UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

Rappresentare le preferenze di selezione dei verbi italiani: un esperimento computazionale

Candidato:
Raffaele Guarasci

Relatore:
Prof. Alessandro Lenci

Correlatore:
Prof. Maria Simi

Anno Accademico 2011-2012

Indice

Introduzione **4**

1. Le Preferenze di Selezione Verbali **6**

1.1.Modello di Resnik 7

1.1.1.Teorizzazione formale del modello 8

1.1.2.Implementazione computazionale 10

1.2.Modello di Schulte Im Walde 15

2. LexIt **16**

2.1.Scelta e preparazione del corpus 17

2.2.Metodologia 18

2.3.Profili distribuzionali dei verbi italiani 21

2.4.Profili semantici 22

3. WordNet e MultiWordNet **25**

3.1.MultiWordNet 27

3.1.1.Implementazione del modello 29

3.1.2.Modello dei dati 30

4. Descrizione dell'esperimento **32**

4.1.Analisi dei risultati 36

4.2.Conclusioni e possibili sviluppi 38

Bibliografia **40**

Introduzione

L'esperimento proposto in questo lavoro si colloca nella serie di metodologie basate su corpora proposte per analizzare le preferenze di selezione verbali. Per preferenze di selezione si intendono i vincoli imposti dai predicati nella realizzazione dei propri argomenti. L'acquisizione di preferenze di selezione da un corpus, proposta inizialmente da Philip Resnik nel 1993[1], si può articolare in due fasi: l'estrazione degli argomenti dai corpora scelti e la generalizzazione delle preferenze di selezione verbali da un'ontologia lessicale.

In questo lavoro viene utilizzato *LexIt*, un lessico di valenza per i verbi della lingua italiana come risorsa lessicale e come ontologia MultiWordNet[2], un database lessicale multilingua strutturato in delle classi semantiche organizzate in modo gerarchico. L'analisi proposta si ricollega ad uno degli ultimi modelli sviluppati per generalizzare le preferenze di

[1] P. Resnik, *Semantic classes and syntactic ambiguity*, in Proceedings of the workshop on Human Language Technology - HLT '93, 278–283, Morristown, 1993.
[2] B. Magnini e C. Strapparava, *Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet*, Atti del XXVIII Congresso della Società di Linguistica Italiana, 415-418, Palermo, 1994

selezione, l'esperimento effettuato da Sabine Schulte Im Walde[3] sui verbi tedeschi, mirato ad ottenere una rappresentazione ad alto livello che generalizzi la preferenze di selezione degli argomenti e fornisca una distribuzione del comportamento dei verbi nella lingua tedesca. Il lavoro effettuato sulla lingua tedesca aveva come obiettivo di fornire una generalizzazione del comportamento dei verbi, considerando soltanto le classi più generiche e astratte in cima alla rete semantica, questo alto livello di generalizzazione limita l'osservazione di alcuni comportamenti specifici di determinati verbi. L'analisi qui proposta mira a superare questo problema e a fornire un livello di rappresentazione più dettagliato, che caratterizzi meglio delle specificità delle preferenze di selezione, considerando nella navigazione della rete semantica di MultiWordNet tutte le classi intermedie, dal momento che le classi generali risultano troppo ampie per caratterizzare in modo differente tutti i verbi. Esprimere le preferenze di selezione nei termini di tutte le classi intermedie della gerarchia permette di far venire alla luce dei comportamenti più specifici nella scelta degli argomenti, che non potevano essere considerati guardando solo le classi generiche.

La tesi è così articolata: il primo capitolo tratta della definizione delle preferenze di selezione, passando in rassegna alcuni dei modelli basati su corpora, dal modello di Resnik a quello di Schulte Im Walde. Nei capitoli successivi vengono elencate le risorse utilizzate per l'analisi, essendo il modello composto di un lessico e di una tassonomia, come detto sopra. In particolare il secondo capitolo presenta le caratteristiche di LexIt, alcune delle metriche di valutazione usate nelle varie analisi e si introduce il conetto di profilo semantico, nel capitolo successivo vengono trattati brevemente WordNet e la sua versione multilingua MultiWordNet. Nel quarto capitolo viene descritto il funzionamento del programma realizzato per la navigazione della rete semantica e per l'estrazione delle preferenze di selezione, si passa poi alle metodologie usate per l'analisi dei dati, dei quali vengono citati alcuni casi rappresentativi emersi dall'analisi.

[3] S. Schulte Im Walde, *Experiments on the Automatic Induction of German Semantic Verb Classes*, Computational Linguistics, 32(2):159–194, 2006.

1. Le Preferenze di Selezione

Si può definire *preferenza di selezione* la proprietà di un verbo di preferire o meno argomenti di un particolare tipo semantico[4]. Dunque si può vedere la preferenza di selezione come una sorta di "vincolo" che opera una restrizione specificando quale siano gli argomenti adatti per un dato predicato.

Il concetto di *preferenze di* selezione o *restrizioni semantiche* ha una lunga storia ed è stato ampiamente trattato sia nella Linguistica Generativa (Katz e Fodor[5], Chomsky[6]) che nella Linguistica Computazionale.

Uno dei primi approcci per caratterizzare le preferenze di selezione è quello di Katz e Fodor, basato sulla più ampia teoria semantica della decomposizione del significato delle parole in *features* lessicali caratterizzanti. Il classico esempio proposto per la lingua inglese è la parola *bachelor* che può indicare un uomo non sposato, scapolo (features: *maschio*

[4] La preferenza di selezione può funzionare anche in senso inverso; un dato argomento seleziona dei predicati, ad esempio: *libro* preferisce il verbo *leggere* al verbo *guidare*. A tale proposito cfr M. Light and W. Greiff, *Statistical models for the induction and use of selectional preferences,*Cognitive Science, 87:1–13, 2002.

[5] J.J. Katz and J.A. Fodor, *The structure of a semantic theory*, Language, 39(2):170–210, 1963.

[6] N. Chomsky, *Aspects of the Theory of Syntax*, The MIT press, Cambridge, MA, 1965.

e *umano*) o un esemplare di foca maschio privo di compagna (features: *maschio* e *animale*).

Applicare questo modello ai predicati significa identificare per i predicati delle condizioni necessarie e sufficienti perché siano semanticamente accettabili per essere associate a quell'argomento. Tali condizioni sono rappresentate come funzioni booleane.

Il limite di questo modello consiste nel fatto che spesso identificare univocamente delle condizioni necessarie e sufficienti valide è un problema insormontabile, inoltre è improprio considerare le preferenze di selezione come delle potenziali risposte a domande si/no. Per questo si parla di *preferenze* e non di *regole*, infatti le preferenze di selezione sono requisiti abbastanza elastici che possono indicare il tipo preferito per un determinato argomento verbale senza tuttavia escludere completamente altre possibilità, come ad esempio usi metaforici; per questi motivi un modello rappresentante preferenze di selezione non può essere formulato nei termini di features binarie (come +/- animato). Inoltre sono stati posti in luce molti altri aspetti problematici di questa metodologia, innanzitutto è difficilmente dimostrabile che i componenti siano unità concettuali non ulteriormente scomponibili, parole che si riferiscono a proprietà percettive come ad esempio "giallo" o "morbido" non si prestano a una scomposizione[7].

1.1. Modello di Resnik

Il modello elaborato da Philip Resnik nel 1993[8] è considerato il più quotato modello computazionale per le preferenze di selezione e il punto di riferimento per i lavori successivi. La strategia adottata da Resnik si compone di una rappresentazione tassonomica dei concetti e di una

[7] Cfr. L. Bentivogli, *Relazioni lessicali e semantiche nella costruzione di un lessico computazionale multilingue: problematiche tecniche e filosofiche*, Bologna 1998.
[8] P. Resnik, *Semantic classes and syntactic ambiguity*, in Proceedings of the workshop on Human Language Technology - HLT '93, 278–283, Morristown, 1993.

formalizzazione probabilistica delle preferenze di selezione definite nei termini di questa tassonomia, che vengono poi computate e analizzate sulla base di frequenze di co-occorrenza tra i predicati e i loro argomenti.

