UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Estensione e valutazione di LexIt, un sistema per l'estrazione automatica di profili distribuzionali di nomi, verbi e aggettivi italiani

Candidata: *Giulia Bonansinga*

Relatore: *Prof. Alessandro Lenci*

Correlatrice: *Prof.ssa Maria Simi*

Anno Accademico 2010-2011

Indice

Introduzione

La sottocategorizzazione è la proprietà dei predicati di selezionare i propri argomenti. Un *frame di sottocategorizzazione* esprime il contesto richiesto da un'entrata lessicale in termini di costituenti sintattici. La disponibilità di informazione di questo tipo è cruciale per molti compiti di *Natural Language Processing* (NLP) che traggono beneficio dall'integrazione di dati sulla struttura argomentale: annotazione morfosintattica, Information Extraction, classificazione semantica dei verbi ecc.
L'obiettivo dell'acquisizione lessicale (*lexical acquisition*) è quello di ricavare i frame di sottocategorizzazione e le preferenze di selezione dei predicati in modo automatico e di esprimere queste informazioni sulla base di dati statistici sulle co-occorrenze delle parole. I lessici di sottocategorizzazione acquisiti in modo automatico, essendo costruiti a partire da dati linguistici reali, hanno il vantaggio di essere facilmente aggiornabili ed estendibili e di includere dati sulla frequenza dei fenomeni linguistici.

Nel Capitolo 1 si discute della necessità di disporre di risorse lessicali e si esplora lo stato dell'arte dei sistemi di acquisizione automatica di informazione di sottocategorizzazione. Infine si introduce LexIt, un sistema per l'acquisizione e la navigazione dei *profili distribuzionali*[1] di nomi, verbi e aggettivi italiani. LexIt è il primo sistema per italiano che si proponga di descrivere la sottocategorizzazione dei predicati su un piano esclusivamente distribuzionale.

Nel Capitolo 2 si discute il problema dell'estrazione delle dipendenze di nomi e aggettivi in LexIt e si illustrano le modifiche approntate a questo scopo ai moduli software per l'acquisizione automatica dai corpora *La Repubblica* e *Wikipedia*. Dalle dipendenze estratte sono ricavati in modo automatico i possibili frame di sottocategorizzazione, di cui si presenta brevemente il processo di acquisizione e selezione.

Nel Capitolo 3 si affronta il tema della valutazione di lessici di sottocategorizzazione sviluppati automaticamente: si presentano le principali modalità di valutazione e si riassumono i risultati ottenuti dai sistemi di acquisizione di frame che costituiscono lo stato dell'arte. Si valutano quindi i frame di sottocategorizzazione acquisiti da LexIt per i verbi, raffrontandoli ai frame attestati in risorse analoghe (gold standard) costruite manualmente e si commentano i risultati ottenuti.

[1] Il profilo distribuzionale di una parola è definito come un *array* di informazione statistica, estratta da un corpus, che ne descrive il comportamento combinatorio. (Lenci et al. 2012).

1. L'acquisizione di informazione di sottocategorizzazione

In questo capitolo si discute la necessità di disporre di risorse lessicali in grado di descrivere il comportamento sintattico e le preferenze semantiche degli argomenti dei predicati sulla base di dati statistici.

Nella sezione 1.1 si motiva l'importanza dell'acquisizione automatica di informazione lessicale con metodi statistico-computazionali e si elencano le applicazioni che possono beneficiare di risorse lessicali di questo tipo. Nella sezione 1.2 si citano alcuni contributi significativi per l'acquisizione automatica di lessici di sottocategorizzazione, che rappresentano lo stato dell'arte per l'inglese e per il francese.

Infine, nella sezione 1.3 si introduce LexIt, un sistema per l'acquisizione automatica di profili distribuzionali di nomi, verbi e aggettivi italiani.

La sezione 1.4 riassume l'ambito in cui si colloca questa tesi e gli obiettivi preposti.

1.1 Metodi statistico-computazionali per l'acquisizione di informazione lessicale

L'acquisizione lessicale (*lexical acquisition*) si occupa di modellare informazioni di sottocategorizzazione e preferenze di selezione semantica delle parole analizzando i dati sulle co-occorrenze delle parole. Con l'ausilio di metodi computazionali è possibile estrarre automaticamente grandi quantità di informazioni di co-occorrenza da corpora, cruciali per numerosi compiti e applicazioni di elaborazione del linguaggio naturale.

Possiamo analizzare i *contesti linguistici*[2] in cui ricorrono i lessemi per descrivere il loro comportamento sintattico e semantico. Consideriamo i contesti d'uso del verbo *legare* (1); la frase (1a), in accordo con la forma transitiva del verbo, non suscita giudizi di agrammaticalità negli italofoni, mentre la frase (1b) viene percepita non valida. Il verbo *legare* ha, per così dire, delle *preferenze sintattiche*, per le quali seleziona un complemento oggetto come argomento.

(1) a. *Ha legato i fiori con un nastro.*

b. **Luigi ha legato con una corda.*

c. **Il sasso lega i polsi.*

La proprietà dei verbi di selezionare i propri argomenti è nota come *sottocategorizzazione*

[2] La nozione di contesto qui assunta non si riferisce alla semplice co-occorrenza delle parole, ma a un'informazione più astratta sui costituenti sintattici, per derivare dai contesti informazioni sulla struttura argomentale dei predicati.

del verbo (*verb subcategorisation*) o, nella grammatica tradizionale, *valenza del verbo*[3]. Un *frame di sottocategorizzazione* rappresenta uno schema combinatorio possibile per un predicato ed è composto dai suoi costituenti sintattici (*slot*). Uno slot è realizzato dai collocati lessicali del predicato, i *filler*, che costituiscono il suo insieme lessicale.

Nella frase (1a), per esempio, si individuano gli slot <soggetto> (sottinteso), <complemento oggetto> ("i fiori") e <complemento preposizionale> ("con un nastro"). Ogni slot seleziona, a sua volta, dei *tipi semantici* prototipici: per esempio, il verbo *legare* tende a selezionare entità animate per il soggetto; per questo motivo la frase 1c non può risultare accettabile per un parlante, che immediatamente la percepisce non valida.

Quest'evidenza empirica viene spiegata dalle *preferenze di selezione*[4], definite come la tendenza di un predicato di ricorrere con parole che appartengono a determinati insiemi lessicali, tale da condizionare il suo comportamento combinatorio[5].

Un *lessico di sottocategorizzazione* elenca per ogni predicato i frame sintattici ad esso associati. La granularità di un lessico dipende dalla teoria linguistica di riferimento e dagli scopi per cui si vuole utilizzare il lessico (Korhonen 2002:38).

Levin (1993:1) afferma che un lessico ideale per i verbi debba fornire entrate lessicali "linguisticamente motivate": dovrebbe perlomeno elencare gli argomenti sintattici dei predicati (distinguendoli dagli aggiunti), *incorporare* una rappresentazione del significato del verbo ed esplicitare la corrispondenza tra i sensi del predicato e le costruzioni sintattiche. Infine, un lessico ideale deve esprimere le preferenze di selezione degli argomenti. Sarebbe di estrema utilità, inoltre, poter disporre di informazione quantitativa, per esempio le probabilità dei possibili frame di un verbo, ricavate da dati sulla frequenza.

La possibilità di avvalersi di uno strumento simile permetterebbe di migliorare le prestazioni di diverse applicazioni per il trattamento del linguaggio. Briscoe e Carroll (1993) hanno rilevato che circa la metà degli errori di analisi sintattica sono da imputare alla mancanza o all'insufficienza di informazioni di sottocategorizzazione.

Risorse lessicali di questo tipo, come ricordato da Korhonen (2002:18), devono però essere quanto più ampie e rappresentative possibili, nonché facilmente producibili ed *estensibili*. Lo sviluppo manuale di un lessico implica un lavoro lungo e laborioso e, purtroppo, intrinsecamente incompleto: una risorsa prodotta manualmente da lessicografi difficilmente tiene conto – o tiene conto in misura ponderata - di ogni varietà del linguaggio. Incorre,

[3] Concetto preso in prestito dalla chimica da Lucien Tesnière nella sua *Teoria della valenza verbale*.

[4] Resnik (1997) definisce le preferenze di selezione come la forza d'associazione tra un predicato e le classi semantiche dei suoi argomenti.

[5] Gli argomenti dei predicati non sono ugualmente probabili tra loro e sono soggetti a *effetti di prototipicità*: per il verbo *mangiare* il complemento oggetto *tipico* è "qualcosa di commestibile". La frase "Lo studente ansioso mangiucchiava la penna" è possibile, ma "penna" è un oggetto certamente meno *tipico* di "pizza".

inoltre, in un rischio concreto di anacronismo: non essendo facilmente aggiornabile, un lessico manuale non può dar conto dei cambiamenti subiti dalla lingua (Schulte Im Walde, 2009:952).
L'utilizzo di dizionari elettronici (*machine-readable dictionaries*, MRD) ha ovviato al problema della costruzione manuale di lessici, ma non al problema cruciale della rappresentatività: Boguraev e Briscoe (1989) notano che le omissioni o la sovrastima della salienza di alcune strutture rispetto ad altre sono frequenti e difficili da individuare; Briscoe (2001) rileva che i lessici costruiti in modo semi-automatico, pur essendo sostanzialmente esatti, tendono ad essere poco esaustivi[6].

Risulta quindi più vantaggioso progettare sistemi automatici per l'acquisizione di informazione da corpora esistenti, con la possibilità di costruire e integrare automaticamente lessici e di distinguere gli apporti di diversi sottodomini linguistici. A dispetto della minore accuratezza, i lessici acquisiti automaticamente si prestano meglio ad applicazioni che operano con dati linguistici "reali" e sono in grado di registrare e associare ai predicati dati statistici sulla frequenza delle strutture linguistiche.

Alcuni compiti linguistici e applicazioni che utilizzano lessici di sottocategorizzazione sono i seguenti:

- modellazione di frame di sottocategorizzazione (Korhonen 2002; Schulte im Walde 2009);
- classificazione semantica dei verbi (Dorr 1997; Schulte im Walde e Brew 2002; Lenci 2012);
- individuazione delle preferenze di selezione di un verbo (Resnik 2003; Erk et al. 2010);
- accuratezza dell'analisi morfosintattica (*parsing*) (Carroll et al. 1998);
- machine translation (Dorr 1997; Hajič et al. 2002);
- in breve, tutte le applicazioni che possono trarre beneficio dall'integrazione di informazione sulla struttura argomentale, come per esempio l'Information Extraction (Surdeanu et al. 2003).

1.2 Lo stato dell'arte dei lessici di sottocategorizzazione

In letteratura si annoverano molti esempi di sistemi per l'acquisizione di informazione lessicale. Schulte im Walde (2009) disegna una panoramica dello stato dell'arte per l'acquisizione automatica di frame di sottocategorizzazione e per l'induzione delle classi

[6] I lessici costruiti in modo semi-automatico hanno cioè alta *precision* ma bassa *recall* (v. Capitolo 3).

semantiche dei verbi da corpora. La tendenza dominante vede l'utilizzo di corpora digitali sempre più estesi e almeno parzialmente annotati[7]. Gli approcci sperimentati differiscono principalmente per la quantità e i tipi di frame considerati: è possibile estrarre un numero di frame prefissato o, al contrario, estrarre tutti i frame che emergono dai dati.

La maggior parte degli approcci, osserva Schulte im Walde, non distingue tra gli argomenti e gli aggiunti di un predicato e generalizza sulle funzioni espresse dagli argomenti o sul tipo di preposizione utilizzata nei complementi indiretti. Questa varietà dipende dagli obiettivi prefissati e dal tipo di applicazioni a cui è destinato il lessico[8]: i sistemi che utilizzano un set di frame prestabilito ottengono risultati in genere più affidabili e coerenti con il corpus in input, mentre approcci più "liberali", pur producendo più rumore, possono dare risultati potenzialmente inaspettati.

La strategia per l'acquisizione di frame è cambiata nel corso degli anni, in ragione della crescente disponibilità di corpora annotati a livello morfo-sintattico. Generalmente il processo si divide nell'estrazione dei frame "candidati" dal corpus e nella selezione dei frame corretti.

Schulte im Walde cita alcuni lavori in cui l'identificazione dei verbi e dei relativi frame si basava su euristiche o su grammatiche a stati finiti, rendendo possibile l'estrazione solo di un numero limitato di frame[9]. Lo studio di Briscoe e Carroll (1997), da cui prendono le mosse molti lavori successivi, è tra i primi ad invertire la tendenza generale: i verbi e le relative dipendenze sono estratti con un parser probabilistico allenato su un corpus annotato a dipendenze. L'output generato comprende tutti i pattern che si accordano alla grammatica utilizzata[10], da cui vengono astratti i frame sintattici, provenienti perlopiù dai dizionari COMLEX (Grishman et al. 1994) e ANLT (Boguraev et al. 1987)[11].

La selezione dei frame è essenziale poiché, potenzialmente, ogni verbo può essere associato ad un frame che in realtà non rispecchia il suo comportamento sintattico; la probabilità che ciò accada aumenta se il verbo è ad alta frequenza. Il metodo tradizionalmente adottato si basa sull'*hypothesis testing*. L'*hypothesis test* cerca di determinare quando un verbo *v*

[7] I primi sistemi per l'estrazione di frame di sottocategorizzazione per i verbi utilizzavano corpora non annotati ed erano in grado di estrarre solo un ridotto numero di frame (Brent 1991; Manning 1993).
[8] Schulte im Walde (2009:955) afferma con convinzione che «non esiste un *optimum* riguardo la quantità e la tipologia di frame. Il lessico di sottocategorizzazione ottimale dipende dal compito o dall'applicazione di NLP che utilizza il lessico». Per esempio, continua Schulte im Walde, la distinzione tra argomenti e aggiunti in un lessico può essere fondamentale in task di machine translation, ma facoltativa in un'applicazione di question answering
[9] Brent (1993) , Ushioda et al. (1993) e Manning (1993).
[10] Una variante della *Definite Clause Grammar* (DCG) di Pereira e Warren (1980), composta da 455 regole.
[11] Per i pattern che non possono essere ricondotti a frame attestati in queste risorse sono stati inclusi ulteriori 30 frame.

occorre con un frame *f* abbastanza frequentemente da escludere che l'associazione <*v*,*f*> sia un errore.

Basandosi su Brent (1993), Briscoe e Carroll utilizzano la variante[12] *binomial hypothesis test*. I 163 frame così ricavati astraggono dalle preferenze di selezione dei verbi e dalle preposizioni dei complementi, ma includono altri tipi di informazione; per esempio, rendono conto del *dative movement* nella lingua inglese[13].

1.2.1 I contributi di Korhonen

Korhonen (2002) osserva che parte del rumore dei lessici acquisiti in modo semi-automatico deriva dai filtri utilizzati nell'ipotesi di selezione (*hypothesis selection*) per escludere i frame scorretti tra quelli estratti. Allo scopo di migliorare la selezione, Korhonen valuta[14] le prestazioni del sistema di Briscoe e Carroll (1997) applicando tre filtri diversi per la selezione dei frame, due varianti dell'hypothesis test e un terzo filtro da lei suggerito:

- *binomial hypothesis test,* BHT (nella versione usata da Briscoe e Carroll);
- *binomial log-likelihood ratio test,* LLR (Gorrell 1999);
- un metodo più semplice, che utilizza la *Maximum Likelihood Estimation* (MLE).

Il terzo metodo assegna un rango ai frame estratti dal classificatore in base alla probabilità che essi hanno di ricorrere con il verbo: *p(frame|verbo)*. Tale probabilità è nota come MLE ed è calcolata come il rapporto tra la frequenza congiunta del frame e del verbo e la frequenza del verbo nel corpus (2).

(2) $$MLE = frequenza\ relativa\ (frame, verbo) = \frac{f(frame,verbo)}{f(verbo)}$$

Viene quindi applicata una soglia ai frame così ordinati per escludere quelli poco probabili. Tale soglia, determinata empiricamente, è risultata essere 0,02. Il sistema di Briscoe e Carroll raggiunge prestazioni migliori con il filtro basato su MLE[15]; BHT e LLR restituiscono risultati meno precisi a basse frequenze e meno completi ad alte frequenze[16].

---

[12] Esistono numerose varianti di questo test, ma la strategia generale prevede che l'ipotesi di correlazione sia verificata quando la frequenza congiunta del verbo e del frame *osservata* nel corpus è superiore alla frequenza *attesa* (cfr. sezione 1.3.1).

[13] Per *dative movement* si intende la trasformazione, reciproca, da una costruzione con oggetto e complemento indiretto ad una costruzione con doppio oggetto.

[14] Cfr. Korhonen (2002:76-78).

[15] Cfr. il Capitolo 3 per la definizione di *Precision, Recall e F-Measure*. Il terzo metodo, basato sulla probabilità condizionata dei frame, ottiene un punteggio di 65.2 in F-Measure, rispetto al 53.3 ottenuto dal BHT e al 45.1 ottenuto dalla LLR. Tuttavia, Korhonen osserva che la MLE è poco robusta alle basse frequenze e rispetto alla sparsità dei dati.

[16] La distribuzione di frame, come qualsiasi distribuzione del linguaggio, è soggetta alla legge di Zipf

Secondo Korhonen, l'errore alla base dell'hypothesis test è assumere che la distribuzione dei frame per tutti i verbi (*incondizionata*, ovvero *p(frame)*) e la distribuzione *condizionata* dei frame rispetto a un verbo (*p(frame|verbo)*) siano assimilabili; sulla falsariga di Levin (1993), che individuava classi semantiche per i verbi interpretabili in termini di sottocategorizzazione, Korhonen suggerisce che probabilmente i gruppi di verbi che hanno un proprio comportamento sintattico necessitino di stime *ad hoc* (2002:85).

Per verificare fino a che punto la classificazione dei verbi su base sintattica o semantica possa essere utile per migliorare l'acquisizione automatica di frame, Korhonen calcola la correlazione, in termini di *distribuzione di frame*, che esiste tra verbi delle stesse classi (semantiche o sintattiche) rispetto a verbi di classi diverse e rileva che la correlazione migliore si ottiene per i verbi classificati semanticamente e associati al loro senso predominante[17] (2002:92-98).
Sulla base di questi risultati, Korhonen sperimenta un **nuovo approccio per la selezione dei frame**, integrando stime probabilistiche semanticamente motivate per guidare il processo di acquisizione dei frame. Tale approccio viene testato sul sistema di acquisizione di frame di Briscoe e Carroll (1997), ma adoperando il chart parser probabilistico di Chitrao e Grishman (1990) su una porzione del *British National Corpus*[18] (BNC).

La valutazione (cfr. Capitolo 3) mette in luce un notevole miglioramento dell'accuratezza del sistema di acquisizione di frame, che si dimostra maggiormente robusto rispetto al problema della sparsità dei dati e rispetto alla predizione di casi non attestati nei corpora di training.

La probabilità condizionata dei frame viene stimata tramite MLE, ma viene corretta con una tecnica di *smoothing*[19], la *linear interpolation*[20], che modifica le distribuzioni condizionate dei frame nei verbi con le stime specifiche della classe di appartenenza del verbo.
Questi accorgimenti migliorano significativamente[21] l'accuratezza dei frame e vengono applicati da Korhonen all'interno di un sistema per l'acquisizione di frame su larga scala.
In sintesi, il nuovo approccio all'ipotesi di selezione proposto da Korhonen si articola nelle fasi seguenti:

(1949): vi sono pochi frame molto frequenti e moltissimi frame a bassa frequenza.
[17] Il grado di correlazione, sia per le classi semantiche che per quelle sintattiche, è stato calcolato con delle *misure di similarità*, la distanza Kullback-Leibler (L) e il coefficiente di correlazione per ranghi di Spearman (RC), che rispettivamente indicano un alto grado di similarità per punteggi superiori a zero e per punteggi vicini agli estremi -1 e 1. Per entrambe le classi, i punteggi di similarità indicano nettamente una correlazione tra i frame dei verbi delle stessi classi rispetto a tutti gli altri verbi.
[18] Burnard (1995).
[19] Cfr. http://it.wikipedia.org/wiki/Smoothing
[20] Chen e Goodman 1996.
[21] Cfr. Korhonen (2002:105;109-110). A livello di classi semantiche si ottiene una performance peggiore per i verbi che richiederebbero una suddivisione in classi più granulari, come i verbi aspettuali (*begin, complete, end…*) e i verbi di comparsa e scomparsa (*arise, emerge, vanish, disappear…*).

1. identificazione di 20 classi semantiche, generalizzate sulla base delle più numerose descritte da Levin, dopo aver verificato la sensatezza dei raggruppamenti con misure di similarità;
2. associazione delle classi semantiche ai verbi in modo automatico, sfruttando la gerarchia di WordNet[22], un database semantico lessicale che organizza le parti del discorso in una rete concettuale;
    a. i gruppi di sinonimi (*synset*) di WordNet sono assegnati alle classi semantiche[23];
    b. per ogni verbo di Levin viene ricercato il senso predominante in WordNet, che viene confrontato con la classe semantica (o le classi semantiche) analoga in Levin;
    c. si naviga la gerarchia di WordNet in profondità, fino ad assegnare ogni synset alla classe semantica che include la maggior parte dei membri del synset;
    d. viene scelta la classe in Levin che comprende il maggiore numero di verbi, che viene assegnata al synset[24];
    e. per ognuna delle 20 classi semantiche così acquisite sono selezionati i verbi da utilizzare per ottenere stime per la classe semantica di appartenenza.
3. selezione di 4-5 verbi rappresentativi per ogni classe, di cui vengono ricercate manualmente occorrenze nel corpus, fino ad ottenere una distribuzione condizionata di frame per ogni verbo;
4. unione delle distribuzioni condizionate ricavate per i verbi della stessa classe;
5. correzione delle distribuzioni con le stime probabilistiche per ogni classe semantica con una tecnica di smoothing;
6. selezione dei frame con probabilità maggiore a una soglia determinata.

In conclusione, lo studio di Korhonen dimostra che una conoscenza probabilistica a priori sulle classi semantiche dei verbi a livello dell'ipotesi di selezione migliora sensibilmente l'acquisizione di frame da corpus.

[22] WordNet (http://wordnetweb.princeton.edu/perl/webwn) rappresenta nomi, verbi, aggettivi e avverbi inglesi in gruppi di sinonimi (synset), in relazione tra loro in una rete concettuale. Per un riferimento consultare Fellbaum (2005).
[23] Dorr (1997) aveva infatti osservato che i verbi sinonimi in WordNet condividono, più o meno, il comportamento sintattico dei verbi raggruppati da Levin, per cui è possibile l'associazione dei verbi "Levin-based" alle classi semantiche di WordNet.
[24] Ancora Dorr (1997) aveva notato una perfetta corrispondenza tra le classi di Levin e i nodi più alti nella gerarchia di WordNet, tale da permettere un abbinamento tra i due sistemi di classificazione.

1.2.2 Lessici di sottocategorizzazione per l'inglese e il francese

VALEX (Korhonen et al. 2006) è un ampio lessico di sottocategorizzazione per l'inglese liberamente scaricabile[25]. VALEX utilizza il sistema di classificazione di frame di Briscoe (2000) e comprende i frame di sottocategorizzazione - e le relative frequenze - di 6.397 verbi, per un totale di 212.741 entrate lessicali.

