impediscono al parser di riconoscere e annotare correttamente la presenza di un verbo, come in (14) dove la forma contratta $I'm$ viene

annotata come sintagma nominale. Questi problemi peró potrebbero, in futuro, essere affontati, con un ampliamento della grammatica.

```
14

I\'mNNP|I-NP  notRB|O sureJJ|I-ADJP  if IN|I-SBAR  itPRP|I-NP  is VBZ|
I-VP  stillRB|I-ADJP currentJJ|I-NP  practice NN|I-NP.
```

2. Presenza di periodi troppo lunghi e complessi, con riferimenti anaforici, come in (15). Tali situazioni non vengono gestite dal prototipo, ma possono essere punto di partenza per una riflessione futura (in (15) compaiono anche problemi di annotazione).

   15.

   Mine antispam filter blocks e-mail that aren't spam and I don't know how to solve this problem.

```
'Mine' 'NNP'| 'I-NP''antispam' 'JJ'| 'I-NP' 'filter' 'NN'|
'I-NP' 'block' 'NNS'| 'I-NP'  'email' 'VBZ'| 'I-VP'  'that'
'IN'| 'I-SBAR' 'aren\'t' 'NN'| 'I-NP'  'spam' 'NN'| 'I-NP'
'and' 'CC'| 'I-NP' 'i' 'NN'| 'I-NP'  'don\'t' 'NN'| 'I-NP'
'know' 'VBP'| 'I-VP' 'how' 'WRB'| 'I-ADVP'  'to' 'TO'| 'I-
VP'  'solve' 'VB'| 'I-VP' 'this' 'DT'| 'I-NP'  'problem'
'NN'| 'I-NP'
```

Precisione

Emergono sei casi di proposizioni non correttamente parafrasate dal prototipo. Di queste tre sono ancora dovute a problemi di parsing, mentre le restanti mostrano problemi nello

stabilire l'accordo temporale fra ausiliare e sintagma verbale. Il prototipo infatti riesce a gestire domande con sintagma verbale della subordinata al tempo presente, ma non è in grado di riconoscere forme passate: di inserire quindi, quando necessario, l'ausiliare *did,* e di modificare la forma verbale al presente. Situazioni come l'esempio (16), non sono quindi gestite dal sistema.

16 . *I don't know if you went there.*

Il prototipo non è in grado di riconoscere la forma passata e di inserire l'ausiliare corretto, trasformando la frase in "did you go?", ma produce un output scorretto "do you went?". La gestione dei sintagmi verbali in tempo passato non emerge dagli esempi tratti dai corpora, per questo non rientra in questa prima formalizzazione di regole di riscrittura, ma è emersa ad una succesiva riflessione sulle possibili situazioni di riscrittura da gestire. Sicuramente può essere un interessante punto di partenza per lavori futuri.

5.5 Conclusioni

Si è qua analizzata la distanza fra il linguaggio naturale e il linguaggio di interrogazione all'ontologia.

Se l'analisi del Capitolo 4 aveva portato risultati incoraggianti, che suggerivano la strada dell'interazione con basi di conoscenza attraverso un input in linguaggio naturale non controllato, nella sezione precedente si sono affrontante in dettaglio le problematiche emerse da tale analisi: problemi di gestione di alcuni costruttori logici e problema dei gestione degli atti indiretti. Per il primo si è descritto un approccio di indebolimento semantico (Carbotta e Calvanese (2007)), mentre per il secondo si è proposto un approccio sul lato linguistico che provvede ad una riscrittura dell'input indiretto in input diretto. Le

soluzioni presentate mostrano come l'interrogazione in linguaggio naturale ad ontologie sia possibile e sia una strada che può aprire prospettive interessanti, che coinvolgono in parallelo il mondo della logica e della linguistica.

Fra i futuri lavori si potrebbe immaginare l'ampliamento dell'indebolimento semantico all'operatore di negazione, e un'implementazione completa ed approfondita del modulo di riscrittura sintattica, ampliando il numero di casi indiretti trattati. Si è inoltre consapevoli che, al fine di avere un riscontro concreto sul livello di naturalezza delle riscritture proposte, sarebbe interessante sottoporre tali parafrasi alla valutazione di parlanti nativi. Anche questo aspetto, sarà oggetto di future ricerche e sprimentazioni.

Capitolo 6

Dal linguaggio naturale a DL-Lite

6.1 Verso Dl-lite

In quest'ultima parte del lavoro si vuole spostare l'attenzione sull'altro aspetto di interazione con l'ontologia, l'arricchimento di conoscenza. Se infatti, fino a questo momento, si è approfondito il rapporto fra il linguaggio naturale e il linguaggio di interrogazione dell'ontologia, adesso è interessante concludere con un'ultima panoramica sul rapporto tra il linguaggio naturale e il linguaggio di definizione della base di conoscenza intensionale, che, in questo contesto, consiste nel frammento di logica descrittiva che va sotto il nome di $DL\text{-}Lite_{core}$.

L'obiettivo è quello di indagare le potenzialità di tale frammento di logica nel supportare un eventuale processo di arricchimento automatico di conoscenza da corpora in linguaggio naturale o da input in linguaggio naturale. Si possono, infatti, immaginare due scenari di applicazione, che vanno dall'inserimento di informazione in linguaggio naturale da parte di un utente umano, all'arricchimento automatico della base di conoscenza attraverso

corpora specifici[1]. Perché questo sia possibile è però necessario che il linguaggio logico riesca a esprimere i costrutti più tipici del linguaggio naturale peculiare di questo tipo di corpora. In altre parole un linguaggio naturale che presenti caratteristiche simili alle asserzioni della T Box, costituita da enunciati come in (1) e in (2), dove non si definiscono fatti su singole istanze, ma asserzioni su intere classi.

1 *a. Every Student is a Person*

b. Every student can borrow something that is a book.

Tali enunciati definiscono le "regole del mondo" rappresentato dall'ontologia[2].

2 *a. External users can access the Internet with their library card.*

b. Applicants must be abide by the Industrial Design laws.

Il focus dell'analisi è indagare se, e in che percentuale, sia possibile sfruttare questo tipo di asserzioni per l'inserimento di conoscenza in un'ontologia, descritta in *DL-Lite*$_{core}$. Si è proceduto quindi ad un'analisi quantitativa sull'intero corpus di asserzioni (si veda Capitolo 4) per l'estrazione di dati riguardanti la frequenza di costrutti logici non supportati da *DL-Lite*$_{core}$, e ad una analisi manuale, più approfondita su un terzo del corpus, indirizzata allo studio di quelle strutture sintattiche problematiche che non potevano emergere dall'analisi automatica. In questa sezione si descrive, quindi, la seconda parte dell'analisi, condotta manualmente, su un sottoinsieme delle asserzioni. In questo modo si vuole dare una visione d'insieme sulle potenzialità linguistiche di questo frammento di logica.

Come vedremo, il risultato emerso dall'analisi mostra i limiti di tale frammento, che, in una eventuale implementazione di un sistema di arricchimento automatico di conoscenza,

[1] Ad esempio, un documento che specifica l'organizzazione e i regolamenti di una biblioteca può essere una fonte di conoscenza perfetta per la costruzione della T-Box di un'ontologia di dominio bibliotecario.

[2] Per la scelta delle asserzioni e dei corpora da analizzare si veda Capitolo 4, paragrafo 4.3

riuscirebbe a gestire, nel migliore dei casi[3], soltanto il 10% delle asserzioni analizzate. Come procedere quindi per cercare di far incontrare, ancora una volta, linguaggio logico e linguaggio naturale? Un approccio possibile consiste nell'uso di un frammento di logica leggermente più ampio di *DL-Lite*$_{core}$, *DL-Lite*$_R$ che, pur mantenendo costi computazionali contenuti (si veda Capitolo 2, paragrafo 2.2.2), copre una sezione di inglese più vasta. Tale frammento risulta capace di gestire il 50% dell'input. Nonostante questo è evidente che ancora una metà dell'input, e quindi metà dell'informazione utile da inserire nella base di conoscenza rimane in una "zona d'ombra" che nessuno dei frammenti analizzati è in grado di gestire. In altre parole, pur utilizzando un frammento di DL-Lite abbastanza espressivo emerge che solo metà dell'input è semanticamente compatibile con esso, mentre l'altra metà rimane al di fuori del frammento condiviso. Inoltre, anche la parte di input semanticamente gestibile non è immediatamente proiettabile sul linguaggio logico, ma necessita di una riscrittura sintattica che costruisca enunciati come in (1).