Resnik formalizza il modello secondo i principi dalla Teoria dell'informazione, ottenendo un'interpretazione delle preferenze di selezione così definita: "*how strongly a predicate selects for an argument is identified with the quantity of information it carries about that argument, where information is interpreted in a strict mathematical sense*[9]*."*

1.1.1. Teorizzazione formale del modello

Dunque la prima componente del modello deve essere una tassonomia concettuale, una rete semantica, nella quale le classi di concetti devono essere così correlate; ad esempio ad esempio *bevanda* è una sottoclasse di *liquidi* e una superclasse di *vino*, *gatto* è una sottoclasse di *felino*, che a sua volta è una sottoclasse di *mammifero*, ma *bevanda* non è solo inteso come significato di un singolo termine, ma come un'etichetta che identifica un set contenente: *acqua*, *vino*, *caffè*, ecc. A sua volta, la classe *liquidi* avrà un proprio superset, contenente non solo il concetto *bevanda* e l'intero set identificato dall'etichetta *bevanda*, ma anche concetti non-bevande, quali *antigelo*, *petrolio*.

La seconda componente del modello deve invece provvedere a fornire una caratterizzazione delle preferenze di selezione nei termini di una relazione probabilistica tra predicati e classi concettuali, basandosi sull'assunzione che un predicato tende ad associarsi prevalentemente con determinate classi di argomenti.

In questo modello le preferenze di selezione sono costituite dalla relazione tra la *prior distribution*, *Pr(c)* e la *posterior distribution*, *Pr(c|p)*, ovvero tra la probabilità del verificarsi di un argomento a prescindere dal predicato e la probabilità condizionata tra i due. La differenza tra le due

[9] P. Resnik, *Selectional constraints: an information-theoretic model and its computational realization*, Cognition 61, 127-159, 1996.

distribuzioni può essere espressa in termini molto precisi usando una misura presa dalla teoria dell'informazione: l'*entropia relativa*[10], definita come segue:

$$D(p||q) = \sum_x p(x) log \frac{p(x)}{p(q)}$$

La distribuzione probabilistica *p* è interpretata come la "vera distribuzione" e la distribuzione *q* come un'approssimazione della distribuzione vera, l'entropia relativa D(p||q) misura l'ammontare di informazione ulteriore necessaria per aggiungere all'approssimazione per raggiungere la vera distribuzione.

Di conseguenza, si può identificare con q la *prior distribution* delle classi, *Pr(c)* in quanto non rappresenta altro che un'approssimazione di come apparirebbe la distribuzione se non si considerassero i predicati. Analogamente la *posterior distribution*, *Pr(c|p)* che è l'effettiva distribuzione delle classi di argomenti rispetto a un particolare predicato si può considerare come p. La differenza tra le due distribuzioni può essere quantificata come segue:

$$S(p) = D(\Pr(c|p)||\Pr(c)) = \sum_c \Pr(c|p) log \frac{\Pr(c|p)}{\Pr(c)}$$

Questa quantità è definita come *selectional preference strength* e determina la forza della preferenza di selezione.

Nel modello di Resnik questa quantità è un numero con uno specifico significato; considerando p(c|p) come distribuzione vera e p(c) come approssimazione, la forza delle preferenze di selezione fornisce il costo, in informazione, di non considerare i predicati, pertanto la forza delle preferenze di selezione di un predicato fornisce l'ammontare di

[10] S. Kullback, R.A. Leibler, *On information and sufficiency*, Annals of Mathematical Statistics, 22, 79-86, 1951; T.M. Cover, J.A. Thomas, *Elements of information theory*. New York, Wiley, 1991.

informazione che questo veicola circa il proprio argomento. p(c|p) e p(c) sono valori che vengono stiamti a partire da un corpus di training.

1.1.2. Implementazione computazionale

Nella sua realizzazione computazionale, il metodo *knowledge-rich*[11] di Resnik utilizza come rappresentazione tassonomica l'ontologia di Wordnet[12]. Inserendo nel modello tutti i synset rappresentati in WordNet (mentre altri approcci selezionano solo alcuni topnodes, in modo da ottenere una rappresentazione più astratta delle preferenze semantiche[13]). L'algoritmo di Resnik funziona nel seguente modo:

1- Assegna il conteggio di co-occorrenza ai synset di WordNet che contengono un nome come testa lessicale; quando una parola ricorre in più di un synset, la sua frequenza viene divisa per il numero di synset.

   Nelle varie simulazioni del modello, Resnik utilizza principalmente il *Brown Corpus of American English*[14], il corpus di riferimento della lingua inglese e CHILDES[15], un corpus comprendente una serie di interazioni dialogiche con bambini.

[11] L'opposizione tra metodi knowledge rich/knowledge poor può essere applicata all'estrazione delle preferenze di selezione. L'approccio knowledge rich prevede di estrarre le preferenze di selezione dalla combinazione di thesauri costruiti manualmente con tecniche stocastiche per l'analisi dei corpora. Il metodo knowledge poor, invece, consiste nell'applicazione di metodi di machine learning non supervisionati senza considerare risorse esterne quali ontologie o corpora annotati manualmente. Cfr. G. Grefenstette, *Explorations in automatic thesaurus discovery*, Springer, 1994; P. Gamallo, A. Agustini, G. Lopes, *Clustering syntactic positions with similar semantic requirements*, Computational Linguistics, 31(1):107–146, 2005.

[12] WordNet è un database semantico contenente 90.000 parole inglesi tra verbi, nomi, aggettivi e avverbi, in cui le parole sono organizzate in classi semantiche. Le parole che condividono un topnode nell'ontologia formano un *synset*. Un *synset* comprennde un insieme di concetti legati da relazioni di sinonimia, meronimia o altro.

[13] Questo è l'approccio proposto da Schulte Im Walde, trattato successivamente.

[14] Cfr. W. Francis e H. Kučera, *Frequency analysis of English usage*, New York 1982.

[15] B. MacWhinney e C. Snow, *The child language data exchange system*, Journal of Child Language, 12, 1985.

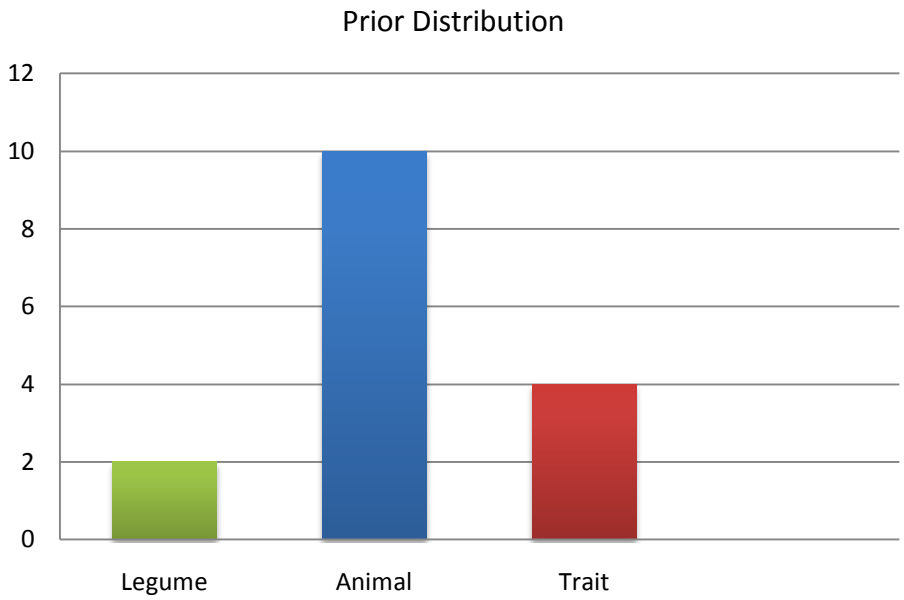
2- Si estende il conteggio di co-occorrenza a tutti i nodi della gerarchia di WordNet

Resnik fornisce l'esempio di due istanze della posizione di oggetto diretto per il verbo *to drink*: "*drink coffee"* e "*drink wine"*. *Coffee* fa parte di 13 classi di WordNet, *wine* di 16; quindi l'istanza di *drink coffee* nel corpus aggiungerà $\frac{1}{13}$ ad ognuna delle classi associate a *coffee*, mentre *drink wine* aggiungerà $\frac{1}{16}$ a tutte le 16 classi associate a *wine*. "Sebbene ognuna delle due parole sia ambigua, solo quelle classi tassonomiche contengono entrambe le parole, come ad esempio *beverage* aggiornano il proprio conteggio per entrambe le istanze. In generale, dal momento che parole diverse possono essere ambigue in modi differenti, il credito (il valore della frequenza) tende ad accumularsi nella tassonomia solo in quelle classi per le quali c'è un effettiva e significativa co-occorrenza, il resto tende a disperdersi senza sistematicità, causando soprattutto rumore. Di conseguenza, nonostante l'assenza di annotazione delle classi nel testo di training, è ancora possibile arrivare a una stima valida di probabilità basata su classi"[16].