Il lessico, essendo acquisito in modo automatico, contiene dichiaratamente molto rumore, ma a seconda degli obiettivi esiste la possibilità di applicare misure di filtro[26] o costruire sotto-lessici; ciò viene fatto nell'ottica di costruire il lessico più adatto al compito di NLP a cui è destinato: ad esempio, per compiti di parsing, è più sensato utilizzare un lessico accurato piuttosto che uno esaustivo, cioè ad alta copertura.

Sull'esempio di VALEX, Preiss et al. (2007) propongono un sistema per l'acquisizione automatica da corpora di frame non solo di verbi, ma anche di nomi e aggettivi, colmando una mancanza importante nel panorama dei lessici di sottocategorizzazione per l'inglese. Preiss et al. presentano un sistema in grado di acquisire 168 frame per i verbi, 31 per i nomi e 37 per gli aggettivi ma, a differenza di altri approcci, che utilizzano alberi sintattici[27], identificano i frame tramite *grammatical relations* ("relazioni grammaticali", definite GR[28]). Un classificatore trasforma le GR in frame, basandosi sui frame attestati da COMLEX, ANLT e NOMLEX[29]; i frame vengono selezionati in base alla loro frequenza relativa rispetto a soglie empiricamente determinate.

LexSchem (Messiant et al. 2008) è il primo lessico di sottocategorizzazione per i verbi per la lingua francese, liberamente consultabile on-line[30]. A differenza dei sistemi citati finora, LexSchem non assume si riferisce ad altre risorse per la lista dei frame da estrarre, ma utilizza il sistema di acquisizione di frame ASSCI (Messiant 2008) per imparare le strutture argomentali dall'input. Sono stati individuali 286 frame, per un totale di 3.267 verbi riportati in LexSchem.

---

[25] http://www.cl.cam.ac.uk/~alk23/subcat/lexicon.html

[26] Per esempio, è possibile applicare una tecnica di *smoothing* per portare alla luce i frame significativi che non sarebbero altrimenti catturati per la sparsità dei dati. Si può anche selezionare un sottoinsieme di frame dal lessico di base tra quelli che, ad esempio, compaiano anche nei gold standard, superino l'hypothesis test oppure abbiano frequenza relativa maggiore ad una soglia determinata.

[27] Un albero sintattico (o *parse* tree) è una rappresentazione in forma di albero della struttura sintattica di una frase.

[28] LE GR sono prodotte dal parser RASP (*Robust Accurate Statistical Parsing*) di Briscoe et al. (2006). Sono organizzate in una gerarchia che mostra i rapporti sintattici tra le teste e i complementi, tale per cui sono ritenute l'input ideale per derivare i frame di sottocategorizzazione

[29] Macleod et al. 1997.

[30] Il lessico è consultabile all'indirizzo http://www-lipn.univ-paris13.fr/lexschem.html ed è disponibile a fini di ricerca sotto la licenza LGPL-LR (*Lesser General Public License For Linguistic Resources*).

ASSCI è stato applicato a un esteso corpus giornalistico (dieci annate del giornale francese *Le Monde*) e, poiché accetta dati "puri", in prima istanza provvede all'annotazione morfosintattica. Il sistema dapprima acquisisce le dipendenze argomentali e da essi ricava dinamicamente tutti i plausibili frame, che vengono sottoposti a un filtro (la MLE proposta da Korhonen et al., 2000). Non viene fatta distinzione tra argomenti e aggiunti, ma Messiant et al. (2008:534) si affidano all'informazione fornita dalla frequenza, assumendo che gli argomenti compaiono in determinate posizioni più frequentemente degli aggiunti; di conseguenza, i frame più frequenti sono più probabilmente corretti.

La Tabella 1.1 confronta gli studi citati in base alle variabili individuate da Schulte im Walde (2009). Le prestazioni dei sistemi di acquisizione sono discusse nel Capitolo 3, che tratta della valutazione del sistema di acquisizione di informazioni di sottocategorizzazione presentato in questo lavoro, LexIt.

Sistema	Input	Metodo di acquisizione	Selezione dei frame
Briscoe e Carroll (1997)	*Susanne corpus* (Sampson 1995) e porzione annotata tratta dal Brown Corpus (138.000 parole)	Briscoe (2000), in grado di estrarre 163 frame Viene fatta distinzione tra argomenti e aggiungi, ma si generalizzano le funzioni dei complementi e non si esprimono le preferenze di selezione.	Binomial hypothesis test (BHT)
Korhonen (2002)	Parte del BNC (20.000.000 di parole)	Briscoe e Carroll (1997) con chart parser Rispetto a Briscoe e Carroll sono espresse le preferenze di selezione.	Statistiche probabilistiche a priori sulle classi semantiche dei verbi
Korhonen et al. (2006)	BNC; North American News Text Corpus (NANT); Guardian corpus; Reuters corpus ; dati usati per I congressi TREC-4 and TREC-5. Occasionalmente si è attinto al Web, per un totale di 904 milioni di parole.	Come Korhonen (2002) Non viene fatta distinzione tra argomenti e aggiunti, sono sottospecificate le funzioni e le preposizioni dei complementi e non sono indicate le preferenze di selezione dei predicati.	E' possibile applicare un filtro per rimuovere rumore con varie opzioni
Preiss et al. (2007)	Non specificato	Acquisizione frame da relazioni grammaticali prodotte dal parser RASP. Non viene fatta distinzione tra argomenti e aggiunti, sono sottospecificate le funzioni e le preposizioni dei complementi e non sono indicate le preferenze di selezione dei predicati.	Esclusi dal lessico i frame acquisiti con frequenza relativa inferiore a una soglia prestabilita.
Messiant (2008)	*Le Monde* (200 milioni di parole) Accetta dati non annotati	ASSCI Non viene fatta distinzione tra argomenti e aggiunti e non sono indicate le preferenze di selezione dei predicati. Sono invece espresse le preposizioni dei complementi.	MLE

Tabella 1.1 – Un confronto tra i sistemi di acquisizione di frame presentati

1.3 LexIt, un sistema per l'acquisizione e la navigazione di profili distribuzionali

LexIt è una risorsa lessicale per l'acquisizione automatica da corpora dei *profili distribuzionali* di verbi, nomi e aggettivi italiani. Il profilo distribuzionale di una parola è definito come un *array* di informazione statistica estratta da un corpus, in grado di descrivere il suo comportamento combinatorio. (Lenci 2012).

LexIt è il primo sistema per l'italiano che si proponga di descrivere le proprietà valenziali dei predicati su un piano esclusivamente distribuzionale, con metodi statistico-computazionali che rappresentano lo stato dell'arte[31].

Figura 1.1 – Le parti di un profilo distribuzionale

Il profilo distribuzionale di una parola si articola in un *profilo sintattico* e in un *profilo semantico*:

- il **profilo sintattico** di una parola indica gli slot sintattici (soggetto, complementi, modificatori, ecc.) e le combinazioni di slot prototipiche (frame) con cui essa può occorrere.
- il **profilo semantico** associato ad ogni slot si realizza su due livelli: l'insieme lessicale (*lexical* set) formato dai filler che realizzano lo slot e le classi semantiche (*semantic* classes) astratte dai filler (*semantic classes*), che rappresentano le preferenze di selezione di quell'argomento.

L'obiettivo che questo lessico si propone è fornire un modello della sottocategorizzazione dei predicati e delle preferenze di selezione semantiche degli argomenti tramite le co-occorrenze

[31] In virtù della disponibilità crescente di corpora, di strumenti per il trattamento linguistico e di metodi statistico-computazionali, le informazioni distribuzionali rappresentano una strategia efficace per descrivere il comportamento dei predicati (Lenci (2012)).

registrate nel corpus[32]. Le co-occorrenze sono combinazioni di parole significativamente legate tra loro: la forza di questo legame può essere catturata da misure statistiche che quantificano il grado di associazione tra le parole (Evert 2008).

Il profilo distribuzionale comprende quindi dati statistici per dare conto dell'associazione che esiste tra la parola target e frame, tra gli slot e i filler che li realizzano e infine tra i filler e la classe semantica d'appartenenza.

I profili distribuzionali presenti in LexIt sono estratti dai corpora *La Repubblica* (Baroni et al. 2004) e *Wikipedia*[33], annotati sintatticamente. Gli strumenti utilizzati per il preprocessing e i metodi per l'estrazione dei profili distribuzionali sono descritti nel Capitolo 2.

Il processo di acquisizione dei profili distribuzionali si articola in quattro fasi (Lenci et al. 2012):

- l'analisi linguistica dei dati con strumenti automatici;
- l'estrazione delle dipendenze dei lemmi (nomi, verbi e aggettivi) dal testo annotato e la loro conversione in frame;
- l'associazione dei collocati lessicali (filler) agli slot che costituiscono gli argomenti;
- la derivazione delle classi semantiche dei collocati.

L'acquisizione di informazione lessicale in LexIt non è supervisionata se non sulla base di criteri statistici. Sono preventivati eventuali errori nei profili distribuzionali, dovuti innanzitutto a limiti per così dire "fisiologici" degli strumenti utilizzati per il trattamento del testo, che pure rappresentano lo stato dell'arte per la lingua italiana. Per la stessa ragione, LexIt non si propone di distinguere gli argomenti dagli aggiunti nei frame.

Nel Capitolo 2 si discute l'estrazione delle dipendenze di nomi e aggettivi. Per un riferimento organico su LexIt si può consultare Lenci et al. (2012).

1.3.1 Le misure di associazione

L'estrazione automatica di informazione lessicale permette l'acquisizione dati quantitativi sulla frequenza dei fenomeni; le misure statistiche possono essere correlate ai dati linguistici per offrire una visione più immediata degli atti linguistici più interessanti, che non necessariamente coincidono con i più frequenti.

Tali misure permettono inoltre di rappresentare più fedelmente proprietà intrinsecamente graduate, quali i tipi di associazione indagati in LexIt (per esempio l'associazione tra un predicato e un frame).

La frequenza assoluta di una coppia di strutture linguistiche (una coppia di parole, una coppia predicato-frame, argomento-filler ecc.) non è, di per sé, indicativa del reale legame

[32] Firth (1957) suggeriva "You shall know a word by the company it keeps".
[33] L'estrazione delle dipendenze degli aggettivi dal corpus *Wikipedia* è in corso.

che sussiste tra loro. Un dato più significativo si ottiene confrontando la frequenza assoluta di due eventi linguistici (la cosiddetta *observed frequency*, "O") con la frequenza attesa (*expected frequency*, "E") degli stessi, definita come la frequenza che avrebbero se fossero statisticamente indipendenti l'uno dall'altro.
La frequenza attesa si calcola come il prodotto della frequenze assolute dei due eventi diviso il numero totale di eventi osservati (3).

(3) $$E < x, y > = \frac{f(x)f(y)}{N}$$

Esistono numerose misure statistiche che rapportano queste due grandezze. In LexIt viene utilizzata la *Local Mutual Information* (LMI, Evert 2008:18), definita come il logaritmo in base 2 del rapporto tra frequenza osservata e frequenza attesa di due eventi, moltiplicato per la frequenza osservata[34] (4). Maggiore è la LMI, più significativo è il legame tra gli eventi, che sono statisticamente dipendenti.

(4) $$LMI = Ox\log_2 \frac{O}{E}$$

Poiché l'acquisizione di profili distribuzionali in LexIt non è supervisionata[35], una misura di associazione statistica si rende necessaria per escludere argomenti erroneamente associati ai predicati a causa di errori di parsing e per qualificare i fenomeni realmente significativi[36].

L'altra misura associata ad ogni coppia predicato-frame è la MLE descritta nella sezione 1.2.1, definita come il rapporto tra la frequenza congiunta del verbo e del frame e la frequenza assoluta del verbo nel corpus.

1.3.2 Un database di profili distribuzionali

L'informazione lessicale e i profili distribuzionali acquisiti sono caricati su un database, sul quale poggia il sito http://sesia.humnet.unipi.it/lexit, che permette di esplorare i profili distribuzionali estratti dai corpora. Il database può essere interrogato secondo diversi criteri. È possibile selezionare la parte del discorso (verbo, nome o aggettivo) di cui si vogliono indagare le dipendenze e il corpus (*La Repubblica* o *Wikipedia*) da cui sono state estratte. Le schede *Lemma, Frame, Slot, Filler e Classe semantica dei filler* propongono una ricerca orientata a ottenere viste diverse del profilo distribuzionale (Figura 1.2).

[34] La moltiplicazione per la frequenza osservata privilegia gli eventi frequenti e ridimensiona il punteggio che sarebbe altrimenti ottenuto dagli eventi rari, molto numerosi (Zipf 1949).
[35] Ad eccezione di criteri di frequenza. Si veda il Capitolo 2.
[36] Come si vedrà nel Capitolo 3, sulla valutazione di LexIt, la LMI aiuta a distinguere, seppur empiricamente e non sistematicamente, i frame che rappresentano i reali argomenti di un predicato da quelli che contengono complementi circostanziali. LexIt, infatti, non distingue nativamente gli argomenti dagli aggiunti, ma la LMI privilegia i legami forti e frequenti (come quelli tra un predicato e i suoi argomenti) rispetto ai legami "casuali", sicuramente meno frequenti nel corpus.

Figura 1.2 – Possibilità di ricerca nel database LexIt: per corpus e POS (a destra) e per parti del profilo distribuzionale (a sinistra).

La scheda *Lemma* consente di scegliere il lemma di cui si vogliono cercare i frame o gli slot. Dalla scheda *Frame* si può condurre la ricerca per frame e, di conseguenza, vedere i verbi (o i nomi, o gli aggettivi) che condividono le stesse strutture argomentali. Analogamente selezionando le schede *Slot* e *Filler* è possibile ricercare rispettivamente per slot sintattico e per collocato lessicale. L'ultima possibilità di navigazione prevista, *Classe semantica del filler*, permette di selezionare una classe semantica e un tipo di slot e di indagare quali filler della classe semantica scelta compaiano nello slot.

La Figura 1.3 mostra i primi frame risultati dalla ricerca per lemma del nome *bontà*, tutti accompagnati dalla frequenza assoluta nel corpus e dal punteggio di forza di associazione ottenuto dalla coppia <parola target; frame>. Tale lista può essere ordinata per nome, frequenza o forza di associazione.

È possibile analizzare i frame ed esplorare gli slot che li compongono, così da visualizzare i filler che li realizzano (Figura 1.4). Espandendo lo slot omonimo del frame comp-di sono elencati, in ordine di frequenza, i filler che più spesso compaiono nello slot insieme alla parola target *bontà*: *scelta*, *prodotto, animo, progetto, operazione, risultato* ecc, per un totale di 380 filler. La possibilità di esplorare i filler dei costituenti sintattici[37] fa di LexIt anche un potenziale dizionario di collocazioni. Quando, come in questo caso, i filler sono sostantivi, viene loro associata una classe semantica (tra quelle dei top-nodes di MultiWordNet[38] (Pianta et al. 2002, cfr. sezione 2.5). In questo modo è possibile ricavare le preferenze di selezione di un argomento; per lo slot comp-di le classi semantiche più tipiche risultano essere *Knowledge* e *Act*, rispettivamente con forza di associazione rispetto allo slot di 419,3166 e 163,2261.

[37] A differenza dei lessici VALEX e LexSchem discussi nella sezione 1.2.2.
[38] MultiWordNet è un database lessicale multilingue che rappresenta i concetti espressi dai lessemi in una rete concettuale. È consultabile all'indirizzo http://multiwordnet.fbk.eu/.

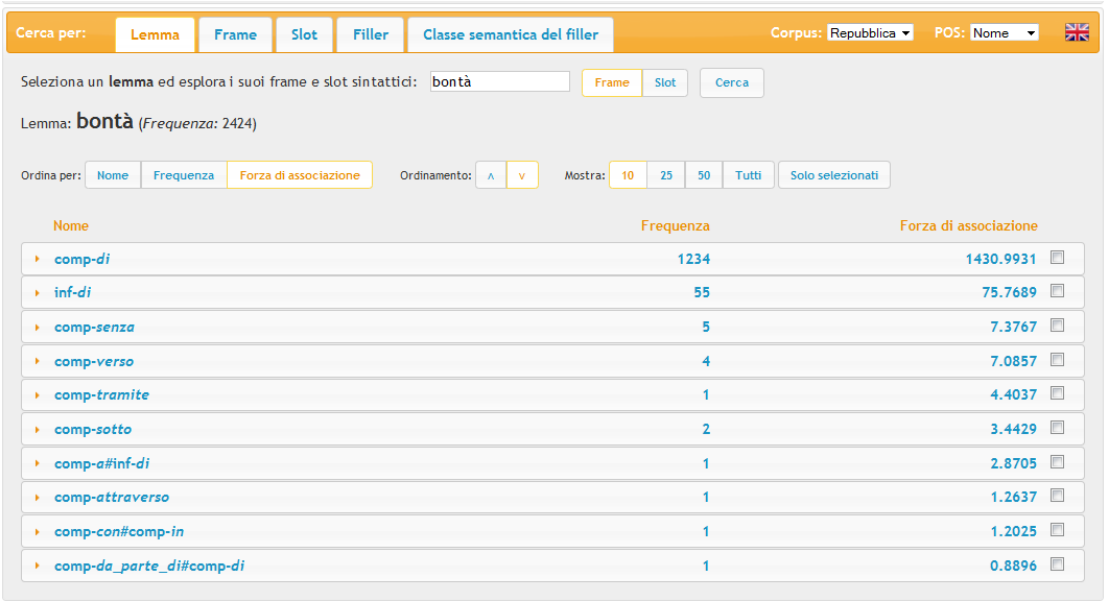


Figura 1.3 – I dieci frame con forza di associazione più alta per *bontà*

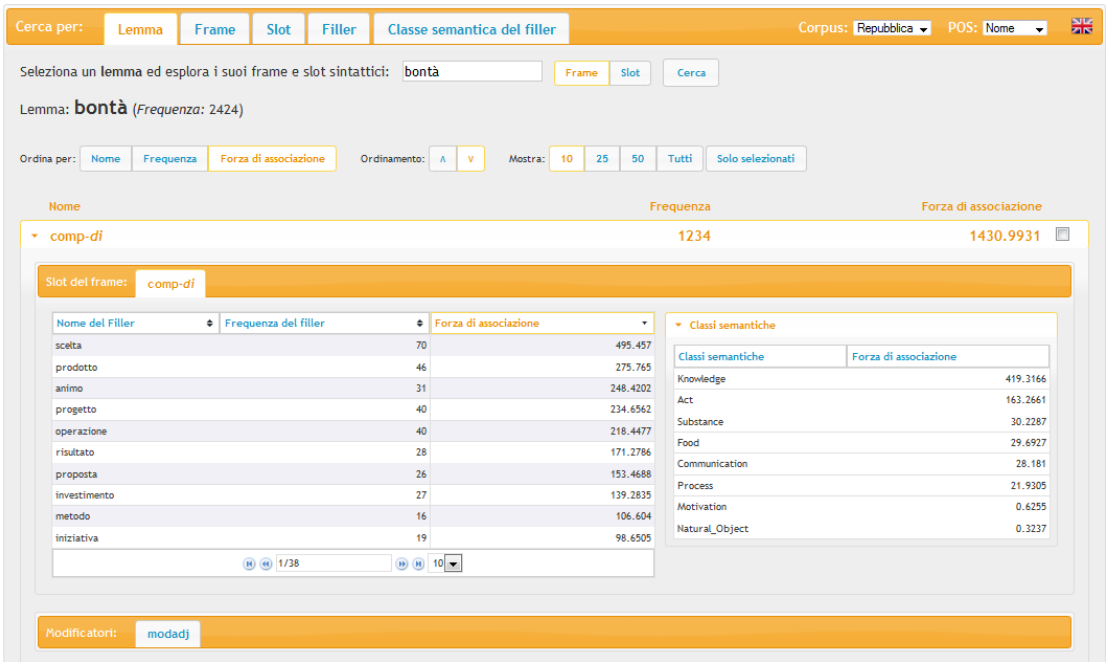


Figura 1.4 – I filler dello slot comp-di per la parola target *bontà* e le relative classi semantiche.

Ad oggi Lexit contiene i profili distribuzionali di 3.873 verbi, 12.766 nomi e 5.559 aggettivi estratti dal corpus *La Repubblica* e di 2.831 verbi e 11.056 nomi estratti dal corpus *Wikipedia*.

I frame estratti sono codificati con etichette sintetiche, nelle quali gli slot sono separati dal simbolo # e sono rappresentati facendo astrazione dalle possibile permutazioni dell'ordine lineare tra gli slot. La tabella 1.2, basata su Lenci et al. (2012), sintetizza il significato degli

slot. Il simbolo * rappresenta le preposizioni o le congiunzioni che possono occorrere nello slot. I frame acquisiti per i verbi, per i nomi e gli aggettivi sono riportati in Appendice A. Per i verbi sono stati selezionati[39] 101 frame univoci (100 frame dal corpus *La Repubblica* e 96 da *Wikipedia* tra quelli estratti automaticamente), per i nomi 124 e per gli aggettivi 44.

A differenza di altri lessici, LexIt non si limita a segnalare la presenza di un complemento preposizionale, ma distingue le preposizioni che possono occorrere, i cui complementi rappresenteranno slot diversi.

Dipendenze comuni a tutte le parti del discorso		
comp-*	Complementi	Vado a pesca Nell'intervista al quotidiano… Sono attivo nel settore
inf-*	infinito introdotto da preposizione	Ho smesso di fumare Ha l'abilità necessaria per riuscire È importante per capire
fin-*	subordinate esplicite	Credo che tu debba stare qui Ho l'impressione che stia fingendo Sarò felice se verrà
Dipendenze dei verbi		
subj	Soggetto	La giornalista ha lanciato il servizio
obj	complemento oggetto	Hanno arrestato il colpevole
subj-0	nessun argomento oltre il soggetto	Gianni piange
si	pronome riflessivo	Gianni si lava
cpred	complemento predicativo	Anna sembra stanca
Dipendenze dei nomi		
0	nessun argomento	Ho un dubbio
Dipendenze degli aggettivi		
pred	contiene il verbo con cui occorre l'aggettivo in funzione predicativa	Ritengo possibile un errore
mod-post	contiene il nome modificato, che appare dopo l'aggettivo	Si esclude un eventuale accordo
mod-pre	contiene il nome modificato, che appare prima dell'aggettivo	Il mercato europeo ci impone di adeguarci

Tabella 1.2 – Etichette utilizzate per gli slot, significato ed esempi.

[39] I frame estratti sono stati sottoposti a una selezione (cfr. Capitolo 2).

1.4 Conclusioni

In questo capitolo introduttivo abbiamo descritto il fenomeno della sottocategorizzazione e abbiamo motivato la necessità di disporre di risorse lessicali che esprimano il comportamento sintattico dei predicati a partire da dati automaticamente acquisiti da corpora, corredati da dati statistici. Abbiamo inoltre presentato LexIt, il primo sistema di acquisizione di profili distribuzionali per l'italiano che si propone di rappresentare le preferenze sintattiche e semantiche dei predicati su un piano esclusivamente distribuzionale.

Nel Capitolo 2 approfondiamo l'estrazione dei profili distribuzionali di nomi e aggettivi italiani, ponendo particolare attenzione ai problemi riscontrati nell'individuazione delle dipendenze sintattiche.