Si aprono quindi due scenari:

1- da una lato la necessità di riscrivere in una sintassi gestita dalla logica, l'input semanticamente accettato. In 6.4 si propone infatti un modulo di semplificazione testuale per tale porzione di input.

2- Dall'altra la necessità di analizzare la porzione di input non semanticamente accettato dalla logica, per capire in che misura e in che direzione la logica stessa potrebbe essere modificata al fine di raggiungere una maggiore espressività (6.5).

[3]Senza calcolare i problemi derivabili da rumori del corpus o da errori di parsing.

6.2 Frammento di logica *DL-Lite*$_{core}$

Come precedentemente accennato nel Capitolo *2*, tutti gli assiomi della base di conoscenza intensionale dell'ontologia sono costituiti da set di asserzioni universali che si esprimono nella forma di: $Cl \subset Cr$, dove Cl è detto "contesto sinistro" e Cr il "contesto destro" (Calvanese, 2006). Tali contesti, come anticipato in 2.2.3, sono definiti rispettivamente da:

1.

$Cl \rightarrow A|\exists R$, e $Cr \rightarrow A|\neg A|\exists R|\neg\exists R$

dove: A indica un concetto atomico, e $\exists R$ un quantificatore esistenziale non qualificato[4], e la relazione di inclusione fra questi due contesti può essere riscritta nella formula FOL : $\forall x.Cl(x) \rightarrow Cr(x)$

notazione con cui si mette in evidenza la natura universale di tali assiomi. Dal punto di vista linguistico, l'universalità viene ad esprimersi attraverso l'uso del quantificatore *everyone* e del determinatore *every*, elementi essenziali di ogni assioma.

Quindi i due contesti andranno a delinearsi nelle seguenti strutture sintattiche:

a. [Every N] SINTAGMA VERBALE

Cl Cr

b. [[Everyone [who SINTAGMA_VERBALE]] SINTAGMA_VERBALE]

Cl Cr

dove, come detto sopra, il quantificatore "everyone" e il determinatore "every" sono

[4]Per la distinzione fra quantificatore esistenziale qualificato e non qualificato si rimanda al Capitolo 2, par. 2.2.1.

parte integrante e fondamentale della sintassi della frase, mentre il contesto sinistro e destro possono rispettivamente essere:

Cl: N | who SINTAGMA VERBALE

Cr: SINTAGMA VERBALE.

Scendendo in maggiore dettaglio, un concetto atomico, che deve essere un predicato unario, può essere espresso con un N (sostantivo) come in (3.a), o con un sintagma verbale con verbo intransitivo (3.b), ma mai con un verbo transitivo con argomento o un sintagma nominale complesso (es: *every nice student*).

3 *a.Every* ***student*** *is a boy*

b Every student ***left***

Una quantificazione esistenziale non qualificata invece potrà essere realizzata attraverso il quantificatore *everyone* seguito dal pronome relativo *who* e da un verbo transitivo (3.c) oppure da un verbo transitivo seguito dal quantificatore esistenziale *something* (3.d), non ulteriormente specificato.

c. Everyone who knows something is a boy

d. Everyone who knows something left.

Nella Tabella 6.1 sono riassunti tutti i costrutti che corrispondono al concetto atomico e al quantificatore esistenziale non qualificato:

Dato $Cr \rightarrow A|\neg A|\exists R|\neg\exists R$, ne deriva che il contesto destro accetta le stesse strutture linguistiche del contesto sinistro, ma anche le negazioni; è quindi possibile esprimere in Cr $\neg A$ come in (3.e) e (3.f), ma anche $\neg\exists R$, come in (3.g).

e. Every student is not a boy

f. Every student does not leave

A	N	*Student*
	SV(intransitivo)	*Left*
$\exists R$	Everyone who SV (transitivo)	*knows something*

Tabella 6.1: Strutture linguistiche corrispondenti a concetto atomico ed esitenziale non qualificato.

g. Every student does not know something.

Oltre ai costrutti logici qua indicati è inoltre possibile inserire l'operatore di disgiunzione in contesto sinistro e di coordinazione in contesto destro, senza perdita di efficienza sul lato logico[5].

Presa in considerazione la classe chiusa dei costrutti finora analizzati, si può affermare che la versione *core* di DL-Lite non accetta le strutture linguistiche presenti in Tabella 6.2.

Alla luce di queste considerazioni si è condotta un'analisi specifica della copertura del frammento su una sezione del corpus (descritto nel Capitolo 4) di 200 asserzioni. [6]

[5]La specifica dell'operatore AND in contesto destro, e dell'operatore OR in contesto sinistro, non costituiscono un problema dal punto di vista dei costi computazionali. (Calvanese *e altri* (2006)). Ma sono di grande aiuto dal punto di vista linguistico perché permettono di ampliare notevolmente la gamma dei costrutti linguistici permessi:

1. *Everyone who is a student or a professor is a Person.*
2. *Everyone who is a student can access the library and can access the lab.*

[6]Se nel Capitolo 4 è stata condotta un'analisi attraverso l'uso di di programmi statistici (*Simple concordance Program),* in questa sezione vengono, invece, descritti i risultati di un'analisi manuale. Si rimanda quindi al Capitolo 4 per ogni dettaglio circa la costruzione dei corpora e la scelta delle asserzioni. Si è scelto di posticipare la descrizione di questa parte dell'analisi, poiché presuppone la conoscenza di

Costruttori problematici per *DL-Lite*$_{core}$	
Contesto sinistro	Negazione Sintagma nominale complesso (“*every nice student*” / “*every student who...*”) Predicato con argomento specificato
Contesto destro	Predicato con argomento specificato. Disgiunzione

Tabella 6.2: Costrutti problematici in *DL-Lite*$_{core}$

In dettaglio, emerge che *DL-Lite*$_{core}$ riuscirebbe a rappresentare semanticamente, nel migliore dei casi, il 10% delle asserzioni, mentre il restante 90% rimarrebbe al di fuori delle sue potenzialità a causa della presenza di strutture non ammesse. In particolare sono state riscontrate: un 5% di frasi negative, un 17% di disgiuntive e un 15% di asserzioni che presentano un sintagma nominale complesso in contesto sinistro. Ma, soprattutto, il 66% delle proposizioni presenti necessita la specifica dell'argomento in contesto destro o in contesto sinistro (entrambi casi non ammessi da *DL-Lite*$_{core}$). Tutti i dati sono riassunti in Tabella 6.3.

dettagli circa i linguaggi logici che viene fornita in questa sezione.

Costrutto	**Percentuale su 200 enunciati**
Gestibili(core)	9,90%
Argomento sinistra	30,00%
Argomento destra	33,00%
N complesso	15,00%
Disgiunzione	17,00%
Negazione sinistra	5.4%

Tabella 6.3: Percentuali di presenza delle strutture problematiche per *DL-Lite*$_{core}$ e *DL-Lite*$_R$

6.2.1 Enunciati che appartengono all'espressività del frammento *DL-Lite*$_{core}$

L'impossibilità di specificare l'argomento nel contesto destro, evidentemente, limita l'espressività del frammento, e possono essere espresse solo asserzioni come (4.a) o (4.b), equivalenti a (4.c)(4.d).

4.

a. All services are integrated at both sites.

b. Patents are legal recognitions.

c. Every service is integrated at both sites.

d. Every patent is a legal_ recognition[7]

Rimangono però escluse tutte le asserzioni più complesse che necessitano della specificazione del quantificatore esistenziale qualificato in contesto sinistro (4.e) o in contesto

[7] Assumendo che "legal_recognition" definisca un concetto unico.

destro (4.f).

e. External users who possess a Library Card may use the Interlibrary Loan service.

f. Inventions must show inventive ingenuity.

La versione core di DL Lite riesce, in altre parole, a gestire efficacemente predicati nominali con verbo essere e un predicativo, ma non risulta sufficiente in predicati verbali, con verbi bivalenti, verbi cioè che ammettono due argomenti. Oltre a predicati nominali, il concetto atomico del contesto destro può essere realizzato da un predicato intransitivo, come in *"Every plane flies"*. Questo tipo di struttura però non risulta dai corpora analizzati, dove ogni occorrenza di un predicato verbale è realizzata da un predicato transitivo con argomento specificato. Quindi si trovano molte espressioni del tipo (4.e) o (4.f), ma nessun verbo intransitivo capace di veicolare un'informazione rilevante per la Tbox, sul 90% dei predicati verbali presenti. D'altra parte, la versione *core* di DL-Lite permette di definire relazioni del tipo: *Every x is y* come in (4.d) *every patent is a legal recognition,* dove si stabilisce l'assioma per cui ogni istanza della classe *Patent,* è anche istanza della classe *Legal_ recognition.*

La percentuale di asserzioni gestibili rimane piuttosto bassa, e, per questo, è difficile immaginare un sistema di arricchimento automatico della conoscenza in un'ontologia formalizzata in *DL-Lite*$_{core}$.