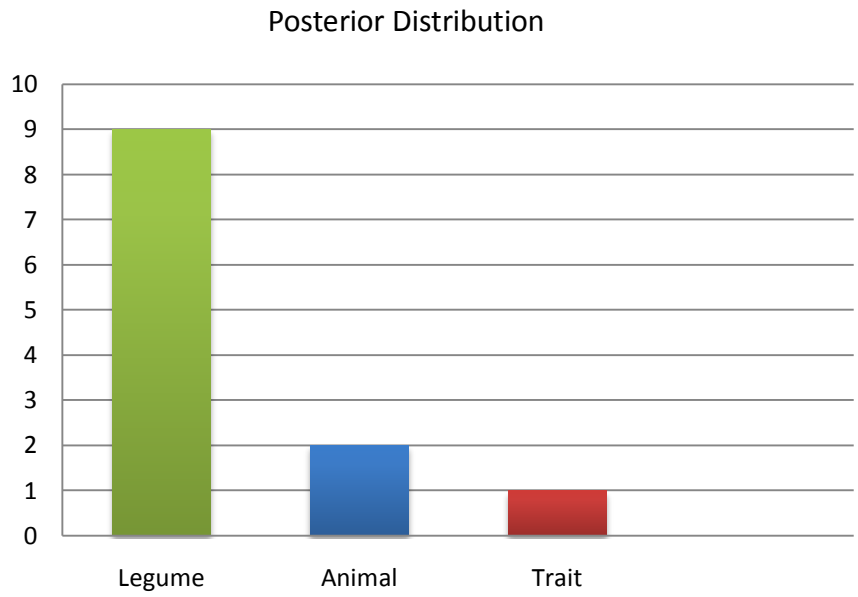
Questo è utilizzato per calcolare la *prior distribution*, *p(classe)* ovvero la probabilità che una classe di WordNet prenda una particolare posizione sintattica, indipendentemente dal predicato, che viene stimata usando soltanto la frequenza della classe costituita a partire dagli argomenti di un dato slot, e la *posterior distribution*, *p(classe|predicato)*, la probabilità di una classe di WordNet nella stessa posizione ma tenendo conto del predicato, stimata tramite la frequenza della coppia predicato-nome. La comparazione tra *prior distribution* e *posterior distribution* serve a quantificare quanto una classe semantica si adatti a uno slot predicato-argomento; le classi semantiche che si adattano meglio a un particolare slot hanno le probabilità più alte di co-occorrere con questi.

[16] P. Resnik, *Selectional preference and sense disambiguation,* Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How, 52–57, 1997.

La figura seguente, presa da Resnik[17], illustra questo approccio: dato un set di classi, *C* (nell'esempio *Legume*, *Animal*, *Trait*), un predicato, *to grow*, e una posizione sintattica, oggetto diretto, la *prior distribution p(classe)* è confrontata con la *posterior distribution p(classe|predicato)* . Tralasciando il predicato, *p(class)*, *Animal* tende a ricorrere più frequentemente come oggetto diretto rispetto a *Legume*; tuttavia se si effettuano i calcoli introducendo i predicati, *Legume* diventa molto più frequente di *Animal*.



[17] P. Resnik, *Selection and information: a class-based approach to lexical relationships*, Phd dissertation, University of Pennsylvania, 1993.



Per pesare la forza di associazione tra un predicato e una classe semantica per una specifica posizione sintattica Resnik applica alle due distribuzioni probabilistiche descritte precedentemente la misura di *entropia relativa* vista precedentemente, qui definita *Selectional Association* e calcolata nel modo seguente:

$$Selectional\ Association = p(classe|predicato)\ log\frac{p(classe|predicato)}{p(classe)}$$

Per quantificare le differenze a livello qualitativo di determinate frasi relativamente a un verbo Resnik fornisce un set di frasi di esempio e la Selectional Association per quel determinato verbo, è bene notare che alcune delle frasi proposte sono semanticamente anomale, perché violano le restizioni imposte dalle preferenze di selezione del predicato :

a- *Mary drank some wine.*
b- *Mary drank some gasoline.*

c- *Mary drank some pencils.*

d- *Mary drank some sadness.*

Selectional Association per il verbo *to drink*

Verbo, Argomento	Classe	Selectional Association
Drink wine	Beverage	0,088
Drink gasoline	Substance	0,075
Drink pencil	Object	0,030
Drink sadness	Psychological Feature	-0,001

Dunque, visto il funzionamento globale del modello di Resnik, si possono così ricapitolare le sue proprietà:

1- Le preferenze di selezione sono espresse in termini probabilistici.
2- Le assunzioni sulla rappresentazione del significato delle parole sono ridotte al minimo, non è necessario esprimere un significato con un set di condizioni booleane necessarie e sufficienti difficili da determinare in ogni circostanza. Il modello necessita di requisiti minimi sulla rappresentazione degli argomenti.
3- Il modello formale e la sua computazionale permetto di specificare preferenze di selezione che possono riferirsi a specifiche proprietà arbitrarie più ampie di un set limitato di primitive semantiche.
4- Il modello permette di avere una precisa misura quantitativa delle associazioni, la *Selectional Association*.

1.2. Modello Schulte Im Walde

Basato sul medesimo approccio proposto da Resnik è il modello elaborato da Sabine Schulte Im Walde[18] per l'assegnazione di tipi semantici agli argomenti dei verbi della lingua tedesca.

Per ogni combinazione verbo-frame-slot, la frequenza dei filler nominali è estesa alla gerarchia dei 15 top-node esclusivi di GermaNet[19] (*Creature*, *Thing*, *Property*, *Substance*, *Food*, *Means*, *Situation*, *State*, *Structure*, *Body*, *Time*, *Space*, *Attribute*, *Cognitive Object*, *Cognitive Process*). La frequenze delle parole collegate a più di un concetto viene divisa uniformemente tra questi . La differenza principale rispetto all'algoritmo di Rensik consiste nel non utilizzare tutti i synset dell'ontologia di WordNet, ma soltanto questi 15 top-node esclusivi. Le frequenze risultanti dei 15 nodi che occorrono in ogni slot sono poi usate per definire una distribuzione probabilistica, normalizzata sulla frequenza totale di co-occorrenza verbo-frame.

Oltre a considerare solo questa selezione di synset di GermaNet, l'agoritmo di Schulte Im Walde non apporta sostanziali differenze a Resnik nel calcolo della frequenza di co-occorrenza tra slot e classi, mentre per effettuare il clustering Schulte Im Walde utilizza il valore p(classe|predicato), diversamente da Resnik che usa una misura proveniente dalla teoria dell'informazione per caratterizzare in modo più ampio le associazioni tra classi semantiche e predicati.

[18] S. Schulte Im Walde, *Experiments on the Automatic Induction of German Semantic Verb Classes*, Computational Linguistics, 32(2):159–194, 2006.
[19] B Hamp e H Feldweg, *GermaNet: a Lexical Semantic Net for German*, Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, 1997.

2. LexIt

LexIt[20] è una risorsa lessicale per lo studio e l'analisi dei verbi della lingua italiana costruita in modo completamente automatico e basata sul corpus *La Repubblica* e Wikipedia. *LexIt* rappresenta di fatto la prima risorsa per la lingua italiana che comprenda informazioni distribuzionali sul comportamento dei verbi italiani[21].