Nel Capitolo 3 confrontiamo i risultati raggiunti dai sistemi di acquisizione automatica descritti in questo capitolo e valutiamo i frame di sottocategorizzazione automaticamente acquisiti da LexIt confrontandoli con i frame attestati in altre risorse lessicali per l'italiano, infine discutiamo l'esito di questa valutazione.

2. Estrazione dei profili distribuzionali di nomi e aggettivi italiani

In questo capitolo viene illustrato il processo di estrazione dei profili distribuzionali in LexIt, con particolare riferimento all'acquisizione dei frame di sottocategorizzazione dei nomi e degli aggettivi. Nella sezione 2.1 sono presentati gli strumenti utilizzati per il preprocessing dei corpora in input.

Nella sezione 2.2 si introduce il sistema di acquisizione delle dipendenze sintattiche di nomi e aggettivi usato in LexIt, approfondito nelle sezioni 2.3 e 2.4.

Nella sezione 2.5 si discutono brevemente le fasi successive all'estrazione delle dipendenze, fino ad arrivare alla definizione dei profili distribuzionali.

Per una trattazione approfondita di LexIt e dell'estrazione dei profili distribuzionali dei verbi italiani si può consultare Lenci (2012) e Lenci et al. (2012).

2.1 Il preprocessing

LexIt contiene i profili distribuzionali di verbi, nomi e aggettivi estratti da corpora annotati con metodi linguistico-computazionali. I corpora al momento utilizzati per l'acquisizione lessicale sono *La Repubblica* (Baroni et al. 2004) e *Wikipedia*[40].

La Repubblica è un corpus giornalistico sviluppato dal centro SSLMIT dell'Università di Bologna, che contiene gli articoli dell'omonimo quotidiano pubblicati dal 1985 al 2000, per circa 331 milioni di token. Il corpus *Wikipedia* raccoglie articoli provenienti dall'enciclopedia libera Wikipedia in lingua italiana, per un totale di 152 milioni di token.

I corpora sono stati tokenizzati, lemmatizzati e annotati morfologicamente usando TANL[41] (*Text Analytics and Natural Language*) e poi annotati sintatticamente con il parser stocastico a dipendenze[42] DeSR[43] (Attardi e Dell'Orletta 2009, Bosco et al. 2009).

Si consideri la frase (5a). L'output dopo la tokenizzazione, la lemmatizzazione e l'analisi morfologica è riportato in (5b). Ad ogni token viene associato un numero identificatore , che sarà utilizzato come riferimento per esprimere le dipendenze sintattiche. Si consideri il token `I` con identificatore `1`: il token viene ricondotto al lemma `il`, riconosciuto innanzitutto come

[40] L'estrazione delle dipendenze degli aggettivi dal corpus *Wikipedia* è attualmente in corso.

[41] TANL è un insieme di strumenti stocastici per il trattamento automatico della lingua italiana sviluppato dall'Università di Pisa e dall'Istituto di Linguistica Computazionale del CNR. Per un riferimento consultare il sito: http://medialab.di.unipi.it/wiki/SemaWiki

[42] L'annotazione sintattica per rappresentazione di dipendenze (contrapposta a quella per rappresentazione di costituenti) descrive una frase in termini di relazioni binarie di dipendenza tra parole. Questo tipo di parsing, poiché si focalizza sulle relazioni grammaticali tra parole, facilita la gestione di dipendenze lunghe.

[43] http://sites.google.com/site/desrparser/

articolo (tag `R` alla voce `PoS`, cioè *part of speech*) e poi, più precisamente, come articolo determinativo (tag `RD` alla voce `PoS` dettagliata). Infine vengono sintetizzati i tratti morfologici, ovvero il numero plurale e il genere maschile (`num=p|gen=m`). Per il verbo (id `5`) sono inoltre espressi il modo (indicativo) e il tempo verbale (presente).

(5) a. *I mercati finanziari europei aprono la settimana all'insegna dell'ottimismo.*

b.

id	token	lemma	PoS	PoS (dett.)	morfologia
1	I	il	R	RD	num=p|gen=m
2	mercati	mercato	S	S	num=p|gen=m
3	finanziari	finanziario	A	A	num=p|gen=m
4	europei	europeo	A	A	num=p|gen=m
5	aprono	aprire	V	V	num=p|per=3|mod=i|ten=p
6	la	il	R	RD	num=s|gen=f
7	settimana	settimana	S	S	num=s|gen=f
8	all'	al	E	EA	num=s|gen=n
9	insegna	insegna	S	S	num=s|gen=f
10	dell'	di	E	EA	num=s|gen=n
11	ottimismo	ottimismo	S	S	num=s|gen=m
12	.	.	F	FS	_

Il parser DeSR completa l'annotazione identificando le dipendenze della frase, facendo uso dei riferimenti ai token e utilizzando il tagset ISST-TANL[44] in Appendice B. L'annotazione completa della frase è riportata in (6). La testa della frase (`ROOT`) è correttamente individuata nel token `5`, che corrisponde al verbo; il verbo, dal momento che non dipende da altri elementi, ha un riferimento fittizio a `0` (`#rif`).

Il token `1`, determinante (`det`), dipende dal token `2`, che è il nome a cui si riferisce (`#rif`). Ad esso puntano anche i modificatori (`mod`) finanziario ed europeo (`#rif 2`).

(6)

id	token	lemma	PoS	…	#rif	relazione
1	I	il	R	…	2	det
2	mercati	mercato	S	…	5	subj
3	finanziari	finanziario	A	…	2	mod
4	europei	europeo	A	…	2	mod
5	aprono	aprire	V	…	0	ROOT
6	la	il	R	…	7	det
7	settimana	settimana	S	…	5	obj
8	all'	al	E	…	7	comp
9	insegna	insegna	S	…	8	prep
10	dell'	di	E	…	9	comp
11	ottimismo	ottimismo	S	…	10	prep
12	.	.	F	…	5	punc

[44] http://medialab.di.unipi.it/wiki/Tanl_Dependency_Tagset

Il nome è in relazione sintattica `subj` (soggetto) con il verbo (`#rif 5`), da cui dipende anche il complemento oggetto (`obj`), nel token `7`. Al nome `settimana` è associato il complemento preposizionale `all'insegna`: tale dipendenza si esprime, come sempre, tramite relazioni binarie: il complemento insegna è in relazione (`#rif 8`) con la sua testa preposizionale (`prep`), che dipende direttamente da `settimana` (`#rif 7`). La relazione intrattenuta tra la testa nominale `settimana` e la testa del complemento preposizionale `al` è etichettata con il tag `comp`.

L'output del parser DEsR è il corpus annotato a dipendenze, frase per frase; questo file di testo viene elaborato con i moduli software che costituiscono LexIt per estrarre le dipendenze sintattiche e, successivamente, i profili distribuzionali di nomi, verbi e aggettivi, completi degli indici statistici che forniscono informazioni sulle più (proto)tipiche e caratteristiche proprietà distribuzionali dei predicati.
Il cuore di LexIt è un insieme di script Perl che, per passi consecutivi, estraggono questo tipo di informazione distribuzionale dall'input annotato a dipendenze.

2.2 Estrazione delle dipendenze di nomi e aggettivi

Dato in input un corpus annotato a dipendenze, si vogliono estrarre gli "slot" sintattici e le combinazioni di slot ("frame") con cui i nomi e gli aggettivi ricorrono, gli insiemi lessicali formati dai collocati ("filler") che appaiono negli slot sintattici e le classi semantiche che descrivono le preferenze di selezione degli slot sintattici (cfr. sezione 1.4).
In prima istanza, è necessario estrarre le dipendenze da cui saranno astratti i frame sintattici.

Si consideri di nuovo la frase (5a) e l'output prodotto dal parser a dipendenze (6). Limitando l'analisi alle dipendenze di nomi e aggettivi, gli argomenti dei nomi che si vorrebbero catturare sono i seguenti:

- gli aggettivi *finanziario* ed *europeo* sono modificatori del sostantivo *mercato;*
- *all'insegna* è un complemento preposizionale che dipende dal sostantivo *settimana*;
- *all'insegna*, a sua volta, è testa sintattica del complemento preposizionale *dell'ottimismo*.

Naturalmente, si vuole catturare che *insegna* e *ottimismo* sono a loro volta sostantivi e si vorrebbe segnalare se i sostantivi sono preceduti da determinanti (articoli, pronomi, predeterminanti) o da aggettivi numerali[45].

[45] I modificatori numerali non sono riportati nei frame di sottocategorizzazione dei nomi (così come gli altri modificatori, che non sono ritenuti argomenti), ma sono comunque catturati nell'estrazione delle dipendenze. Da quest'informazione, in futuro, si potrebbero distinguere i nomi numerabili dai cosiddetti "nomi massa".

Nella fase di estrazione delle dipendenze sintattiche, l'output che si desidera ottenere consiste nell'elenco di tutti i nomi, i verbi e gli aggettivi dell'input (le *entrate lessicali* di LexIt) e nella specificazione delle relative dipendenze sintattiche. Data in input la frase (5a), LexIt produce le dipendenze in (7), organizzate in coppie attributo-valore.

(7)
```
lemma="aprire-v" aux="" si="0" mood="i" subj="mercato-s"
obj="settimana -s"
lemma="mercato-s" det="def"
modadj="europeo-a;finanziario-a"
lemma="settimana-s" det="def" comp_al="insegna-s"
lemma="insegna-s" comp_di="ottimismo-s"
lemma="ottimismo-s"
lemma="finanziario-a" mod-pre="mercato-s"
lemma="europeo-a" mod-pre="mercato-s"
```

L'attributo `lemma` contiene, per l'appunto, il lemma del nome, del verbo o dell'aggettivo. L'attributo `det` indica che `mercato` è preceduto da un determinante definito (`def`), mentre `modadj` indica che il sostantivo è preceduto dagli aggettivi `finanziario` ed `europeo`. I sostantivi `settimana` e `insegna` sono accompagnati dai complementi preposizionali, segnalati dall'etichetta `comp-*`, dove * sta per la preposizione reggente estratta dall'input annotato(`comp-al`, che indica un complemento introdotto dalla preposizione *a*, verrà normalizzato in `comp-a` in un secondo momento), mentre il lemma `ottimismo` è qui registrato senza alcuna dipendenza. Infine l'etichetta `mod-pre` segnala che gli aggettivi `europeo` e `finanziario` sono modificatori del nome `mercato`, che li precede (`mod-post` è l'etichetta per il caso opposto).

L'output dei pattern estratti viene successivamente elaborato con altri script Perl per estrarre i frame, gli slot e i filler e per ricavare le classi semantiche dei filler che sono nomi (cfr. sezione 2.5)

2.3 Le dipendenze argomentali dei nomi

Nell'estrazione delle dipendenze dei nomi, come si è anticipato, i determinanti e i modificatori non sono inclusi nei frame di sottocategorizzazione, in quanto non rappresentano propriamente argomenti del nome.

I modificatori aggettivali, tuttavia, vengono registrati e sono consultabili all'interno delle entrate lessicali di LexIt (Figura 2.1).

Le apposizioni realizzate da nomi o gruppi nominali non sono considerate argomenti del nome, ma vengono registrate a loro volta come lemmi per le rispettive entrate lessicali.

Le dipendenze che, invece, si assumono come argomenti del nome sono i complementi preposizionali (slot di tipo `comp-*`), le proposizioni infinitive (slot di tipo `inf-*`), le completive finite (slot `fin-che`) e il frame `0` (come in Figura 2.1).

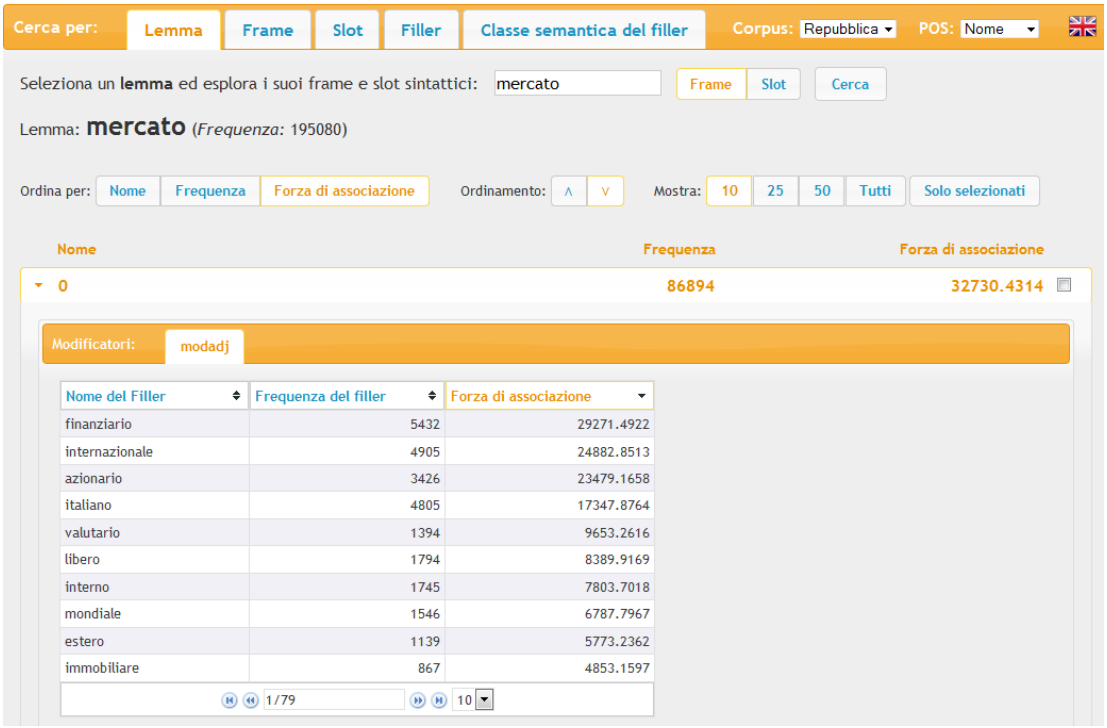


Figura 2.1 – I modificatori aggettivali del sostantivo *mercato* nel frame 0.

Come discusso nella sezione 1.4, LexIt non cerca di distinguere a priori gli argomenti dagli aggiunti, data la difficoltà di stabilire dei criteri. Nel caso dei nomi la distinzione sarebbe ancora più complessa, perché solo un numero relativamente ridotto di nomi richiede effettivamente degli argomenti, ad esempio alcuni deverbali («la conquista *del fortino*»)[46]. Di seguito sono elencati alcuni esempi di dipendenze del nome e l'output prodotto da LexIt.

Complementi preposizionali post-nominali e proposizioni infinitive – Si consideri la frase (8), che contiene un complemento preposizionale ("carcere *di media sicurezza*") e un complemento costituito da un'infinitiva introdotta dalla preposizione *per* ("contratto *per gestire*"). Come visto in (6), l'etichetta per esprimere la relazione tra una testa nominale e un complemento preposizionale è `comp-*`, che esplicita la preposizione o il tipo di complemento espresso (`comp-loc`, `comp-temp`, `comp-a` ecc., cfr. Appendice B).

(8) a. Una *società* è riuscita a procurarsi un *contratto* per gestire un *carcere* di "media sicurezza".

b.
```
lemma="società-s" det="indef"
lemma="contratto-s" det="indef" inf-per="gestire-v"
lemma="carcere-s" det="indef" comp-di="sicurezza-s"
lemma="sicurezza-s" modadj="media-a"
```

[46] Per la descrizione delle dipendenze sintattiche sono state consultate le *Specifiche Linguistiche del Livello Sintattico* di Parole-Simple-Clips (http://www.ilc.cnr.it/clips/SPEC_SINTASSI/SINTASSI_indice.htm), una risorsa lessicale per la lingua italiana che descrive i frame di sottocategorizzazione di nomi, verbi e aggettivi (cfr. sezione 3.3.1).

LexIt individua le dipendenze di tipo `comp` e `prep`, (rispettivamente la preposizione e il nome del sintagma preposizionale), verifica che `prep` dipenda da `comp` e associa l'intero complemento alla testa nominale dell'intero sintagma. Il tipo di complemento estratto dipende della parte del discorso di `prep`: se `prep` è un nome, come in `carcere di "media sicurezza"`, il complemento sarà `comp-di`; se `prep` è un verbo, come in `contratto per gestire`, l'etichetta sarà del tipo `inf-*` e applicherà la preposizione del verbo all'infinito, in questo caso *per*.
In (8b) sono elencate le dipendenze di nomi e aggettivi estratte per la frase (8a).

Un caso particolare: il complemento introdotto da "da parte di" – Si consideri la frase (9°), le cui dipendenze dei nomi sono riportate in (9b): essa contiene un complemento introdotto dalla locuzione preposizionale "dalla parte di", che esprime il ruolo di agente con i nominali. Per catturare questa locuzione non è sufficiente verificare la sequenzialità di una dipendenza di tipo `comp` e di una di tipo `prep`, ma occorre controllare se il token con relazione `prep` è "parte" e se esso è testa di un altro complemento preposizionale ("`del governo`" in 9a). Nello specifico, la sequenza che si ricerca è riportata in Figura 2.2.

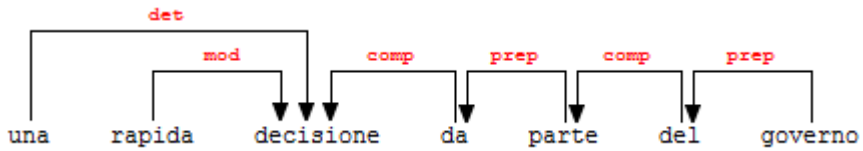


Figura 2.2 – La catena di dipendenze $\texttt{comp}_{(\texttt{da})}\texttt{-prep}_{(\texttt{parte})}\texttt{-comp-prep}$

(9) a. I *sindacati* dei *pensionati* premono per una rapida *decisione* da parte del *governo*.

b.
```
lemma="sindacato-s" det="def" comp-di="pensionato-s"
lemma="pensionato-s"
lemma="decisione-s" det="indef"  comp-da-parte-di="governo-s"
modadj="rapido-a"
lemma="governo-s"
```

Il controllo su `da` e `parte` è necessario per evitare di catturare altre catene preposizionali ("gioco in gomma per bambini"). In particolare, si controlla che la preposizione *da* non sia articolata, per evitare di catturare altre espressioni che non esprimano un complemento d'agente (per es. "dalla parte delle donne").

Completive finite – La relazione ricercata è di tipo `arg`, cioè la dipendenza di un argomento frasale da una testa nominale o verbale. Data la frase (10a), si verifica che il token con relazione di tipo `arg` (`che` nella frase in esame) sia una congiunzione; successivamente si ricerca tra i quindici token seguenti un verbo che si riferisca alla congiunzione (in questo

caso `cedano`, token `#6` riferisce a `che`, token `#3`). Lo spazio di ricerca è ampio per catturare dipendenze lontane nel testo, per esempio a causa di ampi incisi. Le dipendenze della frase (10a) sono elencate in (10b).

(10) a. La *probabilità* che le *parti* sociali cedano è minima.

b.
```
lemma="probabilità-s" det="def" fin-che="cedere-v"
lemma="parte-s" det="def" modadj="sociale-a"
```

Modificatori aggettivali – LexIt cerca di registrare i modificatori aggettivali del nome, ignorando eventuali aggettivi possessivi, pronomi e aggettivi definiti e articoli determinativi, pronomi e aggettivi indefiniti e articoli indeterminativi. Al contempo, si tiene traccia del fatto che il nome possa accompagnarsi a determinanti grazie all'attributo `det`, che può avere valore `def` o `indef` a seconda che il determinante sia definito o indefinito. Nel dettaglio, se la relazione della parola esaminata è `mod` o `det` (modificatore o determinante) e se la parte del discorso è un pronome o aggettivo indefinito o un articolo indeterminativo (rispettivamente `pi, di` e `ri` nel tagset TANL) allora al nome da cui dipendono è associato l'attributo `det="indef`"; se la parte del discorso è invece un pronome o un aggettivo definito o un articolo determinativo (rispettivamente `pd, dd` e `rd`), allora al nome è associato l'attributo `det="def".`

Ugualmente, non sono registrati i numerali, di cui si tiene però conto con l'attributo `modnum`. I numerali sono modificatori (`mod`) con parte del discorso `n` (numero ordinale o cardinale)

La frase (11a) e le relative dipendenze in (11b) forniscono un esempio del modo in cui opera l'algoritmo.

(11) a. Tre *sequestratori* rischiano la *pena* di *morte*, mentre la *posizione* del quarto *imputato* è incerta.

b.
```
lemma="sequestratore-s" modnum="yes"
lemma="pena-s" det="def" comp-di="morte-s"
lemma="morte-s"
lemma="posizione-s" det="def" comp-di="imputato-s"
lemma="imputato-s" modnum="yes"
```

Trattamento dei congiunti e dei disgiunti nei nomi e nei complementi preposizionali – I nomi congiunti vengono identificati e abbinati al lemma del nome a cui si riferiscono. Si consideri la frase (12a), in cui compaiono due coppie di sostantivi congiunti (*carne e cereali, olio d'oliva e conserve*): la congiunzione *e* (etichetta `con`) dipende sintatticamente dal nome che segue; il nome congiunto, che ha dipendenza `conj`, dipende dalla congiunzione. Le dipendenze estratte, di cui è data una rappresentazione schematica in Figura 2.3, sono riportate in (12b).

(12) a. Il *problema* delle eccedenze si ripropone per *carne* e *cereali*, mentre regna il *disordine* in *settori* come *l'olio d'oliva* e le *conserve*.

b. `lemma="problema-s" det="def" comp-di="eccedenza-s"`
`lemma="eccedenza-s"`
`lemma="carne-s" conj="cereale-s"`
`lemma="cereale-s"`
`lemma="disordine-s" det="def" comp-in="settore-s"`
`lemma="settore-s"`
`lemma="olio-s" det="def" comp-d'="oliva-s" conj="conserva-s"`
`lemma="oliva-s"`
`lemma="conserva-s"`

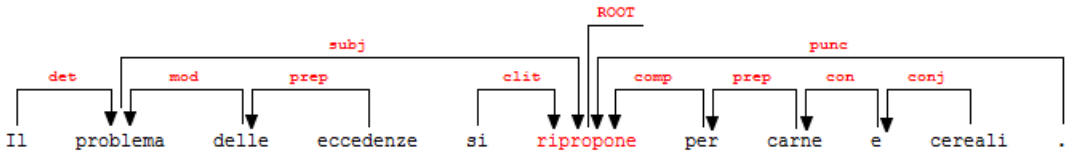


Figura 2.3 – Le dipendenze sintattiche della frase (12a), contenente nomi congiunti.

Lo stesso principio vale per complementi preposizionali in congiunzione tra loro, presenti in (13), (14) e (15).

In (13a) viene dato un esempio di complementi temporali congiunti, pur introdotti da preposizioni diverse. In (13b) si può notare che al primo nome viene abbinato il nome congiunto espresso dallo stesso complemento (`comp-temp-durante="rivolta-s;colpo-s"`).