6.3 Il frammento logico *DL-Lite*$_R$

La versione *DL-Lite*$_R$, leggermente più estesa (Calvanese *e altri* (2006)), consente di ampliare le strutture linguistiche esprimibili, attraverso l'inserimento dell'operatore di congiunzione nel contesto sinistro, che quindi diventa $Cl \rightarrow Cl_1 \cap Cl_2$. Linguisticamente

questo ampliamento si concretizza nella possibilità di utilizzare sintagmi complessi in contesto sinistro come in (5.a),(5.b), (5.c).

5.

a. Every nice student left.

b. Every student who studies left.

c. Everyone who drinks something left.

E' inoltre possibile, come prima, inserire, senza perdita di efficienza, l'operatore OR in contesto sinistro e costruire sintagmi come (5.d).

d. Every canadian or american student speaks english.

Nel contesto destro è, invece, possibile, in questa versione, specificare il quantificatore esistenziale, quindi qualificare l'argomento del verbo che rappresenta la relazione attraverso il determinatore "a", come in (5.e), ma rimane esclusa la possibilità di specificare l'argomento di un verbo transitivo in contesto sinistro (5.f). E' invece possibile coordinare con l'operatore AND più sintagmi in contesto destro (5g).

e. Every student knows a girl.

*f. *Everyone who eats something that is an apple left.*

g. Everyone who eats leaves and goes away.

Riassumendo, dato l'insieme degli operatori descritti, i costrutti non supportati da *DL-Lite*$_R$ sono riportati in Tabella 6.4.

A differenza di quanto accade nella versione *core*, quindi, il grande vantaggio, in termini linguistici, di questo frammento di logica è quello di permettere la specifica dell'argomento

Costruttori non permessi da DL Lite R	
Contesto sinistro	Negazione. Predicato con argomento specificato.
Contesto destro	Disgiunzione.

Tabella 6.4: Costrutti problematici per *DL-Lite*$_R$

del verbo del contesto destro fattore che, da solo, porta ad un incremento del 25% del numero di asserzioni gestibili[8]. Inoltre la possibilità di esprimere sintagmi nominali complessi in contesto sinistro accresce ancora del 15% il totale delle frasi gestibili.

In generale l'estensione del frammento permette di accresce più del 40% la copertura sul corpus, soprattutto grazie alla possibilità di specificare l'argomento del contesto destro. Ne deriva che frasi come (6.a) e (6.b), prima non trattabili, divengono invece esprimibili in *DL-Lite*$_R$

6 *a. Inventions must show inventive ingenuity.*

b. The three-dimensional configuration is a "topography."

c. Every invention must show something that is inventive_ingenuity.

d. Every configuration that is three-dimentional is a "topography".[9]

[8]Non si ha un incremento del 33% come sembrerebbe dalla tabella 24 perché sono gestibili solo argomenti unari (si veda di seguito).

[9]In (6.a) e (6.c) si può notare la presenza dell'argomento del verbo i contesto destro (inventive_ingenuity), mentre in (6.b) e (6.d), si ha un esempio di sintagma nominale complesso formato dal sostantivo "configuration", e dall'aggettivo "three-dimentional", struttura che non sarebbe permessa se non fosse stato introdotto l'operatore di intersezione insiemistica nel contesto sinistro.

Da notare inoltre che, sebbene questa versione di DL-Lite permetta di specificare l'argomento nel contesto destro (6.a), tale argomento deve essere atomico. Non è quindi possibile fare specifiche ulteriori sull'argomento creando una doppia relazione come in:

(6.e)You must provide a primary document that proves the identity,

dove, dato $x$come l'utente, si avrà:

$Provide(x, y) \wedge Document(y) \wedge Prove(y, z) \wedge Identity(z)$

L'istanza di *Document*, in questo contesto, non è atomica, perché coinvolta in due relazioni. Ne deriva che non tutte le asserzioni con un argomento nel contesto destro saranno esprimibili in *DL-Lite*$_R$, ma solo quelle che non presentano ulteriori specificazioni di tale argomento[10]. Linguisticamente, situazioni di doppia relazione su un argomento si possono realizzare attraverso: subordinate relative (7.a), sintagmi preposizionali (7.b), pronomi possessivi (7c).

7 *a. You must provide a document that proves the identity.*

b. You must provide a document with a signature.

c. You must use your card.

Dall'analisi emerge che sul 33% delle asserzioni che presentano la necessità di qualificare l'argomento del contesto destro, solo il 25% è esprimibile in *DL-Lite*$_R$, mentre le restanti presentano strutture come in (7). Nonostante questo limite, il 50.35 % delle asserzioni in linguaggio naturale rientra nel frammento di inglese esprimibile con questa versione di DL-Lite.

[10]Sebbene nella definizione del linguaggio *DL-Lite*$_R$ si imponga una restrizione sulla qualifica dell'esistenziale a sole formule atomiche, si potrebbero accettare anche formule più complesse, vale a dire del tipo $\exists RCr$, perché potrebbero essere riscritte in formule accettabili. In questo contesto però ci si attiene alla definizione formale.

In tabella 6.5 si mettono a confronto le espressività delle due versioni di DL-Lite.

Struttura linguistica	**%**	**Core**	**DL-Lite R**
Gestibili(core)	9,90%	OK	OK
Argomento sinistra	30,00%	NO	NO
Argomento destra atomico	25,45%	NO	OK
Argomento destra non atomico	7.5%	NO	NO
N complesso	15,00%	NO	OK
Disgiunzione	17,00%	NO	NO
Negazione sinistra	5.4%	NO	NO
Totale espressività		9,90%	50.35%

Tabella 6.5: Confronto copertura *DL-Lite*$_R$ e *DL-Lite*$_{core}$

Ne deriva che il 50.35% degli enunciati, semanticamente non problematici, potrebbero essere oggetto di semplificazione testuale, per essere parafrasati in sintassi compatibili con *DL-Lite*$_R$.Come suddetto, si aprono due strade: da un lato lo studio di un modulo di riscrittura che permetta di trattare la porzione di input semanticamente compatibile (il 50.35% di cui sopra). Dall'altra lo studio della porzione di input non compatibile, e delle sue caratteristiche linguistiche, per un ulteriore ampliamento dell'espressività della logica.

6.4 La semplificazione testuale per la porzione di input semanticamente compatibile con *DL-Lite*$_R$

Come detto sopra, anche la porzione di input in linguaggio naturale che risulta semanticamente gestibile attraverso *DL-Lite*$_R$ non è immediatamente proiettabile sul frammento di logica ma necessita di una semplificazione sintattica che ristrutturi la frase in modo da renderla lessicalmente e sintatticamente accettabile dal frammento di inglese controllato dal linguaggio logico. In altre parole un input come (1), semanticamente compatibile con *DL-Lite*$_R$ non è ancora sintatticamente compatibile con esso, ed ha quindi bisogno di un ulteriore passaggio che lo parafrasi in una struttura come (2).

1. *Inventions must show inventive.*

2. *Every invention must show something that is inventive.*

Per gestire questo passaggio si propone in questa sezione una formalizzazione di regole di riscrittura che riprende la formalizzazione prevista per la riscrittura degli atti indiretti (si rimanda alla sezione 5.4.1). Tali regole possono andare a formare una grammatica che sia il nucleo fondamentale di un modulo di riscrittura delle asserzioni. Per dettagli sulla progettazione e sulla contestualizzazione in un processo di analisi linguistica di tale modulo si rimanda al Capitolo 5. Anche in questo caso, infatti, si può immaginare di inserire il momento di riscrittura fra l'analizzatore morfologico e il parser, in modo tale che la modifica avvenga su un enunciato annotato e disambiguato morfologicamente.

6.4.1 Strutture linguistiche da gestire

La formalizzazione di regole di riscrittura deve partire dall'analisi dell'input e delle strutture linguistiche che necessitano di una riscrittura effettiva. Da un lato è necessario modificare l'input e renderlo compatibile con le strutture sintattiche e lessicali ammesse nel contesto sinistro, e dall'altro con quelle ammesse nel contesto destro.

Focalizzando prima di tutto l'attenzione sul contesto sinistro, si richiamano in (3) i termini e le costruzioni ammesse.

3. *Every + sintagma nominale*

   Everyone who + sintagma verbale.