L'obiettivo che questo strumento si pone è quello di descrivere il comportamento dei verbi della lingua italiana sotto il profilo distribuzionale sintattico e semantico, estraendo le informazioni necessarie in modo automatico. il processo per raggiungere questo obiettivo si articola in diversi sottotask: estrazione dei frames di sottocategorizzazione, assegnamento delle preferenze di selezione agli argomenti dei verbi sia come filler lessicali che come classi semantiche, identificazione del ruolo semantico, estrazione automatica delle classi verbali. Il prodotto risultante è una risorsa per lo studio di 3933 verbi italiani, un database comprendente

[20] Cfr. A. Lenci, *Carving Verb Classes from Corpora*, in R. Simone e F. Masini (a cura di) Word Classes, Amsterdam-Philadelphia: John Benjamins. Il progetto LexIt è consultabile all'indirizzo http://sesia.humnet.unipi.it/lexit/.

[21] La risorsa lessicale più simile a *LexIt* è rappresentata dal lessico SIMPLE, il quale contiene informazioni sulle preferenze di selezione verbali per le classi semantiche, ma non per filler lessicali o polisemia degli argomenti; inoltre mentre *LexIt* è costruito in modo completamente automatico, SIMPLE è stato sviluppato manualmente. Cfr A. Lenci et al., *SIMPLE: A general framework for the development of multilingual lexicons*, International Journal of Lexicography, 22(4):489–495, 2000.

frames sintattici, fillers lessicali e classi semantiche per l'analisi statistica dei verbi italiani.

LexIt può anche essere considerato come un lessico di valenza, infatti fornisce, per ciascun verbo, i più significativi pattern sintattici in termini di forza di associazione. La differenza principale tra Lexit e i tradizionali lessici di valenza, quali REDES[22] e il "Wörterbuch der italienischen Verben[23]", consiste nell'essere basato esclusivamente su corpus e costruito in modo completamente automatico, quest'ultima, caratteristica che presenta un punto debole, visto che l'automatizzazione del processo di costruzione introduce nei dati una gran quantità di rumore, tuttavia i risultati costituiscono una valida base per l'analisi del significati dei verbi e altri lavori correlati. Tuttavia, per alcuni aspetti, LexIt va oltre il puro lessico di valenza fornendo per ogni argomento di una costruzione sintattica, i suoi filler prototipici e le sue classi semantiche, inoltre è anche un dizionario elettronico di collocazioni totalmente basato su corpus. Un altro aspetto che lo differenzia dei dizionari tradizionali, invece, è la totale mancanza di frasi di esempio tratte dal corpus di partenza a corredo delle entrate lessicali.

2.1. Scelta e preparazione del corpus

Il primo passo nella costruzione di un lessico è la scelta del corpus di partenza, nella scelta si considerano aspetti quantitativi e qualitativi. Sebbene la dimensione del corpus sia un aspetto fondamentale, le proprietà qualitative risultano essere altrettanto rilevanti e capaci di influenzare sensibilmente il processo di acquisizione dei dati[24].

[22] I. Bosque, *Redes: diccionario combinatorio del espanol contemporáneo*, SM Ediciones, Madrid, 2004.

[23] P. Blumenthal, G. Rovere, *Wörterbuch der italienischen Verben*, Ernest Klettverlag, Stuttgart, 1998. Il *Wörterbuch der italienischen Verben* è l'unico lessico di valenza esistente per i verbi italiani, è basato su un corpus di 50 milioni di parole estratti da *Il Sole 24 Ore* (1989-1990). Il lessico è strutturato in modo che ad ogni entrata sia articolata in una serie di sensi, ciascuna correlata di esempi.

[24] Studi comparativi effettuati confrontando le frequenze dei frame estratte da tipi differenti di corpora dimostrano che fattori come il tipo di discorso e le scelte

Nella costruzione di *Lexit* il corpus scelto è *La Repubblica*[25], sviluppato all'SSLMIT dell'Università di Bologna. Il corpus consta di una collezione di circa 600.000 articoli pubblicati dal 1985 al 2000 dal quotidiano *La Repubblica*, per un totale di 386 milioni di parole divise per tipologia e dominio. La scelta è ricaduta su Repubblica per una serie di motivi: innanzitutto è considerato il corpus di riferimento per la lingua italiana, poi copre un lungo arco di tempo ed il linguaggio usato è quello giornalistico, il quale dovrebbe essere rappresentativo dell'italiano standard, inoltre le categorizzazioni di tipo e dominio si prestano ad analisi ulteriori su specifici sottocorpora.

Nella fase di preparazione vengono effettuate delle analisi automatiche: il corpus viene prima lemmatizzato e sottoposto a un Pos Tagging tramite *ILC-UniPi Tagger*[26], poi ad un parser a dipendenze stocastico, *DeSR*[27]. Dal corpus risultante viene ricavato un profilo distribuzionale per ogni verbo trattato e ogni profilo ottenuto viene organizzato in un profilo sintattico e uno semantico.

2.2. Metodologia

La metodologia usata per costruire LexIt è basata sulle tradizionali misura di associazione applicate allo studio delle collocazioni per valutare la forza di correlazione tra:

semantiche effettuate dai parlanti incidono sui risultati dell'acquisizione. Cfr. D. Roland and D. Jurafsky, *How verb subcategorization frequencies are affected by corpus choice*, Proceedings of the 36th annual meeting on Association for Computational Linguistics, 1122–1128, 1998.

[25] M. Baroni et al., *Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian*, Proceedings of LREC 2004, 1771–1774, Lisboa, 2004.

[26] F. Dell'Orletta et al., *Maximum Entropy for Italian PoS Tagging*, Intelligenza Artificiale, IV(2):10–11, 2007.

[27] C. Bosco et al., *Evalita'09 Parsing Task: comparing dependency parsers and treebanks*, Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, Reggio Emilia, 2009.

1- Verbi e frame sintattici (estrazione di frame di sottocategorizzazione)
2- Argomenti verbali e parole che li compongono (identificazione del set lessciale per lo slot di un frame sintattico)
3- Argomenti verbali e classi semantiche assegnate loro dall'algoritmo

Il primo passo per un'analisi distribuzionale è organizzare le frequenze dei dati in coppie in una matrice di correlazione:

	$w_1 = a$	$w_1 \neq a$	
$w_2 = b$	$a\ AND\ b$	$b\ NOT\ a$	$= R_1$
$w_2 \neq b$	$a\ NOT\ b$	$NOT\ b\ NOT\ a$	$= R_2$
	$= C_1$	$= C_2$	$= N$

La frequenza della combinazione $a\ AND\ b$ quantifica la *Observed Joint Frequency* (O) della coppia di parole, questo è un parametro necessario ma non sufficiente per determinare la forza della possibile associazione tra *a* e *b*. R_1, R_2, C_1 e C_2 rappresentano le frequenze marginali, la cui somma determina la grandezza *N* che rappresenta il conteggio delle possibili combinazioni tra w_1 e w_2 nel corpus. In particolare R_1 rappresenta le istanze di *a* , R_2 le istanze di *b*. Il conteggio della frequenza disposto in una matrice così composta è usato per calcolare la *Expected Joint Frequency* (E) della coppia di parole. Se i due termini sono indipendenti tra loro, è possibile il numero di co-occorrenze applicando la formula seguente:

$$E = \frac{R_1 C_1}{N}$$

Fatto questo, è necessario scegliere un'adeguata misura di associazione in base alla quale valutare e ordinare le coppie di parole. Per stabilire la forza di associazione tra *a* e *b* O non è sufficiente, infatti se ha un alto valore, ma equivalente a E, l'associazione tra a e b potrebbe essere casuale. E è il punto di riferimento, rappresenta in termini statistici il valore che si vorrebbe raggiungere a meno che le due parole siano indipendenti.

Le misure di associazione dovrebbero essere capaci di esprimere quantitativamente il livello di attrazione tra le coppie: dovrebbero quindi assegnare valori alti alle coppie di parole che si attraggono in maniera forte, e valori bassi alle coppie con legami molto deboli tra loro. Proprio sulla base della forza di associazione/repulsione è possibile distinguere tra misure *one-sided*, che distinguono tra correlazioni positive e negative, quantificando sia la forza di attrazione che quella di repulsione, e misure *two-sided*, un'ulteriore divisione è quella operata tra *effect size measures* e *significance measures*[28].