(13) a. *Padre Przekazinski* ha ricordato le *vittime* della *repressione* durante la *rivolta* operaia del '70 e dopo il *colpo* militare del *dicembre* '81.

b. `lemma="padre-s"`
`lemma="vittima-s" det="def" comp_di="repressione-s"`
`lemma="repressione-s" comp-temp-durante="rivolta-s;colpo-s"`
`conjcomp-temp-dopo="colpo-s"`
`lemma="rivolta-s" det="def" modadj="operaio-a"`
`lemma="colpo-s" det="def" comp-di="dicembre-s" modadj="militare-a"`
`lemma="dicembre-s" modnum="yes"`

La frase (14a) fornisce un esempio del trattamento dei congiunti nei complementi argomentali: i verbi *allontanare* e *vivere* sono associati al nome da cui dipendono, *decisione*, e viene segnalata la loro congiunzione.

(14) a. *L'attore* ha preso la *decisione* di allontanarsi dalle *luci* dei *riflettori* e di vivere in *tranquillità*.

b. `lemma="attore-s" det="def"`
`lemma="decisione-s" det="def" inf-di="allontanare-v;vivere-v"`
`lemma="luce-s" comp-di="riflettore-s"`
`lemma="riflettore-s"`
`lemma="tranquillità-s"`

In (15a), infine, si osserva la codifica dei complementi preposizionali in disgiunzione tra loro, del tutto analoga a quella dei nomi disgiunti.

(15) a. Una *cosa* è la *ricerca*, un'altra la rispettiva *evoluzione* tecnologica nel *benessere* o nel *malessere* che promuove.

b.
```
lemma="cosa-s" det="indef"
lemma="ricerca-s" det="def"
lemma="evoluzione-s" det="def" comp-in="benessere-s"
disj_comp-in="malessere-s" modadj="altro-a;rispettivo-a;
tecnologico-a"
lemma="benessere-s"
lemma="malessere-s"
```

2.4 Le dipendenze argomentali degli aggettivi

I frame di sottocategorizzazione degli aggettivi descrivono il loro uso in funzione predicativa o attributiva.
Nel primo caso, l'aggettivo è un complemento predicativo di un verbo copulativo ("Lucia sembra triste"), mentre nel secondo caso modifica un nome.
All'interno della funzione attributiva l'aggettivo si può trovare prima o dopo il nome: in italiano la posizione non marcata dell'aggettivo è post-nominale ("La macchina grigia"). L'aggettivo occupa posizione pre-nominale, solitamente, in espressioni connotate, in cui in giudizio del parlante prevale sullo scopo descrittivo ("pover'uomo" versus "uomo povero", "una gran donna" versus "una donna grande").

In LexIt viene dato conto di questi usi dell'aggettivo negli slot `mod-pre` (il nome occorre prima dell'aggettivo) e `mod-post` (il nome occorre dopo). Se l'aggettivo è in funzione predicativa, lo slot `pred` contiene il verbo.

L'aggettivo, infine, può avere complementi realizzati da proposizioni infinitive ("Questa legge è difficile da interpretare") o da proposizioni finite ("Sono sicuro che tu possa farcela").

Modificatori pre-nominali e post-nominali – La dipendenza degli aggettivi dalla loro testa nominale è espressa nel tagset TANL dall'etichetta `mod`, a cui viene aggiunto il suffisso `pre` o `post` a seconda che il nome occorra prima o dopo l'aggettivo. Quando nell'input si incontra un token con dipendenza `mod`, si controlla se il riferimento al token da cui dipende è un sostantivo e, in questo caso, il nome viene associato al lemma dell'aggettivo.
Nella frase (16a) al sostantivo servizio sono associati un modificatore pre-nominale (`nuovo`)

e uno post-nominale (`segreto`). Ai lemmi degli aggettivi è associato lo stesso nome, abbinato alle sue dipendenze aggettivali come già discusso.

(16) a. Divampa la polemica sull'efficienza dei nostri nuovi *servizi* segreti.

b. `lemma="nuovo-a" mod-post="servizio-s"`
`lemma="segreto-a" mod-pre="servizio-s "`

Nella frase (17a) due aggettivi congiunti, `macabro` e `delinquenziale`, modificano lo stesso nome, a cui è legato solo il primo; il nome associato a due aggettivi congiunti tra loro diviene parte delle dipendenze di entrambi. I controlli che si svolgono sono i seguenti: se un aggettivo ha dipendenza `conj`, si verifica se la sua testa sintattica è un aggettivo e se questo, a sua volta, dipende da un nome. Se le condizioni sono verificate il nome viene registrato anche come dipendenza dell'aggettivo congiunto.

(17) a. Chi ci avrebbe inviato un così macabro e delinquenziale *avvertimento*?

b. `lemma="macabro-a" mod-post="avvertimento-s`
`lemma="delinquenziale-a" mod-post="avvertimento-s`

In (18a) si mostra un esempio di aggettivo plurale che modifica più sostantivi. Nel parsing l'aggettivo viene frequentemente associato a solo uno dei nomi e nell'estrazione delle dipendenze si cerca di rimediare a questo errore. Se l'aggettivo è pre-nominale, come in (18a), si verifica se il nome a cui si riferisce è la testa sintattica di un altro sostantivo ad esso congiunto. In questo caso, entrambi i nomi sono registrati tra gli argomenti dell'aggettivo. Il procedimento è analogo nelle dipendenze post-nominali, salvo il fatto che l'aggettivo si lega al nome congiunto.

(18) a. I volontari devono affrontare *numerosi viaggi e ostacoli*.

b. `lemma="numeroso-a" mod-post="viaggio-s;ostacolo-s"`

Aggettivi in funzione predicativa – La relazione tra una testa e un complemento predicativo è espressa dal tag `pred`. Si consideri la frase (19a): se l'aggettivo ha una relazione di dipendenza `pred` con un verbo, questo viene associato al lemma dell'aggettivo come in (19b).

(19) a. I prezzi garantiti dalla CEE ai produttori sono molto *esigui*.

b. `lemma="esiguo-a" pred="essere"`

In (20a) si dà un esempio di aggettivi congiunti in funzione predicativa. All'aggettivo congiunto viene attribuita la stessa dipendenza sintattica dell'aggettivo da cui dipende (20b).

(20) a. Vallanzasca si è mantenuto piuttosto *calmo* e *tranquillo*, arrendendosi per primo.
b. `lemma="calmo-a" pred="mantenere-v"`
`lemma="tranquillo-a" pred="mantenere-v"`

Complementi preposizionali – La relazione tra un complemento preposizionale e la sua testa aggettivale è espressa da `comp` e varianti (`comp-temp, comp-loc, comp-ind`). Consideriamo il complemento preposizionale `presente nel governo` della frase (21a): il token `nel` dipende dall'aggettivo `presente` (`comp`) ed è a sua volta testa sintattica del nome del sintagma preposizionale, `governo` (relazione `prep`). Al lemma dell'aggettivo `presente` viene abbinato il complemento locativo `comp-loc-in="governo-s".` Il procedimento è identico per il secondo complemento preposizionale nella frase, sorretto dalla preposizione `di`, come si vede in (21b).

(21) a. L'*unico* ministro sick *presente* nel governo è il titolare dell'Agricoltura, Buta Singh, già *responsabile* delle questioni *parlamentari*.
b. `lemma="unico-a" mod-post="ministro-s"`
`lemma="presente-a" mod-pre="ministro-s" comp-loc-in="governo-s"`
`lemma="responsabile-a" comp-di="questione-s"`
`lemma="parlamentare-a" mod-pre="questione-s"`

Proposizioni infinitive – La relazione `comp` si utilizza per segnalare la dipendenza da una testa aggettivale anche di complementi realizzati da proposizioni infinitive. All'interno del complemento preposizionale si controlla che la relazione `prep` sia realizzata da un verbo, invece che da un nome. Si prenda ad esempio la frase (22a): la preposizione `di` è in relazione `comp` con l'aggettivo `capace` ed è a sua volta testa del verbo `ridurre` (relazione `prep`), che viene registrato tra le sue dipendenze come `inf_di="ridurre"` (22b).

(22) a. Ci siamo dimostrati *capaci* di ridurre i costi.
b. `lemma="capace-a" pred="dimostrare-v" inf-di="ridurre-v"`

Completive finite – La dipendenza `arg` esprime la relazione tra un complemento frasale, introdotto dalla congiunzione, e una testa sintattica. Nella frase (23a) la congiunzione subordinante dipende dall'aggettivo `sicuro` (`arg`) ed è testa sintattica del verbo `replicare`, che viene registrato tra le dipendenze dell'aggettivo (23b).

(23) a. Gli organizzatori sono *sicuri* che si replicherà il successo dell'anno *scorso*.
b. `lemma="sicuro-a" pred="essere-v" fin-che="replicare-v"`
`lemma="scorso-a" mod-pre="anno-s"`

2.5 Dalle dipendenze ai profili distribuzionali

Le dipendenze estratte rappresentano tutte potenziali slot per i lemmi relativi (Lenci et al. 2012:3). Per tentare di ridurre il rumore e selezionare le dipendenze realmente significative, i dati estratti sono sottoposti a verifiche semi-automatiche e/o basate su criteri di frequenza.

La lista dei lemmi estratti è sottoposta a un controllo manuale, con l'ausilio di espressioni regolari per individuare lemmi sicuramente scorretti (parole straniere, contenenti punteggiatura ecc.). Vengono mantenuti solo i lemmi (e le relative dipendenze) con almeno 100 occorrenze nel corpus. Per quanto riguarda i lemmi estratti dal corpus *La Repubblica*, questa selezione porta da 278.925 lemmi iniziali per i nomi a 12.765, mentre gli aggettivi si riducono a 5.559 dai 111.069 di partenza.

La lista dei lemmi è stata revisionata manualmente sia per i nomi che per gli aggettivi:

- sono stati esclusi lemmi ad altissima frequenza ma semanticamente poco significativi, per esempio "grazie" per i nomi o "altro" per gli aggettivi (freq. 321386);
- forme alternative dello stesso lemma sono state normalizzate e le frequenze unificate (per esempio *gran* e *grande*);
- sono stati eliminati nomi e aggettivi che il parser non è stato in grado di ricondurre al lemma: parole straniere (*off-shore, live, top-secret, welfare*), plurali di nomi d'uso comune in italiano ma il cui lemma è evidentemente assente nelle risorse lessicali (*juventini*), sigle (*PSI*, *tg, ddl, ct...*) ecc.

Per questo task, tramite espressioni regolari, sono stati cercati pattern che facilmente potevano corrispondere a parole lemmatizzate erroneamente nel parsing o a lemmi poco significativi:

- lemmi terminanti in -a (*antiterroristica*) e in -e (*mazzette*);
- superlativi (i*ngentissime*);
- aggettivi di nazionalità uscenti in -ese, -ano, -ino (*francese, italiano, spezzino*), poco significativi per i nostri scopi;
- aggettivi uscenti in -"enne" (*quarantunenne*); aggettivi numerali (*quinto*);
- lemmi uscenti in consonante, spesso parole straniere;
- lemmi contenenti caratteri non alfabetici (*grand'*, *amore-odio*) o maiuscole, sovente sostantivi di frasi nominali analizzate erroneamente;
- in generale, parole di categorizzazione morfosintattica non immediata che, collocate in frasi complesse, vengono ricondotte alla PoS più probabile date le loro dipendenze (si consideri "rivoltegli" in una frase come "le parole rivoltegli", erroneamente scambiato per un aggettivo).

Le dipendenze sono utilizzate per estrarre, in modo automatico, le sequenze di slot candidate a frame di sottocategorizzazione (Lenci et. al. 2012). LexIt non assume una lista predefinita di frame da cui partire, ma individua le combinazioni di slot più frequenti come potenziali frame per la parte del discorso in esame (verbo, nome o aggettivo).
Nella costruzione dei frame candidati sono escluse, per quanto riguarda i nomi e gli aggettivi, le dipendenze sui congiunti, i numerali e i determinanti. I frame così ricavati sono normalizzati (per esempio comp-al viene ricondotto a comp-a, comp-[sino|fin|sin] sono ricondotti a comp-fino ecc.) e sottoposti a una prima selezione: sono escluse le strutture argomentali troppo complesse (frame con più di quattro slot), quelle poco interessanti (per esempio catene preposizionali dello stesso complemento) e quelle palesemente sbagliate, contenenti per esempio slot frutto di errori di parsing.
Sono quindi selezionati gli *n* frames candidati (complessivamente 101 per i verbi, 124 per i nomi e 44 per gli aggettivi), di cui viene calcolata la frequenza di co-occorrenza con ogni lemma nel corpus annotato.

Gli indicatori statistici associati a ciascun tipo di informazione in *LexIt* permettono di analizzare le proprietà combinatorie delle parole e la struttura argomentale dei predicati, identificandone i tratti distribuzionali più salienti e prototipici. Per ogni *profilo sintattico* (cfr. sezione 1.4) viene calcolata la forza d'associazione tra il lemma e ognuno dei frame, tra il lemma e ognuno degli slot e tra la coppia lemma-slot e ognuno dei filler.

La misura adottata è la Local Mutual Information (LMI) presentata nella sezione 1.4.1, utilizzata per quantificare la salienza statistica di ogni fenomeno linguistico indagato.

Questo approccio mira ad ottenere una distribuzione "condizionata" dei frame e degli slot rispetto al lemma considerato, ma soprattutto tiene conto della rilevanza delle diverse strutture argomentali.

2.6 Conclusioni

In questo capitolo è stato presentato il sistema di estrazione delle dipendenze sintattiche dei nomi e degli aggettivi LexIt. Per un riferimento all'algoritmo di estrazione dei frame si può consultare Lenci et al. (2012).
Nel prossimo capitolo si tratta il problema della valutazione dei lessici acquisiti in modo automatico e si valutano i frame acquisiti per i verbi da LexIt.

3. Valutazione dei profili distribuzionali dei verbi

Fino a questo momento è stata trattata l'acquisizione di informazione lessicale (Capitolo 1) e l'estrazione dei profili distribuzionali di nomi, verbi e aggettivi in LexIt (Capitolo 2). In questo capitolo si discute il problema della valutazione di lessici di sottocategorizzazione acquisiti automaticamente.

Nella sezione 3.1 si presentano le modalità di valutazione più comuni e si definiscono le misure di valutazione utilizzate per giudicare le prestazioni di un sistema. Nella sezione 3.2 si riassumono i risultati ottenuti dai sistemi di acquisizione che rappresentano lo stato dell'arte per l'inglese e il francese.

Nella sezione 3.3 si discute la valutazione dei frame di sottocategorizzazione acquisiti per i verbi da LexIt, si presentano le risorse lessicali (gold standard) che sono state assunte per il confronto. Infine (sezione 3.4) si discutono i risultati della valutazione.

3.1 La valutazione di informazione lessicale automaticamente acquisita

La valutazione di dati lessicali acquisiti in modo automatico si avvale del confronto con una risorsa, detta *gold standard*, che funge da riferimento e che esprime informazione analoga ai dati acquisiti, es. una risorsa lessicale o un dizionario.

Specialmente nell'ambito dei lessici di sottocategorizzazione (e almeno fino a poco tempo fa) non sempre si ha la disponibilità di una risorsa adatta al confronto. In questi casi si opta per lo sviluppo "manuale" di un gold standard, analizzando manualmente un corpus e raccogliendo un numero di attestazioni sufficiente per ogni *test verb* da valutare. Il giudizio, quindi, non è totalmente oggettivo ma ricade su uno o, meglio, più esperti.

Anche la disponibilità di un gold standard preesistente, tuttavia, presenta degli svantaggi: Korhonen (2002:51) osserva che tipicamente un dizionario non elenca tutti i frame che possono essere astratti da un corpus, o viceversa il corpus utilizzato per l'acquisizione può non attestare tutte le strutture argomentali di un verbo riportate invece nel dizionario. Inoltre, il gold standard potrebbe codificare l'informazione di sottocategorizzazione in modo profondamente diverso dalla nostra risorsa, a tal punto da rendere difficoltoso il confronto e da penalizzare, quindi, la valutazione.

Nell'ambito di un lessico di sottocategorizzazione, la valutazione si basa sul conteggio dei frame corretti e di quelli scorretti tra quelli acquisiti dal sistema. Trattandosi di un problema di classificazione binaria, questi dati si possono trasporre in una *matrice di confusione* (Tabella 3.1) in cui si pongono a confronto i frame identificati dal sistema con quelli del gold standard.

Si definiscono *true positive* (TP) i frame correttamente acquisiti dal sistema, che sono attestati anche nel gold standard, mentre si definiscono *false positive* (FP) i frame acquisiti dal sistema che non sono citati nel gold standard, che quindi si assumono scorretti per il verbo considerato.
Infine, si definiscono *false negative* (FN) i frame attestati nella risorsa di riferimento che non sono stati acquisiti, erroneamente, dal sistema, mentre si definiscono *true negative* (TN) i frame scorretti per il verbo in esame, assenti nel gold standard e non recuperati – correttamente - dal sistema.

		Gold Standard	
		Frame attestati	**Frame non attestati**
Risorsa da valutare	**Frame acquisiti**	TP	FP
	Frame non acquisiti	FN	TN

Tabella 3.1 – Matrice di confusione

Si nota che i TP e i FP costituiscono l'insieme dei frame complessivamente acquisiti dal sistema per un verbo, mentre TP e FN rappresentano i frame complessivamente attestati dal gold standard.
Definito il quadro delle possibilità, possono essere utilizzate delle *metriche di valutazione,* mutuate dall'Information Retrieval, per sintetizzare la performance del sistema di acquisizione automatica.
Si definisce *Precision* ("precisione") il rapporto tra i frame corretti predetti del sistema (TP) e il numero totale di frame predetti (TP + FP); si definisce *Recall* ("richiamo") il rapporto tra i frame corretti predetti (TP) e il numero totale di frame corretti (TP + FN).

$$Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN}$$

Infine, la F-Measure è definita come la *media armonica* di Precision e Recall. Questa misura sintetizza le prestazioni del sistema tenendo conto sia della capacità del sistema di dare risultati corretti (Precision), sia della capacità di copertura (Recall).

$$F\text{-}Measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

3.2 Lo stato dell'arte

Le misure di valutazione presentate, in particolare la F-Measure, permettono di confrontare i risultati di sistemi di acquisizione lessicale diversi. Tuttavia, osserva Schulte im Walde (2009), i punteggi ottenuti devono essere interpretati con cautela, specialmente se si confrontano risorse costruite in modo diverso: la valutazione manuale tende ad essere più elastica, mentre il confronto con un gold standard può essere troppo penalizzante. Il gold standard è una risorsa redatta da esperti che aderisce a precisi criteri di codifica e opera a un certo livello di granularità, non necessariamente gli stessi del lessico con cui viene confrontata.

Per dimostrare la validità di un approccio semanticamente motivato per la selezione dei frame (cfr. sezione 1.2.1), Korhonen (2002:134-137) mette a confronto quattro lessici di sottocategorizzazione acquisiti a partire dal sistema di Briscoe e Carroll (1997), ma applicando tecniche diverse per la selezione dei frame:

- LEX-A è la versione del lessico di Briscoe e Carroll, ricavata con il *binomial hypothesis test* (BHT)[47];
- LEX-B utilizza la *maximum likelihood estimate* (MLE)[48];
- LEX-C è costruito applicando la tecnica di smoothing *add-one* (Laplace 1995) e fissando una soglia minima per i frame;
- LEX-D è il lessico ottenuto applicando la *linear interpolation*[49] e considerando le stime delle classi d'appartenenza dei verbi[50].

Korhonen confronta i lessici con un gold standard sviluppato manualmente: LEX-A, che usa il BHT per la selezione dei frame, ha risultati nettamente peggiori rispetto agli altri lessici (52,2 F-Measure per cui è disponibile una classificazione semantica), che ottengono i punteggi di 60,6 (LEX-B), 64,9 (LEX-C) e 78,4 (LEX-D) in F-Measure. LEX-C e LEX-D si dimostrano molto robusti rispetto al problema della sparsità dei dati e hanno, nel complesso, i risultati migliori, ma LEX-D raggiunge prestazioni spiccatamente migliori; in particolare, in 14 delle 18 classi semantiche individuate, LEX-D opera meglio rispetto a LEX-C.
Korhonen valuta i lessici anche nell'ambito di un compito di parsing, ottenendo risultati coerenti[51] con la valutazione rispetto al gold standard costruito.

[47] Cfr. sezione 1.3.
[48] Cfr. sezione 1.3.1.
[49] Chen e Goodman 1996.
[50] Korhonen utilizza 474 verbi, di cui solo 140 sono classificati nei synset di WordNet. Per i verbi per cui non è possibile ottenere stime semanticamente motivate, Korhonen adotta l'approccio usato per LEX-C, che dà i risultati migliori dopo LEX-D.
[51] Cfr. Korhonen 2002:137-144.

Korhonen et al. (2006) optano per una valutazione mista di VALEX (cfr. sezione 1.2.2): 183 verbi sono stati valutati con un confronto manuale con il corpus (per ogni verbo erano richieste almeno 300 occorrenze nel corpus), i frame dei rimanenti 5.659 verbi sono stati confrontati con i frame attestati in COMLEX[52] e ANLT[53]. Per ognuna delle due valutazioni sono stati giudicati i frame di diversi sotto-lessici di VALEX, ottenuti con opportune scelte di filtro. Il sotto-lessico che ha dato miglior punteggio di F-Measure per la valutazione manuale (87,3) è composto dai frame attestati in COMLEX e ANLT e da quelli, tra quelli non attestati nei due dizionari, con frequenza relativa superiore a una soglia determinata. La distribuzione dei frame risultata è stata poi trattata con la linear interpolation.
Nella valutazione con i gold standard ottiene il miglior punteggio un sotto-lessico che comprende i frame che, dopo l'applicazione della linear interpolation, avevano una soglia di frequenza relativa non inferiore a 0,01. Questo lessico ha raggiunto il punteggio di 57,3, decisamente inferiore al punteggio accumulato nella valutazione manuale (69,2). Questa discrepanza si spiega col fatto che i gold standard possono non attestare frame acquisiti da un sistema automatico o, viceversa, possono attestare frame a bassa frequenza assenti nel corpus di input[54].

Preiss et al. (2007) ottengono punteggi di F-Measure che rappresentano lo stato dell'arte per la sottocategorizzazione di verbi, nomi e aggettivi inglesi. La loro valutazione ha preso in esame 183 verbi, 30 nomi e 30 aggettivi che avessero almeno 150 occorrenze nel corpus di provenienza. Il *test corpus* utilizzato si compone di tutte le frasi del *British National Corpus*[55] (BNC) che contengono le parole estratte, elaborate dal sistema di acquisizione di frame. Una parte del test corpus (16.000 frasi circa) è stata esaminata dai linguisti, che hanno individuato manualmente i frame e ne hanno registrato la frequenza. Questi frame, insieme ai frame attestati in COMLEX, ANLT e NOMLEX, costituiscono il gold standard assunto per la valutazione. La Precision e la Recall ottenute sono:

- per i verbi rispettivamente 81,8% e 59,5% (F-Measure 68,9);
- per i nomi 91,2% e 47,2% (F-Measure 62,2);
- per gli aggettivi 95,5% e 57,6% (F-Measure 71,9).