Ma quali sono le strutture con cui si presentano le asserzioni del corpus? Facendo riferimento al contesto sinistro, è possibile distinguere tre classi di asserzioni.

1. Asserzioni introdotte da sostantivi plurali.[11]

   Il sintagma nominale che introduce la frase contiene un sostantivo plurale, non preceduto da alcun determinatore, come in (4). Si può assumere che tale sostantivo indichi una classe di oggetti, quindi un concetto dell'ontologia, su cui è possibile asserire qualcosa.

   4.

   a. *Babies are allowed one bag.*

   b. Every baby is allowed one bag.

[11] I sostantivi plurali inglesi, infatti, possono, a differenza dell'italiano, esprimere genericità; una genericità che può essere associata al concetto di universalità.

2. Asserzioni introdotte da sostantivi plurali preceduti da un determinatore universale diverso da *every.*

   Il sintagma nominale che introduce l'assioma è formato da un determinatore come *all,* o *each* seguito rispettivamente da un sostantivo plurale o singolare. In entrambi i casi l'enunciato si riferisce ad un insieme su cui viene dichiarato l'assioma.[12]

   5.

   a. All pushchairs must be x-ray screened.

   b. Every pushchair must be x-ray screened.

3. Asserzioni introdotte da una costruzione impersonale con *one* o con *you* impersonale, come in 6.

   6.

   a. you can obtain an application form.

   b. one must fill the application form.

Tali strutture impersonali, soprattutto la forma realizzata attraverso il pronome di seconda persona singolare è usata frequentemente nel linguaggio tipico della linee guida[13], e, si può assumere che, in ogni dominio, si riferisca, all'insieme degli utenti del servizio che viene descritto. In altre parole, nel caso delle linee guida del governo canadese si assume che il soggetto di un enunciato come *(6a)* sia l'insieme dei cittadini candesi, nel caso di un'ontologia che descriva un sistema bibliotecario si può assumere che l'utente di riferimento sia lo studente che deve usufruire del servizio, e infine, nelle linee guida di una

[12] In questo caso, un'altra soluzione potrebbe consistere nell'associare *each* e *all* alla stessa entrata lessicale di *every*, vincolandosi peró all'uso di uno specifico dizionario di entrate lessicali.

[13] Il 33% delle asserzioni presenta una struttura del genere.

compagnia aerea si può assumere che il soggetto sia il passeggero tipo. In ogni dominio è quindi possibile recuperare una classe di oggetti a cui associare il riferimento di questa costruzione indiretta, e si può quindi assumere di esplicitarlo in un assioma, riscrivendo 6 in 7.

7.

a. Every citizen can obtain something that is an application form.

E' interessante notare come tale struttura ricorra sempre insieme ad una costruzione con verbo modale, perché in ogni espressione si vuole veicolare un obbligo o una possibilità dell'utente.[14]

Passando invece ad analizzare il contesto destro, esso può accettare sintagmi verbali con argomento specificato, ma tale argomento deve essere necessariamente strutturato nell'esistenziale *something* specificato dalla relativa *that is,* come emerge in (7). Questo argomento può essere sia un complemento oggetto diretto che indiretto, e, per questo motivo, possono emerge due diverse situazioni:

1. Sintagma verbale con complemento oggetto diretto espresso.

   8.

   a. You must provide a personal document.

   b. Every citizen must provide something that is a personal document.

[14] Si riportano alcuni esempi esplicativi:

You can pre-arrange this through BA World Cargo.

You will need to arrange delivery and collection of your bags.

You should contact the responsible foreign government office.

You must provide an original of a primary document.

2. Sintagma nominale con complemento indiretto espresso.

9.

a. Passengers must comply with UK Department for Transport restrictions.

b. Every passenger must comply with something that is UK Department for Transport restrictions.

E' interessante notare, infine, come il 32% delle asserzioni selezionate presenti strutture finali del tipo:

-In order to X, one must y.

-To X, one must y.

-If you want X, you must y.

Tale frequenza è interessante perché rappresenta una percentuale elevata del corpus, e si trova in tutti e tre tipi di corpora (dalle risposte a FAQ, a linee guida), ma d'altra parte, non emergono casi in cui la parafrasi di una struttura del genere non comporti la presenza di un argomento specificato in contesto sinistro, che, quindi, rende l'enunciato ingestibile dal punto di vista semantico.

9.

a. In order to find an article one has to conduct a search in a database.

b. Everyone who to find something that is an article, has to conduct something that is a search and is in a database.[15]

[15]Questo enunciato non è gestibile in *DL-Lite*$_R$a causa della presenza di un argomento specificato in contesto sinistro (that is an article).Ogni asserzione finale, se riscritta in modo da essere più vicina alle strutture gestibili dalla logica, presenta un argomento in Cl.

6.4.2 Le regole

Per arrivare alla riscrittura di tali enunciati si propone un modulo che si basi su formalismi come quelli analizzati nel Capitolo 5. Anche in questo caso, quindi viene adottata la strategia della semplificazione testuale, attraverso una parafrasi che semplifichi l'input e lo renda compatibile con la sintassi *DL-Lite*$_R$[16]. Tale modulo quindi deve prendere in input un testo segmentato e annotato morfologicamente e restituirlo in un formato, tale che sia reinseribile in un processo di parsing automatico a seguito di una parafrasi basata su una grammatica di regole di riscrittura. Le regole a cui si fa riferimento presentano la stessa formalizzazione di quelle viste in 5.4, e di seguito si richiama brevemente la loro struttura.[17]

10.

(Regola 1)

1) Pattern

2) *test(condizioni)*

3) *{ azione};*

1. Pattern. Identifica la sequenza di chunks da recuperare. La sequenza $I-NP I-VP, per esempio, indica la necessità di recuperare un sintagma nominale seguito da un sintagma verbale ad inizio di proposizione.

2. Test: stabilisce ulteriori restrizioni sui chunks recuperati, restrizioni di tipo morfologico o lessicale, come:

   I-NP.lemma= "you",

[16]Si ricorda che l'input a cui si fa riferimento in questo paragrafo è l'input semanticamente compatibile con *DL-Lite*$_R$, che necessita solo di una riformulazione sintattica.

[17]Per maggiori dettagli si rimanda alla sezione 5.4.4

I-VP.categoria= MD, (modale).

3. Azione: fase di modifica vera e propria. La sequenza di sintagmi che corrisponde al pattern, e che rispetta le condizioni specificate nel test, viene modificata in questa fase, con procedimenti diversi (dall'inserimento di termini chiave come "every", alla modifica di sintagmi verbali).

   Ad esempio:

   $I-NP I-NP'

   test (I-NP.lemma= "each"

   I-NP'.cat= NN)

   action { replace(I-NP, every)}[18]

6.4.3 Struttura della grammatica

Una volta richiamato il formalismo scelto per le regole di riscrittura, si va ora ad analizzare il tipo di classi di regole che è possibile distinguere per gestire le varie situazioni linguistiche. Innanzi tutto è necessario dividere le regole per la riscrittura del contesto sinistro e quelle per la riscrittura del contesto destro.

Contesto sinistro:

Si possono distinguere:

1. Asserzioni introdotte da NNS, cioè da sostantivi plurali che indicano una classe di oggetti. La regola deve sostituire il sostantivo ad inizio frase, con la costruzione *every+NN,* cioè determinatore obbligatorio *every* e sostantivo singolare. Inoltre deve

[18]Si formalizza la sostituzione del determinatore "each", con il determinatore "every", dove "every" che viene specificato in action rappresenta il lemma con l'annotazione morfologica.

Asserzione introdotta da NNS	*Babies are allowed one bag*	*Every baby is allowed one bag*
Asserzione introdotta da determinatore *all* + NNS	*All pushchairs must be x-ray screened.*	*Every pushchair must be x-ray screened.*
Asserzione introdotta da determinatore each + NN	*Each service is integrated at both sites*	*Every service is integrated at both sites*
Asserzione introdotta da "you" impersonale	*you must be able to lift your bag unaided*	*Every passenger must be able to lift his bag unaided*
Asserzione introdotta da "one" impersonale.	*One must fill the application form*	*Every citizen must fill the application form.*

Tabella 6.6: Classi di regole per il contesto sinistro

preoccuparsi di modificare il predicato verbale per mantenere l'accordo numerico, quando necessario (se è presente un modale, questa modifica non sarà necessaria).

12.

a. Inventions must show inventive ingenuity.

b. Every invention must show inventive ingenuity.

2. Asserzione introdotta da "all" + sostantivo plurale: come nel caso precedente la regola deve modificare il numero del sostantivo e del verbo (se necessario), e sostituire il determinatore "all" con il determinatore "every".