Le misure *effect size* calcolano quanto le due parole si attraggono rispettivamente, assegnando così valori alti alle coppie la cui frequenza di co-occorrenza è superiore alla frequenza attesa, la misura più conosciuta che rientra in questa tipologia è la *Mutual Information*[29] (MI). Questa misure, proveniente dalla Teoria dell'Informazione, quantifica le parti di informazione condivisa tra due parole co-occorrenti w1 e w2, in particolare confronta la probabilità di co-occorrenza delle due parole con la probabilità di osservarle l'una indipendente dall'altra:

$$MI(w_1, w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

La formula può essere riscritta nei termini della *obseved joint frequency* e *expected joint frequency*:

$$MI = log_2 = \frac{O}{E}$$

Il limite della MI è, che come tutte le *effect size measures*, è sensibile agli eventi rari, pertanto è necessario stabilire una soglia di frequenze per filtrare i dati ed eliminare le coppie con frequenze troppo basse. Una derivata dalla MI molto usata nello studio delle collocazioni e utilizzata anche in LexIt è *la Local Mutual Information*, la LMI. Il vantaggio della LMI rispetto alla MI è quello di evitare la tendenza a sovrastimare gli eventi con frequenze molto basse.

[28] S. Evert, *Corpora and collocations*, Corpus Linguistics. An International Handbook, Mouton de Gruyter, Berlin, 2008.
[29] K. Church and P. Hanks, *Word association norms, mutual information, and lexicography*, Computational linguistics, 16(1):22–29, 1990.

Per quanto riguarda le Significance measures invece, la misura più rappresentativa è la Simple log-likelihood, definita come segue:

$$Simple - ll = 2\left(O\ log\frac{O}{E} - (O - E)\right)$$

La Simple-ll è una *two-sided measure*: assegna la stessa forza di associazione alle coppie associate negativamente o positivamente con la medesima forza.

2.3. Profili distribuzionali per i verbi italiani

Il *distributional profile* di un verbo consiste nell'insieme di informazioni estratte da un corpus per caratterizzare le proprietà distribuzionali del verbo. In letteratura sono stati proposti numerosi metodi per l'acquisizione automatica dei dati, per l'estrazione di frames di sottocategorizzazione, per l'identificazione delle preferenze di selezione verbali[30].

Per quanto riguarda l'estrazione di frames di sottocategorizzazione[31] (SCFs) la caratteristica distintiva di *Lexit* è il suo approccio automatico e non supervisionato al problema: non si fornisce al modello una lista precostituita di frames sintattici, ma vengono identificati automaticamente le costruzioni sintattiche più frequenti nel corpus. Dopo il processo di estrazione si definiscono gli SCFs del modello.

La lista dei suoi frames di sottocategorizzazione (SCFs) di un verbo, ordinati secondo la loro rilevanza statistica, definisce il suo profilo. Ogni SCF corrisponde a uno specifico pattern di dipendenze sintattiche di quello specifico verbo, è formato da un set di slot e identificato da un'etichetta sintetica, ad esempio:

[30] D. Manning e H. Schütze, *Foundations of Statistical Language Processing*, Cambridge Mass.: MIT Press 1999; M. Light, W. Greiff, *Statistical Models for the Induction and Use of Selectional Preferences*, Cognitive Science: 26.269–281.

[31] Per sottocategorizzazione verbale, o valenza verbale si fa riferimento alla capacità dei verbi di scegliere i propri complementi. La struttura di sottocategorizzazione può anche essere definita frame di sottocategorizzazione (SCF), questa fornisce uno strumento per formulare generalizzazioni sul comportamento dei verbi.

soggetto + complemento introdotto dalla *a* + oggetto diretto = *subj#obj#comp-a*
complemento introdotto dalla preposizione *a* + oggetto diretto = *comp-a#obj*

Nel modello viene anche considerato il pronome riflessivo *si* e il caso in cui un verbo appaia nella forma senza dipendenze, ad esempio nella frase "Il vaso si è rotto" (*subj#si#*0). Il processo di definizione del profilo sintattico avviene selezionando un numero di SCFs tra le combinazioni più frequenti, per ogni verbo si confronta la *joint frequency* con ogni SCF, basato su pattern estratti automaticamente dal corpus una volta sottoposto al parser. Dalla frequenza combinata verbo-SCF si ottiene il punteggio di Local Mutual Information[32] (LMI) che restituisce una stima della rilevanza statistica di quel SCF per il verbo dato.

2.4. Profili Semantici

Nella costruzione di *Lexit* si parte dal metodo di Resnik per la procedura del conteggio della frequenza e dell'applicazione di una misura di associazione per valutare la correlazione tra un argomento e la classe semantica che occorre insieme ad esso, ma non si utilizzano tutti i synset di WordNet, bensì un gruppo di top-node selezionati che permettono di avere una visione generalizzata dei set lessicali, analogamente a quanto fatto nel lavoro di Schulte Im Walde[33]. Per ogni verbo si considerano i frames con LL > 0 e per ciascuno slot di questi frame si considerano i filler con LL > 0, usando quindi la LL come una sorta di filtro. Per l'assegnazione di profili semantici a questi slot sono state implementate una serie di variazioni all'algoritmo di Schulte Im Walde, tra le quali la divisione uniforme tra i sensi assegnati ai

[32] La Local Mutual Information è una variante della Pointwise Mutual Information, che risolve alcuni problemi relativi agli eventi poco frequenti che venivano sottostimati. È una misura comunemente usata nell'analisi delle collocazioni lessicali.
[33] S. Schulte Im Walde, *Experiments on the Automatic Induction of German Semantic Verb Classes*, Computational Linguistics, 32(2):159–194, 2006.

nomi nella sezione italiana di MultiWordNet[34] delle frequenze di co-occorrenza di ciascun nome nel ruolo di filler di un dato verbo e il calcolo del valore di associazione della LL effettuato tra ogni combinazione verbo-frame-slot e le top-classi di WordNet.

L'approccio usato nella costruzione di *LexIt* condivide dunque con Schulte Im Walde la selezione di un sottoinsieme di top-node esclusivi usati per un'analisi automatica, mentre le misure di associazione della relazione tra classi semantiche e predicati si basano sul modello precedente di Resnik.

Set Lessicali e Profili Semantici di *leggere*, *sfogliare* e *pubblicare*

Verbo	Set Lessicale	Profilo Semantico
Leggere	libro, giornale, testo, articolo, lettera, romanzo, dichiarazione, pagina.	Communication Artifact Time Substance Motivation
Sfogliare	pagina, margherita, giornale, libro, album, catalogo, rivista, volume, quotidiano, fascicolo.	Communication Artifact Plant Substance Group
Pubblicare	Libro, foto, articolo, lettera, romanzo, notizia, stralcio, intervista, testo, volume. saggio, disco.	Communication Artifact Shape

I profili semantici assolvono una funziona descrittiva e predittiva: da una parte i set lessicali forniscono una panoramica dei nomi che occorrono nel corpus con un verbo in una determinata posizione sintattica, con una valutazione della loro rilevanza statistica. D'altra parte le preferenze di

[34] MultiWordNet è un lessico computazionale basato su WordNet. La sezione italiana è allineata con quella inglese: i synset italiani sono creati come corrispondenti di quelli in lingua inglese e le relazioni semantiche sono importate dal database inglese. Cfr. L. Bentivogli et al., *Multiwordnet: developing an aligned multilingual database*, First International Conference on Global WordNet, number 1996, Mysore, India, 2002.

selezione permettono di generalizzare a partire dalle istanze a delle proprietà astratte degli argomenti dei verbi, consentendo di formulare delle predizioni sugli argomenti non considerati.

Un problema centrale in questa fase della costruzione del lessico è stabilire il livello di granularità dell'informazione semantica associata ai filler da MultiWordNet: nella prima fase si selezionano tutte le parole associate ai 24 top-node selezionati, ma questa operazione può creare delle anomalie nei risultati, derivanti dal fatto che MultiWordNet contiene termini anche molto specifici: ad esempio la parola *libro* non rientra solo nella tipologia *Artifact* e *Communication*, ma designa anche una parte dello stomaco dei ruminanti, quindi *Body Part*. Pertanto *libro* ha come associazione principale il verbo *leggere*, mentre *Body Part* è identificata come seconda classe maggiormente associata nello slot di oggetto diretto. Inoltre dalle analisi qualitative viene evidenziato come l'algoritmo non rappresenta correttamente gli usi metaforici, come *leggere la mano* o *leggere le labbra.*

3. WordNet e MultiWordNet

WordNet[35] è un lessico computazionale semantico per la lingua inglese. Nomi, verbi, aggettivi e avverbi sono raggruppati secondo un criterio di sinonimia. Questo modo permette di rappresentare i concetti lessicali tramite l'insieme delle parole sinonime che lo esprimono. Tali insiemi, detti synset, sono collegati tra loro tramite relazioni che insieme costruiscono una rete semantica.