[52] Grishman, Ralph, Catherine Macleod, and Adam Meyers (1994), COMLEX Syntax: Building a Computational Lexicon. In: *Proceedings of the 15th International Conference on Computational Linguistics*, 268-272. Kyoto, Japan.
[53] Boguraev and Briscoe 1987.
[54] Korhonen et al. offrono la possibilità di creare lessici *ad hoc* nella convinzione che il filtro "migliore" non esiste: esiste una scelta migliore "locale", ma mai "globale", che dipende dall'uso che si intende fare del lessico. A questo proposito, citano un compito di *verb classification* di Korhonen et al. (2003) in cui i risultati migliori sono stati ottenuti con un lessico non filtrato, ricco quindi di frame errati (rumore).
[55] Burnard 1995.

L'alta Precision dei risultati (a discapito di una bassa Recall) si spiega con il fatto che il gold standard contiene frame che non sono stati acquisiti dal classificatore.

Messiant et al. (2008) selezionano 20 verbi francesi per valutare i frame acquisiti da LexSchem con quelli riportati dal *Trésor de la Langue Française Informatisé*[56] (TLFI). Il sistema raggiunge il 79% di Precision e il 55% di Recall, per un punteggio di F-Measure di 65[57].

La Tabella 3.2 riassume i punteggi di F-Measure ottenuti dai sistemi di acquisizione di informazione di sottocategorizzazione presetnati.

Studio	Strategia di valutazione	F-Measure ottenuta
Briscoe e Carroll (1997)	Confronto con Alvey NL Tools Dictionary[58] e COMLEX Syntax Dictionary	46,09
Korhonen (2002)	Analisi manuale: ricerca di 300 occorrenze di ognuno dei *test verbs* nei corpora BNC, LOB[59], SUSANNE[60] e SEC[61].	78,4 * * con linear interpolation per lo smoothing e l'utilizzo di conoscenza probabilistica a priori sulle classi d'appartenenza dei verbi
Korhonen et al. (2006)	Mista: analisi manuale e confronto con COMLEX e ANLT	87,3 nella valutazione manuale 57,30 nella valutazione con i gold standard
Preiss et al. (2007)	Analisi manuale con un *test corpus* estratto dal BNC.	68,9 per i verbi 62,2 per i nomi 71,9 per gli aggettivi
Messiant et al. (2008)	Confronto con TLFI	65

Tabella 3.2 – Un confronto delle prestazioni dei sistemi di acquisizione lessicale presentati

[56] http://atilf.atilf.fr/
[57] Cfr. Poibeau e Messiant (2008) per una trattazione più organica dei problem riscontrati nella valutazione con un gold standard.
[58] Boguraev et al. 1987.
[59] *Lancaster-Oslo-Bergen Corpus* (Garside et al. 1987).
[60] *Susanne Corpus* (Sampson 1995).
[61] *Spoken English Corpus* (Taylor e Knowles 1988).

3.3 La valutazione dei frame acquisiti da LexIt per i verbi

In questa sezione si discute la valutazione dei frame di sottocategorizzazione estratti per i verbi da LexIt dal corpus *La Repubblica*; il procedimento sarebbe del tutto analogo per la valutazione dei frame estratti per i nomi e gli aggettivi, tranne che per la maggiore difficoltà di reperire lessici di sottocategorizzazione che ne trattino le relative strutture argomentali.
La scelta è ricaduta su una valutazione basata sul confronto con tre gold standard (cfr. sezione 3.3.1), dal momento che sono disponibili lessici di sottocategorizzazione per i verbi italiani ragionevolmente confrontabili con LexIt.

Per gli scopi della valutazione sono stati selezionati in modo casuale cento verbi tra i 3.873 estratti dal corpus *La Repubblica*. I verbi estratti sono riportati in Appendice C con le rispettive frequenze di occorrenza. La selezione casuale ha permesso di considerare verbi con frequenze molto diverse: il verbo *dire*, con ben 830.903 occorrenze, e il verbo *miscelare,* appena 429, rappresentano gli estremi di questa distribuzione eterogenea.
In Appendice A sono elencati i frame di sottocategorizzazione acquisiti per i verbi. Si noti che in tutti i frame è indicato lo slot `subj` (soggetto) poiché è un argomento che non deve essere obbligatoriamente espresso in italiano.

3.3.1 I gold standard

I lessici di valenza in lingua italiana non sono numerosi e quelli esistenti, nati con finalità diverse rispetto a LexIt, non annotano tutti lo stesso tipo di informazione lessicale.

Le risorse sviluppate manualmente, ad esempio, non comprendono informazioni sulla frequenza dei frame, che invece sono automaticamente estratte e consultabili in LexIt; per la valutazione sarebbe stato interessante valutare non solo quanti frame di LexIt siano attestati negli altri lessici (e quanti frame sono invece assenti in LexIt), ma anche confrontare il *rango*[62] di questi frame in LexIt e nel gold standard considerato.

Le risorse con cui sono stati confrontati i frame acquisiti da LexIt sono:

- il *PONS Wörterbuch der italienischen Verben* di Peter Blumenthal e Giovanni Rovere (1998), a cui d'ora in poi ci riferiamo come Blumenthal-Rovere;

[62] Per *rango* si intende la posizione occupata da una parola in un ordinamento di frequenza discendente. In una lista di frame, quello con frequenza maggiore ha rango 1.

- Il Sabatini Coletti. Dizionario della Lingua Italiana (Sabatini e Coletti, 2005)[63];
- il lessico PAROLE-SIMPLE-CLIPS (d'ora in poi semplicemente PAROLE), messo a punto dall'Istituto di Linguistica Computazionale (ILC) del CNR di Pisa; è l'unico strumento tra questi che raccoglie le dipendenze, oltre che dei verbi, dei nomi e degli aggettivi.

Blumenthal-Rovere[64] è un dizionario bidirezionale italiano-tedesco dei verbi. Pur poggiando sulla teoria della valenza[65], questo lessico non si limita a descrivere i verbi e le loro reggenze, ma fornisce anche informazioni sul contesto semantico, sul livello stilistico e sugli usi specialistici dei verbi (in particolar modo, data la composizione del corpus, soprattutto in campo economico, tecnico e giuridico).

Le citazioni riportate in qualità di esempi provengono perlopiù da un corpus su supporto informatico molto ampio (50 milioni di parole), che si compone principalmente dei numeri delle annate 1989 e 1990 del quotidiano economico *Il Sole 24 Ore*; compaiono anche, occasionalmente, esempi ricavati da altri quotidiani (*Corriere della Sera* e *la Repubblica*), da pubblicazioni specialistiche (per esempio *Il Foro Italiano* degli anni 1990 e 1991) o da opere letterarie[66]. Circa la metà degli esempi riportati proviene da *Il Sole 24 Ore*.

I verbi attestati, per un totale di 1.729, sono stati estratti sulla base di criteri di frequenza e sono stati confrontati con quelli riportati nel VELI (De Mauro, 1989), che al tempo della redazione costituiva il migliore dizionario italiano di frequenza. I significati di ogni lemma sono descritti in sottosezioni numerate (*sottolemmi*, per un totale di 13753 nel dizionario), ognuno accompagnato dalla formula della struttura sintattica relativa.

Le abbreviazioni utilizzate per descrivere la struttura sintattica sono riportate in Tabella 3.3. La formula tipica presenta in prima posizione il soggetto, se questo è previsto, sotto forma di nome (n), proposizione (S) o infinito (Inf); segue il verbo (V), che se riflessivo viene segnalato con si V.

Alcuni esempi di strutture sintattiche e delle relative formule utilizzate sono riportati in Tabella 3.4.

---

[63] Esiste una versione on-line del dizionario consultabile all'indirizzo http://dizionari.corriere.it/dizionario_italiano/.

[64] Il dizionario è brevemente illustrato in Nied Curcio (2006), pp.64-65.

[65] La teoria della valenza di Lucien Tesnière (1959) assegna al verbo un ruolo centrale nella frase, in quanto centro sintattico della frase da cui dipendono i diversi elementi. La valenza è la proprietà del verbo di richiedere un determinato numero di elementi con i quali combinarsi per formare un concetto di senso compiuto.
L'opera lessicografica di riferimento per Blumenthal-Rovere è Helbig (1992).

[66] Per l'elenco completo consultare Blumenthal e Rovere (1988), pp.XXI-XXIII.

Agg_{pred}	aggettivo in funzione predicativa
Avv	avverbiale in locuzioni verbali (nesso verbo+Avv più o meno lessicalizzato; es. *avanti* in *portare avanti*, *via* in *buttar via*)
Avv_{loc}	avverbiale di luogo (come sottotesto anteposto fino a 2 preposizioni)
Avv_{mis}	avverbiale di misura (misura, prezzo, ecc.)
Avv_{modo}	avverbiale di modo(comprendente anche l'avverbiale strumentale nonché le costruzioni di difficile inquadramento semantico del tipo: "in relazione a + Nome", "in base a + Nome")
Avv_{temp}	avverbiale di tempo
Ger	gerundio
Inf	Infinito
$Inf_{interrog}$	infinito in funzione di interrogativa indiretta
Inf_{pred}	infinito in funzione predicativa
$_{inf}S$	proposizione infinitiva (costruzione dell'infinito con soggetto espresso, riservata in generale allo stile elevato: *Ritengo esser Pietro uno dei nostri validi sostenitori*)
N	(nome o pronome in funzione di) soggetto
N_{plur}	soggetto che indica un plurale o la coordinazione di almeno due N (con N1, N2, N3 $_{plur}$ ha significato analogo)
N_{pred}	sostantivo con funzione predicativa
N1	oggetto diretto
N2	oggetto indiretto (dativo)
N3	tutti gli ulteriori oggetti (preposizionali)
Pred	predicativo (se non compaiono i casi più specifici Agg_{pred} o N_{pred})
Prep	preposizione (specificate nella sezione Gramm. dell'entrata lessicale)
S	preposizione (subordinata) (es. completiva, avverbiale, eventualmente anche relativa)
S_{cong}	subordinata al congiuntivo
$S_{interrog}$	interrogativa indiretta
si V	verbo riflessivo
V	Verbo

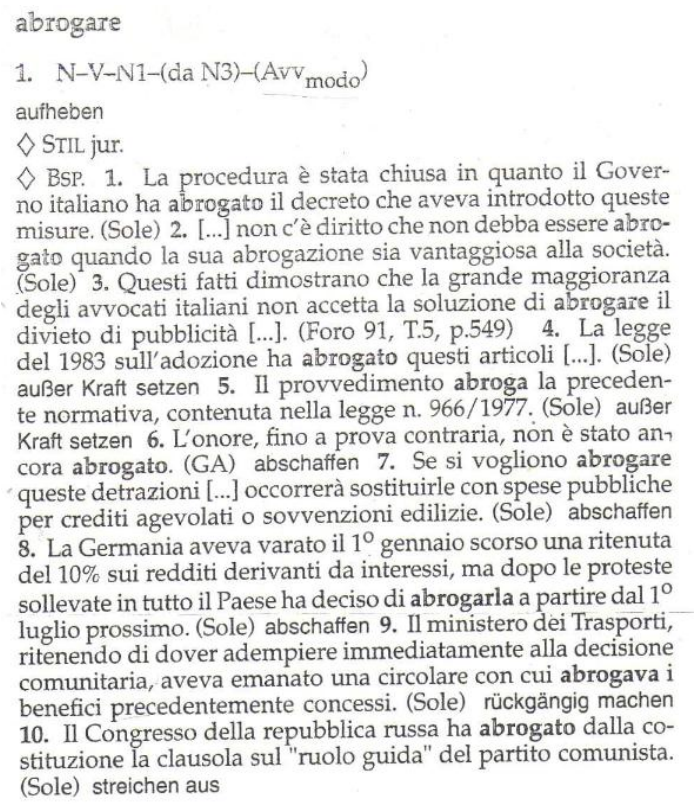
Tabella 3.3 – Abbreviazioni e simboli utilizzati per le strutture sintattiche in Blumenthal-Rovere.

Struttura argomentale	**Esempio**
N-V-N1-(Avv_{temp})	*Ha intimato di sgombrare la piazza entro venerdì.*
N-V-contro/per N3	*La Chiesa predica contro la violenza* *Predicare per i diritti umani non significa indebolire la democrazia.*
N-si V-(Prep N3)	*[...] persone che non si spaventano di fronte a prezzi da record.*
N-(si) V-N1-(in Avv_{loc})	*Rimettersi il cappello in testa.*
N_{plur}-V-(Avv_{mis}/Avv_{modo})	*Abbiamo parlato e discusso, con estrema franchezza.*
N-V-N1-Pred	*Ho trovato irritante la lettera di Luigi.*
N-V-(N1)-a Inf	*Ti sfido a provare quello che dici.*
che S-V-(N2)-Agg_{pred}	*Mi pare difficile che possa riuscire.*
N-V-N1-(Avv_{loc}/in N3)-(Avv_{modo}/Ger)	*In questo filtrato si scioglie agitando il saccarosio.*
N-V-di N3-come (di) N_{pred}	*Si parla di lui come il più probabile successore alla carica.*

Tabella 3.4 – Esempi di codifiche di strutture argomentali in Blumenthal-Rovere.

Per ogni costruzione si riportano uno o più esempi ed eventuali informazioni semantiche, stilistiche (frequenza d'uso, dominio di linguaggio specialistico) e grammaticali di qualsiasi tipo (per esempio l'indicazione che un verbo si trova più frequentemente in forma passiva).

A titolo di esempio, per il verbo *abrogare* in Figura 3.1 Blumenthal-Rovere attesta la struttura sintattica N-V-N1-(da N3)-(Avv_{modo}). Il lemma ha un unico significato ("aufheben", ritirare) e quindi un unico sottolemma. Si trova un'indicazione stilistica ("jur.", linguaggio giuridico) e sono elencati alcuni esempi. Gli elementi tra parentesi sono facoltativi, mentre il complemento diretto N1 è un argomento obbligatorio. Per altri lemmi si possono trovare elementi separati da una barra diagonale: questa convenzione indica elementi posti in alternativa tra loro.



abrogare

1. N–V–N1–(da N3)–(Avv_{modo})

aufheben

◇ STIL jur.

◇ BSP. **1.** La procedura è stata chiusa in quanto il Governo italiano ha **abrogato** il decreto che aveva introdotto queste misure. (Sole) **2.** [...] non c'è diritto che non debba essere **abrogato** quando la sua abrogazione sia vantaggiosa alla società. (Sole) **3.** Questi fatti dimostrano che la grande maggioranza degli avvocati italiani non accetta la soluzione di **abrogare** il divieto di pubblicità [...]. (Foro 91, T.5, p.549) **4.** La legge del 1983 sull'adozione ha **abrogato** questi articoli [...]. (Sole) außer Kraft setzen **5.** Il provvedimento **abroga** la precedente normativa, contenuta nella legge n. 966/1977. (Sole) außer Kraft setzen **6.** L'onore, fino a prova contraria, non è stato ancora **abrogato**. (GA) abschaffen **7.** Se si vogliono **abrogare** queste detrazioni [...] occorrerà sostituirle con spese pubbliche per crediti agevolati o sovvenzioni edilizie. (Sole) abschaffen **8.** La Germania aveva varato il 1° gennaio scorso una ritenuta del 10% sui redditi derivanti da interessi, ma dopo le proteste sollevate in tutto il Paese ha deciso di **abrogarla** a partire dal 1° luglio prossimo. (Sole) abschaffen **9.** Il ministero dei Trasporti, ritenendo di dover adempiere immediatamente alla decisione comunitaria, aveva emanato una circolare con cui **abrogava** i benefici precedentemente concessi. (Sole) rückgängig machen **10.** Il Congresso della repubblica russa ha **abrogato** dalla costituzione la clausola sul "ruolo guida" del partito comunista. (Sole) streichen aus

Figura 3.1 – Entrata lessicale di *abrogare* in Blumenthal-Rovere.

Il dizionario Sabatini-Coletti (2005) include, anche nella sua versione online consultabile sul sito del *Corriere della Sera*, «l'illustrazione, mediante formule intuitive, della proprietà del verbo (valenza) di collegarsi in vari modi (reggenza) agli elementi (argomenti) necessari a formare la struttura portante della frase (il nucleo)»[67]. Gli elementi che possono ampliare il nucleo si distinguono in *circostanti* e in *espansioni*; i circostanti si legano a uno dei costituenti del nucleo, mentre le espansioni si collegano alla frase solo dal punto di vista

[67] http://www.sansoniscuola.it/dizionari/italiano.html

semantico e non da quello sintattico, infatti non pongono vincoli sulla posizione che devono occupare.

Rispetto agli altri lessici, Sabatini-Coletti fornisce un tipo di informazione che si potrebbe definire "di alto livello" e distingue le sole strutture sintattiche in Tabella 3.5. Come si nota, vengono distinti i ruoli sintattici di soggetto, argomento (inteso come complemento oggetto), argomento preposizionale (complemento indiretto introdotto da preposizione) e complemento predicativo.
La funzione logica espressa dal complemento indiretto non è specificata nella formula, ma viene discussa all'interno dell'entrata lessicale.

Verbi transitivi (v. tr.)	Verbi intransitivi (v. intr.)	Verbi riflessivi (v .rifl.)
[sogg-v-arg] [sogg-v-arg-prep.arg] [sogg-v-arg+**c**ompl.pred]	[sogg-v] [sogg-v-prep.arg]	[sogg-v] [sogg-v-arg] [sogg-v-prep.arg] [sogg-v-compl.pred]

Tabella 3.5 - Elenco dei frame attestati nel Sabatini-Coletti per le diverse forme del verbo.

PAROLE (Ruimy et al., 1998) è un lessico italiano costruito nell'ambito di tre diversi progetti. La parte morfologica e sintattica rispetta il modello del progetto europeo LE-PAROLE, che si propone di fornire un modello per la creazione di risorse linguistiche (corpora e lessici) per diverse lingue dell'Unione Europea. Le linee guida specificano che le risorse debbano essere estese, generiche, riusabili e interoperabili tra loro.
PAROLE è costruito nel rispetto di questi principi e si basa sulle raccomandazioni per l'informazione morfosintattica e la sintassi dei verbi di EAGLES (Sanfilippo et al., 1996).
Il livello semantico ha come riferimento il progetto, anch'esso europeo, LE-SIMPLE[68]. Infine, il livello fonologico e l'estensione della copertura lessicale si collocano nell'ambito del progetto italiano *Corpora e Lessici dell'Italiano Parlato e Scritto* (CLIPS).
Il livello sintattico comprende 28.133 lemmi[69], selezionati sulla base di un criterio di frequenza dall'*Italian Reference Corpus* dell'ILC (IRC) (Bindi et al., 1991). Sono presenti i profili sintattici di 3.000 verbi.
A differenza del Blumenthal-Rovere, la varietà di lingua rappresentata non è settoriale e attesta fenomeni linguistici (e lessicali) recenti.
A livello sintattico, una voce lessicale è codificata come una o più unità sintattiche (corrispondenti a elementi XML *SynU*). Un'unità sintattica, in quanto elemento XML, ha

[68] Il sito web di riferimento dei progetti LE-PAROLE e LE-SIMPLE è http://www.ub.edu/gilcub/SIMPLE/simple.html.
[69] http://www.ilc.cnr.it/clips/SPEC_SINTASSI/SINTASSI_1.doc

diversi attributi, uno dei quali è la descrizione (*Description*) che esprime il comportamento sintattico; il valore di questo attributo è l'id dell'oggetto Description associato. Quest'ultimo, a sua volta, è definito dagli attributi *Construction* e *Self:* Construction rimanda all'oggetto omonimo che descrive la struttura argomentale, mentre Self codifica le proprietà dell'unità lessicale considerata nel contesto descritto sintattico da Construction.
Per quanto riguarda la categoria grammaticale de verbi, Self fornisce informazioni sulla sottoclasse (copula, verbo pronominale, verbo riflessivo ecc.), sulla forma passiva (esclusiva o, al contrario, inibita[70]), sull'ausiliare ecc.

Ogni slot in un frame assolve a una funzione sintattica, che può essere ugualmente realizzata da diversi filler in relazione paradigmatica tra loro. La Figura 3.2 (tratta da Ruimy et al., 1998: 4) mostra un esempio di forme alternative per le funzioni soggetto e oggetto nel verbo *chiarire*, concettualmente raccolte in un'unica Description.

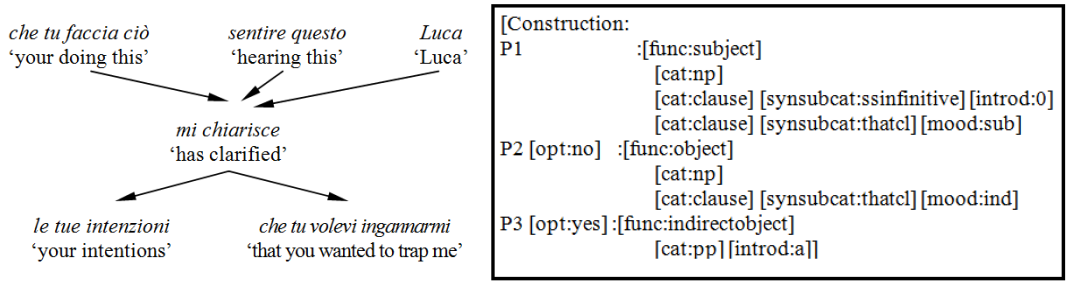


Figura 3.2: Realizzazioni di soggetto e oggetto per il verbo *chiarire* e relativa rappresentazione concettuale.

Nella Figura 3.3 si riporta la SynU - e la Description ad essa collegata - che esprime il comportamento sintattico di Figura 3.2. Per una trattazione esaustiva delle specifiche linguistiche del livello sintattico si può consultare il sito web di CLIPS[71].

```
<SynU
  id="SYNUchiarireV2"
  naming="chiarire"
  comment="trsclocl"
  example="chiarire qlco a qlcu/   ti
ho già chiarito che volevo fare questo
/ ha chiarito di voler continuare in
tal senso"
  description="t8thindorinfdiCsC-
indopt-xa">
</SynU>
```

```
<Description
  id="t8thindorinfdiCsC-indopt-xa"
  example="Luca comunica a qlcu una
notizia; - che e' arrivato; - di aver
finito"
  self="SELFVxa"
  construction="t8thindorinfdiCsC-
indopt"/>
```

Figura 3.3 – Un'unità sintattica del verbo *chiarire* (identificatore SYNUchiarireV2) e relativa Description associata (identificatore `t8thindorinfdiCsC-indopt.xa`).

[70] Il verbo *fruttare*, per esempio, permette solo forma attiva (**questo terreno è fruttato molto*), mentre il verbo *costituire* solo forma passiva (**la casa costituisce 3 vani*).
[71] http://www.ilc.cnr.it/clips/SPEC_SINTASSI/SINTASSI_indice.htm. Di particolare aiuto risultano le Appendici A e B, rispettivamente una guida alla lettura degli identificatori delle strutture sintattiche e l'elenco degli stessi per ogni categoria grammaticale.