13.

a. All services are integrated at both sites.

b.Every service is integrated at both sites.

3. Asserzione introdotta da "each"+ sostantivo singolare: la regola deve provvedere semplicemente alla modifica del determinatore.

4. Asserzione introdotta da "you" impersonale: la regola deve provvedere alla modifica del soggetto della frase, sostituendolo con un sostantivo che si riferisca alla classe su cui viene richiamata l'asserzione. Ad esempio la frase "*you can obtain an application form*", nelle linea guida del sito del governo canadese che dà indicazioni circa le procedure che il cittadino è tenuto a seguire per accedere a determinati servizi, avrà come classe di rifermento quella dei cittadini stessi. Per tale ragione, in questo determinato dominio, può essere riscritta in:

14. Every citizen can obtain an application form.

Questo gruppo di regole sarà quindi dipendente dal dominio dell'ontologia, e dovrà andare a modificare non solo il soggetto della frase, ma anche eventuali elementi che si riferiscano ad esso, con strutture anaforiche (ad esempio pronomi personali).

15.

a. You must be able to lift your bag.

b. Every passenger must be able to lift his bag.

5. Asserzioni con struttura impersonale "one". Come in 4), la regola deve gestire la sostituzione di "one", e in questo caso non sarà necessario sostituire altri elementi che si riferiscano al soggetto.

16.

a. One has to fill the application form.

b. Every citizen has to fill the application form.

Contesto destro:

Per quanto riguarda il contesto destro, esso deve essere modificato in modo tale da permettere l'inserimento di un elemento esistenziale *something,* specificato successivamente da una relativa. Le regole per la gestione del contesto destro devono modificare quindi la sequenza costituita da un sintagma verbale e da un suo argomento; a differenza delle regole riguardanti il contesto sinistro, il pattern non andrà a recuperare sintagmi che si trovino ad inizio frase. Le classi di regole sono fondamentalmente due:

1. Sintagma verbale seguito da complemento oggetto:

risponde alla sequenza I-VP I-NP. La regola deve inserire l'esistenziale *something* e la relativa, in modo tale che il sostantivo del sintagma nominale diventi parte di un predicato nominale della relativa (*something that is I-NP*).

17.

a. You must provide a document.

b. You must provide something that is a document.[19]

2. Sintagma verbale seguito da complemento indiretto:

risponde alla sequenza *I-VP I-PP I-NP*[20]. Come in 1) la regola deve provvedere all'inserimento dell'esistenziale con la struttura relativa, ma inserirla dopo il sintagma preposizionale.

18.

a.Pets will be carried in the aircraft hold.

b. Pets will be carried in something that is the aircraft hold.

6.4.4 Valutazione

L'analisi linguistica della porzione di corpus gestibile da *DL-Lite*$_R$ (50.35% del sottocorpus analizzato manualmente) ha fatto emergere le situazioni ricorrenti che, in fase di parafrasi automatica, è necessario andare a modificare. Queste situazioni sono formalizzate in regole che, al momento dell'implementazione permetterebbero di andare a trascrivere tale porzione di input in un input sintatticamente compatibile con il linguaggio

[19]Per facilitare la lettura, si riporta, in questo contesto, solo la modifica del contesto destro.

[20]Sintagma verbale seguito da un sintagma preposizionale e un sintagma nominale: *carried I-VP inI-PP the aircraftI-NP*

logico. Pur consapevoli di dover rimandare un'effettiva valutazione ad un secondo momento, in seguito ad un'implementazione concreta delle regole qui presentate, può essere comunque interessante andare a valutare informalmente la soluzione proposta indagando il livello di copertura di tali regole. In altre parole, calcolare quale percentuale dei asserzioni dell'input potenzialmente riscrivibile sarebbe, nel migliore dei casi, effettivamente gestito. Analizzando, quindi, la porzione di input potenzialmente riscrivibile (cioè, come suddetto, il 50.35% del corpus totale) emerge che: l'82% di esso sarebbe parafrasabile dalle regole precedentemente descritte, mentre il 12% degli enunciati rimarrebbero al di fuori delle strutture gestibili dalla grammatica. Questo 12% di enunciati presenta infatti costruzioni sintattiche molto complesse, che non vengono coperte dalle regole della grammatica; un esempio è dato dalla frase in *(19)*, che necessita del riconoscimento di sequenze sintagmatiche molto lunghe.

19

a. Non-profit libraries, archives and museums may copy published and unpublished works protected by copyright in order to maintain and manage their collections.

Concludendo, prendendo in considerazione il 50.35% dell'intero sottocorpus, cioè la porzione di input semanticamente accettata da *DL-Lite*$_R$, e sottoponendolo ad una, necessaria, operazione di parafrasi, affinché sia sintatticamente ben formata secondo il linguaggio controllato dallo stesso *DL-Lite*$_R$, sarebbe effettivamente riscrivibile circa il 42% (dell'intero corpus).

Si possono, a questo punto, suggerire due strade da percorrere. Da un lato lavorare per una definizione migliore del modulo di riscrittura per tendere ad una copertura del 100% dell'input potenzialmente riscrivibile (quindi lavorare perché quel 42% attuale, raggiunga il totale delle frasi potenzialmente riscrivibili costituito dal 50.35% del corpus). Dall'altro

cercare di lavorare sul versante del linguaggio logico, per innalzare la percentuale di strutture linguistiche potenzialmente gestibili dalla logica stessa, senza incrementare troppo la complessità computazionale. E' sotto questo profilo che si delinea la riflessione sulle strutture linguistiche non gestite dal frammento di logica.

6.5 Strutture linguistiche al di fuori di entrambi i frammenti

Sebbene sia possibile, quindi, con un sistema di semplificazione testuale, gestire il 42% dell'input, in realtà questa percentuale rimane ancora troppo bassa per poter pensare di automatizzare completamente il processo di arricchimento di una base di conoscenza. Ma come risolvere il problema? In questo caso il ponte fra il linguaggio naturale e il linguaggio logico non è un obiettivo semplice, perché il linguaggio naturale da gestire risulta ricco e sintatticamente complesso, tanto che metà dell'input rimane al di fuori del frammento più ampio di logica. Riprendendo la metafora del fiume è come se fosse necessario costruire il ponte tra i due linguaggi in un punto in cui gli argini sono più distanti l'uno dall'altro. Costrutti come la disgiunzione la negazione e l'esistenziale qualificato in contesto sinistro rimangono al di fuori delle possibilità espressive di questi due frammenti. Se da un lato il problema della negazione non interessa più del 5.40% del sotto-corpus, la disgiunzione è invece un costrutto molto frequente. Questo emerge sia nell'analisi su larga scala del Capitolo 4 (dove risulta che la disgiunzione compare con una frequenza relativa del 2.46% dell'intero corpus), sia nell'analisi dettagliata del sotto-corpus, dove le asserzioni disgiuntive coprono il 17%. D'altra parte il costo computazionale dell'operatore di disgiunzione, rende difficile e sconveniente un suo inserimento nel frammento di logica ridotta.

Ancora più frequente però è il problema della qualificazione dell'argomento in contesto sinistro. Il 30% del corpus infatti é costituito da costrutti che, esplicitamente (8.a) o implicitamente (8.b, trascritta in 8.c), presentano un argomento in contesto sinistro.

8. a. External users who possess a Library Card may use the Interlibrary Loan service.

b. In order to find single journal articles one has to conduct a search in a database

c. Everyone who want to find a single journal articles has to conduct a search in a database.

Ciò che si vuole mettere in evidenza è quindi che la presenza di un quantificatore esistenziale qualificato in contesto sinistro, e la conseguente possibilità linguistica di specificare l'argomento in contesto sinistro, incrementerebbe sensibilmente la copertura linguistica del frammento. Anche lasciando al di fuori dell'espressività del frammento la negazione e la disgiunzione, la percentuale degli enunciati condivisi salirebbe all'80%.

Ne deriva che, alla luce dell'analisi portata avanti in questo contesto, un eventuale, futuro, tentativo di ampliare il frammento di logica al fine di modellare tale linguaggio logico sul linguaggio naturale, potrebbe prendere le mosse dal problema dell'ampliamento dell'espressività del contesto sinistro. Solo questa modifica, infatti, porterebbe ad un sensibile incremento di espressività. Gli operatori logici di negazione e disgiunzione, d'altra parte, non sono così significativi in termini di copertura, quindi un loro inserimento non porterebbe un sensibile giovamento sul piano linguistico.