Per far fronte al problema della polisemia, WordNet distingue in modo netto i significati delle parole, i concetti lessicali, dalle forme ad essi associate, i significanti, tramite i quali i concetti vengono espressi. Essendo questa corrispondenza un rapporto molti a molti, essa viene rappresentata con un matrice bidimensionale:

	F_1	F_2	F_n
S_1		$E_{1,2}$	
S_2	$\{E_{2,1}$	$E_{2,2}$	$E_{2,n}$
S_m		$E_{m,2}$	

[35] G. A. Miller et al., *WordNet: An online lexical database*, Int. J. Lexicograph. 3, 4, 235–244, 1990; Fellbaum, Christiane, *WordNet and wordnets*, Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670, 2005.

Ogni colonna contiene i vari significati (S_1, S_2,…,S_m) di una stessa forma F, rappresentando così il fenomeno della polisemia delle parole. In ogni riga compaiono invece tutte le forme di parola (F_1, F_2,…,F_n) sinonime, con lo stesso significato S. Ogni cella della matrice raffigura la corrispondenza, E, tra una forma di parola e un significato (unità lessicale). Esempio:

animal	{dog}		
hinged device	{dog	pawl	click}

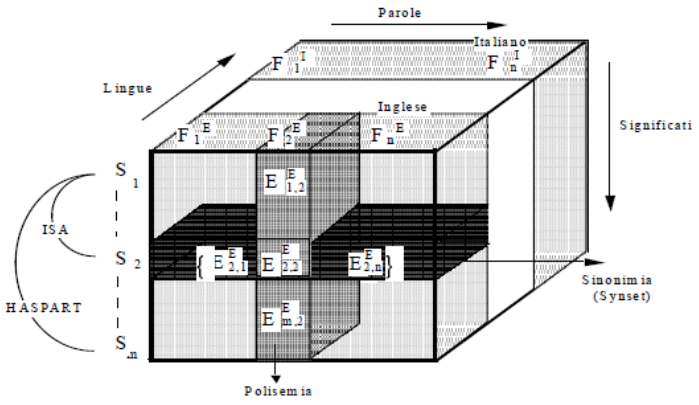
Rappresentare un concetto tramite synset rende necessario mantenere separati nomi, verbi, aggettivi e avverbi, visto che parole di categorie sintattiche diverse non possono essere sinonime e devono essere organizzati e strutturati in modo diverso nel database lessicale. Pertanto i nomi sono organizzati come gerarchie ereditarie, gli aggettivi in base all'opposizione semantica, antonimia, e i verbi secondo una serie di relazioni di implicazione. Le relazioni semantiche valgono tra i concetti, quindi tra i synset presi come unità, mentre le relazioni lessicali valgono tra le parole contenute nei synset stessi. La relazione lessicale più comune che si ritrova in tutte le categorie lessicali è quella di sinonimia, usata anche come criterio costruttivo dei synset, le altre relazioni sia lessicali che semantiche sono specifiche per le diverse categorie lessicali, come si può vedere nella tabella seguente:

Categoria	Relazione	Tipo	Esempio
Nomi	ipo/iperonimia meronimia	Semantica	dog IS A KIND OF animal arm IS A PART OF body
Verbi	Implicazione: Causa Troponimia opposizione	Semantica	 to kill CAUSES to die to limp IS ONE WAY to walk to die ANTONYM to be born
Aggettivi	antonimia	Lessicale	hot
Avverbi	aggettivo da cui deriva antonimia	Lessicale	quickly DERIVED FROM quick quickly ANTONYM slowly

3.1. MultiWordNet

MultiWordNet[36] è un database lessicale multilingue contenente informazioni sui termini di numerose lingue, tra cui l'italiano e può essere considerato un'estensione di WordNet. L'ipotesi teorica su cui si basa MultiWordNet è quella secondo cui le strutture sul livello lessicale di lingue diverse siano confrontabili e in gran parte sovrapponibili, dunque che i parlanti di lingue diverse possano condividere gran parte dei concetti lessicali e delle relazioni che intercorrono tra questi concetti.

Partendo da questa ipotesi di base, è stata elaborata una metodologia per costruire una rete semantica di qualsiasi lingua utilizzando come base di partenza la rete concettuale già esistente per la lingua inglese. Per raggiungere questo scopo si è realizzata una matrice lessicale multilingue che permettesse di estendere la matrice bidimensionale implementata in WordNet, aggiungendo una terza dimensione sulla quale è possibile considerare le diverse lingue. Tale matrice può essere rappresentata con la figura seguente:



[36] B. Magnini e C. Strapparava, *Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet*, Atti del XXVIII Congresso della Società di Linguistica Italiana, 415-418, Palermo, 1994; Artale et al., *Lexical Discrimination with the Italian Version of WordNet*, Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid; *WordNet for Italian and Its Use for Lexical Discrimination*, Atti del Quinto Congresso della Associazione Italiana per l'Intelligenza Artificiale, Roma 1997.

Ogni strato verticale della matrice corrisponde a una lingua diversa, il primo strato rappresenta la matrice lessicale bidimensionale di WorNet in cui, ogni colonna rappresenta i vari significati (S_1, S_2,..,S_m) di una stessa forma F e ogni riga rappresenta tutte le forme di parola (F_1, F_2,...,F_n) che hanno lo stesso significato S. La costruzione della matrice si effettua trovando le corrette corrispondenze tra i synset della lingua presa in considerazione e i synset inglesi già esistenti, ogni nuovo synset viene creato in corrispondenza con un synset inglese, creando così automaticamente una relazione di equivalenza tra i synset delle due lingue. Stabilita questa corrispondenza, ogni coppia di synset della nuova lingua eredita le relazioni semantiche valide per i synset inglesi equivalenti. L'estensione delle relazioni semantiche dell'inglese alle altre lingue si basa sull'equivalenza dei synset in lingue diverse e quindi l'assunzione che le relazioni semantiche siano costanti rispetto alle lingue. Al contrario, le relazioni lessicali vanno ridefinite per ciascuna lingua, dal momento che dipendono dalla lingua presa in analisi.

Il vantaggio dell'approccio appena descritto risiede nella possibilità di "riutilizzare" gran parte delle informazioni già disponibili in WordNet. In particolare, esso permette di acquisire i synset ed ereditarne le relazioni che li strutturano in modo semi-automatico. La scelta di acquisire le relazioni automaticamente dalla lingua inglese costituisce un grande vantaggio pratico nella creazione della rete semantica.

3.1.2. Implementazione del modello

Il modello proposto da MultiWordNet per la costruzione di un WordNet multilingue consiste nella costruzione di una rete semantica che mantenga il più possibile le relazioni semantiche preesistenti in WordNet. Questo si realizza creando i nuovi synset in perfetta corrispondenza con i synset esistenti per la lingua inglese e importandone le relazioni semantiche[37]. Questo approccio permette di minimizzare tutte quelle differenze dipendenti dalle varie lingue, inoltre permette di restare aderente il più possibile ai criteri costruttivi e alle scelte del WordNet originale. Il potenziale risvolto negativo è quello di una dipendenza forzata sia dal lessico che dalle strutture concettuali della lingua inglese. Tale rischio è tuttavia ridotto dalla possibilità, durante la costruzione delle nuove strutture, di divergere dalla struttura originaria di WordNet, quando necessario.