3.3.2 Criteri di confronto

In questa sezione vengono illustrate le linee guida che sono state adottate per confrontare i frame riportati dai gold standard con quelli di LexIt.
Come discusso nella sezione 3.3.1, i gold standard utilizzati hanno livelli diversi di granularità; è frequente che, per lo stesso verbo, Sabatini-Coletti elenchi un numero ridotto di frame, mentre Parole e Blumenthal-Rovere forniscano invece una lista ben più lunga e articolata delle strutture argomentali in cui il verbo può occorrere.
I frame di LexIt, a loro volta, differiscono dalle strutture argomentali di ognuno dei gold standard per livello di astrazione e pretese di completezza. Rispetto a Sabatini-Coletti, LexIt riporta frame maggiormente dettagliati. Gli stessi frame, tuttavia, saranno talvolta meno dettagliati e informativi rispetto a quelli di Blumenthal-Rovere o di PAROLE, che sono stati realizzati manualmente da lessicografi.

Di conseguenza, per ogni coppia LexIt-gold standard è necessario elaborare una sorta di *mappatura* che illustri i criteri con cui si fanno corrispondere i frame del gold standard a quelli di LexIt. Questo passaggio è necessario per permettere un confronto tra risorse che utilizzano codifiche sintattiche molto diverse tra loro.

Di seguito elenchiamo alcuni dei problemi riscontrati nel confronto dei frame dei gold standard con i frame di LexIt e le soluzioni che sono state adottate, che esprimono i "criteri di equivalenza" usati.

Distinzione tra argomenti e aggiunti - Un problema fondamentale nella classificazione dei frame risiede nella natura stessa di LexIt: trattandosi di un sistema di estrazione automatica, LexIt estrae per ogni verbo tutte le possibili costruzioni sintattiche che ritrova attestate nel corpus, siano esse significative o meno per quel verbo, e le riporta tutte indistintamente. In altre parole, non viene svolta alcuna selezione ad hoc dei frame rispetto al verbo *v* considerato.
I gold standard, costruiti manualmente, riportano invece solo le strutture argomentali "di base" in cui compare il verbo in esame.

La differenza che intercorre tra i gold standard e LexIt può essere chiarita con un esempio: per il verbo bivalente *accontentare* tutti i gold standard riportano il frame che in LexIt corrisponderebbe a `subj#obj`, in cui tutti gli argomenti sono obbligatori per il significato del verbo. Consideriamo la frase «Ho accontentato mio figlio per il suo compleanno», la cui struttura argomentale si potrebbe esprimere in LexIt con il frame `subj#obj#comp-per`: lo

slot `comp-per` corrisponde a un aggiunto per il verbo *accontentare*, ma LexIt non dispone di regole per distinguere a monte, fin dall'estrazione dei frame, gli argomenti dagli aggiunti[72], quindi estrae anche il frame `subj#obj#comp-per` per *accontentare*.
Nei gold standard, naturalmente, i frame riportati includono solo gli argomenti necessari per la compiutezza del verbo, ma è implicito che i frame possano essere arricchiti con aggiunti.
In LexIt, invece, il frame `subj#obj` indica una struttura argomentale in cui compaiono esattamente un soggetto e un oggetto *e nient'altro*. Le "estensioni" dei frame che possono essere estratte dal corpus in LexIt sono a loro volta frame, che ai fini della valutazione vengono registrati come TP.
Il primo criterio utilizzato è basato sulla seguente assunzione: frame "ricchi" ma chiaramente errati (che hanno, per esempio, filler improbabili) o statisticamente irrilevanti sono registrati come FP. Il più delle volte i frame errati sono dovuti a un parsing scorretto; in frasi molto complesse, per esempio, è frequente che un sintagma preposizionale che dipende da un nome venga invece riferito, erroneamente, al verbo.

Sottospecificazione (e "sovraspecificazione") delle preposizioni – Sabatini-Coletti sottospecifica le preposizioni ammesse nei complementi preposizionali, mentre LexIt estrae i frame direttamente dall'input linguistico e fornisce, quindi, un'informazione più granulare. Non esistendo una perfetta corrispondenza tra le due "codifiche sintattiche", si è deciso di considerare TP quei frame di LexIt che, oltre a poter essere generalizzati nei frame del Sabatini-Coletti, siano anche *plausibili* per il verbo in esame. Si considerino per esempio alcuni dei frame di LexIt per il verbo *distare* (24). Il Sabatini-Coletti attesta un generico frame `[sogg-v-arg-prep.arg]` ("*La casa dista un chilometro dal paese*"), che corrisponde perfettamente al frame 24a, ma che comprenderebbe anche i meno frequenti (e soprattutto meno sensati: *"*La casa dista per 8 km"*) frame 24b, 24c e 24d.
Questi ultimi, non essendo ritenuti dei frame validi per *distare*, sono registrati come FP.

(24)
```
a. subj#obj#comp-da       freq. 195
b. subj#obj#comp-verso    freq. 1
c. subj#obj#comp-su       freq. 1
d. subj#obj#comp-per      freq. 2
```

Blumenthal-Rovere in certi casi fornisce un elenco molto ampio delle preposizioni ammesse, talvolta poco frequenti o addirittura desuete nell'italiano corrente (perlopiù afferenti ad un sottodominio linguistico, come quello giuridico) che, non essendo attestate nel corpus, non

[72] Ad ogni frame, però, è associato un valore (LMI, cfr. sezione 1.3.1) che dà una misura della salienza statistica del frame per il verbo, che privilegia i frame che contengono argomenti tipici per il verbo.

compaiono in LexIt: i frame relativi sono registrati come FN[73].
In altri casi, Blumenthal-Rovere omette di elencare le preposizioni lecite per i complementi, limitandosi a indicare la funzione logica; per il verbo *circolare* Blumenthal-Rovere attesta le strutture argomentali in (25).

(25) a. *Al Senato circolano ipotesi su una possibile astensione del gruppo socialista.* N-V-(Avv_{loc})
b. *Greenspan ha confermato quanto già circolava nei mercati.*

c. *Venerdì […] si circolerà in tutta la Lombardia a targhe alterne.* N-V-(Avv_{modo})
d. *La notizia circolava con insistenza all'Associazione industriali.*

Come per Sabatini-Coletti, il criterio adottato è quello di considerare TP i frame di LexIt sensati per il verbo e compatibili con quelli attestati nel gold standard.

PAROLE, come si è accennato, utilizza una codifica ancora più complessa e rigorosa. Nelle Description dei verbi sono elencate le sole preposizioni ammesse (mai più di tre nel caso dei verbi selezionati per la valutazione) per una data posizione sintattica.
Consideriamo il verbo *dislocare*, a cui PAROLE associa la Description `t-adjppinVsu-xa`, in cui:

- `t` indica un verbo transitivo;
- `xa` indica l'ausiliare avere;
- `adjppinVsu` indica che il secondo complemento ha la funzione grammaticale di aggiunto (`adj`), realizzato da un sintagma preposizionale (`pp`) introdotto dalla preposizione *in* o *su*.

Questa Description corrisponde, in senso stretto, ai frame di LexIt `subj#obj#comp-in` e `subj#obj#comp-su`.

Come anche Blumenthal-Rovere, LexIt attesta diverse altre strutture argomentali plausibili per il verbo in esame, che verrebbero però considerate FP se si adottassero regole di corrispondenza drastiche; per esempio, sarebbe classificato come FP il frame `subj#obj#comp-a` ("*Abbiamo dislocato le truppe al confine"*).
Questo frame non è altro che una variante del complemento locativo espresso dall'aggiunto; poiché assolve bene alla funzione sintattica descritta in PAROLE, viene considerato TP.

[73] Per il verbo *gravare*, per esempio, Blumenthal-Rovere attesta la frase "Pertanto il Negidi, con atto d'appello […] si gravava avverso la sentenza del tribunale" (AA.VV, Formulario della procedura civile, Milano 1988), il cui frame in LexIt equivarrebbe a subj#si#comp-avverso; la preposizione è desueta e appartiene a una varietà del linguaggio di dominio, perciò è prevedibile che non compaia nel corpus *La Repubblica* utilizzato per l'estrazione dei frame; il corpus è giornalistico e in quanto tale riflette una lingua viva.

Polisemia del frame – In Blumenthal-Rovere all'annotazione sintattica si combina quella semantica. Nel caso una struttura argomentale si presti a veicolare più di un senso, sono attestati tanti frame quanti sono i significati che il verbo può assumere.
Per chiarezza, si consideri il frame `N-V-N1` (corrispondente a `subj#obj` in LexIt) del verbo *degradare* (26). LexIt dispone di informazioni sulle classi semantiche dei filler, ma non si propone di distinguere i diversi sensi che possono essere assunti da uno stesso frame; i frame polisemici in (26) vengono tutti ricondotti al frame `subj#obj`.

(26)	a.	`N-V-(N1) entehren`	`disonorare`	*Questa azione ti degrada.*
	b.	`N-V-N1 zerstören`	`distruggere`	*Degradare l'ambiente.*
	c.	`N-V-N1 abbauen`	`ridurre`	*La flora batterica degrada l'amido.*
	d.	`N-V-N1 abtönen (raro)`	`colorare`	*Degradare i colori.*

Occorrenza significativa dei frame – Nel confronto tra PAROLE e LexIt, considerando il rigore delle Description del primo, i frame di LexIt sono stati categorizzati in base a principi più restrittivi: per essere ritenuto TP, un frame deve non solo essere sensato per il verbo, ma anche occorrere nel corpus in misura significativa; in altre parole, nel caso di una struttura argomentale incerta da valutare, è una sua alta frequenza a sciogliere il dubbio e a fornire una prova della sua "validità" per quel verbo. Dal momento che LexIt permette l'esplorazione dei frame, nel caso di frequenze medio-basse sono stati esaminati uno per uno i filler che riempiono gli slot; in questo modo sono anche stati individuati i casi in cui lampanti errori di parsing avevano portato all'estrazione di un frame non corretto. Si consideri il verbo *evitare*, per cui PAROLE riporta le Description `t-xa`, `t4thsuborfinf058infdiCoC-ind-xa`, `t8thsubCnotsCorinfdiCsC-xa` e `rr-xerec`, rispettivamente equivalenti, in senso stretto, ai frame di LexIt `subj#obj`, `subj#obj#inf-di` e `subj#si#0`. LexIt attesta anche il frame `subj#obj#comp-a`, che ha frequenza alta (1904) e filler come *momento, costo, causa* (27).

(27) *Lo evito a costo di cambiare bar.*

In questo caso, poiché il frame `subj#obj#comp-a` "esteso"da `subj#obj` risulta sensato per il verbo, esso viene considerato TP per PAROLE. Ciò non avviene, ad esempio, per il frame `subj#obj#comp-sotto` dello stesso verbo, che oltre alla frequenza bassissima (8) ha filler improbabili (*peso, influsso, sole, bomba*) che fanno propendere per occasionali errori del parsing.

Soggetto frasale – PAROLE e Blumenthal-Rovere distinguono diverse realizzazioni del soggetto e hanno dei frame dedicati per i soggetti frasali. Il verbo *costare*, per esempio, prevede che la funzione di soggetto possa essere assolta, oltre che da un sintagma nominale,

anche da una proposizione soggettiva all'infinito (28a, 28b) oppure da una completiva al congiuntivo (28c).

(28)	a. *Non ti costa nulla essere gentile* (BR)	`Inf-V-(N1)-N2`
	b. *Partire mi è costato tanto* (P)	`i4thsubCnotIcorinfOCiC5-ind-xe`
	c. *Che se ne sia andato così mi è costato* (P)	`i4thsubCnotIcorinfOCiC5-ind-xe`

La Description di Parole (28b e 28c) specifica che per questo verbo intransitivo (`i`) il soggetto (tra `4` e `5`) può essere realizzato con una *that clause* (`thsub`) non coreferente con l'oggetto indiretto della principale (`CnotiC`), oppure (`or`) con una soggettiva all'infinito (`inf`) coreferente con l'oggetto indiretto (`CiC`).
LexIt non distingue tra le funzioni soggettive ed oggettive delle proposizioni completive perciò questi frame non sono stati computati agli scopi della valutazione, o meglio sono stati tutti ricondotti indistintamente al frame `subjobj#comp-a`.

Forma riflessiva e forma pronominale – Un verbo ha forma riflessiva quando il soggetto compie l'azione e allo stesso tempo la subisce. La forma riflessiva di distingue in:

- forma riflessiva propria: l'azione si riflette sul soggetto e le particelle pronominali hanno funzione di complemento oggetto ("Io mi lavo");
- forma riflessiva apparente: le particelle pronominali svolgono la funzione di complemento di termine ("Io mi lavo i capelli");
- forma riflessiva reciproca: l'azione compiuta da due soggetti ricade su di essi ("Le sorelle si abbracciano").

I verbi detti *pronominali* (accompagnati da una particella pronominale: *accanirsi, pentirsi, ribellarsi, vergognarsi…*) hanno una forma simile ai verbi riflessivi, ma essendo intransitivi non hanno alcun valore riflessivo. Le particelle pronominali sono parte integrante del verbo e non corrispondono ad alcun complemento: "Io mi pento".

Sabatini-Coletti adotta una concezione unitaria della riflessività dei verbi: non distingue le forme riflessive e non include il pronome riflessivo tra gli argomenti previsti per un verbo, perché «il pronome riflessivo non indica un argomento diverso dal soggetto»[74]. Allo stesso modo, non sono segnalati i verbi che indicano reciprocità (almeno nella versione on-line del dizionario).

Come Sabatini-Coletti, Blumenthal-Rovere non distingue le diverse forme di riflessività (tutte rappresentate con l'etichetta `si V`). La reciprocità del verbo non viene esplicitamente dichiarata, ma è segnalata dall'abbreviazione `Nplur`, che indica un soggetto plurale.

[74] Sabatini e Coletti 2005, pp. X-XI.

PAROLE codifica le forme della forma riflessiva - e della forma pronominale in genere - con estrema precisione. Le Description utilizzate sono le seguenti:

- `r-xeref` esprime la forma riflessiva propria ("Io mi lavo");
- `rimp-xeref` esprime la forma riflessiva impropria ("Io mi lavo i capelli")[75];
- `ip-xepro` esprime la forma intransitiva pronominale ("Il vaso si è rotto")[76].
- `tp-xepro` esprime la forma transitiva pronominale per i casi in cui il soggetto sembra avere, piuttosto che il ruolo semantico di agente, quello di paziente (*"Ho rotto il naso a me stesso" vs "Mi sono rotto il naso").
- `rr-xerec` indica la struttura reciproca, in cui compaiono «un soggetto plurale o una coordinazione di soggetti appartenenti alla stessa classe semantica e una forma verbale transitiva preceduta dal pronome clitico riflessivo» ("Luca e Piero si aiutano").

Queste distinzioni applicate in PAROLE non sono catturate in LexIt, che codifica tutte le situazioni elencate con il frame `subj#si#0`. Infatti LexIt fa intrinsecamente astrazione rispetto alle diverse funzioni del pronome riflessivo.
Si considerino le strutture sintattiche del verbo *accordare* in PAROLE (29). Le forme transitive 29a, 29b e 29c si riconducono al frame `subj#obj` di LexIt e a sue estensioni; le strutture sintattiche rimanenti vengono astratte e unificate in LexIt nel frame `subj#si#0` (o in sue estensioni).

(29)	a. *Accordare il violino*	`t-xa`
	b. *Accordare il colore dell'abito a quello degli accessori*	`t-ppa-xa`
	c. *Accordare la fede con la ragione*	`t-ppcon-xa`
	d. *Mi sono accordato con Carlo*	`ip-ppcon-xepro`
	e. *Che tu faccia questo non si accorda con le tue promesse*	`ip4thsuborinf05-ppcon-xepro`
	f. *I tuoi accessori si accordano perfettamente al tuo vestito*	`ip-ppa-xepro`
	g. *L'aggettivo si accorda con il sostantivo*	`ip-ppcon-xepro`
	h. *Marco e Carlo si sono accordati sul prezzo*	`rr-ppsuopt-xerec`
	i. In italiano l'aggettivo e il sostantivo si accordano	`rr-xerec`

Riassumendo, i criteri utilizzati per classificare i frame di LexIt come TP, FP o FN sono i seguenti:

[75] Ben 63 altre description danno conto di strutture riflessive con più di due argomenti (complemento preposizionale, complemento predicativo del soggetto, complemento frasale).
[76] Come sopra: 133 description codificano costruzioni più ricche.

1) affinché un frame venga registrato come TP, è condizione necessaria ma non sufficiente che esso sia sensato per il predicato e, in situazioni d'incertezza, che sia ricorrente in modo statisticamente significativo;
2) se un frame che contiene un complemento è giudicato TP, si considerano TP anche sue eventuali varianti in LexIt che assolvono alla medesima funzione logica (v. complemento locativo espresso da più preposizioni);
3) sono registrati come TP i frame di LexIt che estendono con aggiunti (argomenti circostanziali) un frame già giudicato TP;
4) i frame polisemici vengono computati una sola volta, considerando la sola struttura argomentale;
5) le possibili realizzazioni del soggetto sono sottospecificate e ricondotte allo slot `subj`;
6) le diverse forme riflessive del verbo (propria, apparente, reciproca) e la forma intransitiva pronominale sono sottospecificate e ricondotte al frame `subj#si#0`.

3.3.3 La valutazione

La valutazione di LexIt, come quella di altri sistemi di acquisizione lessicale, si basa sulla selezione di frame che rispettano certi indici statistici, allo scopo di escludere fonti di rumore dovute a errori di parsing (Korhonen 2002, Lenci et al. 2012).
Ad ognuno dei frame estratti da LexIt sono state associate le seguenti informazioni:

- frequenza relativa del frame per il verbo considerato, ovvero la MLE. È la stessa misura discussa da Korhonen (2002) e applicata da Preiss et al. (2007) e da Messiant et al. (2008), definita come:

$$MLE = \frac{frequenza < verbo; frame >}{frequenza\ verbo}$$

- valore di forza d'associazione tra il verbo e il frame, misurato con LMI (cfr. sezione 1.4.1)[77]. Sono valutati solo i frame con LMI positiva.

[77] Ricordando che la LMI correla la frequenza attesa (la frequenza che avrebbero verbo e frame se fossero statisticamente indipendenti) con quella osservata, si può dire che il verbo e il frame manifestano un'associazione se la frequenza osservata è sensibilmente diversa di quella attesa; l'associazione sarà positiva se la frequenza osservata è significativamente maggiore di quella attesa, negativa altrimenti.
Si consideri il frame `subj#obj` del verbo *accontentare*, che ha frequenza relativa molta alta (19%), ma LMI negativa. Se, come discusso nella sezione 3.3.2, venissero esclusi i frame di LexIt che rappresentano estensioni dei frame dei gold standard, il frame `subj#obj` per *accontentare* non risulterebbe affatto catturato da LexIt: poiché questo accadrebbe spesso con frame `subj#obj`, molto frequenti nei verbi transitivi, si ritiene sensato considerare validi i frame "estesi" con complementi circostanziali, purché validi per il verbo in esame.

Per ogni verbo sono stati confrontati, secondo i criteri descritti, i frame estratti da LexIt rispetto ai frame attestati in ognuno dei gold standard, per un totale di oltre 6000 frame confrontati. La Figura 3.4 mostra il giudizio assegnato ai 39 frame del verbo *abrogare* (freq. 1409) acquisiti da LexIt. Per ogni frame si riporta la frequenza assoluta (FREQLEX) e la frequenza relativa (FREQREL) della coppia verbo-frame, il punteggio della misura di associazione (ASSOFLEXIT) e valori binari (0/1) per segnalare l'assenza o la presenza del frame nei gold standard (BR per Blumenthal-Rovere, SC per Sabatini-Coletti e infine PAROLE) o in LexIt (per quanto riguarda i FN).

VERBO	FREQV	FRAME	FREQFLEX	FREQREL	ASSOFLEXIT	LEXIT	BR	SC	PAROLE
abrogare	1409	subj#obj	817	0,5798	848.0333	1	1	1	1
abrogare	1409	subj#obj#comp-su	29	0,0206	52.7967	1	1	1	1
abrogare	1409	subj#0	301	0,2136	38.7967	1	0	0	0
abrogare	1409	subj#obj#comp-per	21	0,0149	26.6439	1	1	1	1
abrogare	1409	subj#obj#comp-di	15	0,0106	18.2619	1	1	1	1
abrogare	1409	subj#obj#comp-con	18	0,0128	6.7158	1	1	1	1
abrogare	1409	subj#si#obj#fin-che	1	0,0007	3.3859	1	0	0	0
abrogare	1409	subj#si#inf-0	4	0,0028	2.8871	1	0	0	0
abrogare	1409	subj#si#comp-attraverso	1	0,0007	2.8719	1	0	0	0
abrogare	1409	subj#obj#comp-senza	2	0,0014	2.443	1	1	1	1
abrogare	1409	subj#obj#inf-da	2	0,0014	2.0712	1	0	0	0
abrogare	1409	subj#obj#comp-tra	2	0,0014	1.7246	1	1	1	1
abrogare	1409	subj#si#obj#comp-per	1	0,0007	1.6266	1	0	0	0
abrogare	1409	subj#obj#comp-attravers	1	0,0007	1.458	1	1	1	1
abrogare	1409	subj#si#obj#comp-da	1	0,0007	1.4357	1	0	0	0
abrogare	1409	subj#obj#comp-contro	1	0,0007	0.4937	1	1	1	1
abrogare	1409	subj#obj#fin-che	2	0,0014	0.481	1	1	1	1
abrogare	1409	subj#si#obj#comp-con	1	0,0007	0.4207	1	0	0	0
abrogare	1409	subj#comp-per	12	0,0085	0.0729	1	0	0	0
abrogare	1409	subj#comp-senza	1	0,0007	-0.1055	1	0	0	0
abrogare	1409	subj#obj#comp-da	8	0,0057	-0.1705	1	1	1	1
abrogare	1409	subj#obj#inf-di	1	0,0007	-0.5171	1	0	0	0
abrogare	1409	subj#obj#inf-0	2	0,0014	-1.2998	1	0	0	0
abrogare	1409	subj#si#comp-per	1	0,0007	-1.3014	1	0	0	0
abrogare	1409	subj#obj#comp-come	1	0,0007	-1.343	1	0	0	0
abrogare	1409	subj#si#obj	19	0,0135	-1.5777	1	0	0	0
abrogare	1409	subj#comp-come	2	0,0014	-2.0786	1	0	0	0
abrogare	1409	subj#obj#inf-a	4	0,0028	-3.9689	1	1	1	1
abrogare	1409	subj#comp-con	17	0,0121	-5.1538	1	0	0	0
abrogare	1409	subj#obj#comp-in	32	0,0227	-7.5844	1	1	1	1
abrogare	1409	subj#inf-a	2	0,0014	-7.9493	1	0	0	0
abrogare	1409	subj#comp-da	21	0,0149	-9.1456	1	0	0	0
abrogare	1409	subj#comp-di	6	0,0043	-13.1074	1	0	0	0
abrogare	1409	subj#cpred	5	0,0035	-13.1818	1	0	0	0
abrogare	1409	subj#inf-0	5	0,0035	-13.3315	1	0	0	0
abrogare	1409	subj#comp-in	7	0,0050	-18.6382	1	0	0	0
abrogare	1409	subj#comp-a	5	0,0035	-20.0988	1	0	0	0
abrogare	1409	subj#si#0	20	0,0142	-24.9285	1	0	0	0
abrogare	1409	subj#obj#comp-a	18	0,0128	-28.6311	1	1	1	1

Figura 3.4 – Il confronto tra LexIt e i gold standard per il verbo *abrogare*

Con l'ausilio di uno script Perl, per ogni verbo sono state calcolate Precision, Recall e F-Measure, considerando soglie crescenti di MLE e LMI rispetto alle quali selezionare i frame da conteggiare nel computo dei TP e dei FP.