6.6 Conclusioni

A conclusione di questa sezione, riguardante l'analisi del trade-off fra il linguaggio logico di descrizione dell'ontologia e il linguaggio naturale ad esso più simile, si può notare come

l'argomento suggerisce molte riflessioni. Se la gestione delle queries poteva nel complesso contare su una copertura pressoché totale da parte del frammento di logica, la distanza fra il linguaggio naturale usato nei testi di norme o regolamenti e il frammento DL-Lite (in entrambe le versioni oggetto di analisi) sembra essere molto maggiore. Andando ad isolare solo gli enunciati all'interno del corpus accettati da *DL-Lite*$_R$ queste costituiscono poco più del 50%.

Quindi si sono analizzati i due aspetti del problema: da un lato la parafrasi, possibile attraverso un approccio di semplificazione testuale, dell'input semanticamente gestibile, volta a raggiungere la compatibilità sintattica fra questa porzione di input e il frammento di logica. Dall'altra l'analisi dell'input non gestibile, volta a capire quali strutture linguistiche costituiscono l'ostacolo maggiore verso una copertura soddisfacente del frammento *DL-Lite*$_R$.

Anche in questo caso emerge come il raggiungimento di un buon compromesso tra l'espressività della logica e il linguaggio naturale passi attraverso lo sforzo congiunto di approcci linguistici e logici.

Da un lato la semplificazione testuale, dall'altro eventuali modifiche della logica, studiate secondo le caratteristiche del linguaggio naturale.

Capitolo 7

Conclusioni

In questo lavoro si è cercato di muovere qualche passo verso la definizione di un frammento di linguaggio condiviso tra la logica dell'ontologia e l'utente, attraverso lo sforzo congiunto di approcci logici e linguistici, proponendo la semplificazione testuale come strategia per la soluzione di molti problemi emersi dall'analisi dei corpora. In altre parole, volendo fornire al lettore un' immagine chiara ed esplicativa, si è cercato di dare un contributo alla costruzione di quel ponte che permetta di collegare le sponde del fiume che divide utente ed ontologia, senza costringere l'utente ad esprimersi attraverso un linguaggio controllato. Ma è possibile costruire questo ponte?

I risultati emersi aprono sicuramente problematiche da affrontare, ma, nello stesso tempo, sono decisamente incoraggianti.

Si sono analizzati due scenari di applicazione: l'una è l'interrogazione dell'ontologia,in cui le domande dell'utente devono essere proiettate sulla base di conoscenza, e l'altra l'arricchimento automatico di conoscenza dell'ontologia, dove sono invece le definizioni e le asserzioni dell'utente a dover essere catturate e proiettate nella base di conoscenza

terminologica dell'ontologia.

Il primo contributo che si è cercato di dare per la ricerca di tale frammento è stata un'approfondita analisi dei corpora che mettesse in luce i problemi legati all'espressività del frammento attualmente esistente. Partendo dai limiti conosciuti della logica in esame si è voluto esaminare quanto le strutture non trattate da tale logica fossero frequenti in linguaggio naturale. Tale analisi è stata effettuata sia su corpora di domande che su corpora di asserzioni per avere una panoramica di entrambi i frammenti di logica coinvolti (DL-Lite e CQs).

Nel primo caso il risultato dell'analisi ha fatto emergere un dato interessante, cioè la relativa rarità dei costrutti logici non accettati, costrutti che, inoltre, possono essere parzialmente gestiti attraverso un approccio di indebolimento semantico (Carbotta e Calvanese (2007)). D'altra parte però, tale analisi ha portato alla luce anche un altro fenomeno: l'esistenza, nel linguaggio tipicamente usato dall'utente per interrogare ontologie, di atti indiretti, cioè di quelle situazioni computazionalmente problematiche in cui l'intenzione del parlante non corrisponde all'intenzione letteralmente veicolata dal tipo di frase pronunciata. La gestione di tale fenomeno è stata affrontata proponendo un modulo di semplificazione testuale attraverso la parafrasi automatica dell'atto indiretto in atto diretto. In entrambi i problemi e in entrambe le soluzioni descritte o proposte, vi è stata un'attenzione particolare a rispettare la libertà espressiva dell'utente nel formulare la richiesta; in altre parole, sia nel caso di indebolimento semantico descritto, che nel caso della semplificazione testuale proposta, si è cercato di tutelare la possibilità dell'utente di inserire un input libero, andando a semplificare tale input automaticamente in una fase successiva.[1] Tutto questo per sfruttare i vantaggi di un linguaggio controllato, che però

[1]Nel caso dell'indebolimento semantico, che coinvolge la modifica del significato dell'input originario,

sia, in un certo senso, generato automaticamente, senza costringere l'utente ad imparare ostiche regole di grammatiche controllate.

In secondo luogo l'analisi ha riguardato il versante delle asserzioni per cercare di far emergere la distanza fra il frammento di linguaggio espresso dalla logica di definizione dell'ontologia, e il tipo di linguaggio tipicamente usato dall'utente per esprimere assiomi o definizioni. Quello che ne è emerso è la possibilità di esprimere una porzione di input (pari circa alla metà del corpus) a condizione che venga riscritta in una sintassi controllata accettabile dall'ontologia. A tal fine si è ricorsi ancora una volta alla strategia della semplificazione testuale, proponendo la progettazione di un modulo di riscrittura . Inoltre si sono analizzati i costrutti linguistici più frequenti nella porzione di input non accettata, per capire quali potrebbero essere i primi passi in un eventuale tentativo di ampliamento dell'espressività del frammento di logica.

Tutte le proposte e le analisi effettuate mettono in evidenza la necessità e l'intenzione di operare in parallelo sia dal punto di vista linguistico che logico, nella consapevolezza che il raggiungimento di un buon compromesso non può che essere il risultato del contributo congiunto di entrambe le discipline

Ed è da entrambi i versanti che si può pensare di proseguire su questa strada cercando di affrontare le problematiche aperte e sollevate anche da questo lavoro. Fra i possibili contributi che potranno essere oggetto di futuri lavori, è interessante ricordare la già accennata possibilità di effettuare un test di gradimento da parte dell'utente dell'output del processo di indebolimento semantico. D'altra parte, sul versante dell'arricchimento della conoscenza dell'ontologia, il primo passo dovrà essere l'implementazione effettiva del modulo di semplificazione testuale, che al momento è stato oggetto solo di una for-

l'utente viene, ovviamente, messo a conoscenza di tale modifica.

malizzazione astratta. In seguito, peró, potrebbe essere interessante lavorare ancora sulla definizione del frammento di logica, per far sí che risponda sempre più alle caratteristiche del linguaggio naturale. Un obiettivo che può essere raggiunto solo attraverso il duplice approccio linguistico-logico.

Appendice A

Formalizzazione

Si riporta in appendice un campione delle formalizzazioni per la parafrasi degli atti indiretti in atti diretti.

Tali regole sfruttano l'annotazione morfologica del tagger della catena C&C.

Per quanto riguarda i pronomi e gli aggettivi interrogativi l'annotazione usata è la seguente[1] :

what: I-NP-WP

who: INP-WP

which: I-NP-WP,

when: I-ADVP-WRB

[1]Legenda nomi sintagmi:

I-NP: sintagma nominale,

I-ADVP: sintagma avverbiale

I-VP: sintagma verbale

I-SBAR: congiunzione

where:I-ADVP-WRB

if: I-SBAR- IN

In ogni regola è possibile trovare variabili (X, Y), a cui è attribuito un valore in fase di test, e che permettono di individuare il sintagma specifico. Ogni pattern inoltre è preceduto dal simbolo $, per indicare "inizio linea", e può presentare quantificatori classici delle espressioni regolari (*|+|?).