Un altro importante vantaggio del modello è l'automatizzazione della procedura di creazione dei nuovi synset in corrispondenza a quelli esistenti di WordNet e l'individuazione delle divergenze tra i synset esistenti e i nuovi durante la costruzione. Le procedure automatiche su cui si basa la costruzione di WordNet per la lingua italiana sono due[38]. La prima è la *Assign-procedure* che permette, dato un concetto in italiano, di selezionare una lista dei corrispondenti synset esistenti in WordNet più probabili, questa lista è usata per creare in modo semi-automatico i synset italiani. La seconda procedura, la *LG-procedure*[39], permette di individuare *lexical-gaps*, ovvero i casi in cui un determinato concetto di un linguaggio è espresso tramite una combinazione di parole in un altra lingua. Entrambe

[37] Un altro metodo possibile per la costruzione di un database semantico-lessicale multilingue basato su WordNet è quello proposto nella costruzione di EuroWordNet e consiste nella costruzione di più WorNet specifici per ciascun linguaggi e tra loro indipendenti, successivamente si provvede a ricercare le corrispondenze tra i vari modelli costruiti. Cfr. P. Vossen, *Categories and classifications in EuroWordNet*, Proceedings of the First International Conference on Language Resources and Evaluation. Granada, 399-407, 1998; *Special Issue on EuroWordNet*, Computer and the humanities, 2-3, 73-251, 1998.

[38] cfr. Pianta et al., *MultiWordNet: developing an aligned multilingual database*, Proceedings of the First International Conference on Global WordNet, Mysore, India, 2002.

[39] cfr. L. Bentivogli, E. Pianta, *Looking for lexical gaps*, Proceedings of the ninth EURALEX International Congress, Stuttgart, 2000.

queste procedure usano come risorsa linguistica la versione elettronica del dizionario bilingue Collins.

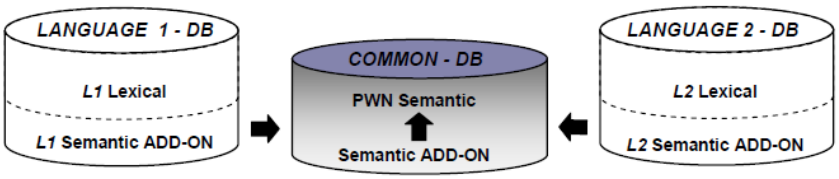
3.1.3. Modello dei dati

Il modello dei dati risultante di MultiWordNet riflette ovviamente l'assunzione secondo la quale esiste un insieme di informazioni comuni a "tutte le lingue, e altri dati specifici per ogni lingua.

Le relazioni semantiche (iperonimia, meronimia, ecc...) sono condivise, mentre le relazioni lessicali dipendono da una particolare lingua. Pertanto nell'implementazione del modello italiano esiste un modulo comune, COMMON-DB, contenente le relazioni semantiche, e due moduli distinti per le relazioni lessicali di ciascuna lingua, ITALIAN-DB e ENGLISH-DB. Un altro elemento focale è la corrispondenza incrociata tra i synset, realizzata usando lo stesso identificatore dei synset nelle diverse lingue; tutti i synset con lo stesso identificatore appartenenti a lingue diverse puntano allo stesso *multisynset*, il quale non riguarda più le relazioni tra parole sinonime della stessa lingua, ma individua una relazione di sinonimia più ampia tra synset equivalenti in lingue diverse. COMMON-DB descrive le relazioni tra i *multisynset* di MultiWordNet, di fatto tutta l'informazione indipendente dal linguaggio viene aggiunta al modulo comune.

Il modello dei dati di MultiWordNet intende rappresentare il fatto che lingue differenti condividono una grande quantità di informazione a livello concettuale, tuttavia il modello dei dati deve anche rappresentare le divergenze concettuali tra le lingue, pertanto, nonostante le relazioni semantiche costituiscano il database comune, è possibile aggiungere nuove tramite il modulo COMMON-ADD-ON o i moduli aggiungiti per ciascuna lingua che possono sovrascrivere i dati originali.

La seguente figura fornisce una panoramica delle principali caratteristiche del modello dei dati di MultiWordNet, le frecce rappresentano le relazioni che vengono sovrascritte.



I principali problemi ricorrenti nella costruzione di MWN italiano possono essere i lexical-gaps citati sopra, che si verificano quando una singola unita lessicale di una lingua viene espressa con una serie di parole nell'altra lingua e le differenze di denotazione del significato, nel caso in cui esistano corrispondenti tra le lingue, ma hanno una denotazione differente (più generico, più specifico). Per ovviare a questi problemi, nell'architettura di MultiWordNet vengono inseriti degli speciali nodi vuoti nel caso in cui un concetto di una lingua non abbia corrispondenti nell'altra.

4. Descrizione dell'esperimento

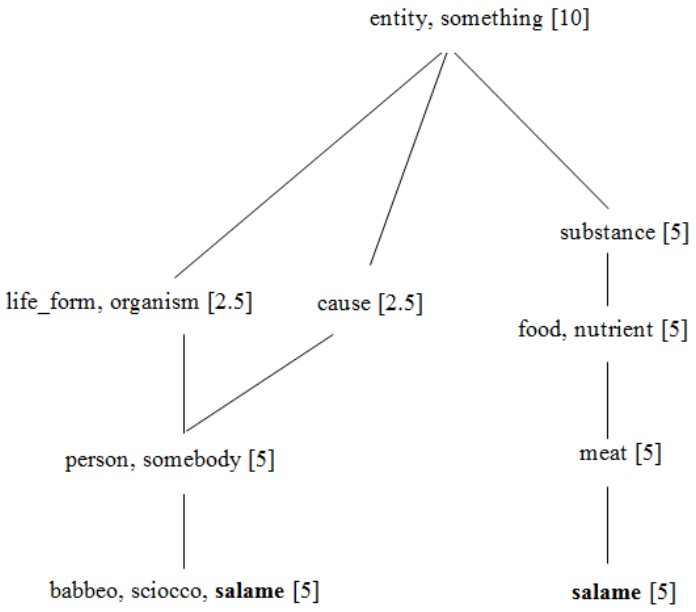
Il programma realizzato mira a estrarre le preferenze di selezione dei verbi italiani basandosi sulle informazioni estratte da LexIt e utilizzando la gerarchia di MultiWordNet.

LexIt, come detto precedentemente, contiene i profili distribuzionali dei verbi italiani estratti dal corpus *La Repubblica* e Wikipedia. Il formato dei dati di output estratti da LexIt utilizzati in questo lavoro è così composto, per ogni verbo sono specificati i frames di sottocategorizzazione, gli slot del frame, i filler nominali e le frequenza per ogni filler. Il formato dei dati estratti da LexIt è dunque composto dalla combinazione verbo-frame-filler-frequenza, come mostrato nell'esempio seguente.

testa verbale + frame	filler nominale	frequenza
avanzare-v%obj%obj	comitiva-s	1
avanzare-v%obj%subj	invasione-s	2
avanzare-v%si#0%subj	candidatura-s	5

Tali dati estratti da LexIt forniscono precise informazione lessicale, ma è necessaria una risorsa per generalizzare l'informazione relativa alle preferenze di selezione e associarla alla descrizione dei verbi rispetto a

determinati argomenti. WordNet è stato ampiamente utilizzato come fonte per un'informazione a livello abbastanza dettagliato sulle preferenze di selezione[40] Pertanto come fonte per l'informazione sulle preferenze di selezione viene utilizzato MultiWordNet che contiene anche la lingua italiana. Come detto prima, MultiWordNet è organizzato in una gerarchia di synset sinonimi. La figura seguente fornisce una rappresentazione semplificata della gerarchia di MultiWordNet per il nome *salame.*



Come si vede dalla figura, il nome *salame* è associato a due sensi: salame come sinonimo di persona sciocca e salame come tipo di cibo. La gerarchia di MultiWordNet per ogni nome contenuto nei filler dei file estratti da LexIt,

[40] F. Ribas, *On Learning More Appropriate Selectional Restrictions*, Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, Dublin, Ireland 1995; P. Resnik, *Selectional Preference and Sense Disambiguation*, Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, 1997; S. Clark e D. Weir, *Class-Based Probability Estimation using a Semantic Hierarchy*, Computational Linguistics, 28, 2002. Anche WordNet per la lingua tedesca GermaNet è stata utilizzata come fonte per le preferenze di selezione. Cfr. A. Wagner, *Enriching a Lexical Semantic Net with Selectional Preferences by Means of Statistical Corpus Analysis*, Proceedings of the ECAI Workshop on Ontology Learning, 37–42, Berlin, Germany 2000.

viene utilizzata per la costruzione delle preferenze di selezione per quella determinata combinazione verbo-frame-filler a quello specifico livello della gerarchia.