La media dei punteggi ottenuti dai singoli verbi costituisce i valori di Precision, Recall e F-Measure a cui fare riferimento per stabilire il valore del sistema di acquisizione di frame di LexIt.

Le soglie considerate per la MLE vanno da 0 (nessun filtro) a 0,1 con incremento di 0,001. Le soglie considerate per la LMI vanno da 0 (frame con LMI positiva) a 10.000: il grande dislivello tra i due estremi dipende dai valori molto alti che può assumere la LMI.

3.4 Risultati e conclusioni

Nell'Appendice D sono riportati i punteggi ottenuti per ognuno dei gold standard alle diverse soglie di MLE. La soglia di MLE empiricamente determinata che permette di raggiungere la F-Measure più alta è 0,018, mentre la soglia di LMI oscilla tra 26 e 29 nei diversi gold standard.

Nelle Tabelle 3.6 e 3.7 sono riportate le medie dei valori di Precision, Recall e F-Measure dei cento *test verbs* a queste soglie (0,018 per MLE, 26 per LMI) rispetto ad ognuno dei gold standard. Si nota, per ogni gold standard, una Recall molto alta (addirittura oscillante tra 96 e 97% per PAROLE con entrambe le soglie). Riferendoci alla MLE, la Precision (nettamente) migliore si ottiene a confronto con il Blumenthal-Rovere, che non a caso elenca quasi sistematicamente un numero molto più alto di frame rispetto a Sabatini-Coletti e a PAROLE.

Gold Standard	Precision	Recall	F-Measure
Blumenthal-Rovere	0,78	0,91	0,82
Sabatini-Coletti	0,69	0,95	0,78
PAROLE	0,69	0,97	0,78

Tabella 3.6 – I migliori punteggi di Precision, Recall e F-Measure ottenuti con soglia di MLE

Gold Standard	Precision	Recall	F-Measure
Blumenthal-Rovere	0,82	0,92	0,85
Sabatini-Coletti	0,80	0,95	0,85
PAROLE	0,77	0,96	0,84

Tabella 3.7 – I migliori punteggi di Precision, Recall e F-Measure ottenuti con soglia di LMI

Per quanto riguarda i punteggi ottenuti considerando i frame con LMI maggiore a soglie crescenti, la Precision qui ottenuta migliora di circa dieci punti percentuali rispetto alla soglia di MLE che dà migliori risultati, a dimostrazione del fatto che questa misura permette di escludere del rumore e, privilegiando i frame frequenti, aiuta a distinguere sommamente, a posteriori, gli argomenti dagli aggiunti.

I verbi che hanno totalizzato i punteggi più bassi sono *corteggiare* (F-Measure attestata intorno a 0,57 in ognuno dei gold standard) e *vergognarsi* (F-Measure oscillante tra 0,47 e 0,53). In entrambi i casi la causa è da imputare ad una Precision molto bassa, che non supera il 4% . I verbi per cui, invece, i frame acquisiti sono paragonabili a quelli dei gold standard sono *costare*, *peggiorare*, *recitare*, *sfondare*, *terminare*, che alla soglia 0,018 di MLE hanno tutti F-Measure pari a 1; in altre parole, i frame estratti per questi verbi sono stati tutti giudicati TP e hanno quindi raggiunto il punteggio massimo di Precision e Recall.

Le figure 3.5 e 3.6 mostrano l'andamento della F-Measure nel confronto con i tre gold standard, rispettivamente per le soglie di MLE e di LMI.

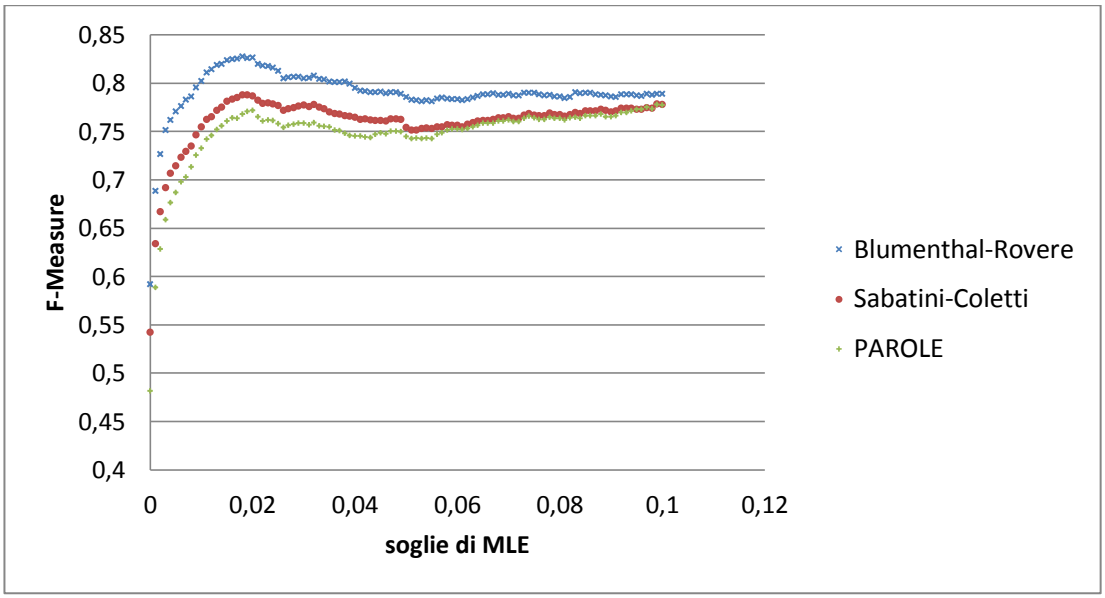


Figura 3.5 – Andamento di F-Measure per soglie crescenti di MLE

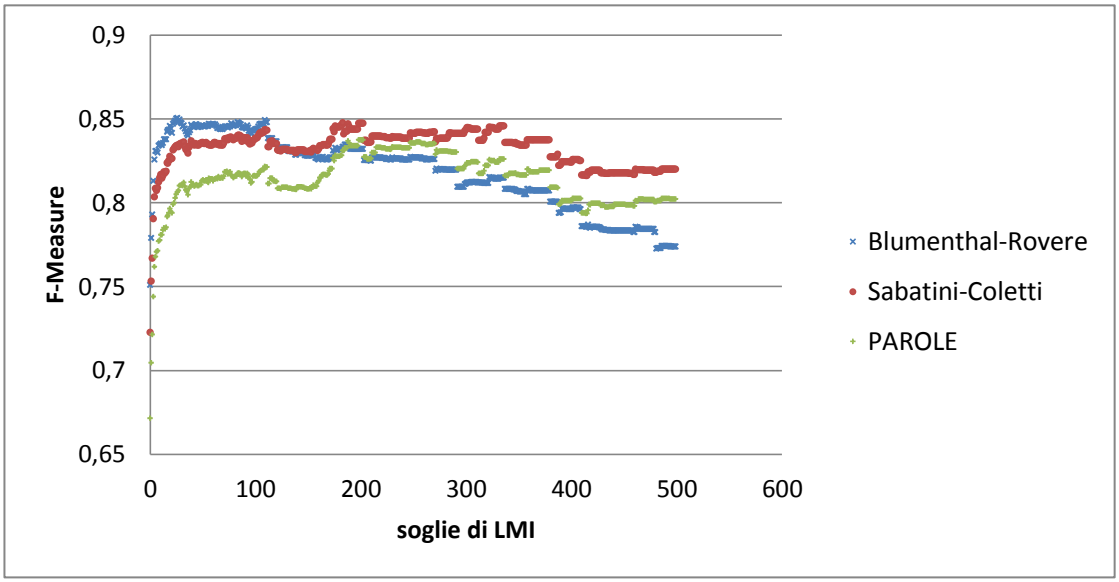


Figura 3.6 – Andamento di F-Measure per soglie crescenti di LMI

Conclusioni

I sistemi per l'acquisizione automatica di informazione lessicale possono essere di estrema utilità per numerose applicazioni di NLP. Come si è visto, l'estrazione automatica di informazione lessicale consente, in parallelo, di ottenere dati quantitativi sulla salienza statistica dei fenomeni. Il principio su cui si sorregge LexIt, il sistema di acquisizione di frame sintattici presentato, si è dimostrato corretto: è possibile descrivere le proprietà di sottocategorizzazione dei predicati e le loro preferenze di selezione in termini distribuzionali.
La valutazione dei frame di sottocategorizzazione acquisiti per i verbi da LexIt ha dato risultati che si allineano allo stato dell'arte e che confermano, dati gli alti valori di Precision, Recall e F-Measure, le potenzialità di sistemi automatici di acquisizione lessicale.
LexIt, tuttavia, non è esente da errori, in quanto si poggia su strumenti di lemmatizzazione e annotazione morfosintattica che, inevitabilmente, non sono sempre in grado di riconoscere l'input e producono dipendenze errate a livello sintattico. Poiché l'acquisizione dei profili distribuzionali di LexIt è totalmente automatica, gli errori prodotti si propagano ai livelli successivi dell'estrazione; l'ausilio dei dati statistici e di misure di associazione riesce a contenere almeno parte del rumore generato, privilegiando i dati linguistici significativamente frequenti.
La valutazione qui proposta per i verbi è sostanzialmente replicabile per i frame di sottocategorizzazione acquisiti per i nomi e gli aggettivi[78].

Ad oggi è in corso l'estrazione dei profili distribuzionali degli aggettivi dal corpus *Wikipedia*. Data la crescente – inarrestabile – disponibilità di corpora, si prospetta la possibilità di estrarre i profili distribuzionali da corpora di dominio, favorendo così un interessante paragone con i frame acquisiti da corpora generalisti come *La Repubblica* e *Wikipedia*.
Per quanto riguarda gli sviluppi futuri di LexIt[79], come auspicato da Korhonen et al. (2006), «è necessaria una rappresentazione degli aspetti semantici della sottocategorizzazione»: le linee di sviluppo del progetto LexIt comprendono il miglioramento del sistema di estrazione dei frame sintattici e della rappresentazione delle preferenze di selezione[80][81], l'induzione di classi verbali su base distribuzionale (Lenci 2012), l'integrazione dei profili semantici con informazioni sulla polisemia degli argomenti e sui ruoli semantici (Lenci et al. 2012)

[78] Ciò permetterebbe un confronto a tutto tondo con il sistema di acquisizione per i nomi, i verbi e gli aggettivi inglesi di Preiss et al. (2007) presentato nei capitoli 1 e 3.
[79] Di cui si trovano costanti aggiornamenti sul sito web: http://sesia.humnet.unipi.it
[80] http://sesia.humnet.unipi.it/lexit/faq.php
[81] Schulte im Walde (2006) suggerisce che un'applicazione di parsing darebbe risultati migliori se potesse riferire a un repertorio di preferenze di selezione, perché potrebbe generalizzare l'analisi sintattica in base alla classe semantica della parola di dubbia classificazione.

Appendici

Appendice A - Frame estratti da LexIt per i verbi, i nomi e gli aggettivi

Frame estratti per i verbi dai corpora *La Repubblica* e *Wikipedia*		
subj#0	subj#obj#comp-con	subj#si#comp-tra
subj#comp_contro	subj#obj#comp-contro	subj#si#comp-verso
subj#comp-a	subj#obj#comp-da	subj#si#cpred
subj#comp-a#comp-da	subj#obj#comp-di	subj#si#fin-che
subj#comp-a#cpred	subj#obj#comp-in	subj#si#fin-come
subj#comp-a#fin-che	subj#obj#comp-per	subj#si#fin-perché
subj#comp-a#inf-di	subj#obj#comp-senza	subj#si#fin-se
subj#comp-attraverso	subj#obj#comp-sotto	subj#si#inf-0
subj#comp-come	subj#obj#comp-su	subj#si#inf-a
subj#comp-con	subj#obj#comp-tra	subj#si#inf-da
subj#comp-contro	subj#obj#comp-verso	subj#si#inf-di
subj#comp-da	subj#obj#cpred	subj#si#obj
subj#comp-di	subj#obj#fin-che	subj#si#obj#comp-a
subj#comp-in	subj#obj#inf-0	subj#si#obj#comp-a#comp-da
subj#comp-per	subj#obj#inf-a	subj#si#obj#comp-attraverso
subj#comp-senza	subj#obj#inf-da	subj#si#obj#comp-come
subj#comp-sotto	subj#obj#inf-di	subj#si#obj#comp-con
subj#comp-su	subj#si#0	subj#si#obj#comp-contro
subj#comp-tra	subj#si#comp-a	subj#si#obj#comp-da
subj#comp-verso	subj#si#comp-a#comp-da	subj#si#obj#comp-di
subj#cpred	subj#si#comp-a#cpred	subj#si#obj#comp-in
subj#fin-che	subj#si#comp-a#fin-che	subj#si#obj#comp-per
subj#fin-come	subj#si#comp-a#inf-di	subj#si#obj#comp-senza
subj#fin-perché	subj#si#comp-attraverso	subj#si#obj#comp-sotto
subj#fin-se	subj#si#comp-come	subj#si#obj#comp-su
subj#inf-0	subj#si#comp-con	subj#si#obj#comp-tra
subj#inf-a	subj#si#comp-contro	subj#si#obj#comp-verso
subj#inf-da	subj#si#comp-da	subj#si#obj#cpred
subj#inf-di	subj#si#comp-di	subj#si#obj#fin-che
subj#obj	subj#si#comp-in	subj#si#obj#inf-0
subj#obj#comp-a	subj#si#comp-per	subj#si#obj#inf-a
subj#obj#comp-a#comp-da	subj#si#comp-senza	subj#si#obj#inf-da
subj#obj#comp-attraverso	subj#si#comp-sotto	subj#si#obj#inf-di
subj#obj#comp-come	subj#si#comp-su	

Frame estratti per i nomi dai corpora *La Repubblica* e *Wikipedia*			
0	comp-con#comp-in	comp-di#comp-tra	comp-per#comp-per
comp-a	comp-con#comp-per	comp-di#comp-verso	comp-per#comp-su
comp-a#comp-come	comp-con#comp-su	comp-di#fin-che	comp-per#comp-tra
comp-a#comp-con	comp-con#inf-per	comp-di#inf-a	comp-per#inf-di
comp-a#comp-contro	comp-contro	comp-di#inf-da	comp-per#inf-per
comp-a#comp-da	comp-contro#comp-di	comp-di#inf-di	comp-presso
comp-a#comp-da_parte_di	comp-contro#comp-in	comp-di#inf-in	comp-secondo
comp-a#comp-di	comp-contro#comp-per	comp-di#inf-per	comp-senza
comp-a#comp-in	comp-da	comp-di#inf-senza	comp-sopra
comp-a#comp-per	comp-da#comp-da	comp-dietro	comp-sotto
comp-a#comp-presso	comp-da#comp-di	comp-dopo	comp-su
comp-a#comp-secondo	comp-da#comp-in	comp-durante	comp-su#comp-a
comp-a#comp-senza	comp-da#comp-per	comp-fino	comp-su#comp-di
comp-a#comp-sotto	comp-da#comp-su	comp-fuori	comp-su#comp-in
comp-a#comp-su	comp-da#comp-tra	comp-in	comp-su#comp-su
comp-a#comp-tra	comp-da#inf-per	comp-in#comp-a	comp-su#comp-tra
comp-a#fin-che	comp-da_parte_di	comp-in#comp-con	comp-su#inf-per
comp-a#inf-a	comp-da_parte_di#comp-di	comp-in#comp-da	comp-tra
comp-a#inf-da	comp-davanti	comp-in#comp-per	comp-tra#comp-in
comp-a#inf-di	comp-dentro	comp-in#comp-senza	comp-tramite
comp-a#inf-per	comp-di	comp-in#comp-su	comp-verso
comp-attraverso	comp-di#comp-a	comp-in#comp-tra	fin-che
comp-attraverso#comp-di	comp-di#comp-in	comp-in#fin-che	inf-a
comp-circa	comp-di#comp-lungo	comp-in#inf-a	inf-come
comp-come	comp-di#comp-oltre	comp-in#inf-da	inf-con
comp-come#comp-di	comp-di#comp-per	comp-in#inf-di	inf-da
comp-come#comp-in	comp-di#comp-presso	comp-in#inf-per	inf-di
comp-come#comp-per	comp-di#comp-secondo	comp-lungo	inf-fino
comp-con	comp-di#comp-senza	comp-mediante	inf-in
comp-con#comp-da	comp-di#comp-sotto	comp-oltre	inf-per
comp-con#comp-di	comp-di#comp-su	comp-per	inf-senza

Frame estratti per gli aggettivi dal corpus *La Repubblica*			
mod-post	mod-pre#comp-su	pred#comp-a	pred#comp-su
mod-pre	mod-pre#comp-tra	pred#comp-attraverso	pred#comp-tra
mod-pre#comp-a	mod-pre#comp-verso	pred#comp-come	pred#comp-verso
mod-pre#comp-come	mod-pre#fin-che	pred#comp-con	pred#fin-che
mod-pre#comp-con	mod-pre#inf-a	pred#comp-contro	pred#fin-se
mod-pre#comp-da	mod-pre#inf-da	pred#comp-da	pred#inf-a
mod-pre#comp-di	mod-pre#inf-di	pred#comp-di	pred#inf-da
mod-pre#comp-in	mod-pre#inf-in	pred#comp-in	pred#inf-di
mod-pre#comp-per	mod-pre#inf-per	pred#comp-per	pred#inf-in
mod-pre#comp-senza	mod-pre#inf-senza	pred#comp-senza	pred#inf-per
mod-pre#comp-sotto	pred	pred#comp-sotto	pred#inf-senza

Appendice B - Tagset dipendenze ISST-TANL

Tag	Relazione	Descrizione	Esempi
arg	argomento	Relazione tra una testa nominale o verbale e un argomento frasale (non in funzione di soggetto).	*Il 63% dei francesi ha imposto al presidente di rinunciare alla sua bomba* *È giunto il momento di creare un'area denuclearizzata* *Le autorità hanno annunciato che il blitz è concluso* *La decisione di continuare...* *escludendo che il militare volesse veramente mettere in pericolo...* *si sono rifiutati di fornire informazione*
aux	ausiliare	Relazione tra un verbo e il suo ausiliare	*Il corazziere è stato individuato* *Ha dichiarato di aver pagato i terroristi*
clit	clitico	Relazione tra un pronome clitico e una testa verbale usata in forma pronominale	*La sedia si è rotta* *Non ci rendiamo conto* *Si tratta della scoperta*
comp	complemento	Relazione tra una testa e un complemento preposizionale	*Fu assassinata da un pazzo* *E' più interessante del libro* *Oggi come allora* *Più di quattrocento esemplari* *Osteggiata dal governo di Berna* *Grande quanto mezza Italia*
comp-ind	complemento indiretto	Denota il partecipante interessato in un evento	*Ho dato il libro a lui* *I carabinieri gli hanno recapitato il decreto*
comp-loc	complemento locativo	Esprime la direzione o il luogo del movimento di un'azione	*Si trovava in un parco* *Era uscito di casa alle 10*
comp-temp	complemento temporale	Relazione tra un complemento temporale e una testa verbale	*Nel 1985 è stata uccisa un'antropologa* *L'allarme è scattato la scorsa*

			settimana
con	congiunzione copulativa	Relazione tra una congiunzione copulativa e il primo congiunto	*Una ragazza violentata e sequestrata da due slavi* *Gabriella e Paolo sono partiti* *Hanno riarmato, addestrato e preparato l'esercito* *Hanno riarmato, addestrato e preparato l'esercito* *Scontri, assalti e centinaia di feriti* *Scontri, assalti e centinaia di feriti*
concat	concatenazione	Relazione tra i token che formano una multiword expression, nomi propri complessi ecc.	*Il segretario di De Michelis* *L'enciclica "Mulieris dignitatem"* *La International Public Sport* *La International Public Sport*
conj	congiunto collegato da una congiunzione copulativa	Relazione tra congiunti, dove il primo è la testa dell'intera struttura	*Una ragazza violentata e sequestrata da due slavi* *Gabriella e Paolo sono partiti* *Hanno riarmato, addestrato e preparato*
det	determinante	Relazione tra una testa nominale e il suo determinante	*Una sala* ha dovuto essere sgomberata Rilevata *la presenza* di gas
dis	congiunzione disgiuntiva	Relazione tra una congiunzione disgiuntiva e il primo congiunto	*Cassonetti dell'immondizia* rovesciati o *incendiati* *Partecipa* a *manifestazioni politiche* o *a dibattiti*
disj	congiunto in un composto disgiuntivo collegato da una congiunzione disgiuntiva	Relazione tra un congiunto (o più) e il primo congiunto, testa dell'intera struttura.	*Cassonetti dell'immondizia* rovesciati *o* incendiati *Partecipa* a *manifestazioni politiche o* a *dibattiti*
mod	modificatore	Relazione tra una testa e il suo modificatore (frasale, aggettivale, avverbiale, apposizione)	*I colori* sono sempre *gli stessi* Colori intensi *Trionfo di Didoni nei* 20 km *di marcia* Cesare *l'*Imperatore Per *arrivare in tempo, sono* partito *molto presto* Quando *la campanella suona, i bambini* escono *da scuola*

mod_loc	modificatore locativo	Relazione tra una testa e il suo modificatore che esprime un luogo.	*Non* so dove *Tutto* cominciò *proprio* lì *Avrei voluto* fermar*mi* qui *più a lungo*
mod_rel	modificatore relativo	Relazione tra la testa verbale di una frase relativa e la sua testa nominale nella frase principale	*Box* che è stato *trovato* nel pomeriggio Quell'*ordine* che i due Stranamore pentiti avevano *imposto* per cinquant'anni Non è mai stato accertato *chi volle* la sua morte
mod-temp	modificatore temporale	Una relazione temporale tra una testa e il suo modificatore	*Ieri* hanno *dormito* all'aperto Scoperto 75 *anni fa* Non *superano mai* gli 8 milioni
modal	verbo modale	Relazione tra un verbo modale e la sua testa verbale	*Una sala ha* dovuto *essere* sgomberata *Avrebbe* potuto ripeter*si*
neg	negazione	Modificatore negativo	*A volte* non dormo
obj	complemento oggetto	Relazione tra un verbo e il suo oggetto diretto	*Hanno* un *modo* di ragionare rozzo *Centellinando* le *informazioni*
pred	complemento predicativo	Relazione tra una testa e il suo complemento predicativo	*L'incontro è* stato fatale *Questo* è *il* messaggio *finale*
pred_loc	predicativo locativo	Esprime una proprietà spaziale del soggetto, dopo il verbo di collegamento	*Il presidente non* era in *casa*
pred-temp	predicativo temporale	Esprime una proprietà temporale del soggetto, dopo il verbo di collegamento	*La riunione* è alle *5*
prep	preposizione	Relazione tra una testa preposizionale e il suo complemento	*Un contributo* alla lotta *contro la criminalità* *Prima* di partire *ho telefonato*
punc	punteggiatura	Relazione tra un token e un simbolo di punteggiatura	*Teatro* della tragedia *,* ...
ROOT	radice della frase	Testa della frase	*Desidero dormire* *(solo il dipendente è sottolineato, perché la testa è un nodo fittizio)*
sub	frase subordinata	Relazione tra una subordinata e la testa verbale	*Ha detto* che *non* intendeva *fare nulla* *Le autorità hanno annunciato* che *il blitz è* concluso *Venne ucciso* mentre cercava *di difendere la ragazza*

subj	soggetto	Relazione tra un verbo attivo e il soggetto	*il* testimone *ha* parlato *subito* *le* vittime seguivano *gli aiuti*
subj_pass	soggetto passivo	Relazione tra un verbo passivo e il suo soggetto	*I* missionari *erano stati* rapiti *la mattina presto* *Circa 83.000* franchi *furono* spesi
voc	vocativo	Espressione vocativa	*Signor* presidente*,* chiedo *la parola*

Appendice C – Verbi estratti per la valutazione di LexIt

Verbo	Frequenza
abrogare	1409
accomodare	1234
accontentare	10844
accordare	4406
affollare	4210
affrettare	4829
allegare	1621
animare	5593
apprendere	12651
arrabbiare	4982
attenere	3880
avvertire	34751
avviare	32495
azzerare	2852
caratterizzare	10090
catalogare	1041
chiamare	104395
circolare	14780
collaudare	1389
contemplare	3033
coronare	1465
corteggiare	1794
costare	30754
cucinare	1618
decretare	4099
degradare	1210
deprimere	2244
destabilizzare	1901
difendere	51312
dilatare	1835
dire	830903
dislocare	805
disporre	37068
dissipare	1398
distare	836
educare	2882
esportare	4559
evitare	68251
gratificare	1255
gravare	4565
incaricare	5892
incrinare	2349
indurre	14127
ingaggiare	3547
intercettare	3073
investire	26607
ipotizzare	10776
limitare	38185
litigare	5809
logorare	1390

Verbo	Frequenza
mascherare	2150
meritare	18015
miscelare	429
noleggiare	661
omettere	1538
ovviare	1007
parlare	285423
peggiorare	5034
possedere	16550
predicare	3023
prefiggere	1181
raddoppiare	7337
ratificare	3380
razionalizzare	1607
recitare	14810
regolamentare	1952
riaffermare	3048
riaprire	12424
rimettere	18419
rimpiangere	3406
rimproverare	6593
rimuovere	6587
ringraziare	9024
ritagliare	1967
scadere	10325
schierare	16483
sciogliere	14066
scivolare	7781
selezionare	3063
sfidare	6754
sfiorare	11759
sfondare	4742
Sfornare	2170
sgombrare	1581
siglare	3852
simpatizzare	474
spaventare	7471
sponsorizzare	1834
stabilire	39015
staccare	7065
stancare	2871
strumentalizzare	2036
stupire	10993
svegliare	6057
sviluppare	16344
terminare	8923
trainare	891
travolgere	8483
vergognarsi	3302
viaggiare	15955

Appendice D – Risultati della valutazione rispetto a soglie di MLE

Punteggi di Precision (P), Recall (R) e F-Measure (F-M) ottenuti a confronto dei diversi gold standard per soglie crescenti di MLE (da 0 a 0,1 con incremento di 0,001).