I valori in caratteri maiuscoli si riferiscono a macro che possono racchiudere insiemi di elementi (come l'insieme dei verbi sensibili, usati per l'individuazione della domanda indiretta).

```
Regole wh -soggetto

/*regola per la gestione del :  wh- soggetto.  Casi :  who,which, what*/

$ X* I-VP Z I-VP

test  (

Z= I-NP,

I-VP.lemma= VERBI_SENSIBILI,

Z.cat= WP|WDT )

action=> {  deleteUntil (Z)}


/*regola per la gestione delle interrogative wh, con inversione.  (I don't
know what you should do).  Cancella tutti i termini fino a Z (wh-) escluso
e muove il modale

dopo Z.
```

```
 Casi:  who, which, what*/

 $ X* I-VP Z I-NP Y I-VP

 test (
 I-VP.lemma= VERBI-SENSIBILI,
 Z.cat= WP|WDT,
 Y= I-VP,
 Y.cat= MD
 )= > action{deleteUntil(Z)
 move (Y, (Zindex+1)) };


 /*regola per la gestione delle interrogative wh, con inversione.  (I don't
know what you are  doing).
 Cancella tutti i termini fino a Z (wh-) escluso e muove l'ausiliare.
 dopo Z.
 Casi:  who, which, what*/

 $ X* I-VP Z I-NP Y I-VP

 test (
 I-VP.lemma= VERBI-SENSIBILI,
 Z=I-NP,
 Z.cat= WP|WDT,
 Y= I-VP,
```

```
 Y.lemma="be"|"have"
 )= > action{deleteUntil(Z)
 move (Y, (Zindex+1)) };

 /*regola per la gestione delle interrogative wh, con inversione.  Caso modale:
I don't know what you should do.
 Cancella tutti i termini fino a Z (wh-) escluso e muove il modale dopo Z.
Casi 'when' e 'where'*/

 $ X* I-VP Z I-NP Y I-VP

 test (
 I-VP.lemma= VERBI-SENSIBILI,
 Z=I-ADVP
 Z.cat= WRB,
 Y= I-VP,
 Y.cat= MD
 )= > action{deleteUntil(Z)
 move (Y, (Zindex+1)) };

 /*regola per la gestione delle interrogative wh, con inversione.Caso ausiliare
essere o avere:  I don't know what you are doing.
 Cancella tutti i termini fino a Z (wh-) escluso e muove l'ausiliare.  Casi
'where' e 'when'
 dopo Z.*/
```

```
 $ X* I-VP Z I-NP Y I-VP

 test (

 I-VP.lemma= VERBI-SENSIBILI,

 Z=I-AVDP

 Z.cat= WRB

 Y= I-VP,

 Y.lemma="be"|"have"

 )= > action{deleteUntil(Z)

 move (Y, (Zindex+1)) };



 /*regole wh:  inserimento do, controlla la persona del verbo, e inserisce
l'ausiliare do in posizione (Z+1), cioè dopo l'elemento wh.*/

 //Casi:  what, which, who

 $X* I-VP Z Y I-VP

 test   (

 X=ALL,

 I-VP.lemma= VERBI-SENSIBILI,

 Z= I-NP,

 Z.cat= WP|WDT,

 Y=I-NP,

 Y.lemma= "I"|"you"|"we"|"they",
```

```
 )=> action { deleteUntil(z),
 insert (do, (Zindex+1))};


 /*regole wh:  inserimento does, controlla che il soggetto sia diverso da "you,
I, we, They", e che il verbo sia all terza persona singolare.
 Quindi modifica la categoria del verbo, inserisce l'ausiliare does in posizione
(Z + 1) cioè dopo l'elemento wh.*/

 $ X* I-VP Z Y W

 test   (
 X=ALL,
 I-VP.lemma= VERBI-SENSIBILI,
 Z= I-NP,
 Z.cat= WP|WDT,
 Y=I-NP,
 Y.lemma != "I"|"you"|"we"|"they",
 W= I-VP,
 W.cat=VBZ
 )=> action { deleteUntil(Z),
 insert (does, (Zindex+1)),
 set(W.cat = VB )};


 //*wh:  inserimento do, controlla la persona del verbo, e inserisce l'ausiliare
do in posizione (Z+1), cioè dopo l'elemento wh.
```

```
 Casi:  when, where.*/

 $ X* I-VP Z Y I-VP

 test   (

 X=ALL,

 I-VP.lemma= VERBI-SENSIBILI,

 Z= I-ADVP,

 Z.cat= WRB,,

 Y=I-NP,

 Y.lemma= "I"|"you"|"we"|"they",

 )=> action { deleteUntil(z),

 insert (do, (Zindex+1))};

 /*regole wh:  inserimento does, controlla che il soggetto sia diverso da "you,
I, we, They", e che il verbo sia alla terza persona singolare.

 //Quindi modifica la categoria del verbo, inserisce l'ausiliare does in posizione
(Z + 1) cioè dopo l'elemento wh.*/

 //Casi:  when, where

 $ X* I-VP Z Y W

 test   (

 X=ALL,

 I-VP.lemma= VERBI-SENSIBILI,

 Z= I-ADVP,

 Z.cat= WRB,
```

```
 Y=I-NP,
 Y.lemma != "I"|"you"|"we"|"they",
 W= I-VP,
 W.cat=VBZ
 )=> action { deleteUntil(Z),
 insert (does, (Zindex+1)),
 set(W.cat = VB )};

 /*REGOLE DOMANDE BOOLEANE*/

 /*Inversione*/

 /*regola per la gestione delle interrogative booleane, con inversione.  (
 Cancella tutti i termini fino a Z (if) compreso e muove il modale in prima
posizione.*/

 $ X* I-VP Z I-NP Y I-VP

 test (
 I-VP.lemma= VERBI-SENSIBILI,
 Z= I-SBAR,
 Z.cat= IN,
 Y= I-VP,
 Y.cat= MD
 )= > action{delete(Z)
```

```
 move (Y, 0 )};

 /*regola per la gestione delle interrogative booleane, con inversione.
 Cancella tutti i termini fino a Z (if) compreso e muove l'ausiliare in prima
posizione.  */

 $ X* I-VP Z I-NP Y I-VP

 test (
 I-VP.lemma= VERBI-SENSIBILI,
 Z= I-SBAR,
 Z.cat= IN,
 Y= I-VP,
 Y.cat= "be"|"have"
 )= > action{delete(Z)
 move (Y, 0 )};

 /*Regole per la gestione interrogative booleane, inserimento do:

 //regole wh:  inserimento do, controlla la persona del verbo, e inserisce
l'ausiliare do in posizione iniziale.*/

 $ X* I-VP Z Y I-VP

 test   (
```

```
 X=ALL,
 I-VP.lemma= VERBI-SENSIBILI,
 Z= I-SBAR
 Z.cat= IN,,
 Y=I-NP,
 Y.lemma= "I"|"you"|"we"|"they",
 )=> action { delete(Z),
 insert (do, (Zindex+1))};

 /*regole wh:  inserimento does, controlla che il soggetto sia diverso da "you,
I, we, They", e che il verbo sia alla terza persona singolare.
 Quindi modifica la categoria del verbo, inserisce l'ausiliare does in posizione
iniziale.*/

 $ X* I-VP Z Y W

 test   (
 X=ALL,
 I-VP.lemma= VERBI-SENSIBILI,
 Z= I-SBAR
 Z.cat= IN,
 Y=I-NP,
 Y.lemma != "I"|"you"|"we"|"they",
 W= I-VP,
 W.cat=VBZ
```

```
 )=> action { delete(Z),
 insert (does, (Zindex+1)),
 set(W.cat = VB )};


 /*REGOLE PER LA GESTIONE DELLE ASSERZIONI*/

 /*CONTESTO SINISTRO*/

 /*regola per la modifica del contesto sinistro introdoto da un sostantivo
plurale.
 inserimento di "every" e modifica del numero del sostantivo e dell'accordo
con il verbo*/

 $ X [Y] Z
 test (
 X= I-NP,
 X.cat= NNS,
 Y=I-NP|I-PP,
 ) action => ##{ insert (EVERY, 0),
 set(X.cat= NN),
 set(Z.cat= VBZ)
 }


 /*regola per la modifica del contesto sinistro in asserzione introdotta da
determinatore "all"*/
```

```
 $ I-NP X [Y] Z

 test (

 I-NP.lemma="all",

 X= I-NP,

 X.cat= NNS,

 Y=I-NP|I-PP,

 ) action => ##{

 delete(I-NP),

 insert (EVERY, 0),

 set(X.cat= NN),

 set(Z.cat= VBZ)

 }


 /*regola per la modifica del contesto sinistro in asserzione introdotta da

"each".  Semplice sostituzione*/

 $ I-NP X [Y] Z

 test (

 I-NP="each",

 X= I-NP,

 Y=I-NP|I-PP,

 ) action => ##{ delete(I-NP)

 insert (EVERY, 0),

 }
```

```
 /*regola per la modifica del contesto sinistro.  Asserzione introdotta da
"you" impersonale*/

 $ I-NP [X] I-VP
 test (I-NP.cat= PRP,
 I-NP.lemma= "you",
 )=> action{ delete(I-NP),
 insert(every,0)
 insert(UTENTE_TIPO, 1)
 //utente tipo sarà il concetto che ha più porbabilità di essere soggetto
 //di un'asserzione della tbx, dipendentemente dal dominio.  In un dominio
di regole comportamnetali
 // di una compagnia aerea e' presumibilmente il //passeggero
 set(I-VP.cat= VBZ)}



 /*regola per la modifica del contesto sinistro.  Asserzione introdotta da
"one" impersonale; come sopra ma senza la modifca del verbo.*/

 $ I-NP [X] I-VP
 test (I-NP.cat= PRP,
 I-NP.lemma= "one",
 )=> action{ delete(I-NP),
 insert(every,0)
 insert(UTENTE_TIPO, 1)
```

```
 //utente tipo sarà il concetto che ha più porbabilità di essere //soggetto

 //di un'asserzione della tbx, dipendentemente dal //dominio.  In un dominio
di regole comportamnetali
 //di una compagnia aerea e' presumibilmente il //passeggero
 }

 /*Regole per la modifica del contesto destro*/

 /*riscrittura oggetto diretto.  Caso "provide [a] document"*/

 IVP [X] Y

 test(
 X.cat = DT,
 Y= I-NP,
 Y.cat= NN|NNS,
 )= > action {insert(SOMETHINGTHAT, (I-VPindex+1))}


 /*riscrittura oggetto indiretto.  Caso:  comply with [the] departement.*/

 IVP I-PP [X] Y

 test(
 X.cat = DT,
 Y= I-NP,
```

```
Y.cat= NN|NNS,

)= > action {insert(SOMETHINGTHAT, (I-PPindex+1))}
```

.

Bibliografia

Allen J. (1995). *Natural Language Understanding.* Benjamin Cummings, Menlo Park, CA.

Andreolli; Fioretto; Gario (1996). *English grammar.* Petrini Editore, Torino, Italy, seconda edizione.

Atzeni; Ceri P. a. (2002). *Basi di dati. Modelli e linguaggi di interrogazione.* McGraw-Hill.

Austin J. L. (1962). How to do thing with words.

Baader F.; Horrocks I.; Sattler U. (2003). Description logics as ontology languages for the semantic web. In *D. Hutter and W. Stephan, editors, Festschrift in honor of J org Siekmann, Lecture Notes in Artificial Intelligence.*

Bernardi R.; Bonin F.; Calvanese D.; Carbotta D.; Thorne C. (2007a). English querying over ontologies: E-quonto. In *AI*IA*, pp. 170–181.

Bernardi R.; Calvanese D.; Thorne C. (2007b). Lite natural language. In *7th Int. Workshop on Computational Semantics (IWCS-7).*

Berruto G. (1998). *Sociolinguistica dell'italiano contemporaneo.* Carrocci Editore, Roma.

Berruto G. (2004). *Corso elementare di linguistica generale*. UTET, Torino, Italy.

Boella G.; Lesmo L.; Damiano R. (2004). On the ontological status of plans and norms. *Artif. Intell. Law*, **12**(4), 317–357.

Borgida A.; Brachman R.; McGuinness D.; Resnick (1989). Classic: A structural data model for objects. In *ACM SIGMOD*, pp. 59–67.

Bos J.; Clark S.; Steedman M.; Curran J. R.; Hockenmaier J. (2004). Wide-coverage semantic representations from a ccg parser. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, p. 1240, Morristown, NJ, USA. Association for Computational Linguistics.

Calvanese D. (1996). *Rappresentazione della conoscenza basata su classi: ragionamento in modelli artitrari e modelli finiti. Tesi di dottorato*. Tesi di Dottorato di Ricerca, Univerita' degli Studi di Roma La Sapienza.

Calvanese D.; Giacomo G. D.; Lembo D.; Lenzerini M.; Rosati R. (2005). Dl-lite: Tractable description logics for ontologies. In *AAAI*, pp. 602–607.

Calvanese D.; De Giacomo G.; Lembo D.; Lenzerini M.; Rosati R. (2006). Data complexity of query answering in description logics. In *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2006)*, pp. 260–270.

Carbotta D.; Calvanese D. (2007). *Towards Natural Language Question Answering Over DL-Lite Knowledge Bases*. Tesi per Master, Free University of Bolzen. Unpublished manuscript.

Carroll J.; Minnen Y.; Canning S.; Devlin J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology,1998.*

Chandrasekar R.; Srinivas B. (1997). Automatic induction of rules for text simplification. In *Proceedings of the 16th conference on Computational linguistics*, Morristown, NJ, USA. Association for Computational Linguistics.

Chandrasekar R.; Doran C.; Srinivas B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics*, pp. 1041–1044, Morristown, NJ, USA. Association for Computational Linguistics.

Cioffi; Luppi; Vigorelli; Zanette (1991). *I testo filosofico.* Edizioni scolastiche Bruno Mondadori, Milano.

Coppola D. (2004). *Dal formato didattico allo scenario. Interagire e comunicare in lingue e culture altre.* Edizioni ETS, Pisa.

Curran J.; Clark S.; Bos J. (2007). Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 33–36, Prague, Czech Republic. Association for Computational Linguistics.

Dongilli P.; Franconi E.; Tessaris S. (2004). Semantics driven support for query formulation. In *Proceedings of the 2004 International Workshop on Description Logics (DL-04). Volume 104 of CEUR Workshop Proceedings. (2004).*

Fuchs N. E.; Schwertel U.; Schwitter R. (1999). Attempto Controlled English — not just another logic specification language. *Lecture Notes in Computer Science*, **1559**, 1–20.

Gangemi A.; Guarino N.; Masolo C.; Oltramari A.; Schneider L. (2002a). Sweetening ontologies with dolce. In *In 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), volume 2473 of Lecture Notes in Computer Science, page 166 ff, Sig uenza, Spain, Oct. 1–4, 2002.*

Gangemi A.; Guarino N.; Masolo C.; Oltramari A.; Schneider L. (2002b). Sweetening ontologies with dolce.

Gordon D.; Lakoff G. (1971). Conversational postulates. *CLS-71*, pp. 200–213.

Gruber T. R. (1993). Towards principles for the design of ontologies used for knowledge sharing In *Formal Ontology in Conceptual Analysis and Knowledge Representation.* A cura di Guarino N., Poli R. KAP.

Guarino N. (1998). Formal ontology and information systems. In *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98, Trento, Italy, pages 3– 15. IOS Press, June 1998.*

Jurafsky D.; Martin J. H. (2000). *Speech and Language Processing.* Prentice Hall, Upper Saddle River, NJ 07458.

Kamp H.; Reyle U. (1993). *From Discourse to Logic: Introduction to Model-theoretic Semantics, Logic and Discourse Representation Theory.* Kluwer Academic Publishers.

Lenci A. (1992). *Linguaggio e comunicazione. Una teoria degli atti linguistici.* ETS, Pisa.

Lenci A.; Montemagni S.; Pirrelli V. (2005). *Testo e computer. Elementi di linguistica computazionale.* Carrocci, Roma.

Lepschy G. (1992). *La linguistica del novecento.* Il Mulino, Bologna.

Levinson S. (1983). *La pragmatica*. Il Mulino, 2 edizione.

L.Renzi; G.Salvi; Cardinaletti A. (2001). *Grande grammatica italiana di consultazione*. Il Mulino, Bologna, Italy.

McEnery T.; Wilson A. (1996). *Corpus Linguistics*. Edinburgh University Press.

Mitamura T.; Nyberg E. (2001). Automatic rewriting for controlled language translation.

Moldovan D.; Harabagiu S.; Girju R.; Morarescu P.; Lacatusu F.; Novischi A.; Badulescu A.; Bolohan O. (2002). Lcc tools for question answering.

N.Calzolari; Gangemi A.; Huang C.-R.; Lenci A.; Oltramani A.; Prevot L. (2008). Ontologies and the lexicon,. to be published in Cambridge Studies in Natural Language Processing.

Quillian (1967). Word concepts: a theory and simulation of contemporary models of semantic capabilities.

Sabatini F. (1994). *La comunicazione e gli usi della lingua*. Loescher Editore, Torino, Italy.

Schwitter R.; Tilbrook M. (2006a). Annotating websites with machine-processable information in controlled natural language. In *AOW '06: Proceedings of the second Australasian workshop on Advances in ontologies*, pp. 75–84, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Schwitter R.; Tilbrook M. (2006b). Annotating websites with machine-processable information in controlled natural language. In *AOW '06: Proceedings of the second*

Australasian workshop on Advances in ontologies, pp. 75–84, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Searle J. (1975). *Indirect Speech acts in Cole, P. and Morgan„ Speech acts : Syntax and Semantics*, volume 3, pp. 59–82. Academia Press.

Siddharthan A. (2003). *Syntactic simplification and Text Cohesion*. Tesi di Dottorato di Ricerca, University of Cambridge, UK., 2003.

.