	Blumenthal Rovere			Sabatini Coletti			PAROLE		
soglia	**P**	**R**	**F-M**	**P**	**R**	**F-M**	**P**	**R**	**F-M**
0	0,44	0,97	0,59	0,38	0,99	0,54	0,33	1	0,48
0,001	0,55	0,96	0,69	0,48	0,99	0,63	0,43	1	0,59
0,002	0,6	0,96	0,73	0,52	0,99	0,67	0,48	1	0,63
0,003	0,64	0,95	0,75	0,55	0,99	0,69	0,51	1	0,66
0,004	0,66	0,95	0,76	0,57	0,99	0,71	0,53	1	0,68
0,005	0,67	0,95	0,77	0,58	0,99	0,71	0,55	0,99	0,69
0,006	0,68	0,94	0,78	0,59	0,99	0,72	0,56	0,99	0,7
0,007	0,7	0,94	0,78	0,6	0,99	0,73	0,57	0,99	0,7
0,008	0,7	0,93	0,79	0,61	0,99	0,73	0,58	0,99	0,71
0,009	0,72	0,93	0,8	0,62	0,99	0,75	0,6	0,99	0,73
0,01	0,73	0,93	0,8	0,63	0,99	0,75	0,61	0,99	0,73
0,011	0,74	0,93	0,81	0,64	0,99	0,76	0,62	0,99	0,74
0,012	0,75	0,93	0,81	0,65	0,99	0,76	0,62	0,99	0,75
0,013	0,76	0,92	0,82	0,66	0,99	0,77	0,63	0,99	0,75
0,014	0,76	0,92	0,82	0,66	0,98	0,78	0,63	0,99	0,76
0,015	0,77	0,92	0,82	0,67	0,98	0,78	0,64	0,99	0,76
0,016	0,77	0,92	0,82	0,67	0,98	0,78	0,64	0,99	0,76
0,017	0,78	0,91	0,82	0,67	0,98	0,78	0,64	0,99	0,76
0,018	0,78	0,91	0,83	0,68	0,98	0,79	0,65	0,99	0,77
0,019	0,78	0,91	0,83	0,68	0,98	0,79	0,65	0,99	0,77
0,02	0,79	0,91	0,83	0,68	0,98	0,79	0,65	0,99	0,77
0,021	0,78	0,91	0,82	0,67	0,98	0,78	0,65	0,99	0,76
0,022	0,78	0,9	0,82	0,67	0,98	0,78	0,64	0,99	0,76
0,023	0,78	0,9	0,82	0,67	0,98	0,78	0,64	0,99	0,76
0,024	0,78	0,9	0,82	0,67	0,98	0,78	0,64	0,99	0,76
0,025	0,78	0,9	0,81	0,67	0,98	0,78	0,64	0,99	0,76
0,026	0,77	0,89	0,81	0,66	0,98	0,77	0,63	0,99	0,75
0,027	0,77	0,89	0,81	0,67	0,98	0,77	0,64	0,99	0,76
0,028	0,78	0,89	0,81	0,67	0,98	0,77	0,64	0,99	0,76
0,029	0,78	0,89	0,81	0,67	0,98	0,78	0,64	0,99	0,76
0,03	0,78	0,88	0,8	0,67	0,97	0,78	0,64	0,99	0,76
0,031	0,78	0,88	0,81	0,67	0,97	0,78	0,64	0,99	0,76
0,032	0,78	0,88	0,81	0,67	0,97	0,78	0,64	0,99	0,76
0,033	0,78	0,88	0,8	0,67	0,97	0,78	0,64	0,99	0,76
0,034	0,78	0,88	0,8	0,67	0,97	0,77	0,64	0,99	0,76
0,035	0,78	0,88	0,8	0,66	0,97	0,77	0,64	0,99	0,75
0,036	0,78	0,88	0,8	0,66	0,97	0,77	0,63	0,99	0,75
0,037	0,78	0,87	0,8	0,66	0,97	0,77	0,63	0,99	0,75
0,038	0,78	0,87	0,8	0,66	0,97	0,77	0,63	0,99	0,75
0,039	0,78	0,87	0,8	0,66	0,97	0,77	0,63	0,99	0,75
0,04	0,78	0,87	0,79	0,66	0,97	0,76	0,62	0,99	0,74
0,041	0,77	0,87	0,79	0,65	0,97	0,76	0,63	0,99	0,75
0,042	0,77	0,87	0,79	0,65	0,97	0,76	0,62	0,99	0,74
0,043	0,77	0,87	0,79	0,65	0,97	0,76	0,62	0,99	0,74
0,044	0,77	0,87	0,79	0,65	0,97	0,76	0,63	0,99	0,75
0,045	0,78	0,86	0,79	0,65	0,97	0,76	0,63	0,99	0,75
0,046	0,78	0,86	0,79	0,65	0,97	0,76	0,63	0,99	0,75
0,047	0,78	0,86	0,79	0,66	0,97	0,76	0,63	0,99	0,75

0,048	0,78	0,86	0,79	0,66	0,97	0,76	0,63	0,99	0,75
0,049	0,78	0,86	0,79	0,65	0,97	0,76	0,63	0,99	0,75
0,05	0,77	0,86	0,79	0,65	0,96	0,75	0,63	0,98	0,74
0,051	0,77	0,85	0,78	0,65	0,96	0,75	0,63	0,98	0,74
0,052	0,77	0,85	0,78	0,65	0,96	0,75	0,63	0,98	0,74
0,053	0,77	0,85	0,78	0,65	0,96	0,75	0,63	0,97	0,74
0,054	0,78	0,85	0,78	0,65	0,96	0,75	0,63	0,97	0,74
0,055	0,77	0,85	0,78	0,65	0,96	0,75	0,63	0,97	0,74
0,056	0,78	0,85	0,78	0,65	0,96	0,75	0,63	0,97	0,75
0,057	0,78	0,85	0,78	0,65	0,96	0,75	0,64	0,97	0,75
0,058	0,78	0,85	0,78	0,65	0,96	0,76	0,64	0,97	0,75
0,059	0,78	0,85	0,78	0,65	0,96	0,76	0,64	0,97	0,75
0,06	0,78	0,85	0,78	0,65	0,96	0,76	0,64	0,97	0,75
0,061	0,78	0,85	0,78	0,65	0,96	0,75	0,64	0,97	0,75
0,062	0,79	0,85	0,78	0,66	0,96	0,76	0,64	0,97	0,75
0,063	0,79	0,85	0,78	0,66	0,96	0,76	0,65	0,97	0,75
0,064	0,79	0,84	0,79	0,66	0,96	0,76	0,65	0,97	0,76
0,065	0,8	0,84	0,79	0,66	0,96	0,76	0,65	0,97	0,76
0,066	0,8	0,84	0,79	0,66	0,96	0,76	0,65	0,97	0,76
0,067	0,8	0,84	0,79	0,66	0,96	0,76	0,65	0,97	0,76
0,068	0,8	0,84	0,79	0,67	0,96	0,76	0,66	0,97	0,76
0,069	0,8	0,84	0,79	0,67	0,96	0,76	0,66	0,97	0,76
0,07	0,8	0,84	0,79	0,67	0,96	0,77	0,66	0,97	0,76
0,071	0,8	0,84	0,79	0,67	0,96	0,76	0,65	0,97	0,76
0,072	0,8	0,84	0,79	0,67	0,96	0,76	0,65	0,97	0,76
0,073	0,8	0,84	0,79	0,67	0,96	0,77	0,66	0,97	0,76
0,074	0,8	0,84	0,79	0,67	0,96	0,77	0,66	0,97	0,77
0,075	0,8	0,84	0,79	0,67	0,96	0,77	0,66	0,97	0,76
0,076	0,8	0,84	0,79	0,67	0,95	0,77	0,66	0,97	0,76
0,077	0,8	0,84	0,79	0,67	0,95	0,77	0,66	0,97	0,76
0,078	0,8	0,84	0,79	0,68	0,95	0,77	0,66	0,97	0,76
0,079	0,8	0,84	0,79	0,67	0,95	0,77	0,66	0,97	0,76
0,08	0,8	0,84	0,79	0,67	0,95	0,77	0,66	0,97	0,76
0,081	0,8	0,83	0,78	0,67	0,95	0,77	0,66	0,97	0,76
0,082	0,8	0,83	0,79	0,68	0,95	0,77	0,66	0,97	0,76
0,083	0,81	0,83	0,79	0,68	0,95	0,77	0,67	0,97	0,76
0,084	0,81	0,83	0,79	0,68	0,95	0,77	0,66	0,97	0,76
0,085	0,82	0,83	0,79	0,68	0,95	0,77	0,67	0,97	0,77
0,086	0,82	0,83	0,79	0,68	0,95	0,77	0,67	0,97	0,77
0,087	0,81	0,83	0,79	0,68	0,95	0,77	0,67	0,97	0,77
0,088	0,81	0,83	0,79	0,68	0,95	0,77	0,67	0,97	0,77
0,089	0,81	0,83	0,79	0,68	0,95	0,77	0,67	0,97	0,77
0,09	0,81	0,83	0,79	0,68	0,95	0,77	0,67	0,97	0,77
0,091	0,81	0,83	0,79	0,68	0,95	0,77	0,67	0,97	0,77
0,092	0,82	0,83	0,79	0,69	0,95	0,77	0,67	0,97	0,77
0,093	0,82	0,83	0,79	0,69	0,95	0,77	0,67	0,97	0,77
0,094	0,82	0,83	0,79	0,69	0,95	0,77	0,67	0,97	0,77
0,095	0,82	0,83	0,79	0,69	0,95	0,77	0,68	0,97	0,77
0,096	0,82	0,83	0,79	0,69	0,95	0,77	0,68	0,97	0,77
0,097	0,82	0,83	0,79	0,69	0,95	0,77	0,68	0,97	0,77
0,098	0,82	0,83	0,79	0,69	0,95	0,77	0,68	0,97	0,77
0,099	0,82	0,83	0,79	0,7	0,95	0,78	0,69	0,97	0,78
0,1	0,82	0,83	0,79	0,69	0,95	0,78	0,69	0,97	0,78

Ringraziamenti

Ringrazio la Dott.ssa Nilda Ruimy dell'Istituto di Linguistica Computazionale del CNR di Pisa per avermi dedicato del tempo per introdurmi alla consultazione del lessico computazionale PAROLE-SIMPLE-CLIPS, per avermene fornito l'accesso e per essersi sempre resa disponibile nell'assistermi.

Questa tesi è dedicata ai miei nonni.

Bibliografia

Attardi, Giuseppe e Felice Dell'Orletta. 2009. *Reverse revision and linear tree combination for dependency parsing*. In: Proceedings of NAACL-HLT 2009, pp. 261–264, Boulder, USA.

Baroni, Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston e Marco Mazzoleni. 2004. *Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian*. In: Proceedings of LREC 2004. Lisboa, pp. 1771- 1774.

Bindi, Remo, Monica Monachini, Paola Orsolini. 1991. *Italian Reference Corpus. General Information and Key for Consultation*. ILC-TLN-1991-1. ILC-CNR, Pisa.

Blumenthal, Peter e Giovanni Rovere. 1998. Wörterbuch der italienischen Verben. Ernest Klettverlag, Stuttgart.

Boguraev, Branimir, Ted Briscoe, John Carroll, David Carter e Claire Grover. 1987. *The Derivation of a Grammatically-Indexed Lexicon from the Longman Dictionary of Contemporary English*. In: Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics. Stanford, pp. 193-200.

Boguraev, Branimir, e Ted Briscoe (a cura di). 1989. *Computational Lexicography for Natural Language Processing*. Longman, London.

Bosco Cristina, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Felice Dell'Orletta e Alessandro Lenci. 2009. *Evalita'09 parsing task: comparing dependency parsers and treebanks*. In: Proceedings of EVALITA 2009, Reggio Emilia, Italia.

Brent, Michael R. 1991. *Automatic Semantic Classification of Verbs from their Syntactic Contexts: an Implemented Classifier for Stativity*. In: Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics. Berlino, pp. 222-226.

Brent, Michael R. 1993. *From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax*. In «Computational Linguistics», 19(2), pp. 243-262.

Briscoe, Ted e John Carroll. 1997. *Automatic Extraction of Subcategorization from Corpora.* In: Proceedings of the 5th ACL Conference on Applied Natural Language Processing. Washington, DC, pp. 356-363.

Briscoe, Ted e John Carroll. 1993. *Generalized probabilistic LR parsing for unification-based grammars*. Computational Linguistics 19.1, pp. 25-60.

Briscoe, Ted. 2000. *Dictionary and System Subcategorisation Code Mappings*. Unpublished Manuscript. University of Cambridge Computer Laboratory.

Briscoe, Ted. 2001. *From dictionary to corpus to self-organizing dictionary: learning valency associations in the face of variation and change*. In: Proceedings of the Corpus Linguistics, Lancaster University, UK.

Briscoe, Ted, Joe Carroll e Rebecca Watson. 2006. *The second release of the RASP system*. In: Proc. of the COLING/ACL 2006 Interactive Presentation Sessions. Sydney, Australia.

Burnard, Lou. 1995. *The BNC Users Reference Guide*. British National Corpus Consortium, Oxford, May.

Carroll, John, Guido Minnen e Ted Briscoe. 1998. *Can Subcategorisation Probabilities help a Statistical Parser?* In: Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora. Montreal, Canada, pp.118-126.

Chen, Stanley. F. e Joshua Goodman. 1996. *An empirical study of smoothing techniques for language modeling*. In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics. Santa Cruz, California, pp. 310-318.

Chitrao, M. and Grishman, R. 1990. Statistical parsing of messages. In Proceedings of the Darpa Speech and Natural Language Workshop, Hidden Valley, PA. 263-266.

Dorr, Bonnie J. 1997. *Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation*. In: «Machine Translation», 12(4), pp.271-322.

Erk, Katrik e Sebastian Padó e Ulkrike Padó. 2010. *A flexible, corpus-driven model of regular and inverse selectional preferences.* In «Computational Linguistics», 36(4), pp. 723–763.

Evert, Stefan. 2008. *Corpora and collocations*. In: Lüdeling Anke e M erja Kytö (a cura di). *Corpus Linguistics: An International Handbook*, capitolo 58. Mouton de Gruyter, Berlin.

De Mauro, Tullio (a cura di). 1989. *Vocabolario Elettronico della Lingua Italiana*. IBM Italia

Fellbaum, Christiane. 2005. *WordNet and wordnets*. In: Brown, Keith et al. (a cura di), *Encyclopedia of Language and Linguistics*, 2ª edizione. Oxford, Elsevier, pp.665-670.

Firth, John Rupert. 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Garside, Roger, Geoffrey Leech e Geoffrey Sampson. 1987. *The Computational Analysis of English: A Corpus-Based Approach*. Longman, London.

Gorrell, Genevieve. 1999. *Acquiring Subcategorisation from Textual Corpora*. Tesi del Master in Filosofia. University of Cambridge, UK.

Grishman, Ralph, Catherine Macleod e Adam Meyers. 1994. *COMLEX Syntax: Building a Computational Lexicon*. In: Proceedings of the 15th International Conference on Computational Linguistics. Kyoto, Giappone, pp. 268-272.

Hajič, Jan, Martin Čmejrek, Bonnie Dorr, Yuan Ding, Jason Eisner, Daniel Gildea, Terry Koo, Kristen Parton, Gerald Penn, Dragomir Radev e Owen Rambow. 2002. *Natural language generation in the context of machine translation*. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore. Summer Workshop Final Report.

Helbig Gregory M., 1992. *Probleme der Valenz- und Kasustheorie*, Tübingen.

Korhonen, Anna. 2002. *Subcategorization Acquisition.* Tesi di dottorato. Computer Laboratory, University of Cambridge.

Korhonen Anna, Yual Krymolowski e Ted Briscoe. 2006. *A large subcategorization lexicon for natural language processing applications*. In: Proceedings of LREC 2006, Genova, Italy.

Laplace, Pierre Simon. 1995. *Philosophical Essay On Probabilities*. Springer-Verlag.

Lenci Alessandro. 2012. *Carving Verb Classes from Corpora*. In: Raffaele Simone e Francesca Masini (a cura di) *Word Classes*. Amsterdam - Philadelphia: John Benjamins.

Lenci Alessandro, Gabriella Lapesa e Giulia Bonansinga (2012). *LexIt: A Computational Resource on Italian Argument Structure*. In: Proceedings of LREC 2012, Istanbul.

Levin, Beth. 1993. *English Verb Classes and Alternations.* The University of Chicago Press.

Macleod Catherine, Adam Meyers, Ralph Grishman, Leslie Barrett e Ruth Reeves. 1997. *Designing a dictionary of derived nominals.* In: Proc. of RANLP, Tzigov Chark, Bulgaria.

Manning, Christopher D. 1993. *Automatic Acquisition of a Large Subcategorization Dictionary from Corpora*. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. Columbus, OH, pp. 235-242.

Messiant, Cédric. 2008. *ASSCI : A subcategorization frames acquisition system for french.* In: Proceedings of the Association for Computational Linguistics (ACL) Student Research Workshop, Colombus, Ohio. Association for Computational Linguistics.

Messiant Cédric, Anna Korhonen e Thierry Poibeau. 2008. *LexSchem: A large subcategorization lexicon for French verbs*. In: Proceedings of LREC 2008, Marrakech, Morocco.

Nied Curcio, Marina. 2006. *La lessicografia tedesco-italiana: storia e* tendenze, in F.San Vincente (a cura di). *Lessicografia bilingue e traduzione*. *Metodi, strumenti, approcci attuali*. Polimetrica Publisher, Monza, pp. 64-65

Pereira, Fernando e David H.D. Warren. 1980. *Definite clause grammars for language analysis - a survey of the formalism and a comparison with augmented transition networks*. In: «Artificial Intelligence» 13.3, pp. 231-278.

Pianta, Emanuele, Luisa Bentivogli e Christian Girard. 2002. *MultiWordNet: developing an aligned multilingual database*. In: Proceedings of the first International Conference on Global WordNet. Mysore, India.

Poibeau Thierry e Cédric Messiant. 2008. *Do we still need gold standard for evaluation?* In: Proceedings of the Language Resources and Evaluation Conference (LREC), Marrakech.

Preiss Judita, Ted Briscoe e Anna Korhonen. 2007. *A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora.* In: Proceedings of ACL 2007. Prague, Czech Republic, pp. 912–919.

Resnik, Philip. 1997). *Selectional Preference and Sense Disambiguation*. In: Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How. Washington DC, pp. 52-57.

Ruimy, Nilda, Ornella Corazzari, Elisabetta Gola, Antonietta Spanu, Nicoletta Calzolari e Antonio Zampolli. 1998. *The European LE-PAROLE Project: the Italian Syntactic Lexicon*. In: Proceedings of LREC1998, pp. 241–248.

Sabatini, Francesco e Vittorio Coletti. 2005. *Il Sabatini Coletti: dizionario essenziale della lingua italiana*. Rizzoli-Larousse. Milano.

Sampson, Geoffrey. 1995. *English for the Computer.* Oxford University Press, Oxford, UK.

Sanfilippo, Antonio. et al. 1996. *Subcategorization Standards, Report of the Eagles/Lexicon/Syntax Group.*

Schulte im Walde, Sabine e Chris Brew. 2002. *Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information.* In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp.223-230.

Schulte im Walde, Sabine. 2009. *The Induction of Verb Frames and Verb Classes from Corpora* In: Anke Lüdeling e Merja Kytö (a cura di) *Corpus Linguistics. An International Handbook.* Mouton de Gruyter, Berlin.

Surdeanu Mihai, Sanda Harabagiu, John Williams e Paul Aarseth. 2003. *Using predicate-argument structures for information extraction*. In: Proc. of the 41st Annual Meeting of ACL, Sapporo.

Taylor, Lolita e Gerry Knowles. 1988. *Manual of Information to Accompany the SEC*

Corpus: the Machine-Readable Corpus of Spoken English. University of Lancaster, UK, Ms.

Ushioda, Akira, David A. Evans, Ted Gibson e Alex Waibel. 1993. *The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora*. In: Proceedings of the Workshop on the Acquisition of Lexical Knowledge from Text. Columbus OH, pp.95-106.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.