L'approccio seguito nella costruzione del programma è simile a quello proposto da Schulte Im Walde[41], e procede in questo modo. Per ogni nome in una combinazione verbo-frame-filler la frequenza è divisa per il numero di sensi del nome e propagata per tutta la gerarchia. Se un synset è collegato a iperonimi multipli, la frequenza viene divisa per il numero di iperonimi, se più synset puntano ad un unico iperonimo, la frequenza viene sommata. Ovviamente la somma della frequenza dei top-nodes sarà uguale alla frequenza iniziale della combinazione verbo-frame-filler. Ad esempio se la frequenza del nome *salame* rispetto al verbo *mangiare* in posizione di oggetto diretto è uguale a 10, si assegna a entrami i synset contenenti il nome un valore pari a 5. I valori di questi nodi si propagano per tutta la gerarchia, dividendosi in caso di nodi multipli e sommandosi in caso più nodi puntino a uno solo. Ripetere l'assegnamento della frequenza per tutti i nomi contenuti negli slot, fornisce una distribuzione della frequenza di tutte le combinazioni verbo-frame-filler estratte da LexIt su tutti i synset di MultWordNet.

La novità nell'approccio proposto rispetto all'algoritmo di Schulte Im Walde consiste nel non esprimere le preferenze di selezione soltanto nei termini di una selezione di top-nodes mutualmente esclusivi, ma per tutti i nodi intermedi appartenenti alla gerarchia.

Ricapitolando i passi seguiti nell'esecuzione del programma sono i seguenti:

- Per ogni nome in una combinazione verbo-frame-filler la frequenza del nome viene divisa per il numero dei suoi sensi
- Questa operazione viene propagata a tutta la gerarchia dei synset, se un synset ha più iperonimi, la frequenza viene divisa per il numero

[41] S. Schulte Im Walde, *GermaNet Synsets as Selectional Preferences in Semantic Verb Clustering*, Proceedings of the GermaNet Workshop, 2003; *Experiments on the Automatic Induction of German Semantic Verb Classes*, Computational Linguistics 32(2):159-194, 2006.

di iperonimi, se più synset puntano a un solo iperonimo, la loro frequenza si somma.

- L'operazione viene ripetuta per tutti i dati estratti da LexIt.
- Viene calcolata la forza di associazione usando la Local Mutual Information per ogni coppia verbo-frame rispetto a un dato synset.

Il formato di output dei dati è coerente con quello utilizzato da LexIt, ogni filler della sequenza verbo-frame-filler contenente un argomento nel file di partenza viene sostituito da tutti i synset che incontra nel risalire la gerarchia, ad ognuno di questi viene associato il valore di frequenza corrispondente.

testa verbale + frame	synset	frequenza
avanzare-v%obj%obj	accomplishment	7.616
avanzare-v%obj%obj	accusation;accusal	23.375
avanzare-v%obj%obj	activity	1004.9
avanzare-v%si#0%subj	abstraction	70.993
avanzare-v%si#0%subj	act;human_action	218.25
avanzare-v%si#0%subj	affair;occasion	0.3333

Un elemento importante da considerare è che la lingua di partenza dei synset è l'italiano, in modo da potersi interfacciare con i dati estratti da LexIt, nella navigazione dell'albero si passa alla gerarchia dei synset inglesi, questa scelta è motivata dal voler usare la lingua inglese come metalinguaggio per rappresentare le classi, questo è reso possibile dal fatto che in MultiWordNet i concetti rappresentati sono organizzati in maniera parallela tra le varie lingue, come detto precedentemente.

4.1 Analisi dei risultati

Una volta navigata l'intera gerarchia di MultiWordNet i dati risultanti composti da verbo-frame-synset e rispettiva frequenza, vengono valutati utilizzando la Local Mutual Information per stabilire la forza di associazione tra quella coppia verbo-frame e quello specifico synset. Ottenuti i dati ordinati, si procede ad un'analisi qualitativa al fine di mettere in luce fenomeni che non sarebbe stato possibile osservare esprimendo le preferenze di selezione solo in termini dei top-node, come viene fatto attualmente in LexIt. Analizzare le preferenze di selezione per ogni synset della gerarchia permette infatti di avere una rappresentazione più specifica del comportamento dei verbi nella selezione degli argomenti, basandosi sull'assunzione che le classi ai livelli intermedi della gerarchia possano caratterizzare meglio le specificità del comportamento dei verbi, impossibili da vedere considerando solo le classi in cima alla gerarchia (ovvero i top-nodes), troppo ampie e generiche.

Un esempio che dimostra i risultati ottenuti dall'analisi riguarda il verbo *abbagliare* con frame di sottocategorizzazione #obj (oggetto diretto), il synset con un più alto maggiore di LMI risulta essere *radiation* nello slot. La tabella seguente e le successive rappresentano i dati nel formato del database di LexIt: verbo % frame di sottocatorizzazione % slot sintattico – synset associato alla fine dell'analisi e forza di associazione.

verbo-frame	synset	LMI
`abbagliare-v%obj%subj`	`radiation`	`10.586`

Se si guardasse soltanto ai top-nodes della gerarchia di MultiWordNet, il verbo andrebbe associato a *natural_phenomenon*, un concetto molto più generico e astratto che può esprimere un insieme di concetti anche molto diversi da *radiation*. È dunque un fenomeno rilevante rispetto all'analisi effettuata solo sui top-nodes e caratterizza meglio il comportamento del verbo per quel determinato frame.

Per lo stesso frame si sono riscontrati altri casi degni di nota, alcuni esempi significativi possono essere: il verbo *abbandonare* il cui valore di LMI per il synset *housing;lodging* è molto più alto di quello per *artifact* o *physical_object* (1.341) e lascia intendere una specificazione maggiore del verbo sugli argomenti preferiti, riferiti a contesti specifici.

verbo-frame	synset	LMI
abbandonare-v%obj%obj	housing;lodging	6.305
abbandonare-v%obj%obj	work	5.696
abbandonare-v%obj%obj	duty	4.091
abbandonare-v%obj%obj	energy	3.768

Negli esempi seguenti, verbo *avanzare* e *fingere*, si vede chiaramente come i synset col più altro valore di LMI cui si associa il verbo risultano molto più caratterizzanti dei corrispettivi top-nodes. In questo caso l'unico top-node cui convergono *proposition*, *inactiveness*, *obedience* e *submissiveness* è *abstraction*, che avrebbe fatto perdere molta informazione sulle preferenze dei due verbi.

verbo-frame	synset	LMI
avanzare-v%obj%obj	proposition	10.421
fingere-v%obj%obj	inactiveness;inactivity	371
fingere-v%obj%obj	obedience	221
fingere-v%obj%obj	submissiveness	217

Cambiando il frame si osservano altri risultati interessanti, ad esempio restringendo il campo di analisi al frame subj#0 si possono notare altri risultati interessanti, come ad esempio il verbo *avanzare*, il cui secondo synset con maggior valore di LMI, dopo *person,individual,someone* risulta essere *leader*.

verbo-frame	synset	LMI
avanzare-v%0%subj	leader	253,24

Guardando il comportamento del verbo in LexIt, si nota che la classe semantica con cui ha maggior forza di associazione è *Person* (73.7787) e l'iperonimo top-node di *leader* in MWN è *life_form,oganism*, quindi il valore alto di LMI risultante da questa analisi restringe il campo e specifica meglio il concetto all'interno di una classe molto più ampia e varia.

4.2 Conclusioni e possibili sviluppi

L'analisi effettuata ha permesso di ottenere un livello di generalizzazione più specifico per descrivere il comportamento dei predicati nella selezione dei loro argomenti, effettuare l'analisi e calcolare i risultati per tutte le classi della gerarchia ha consentito di evidenziare delle specificità nelle preferenze di selezione messe in luce dalle classi intermedie di MultiWordNet.

L'analisi, come detto sopra, è effettuata navigando la gerarchia di MWN a partire dai synset italiani estratti da LexIt e passando ai livelli successivi con le classi della gerarchia inglese, perfettamente corrispondente alla gerarchia dei synset italiani. L'esperimento è stato implementato lavorando sui file di testo estratti da LexIt nel formato sopra desctitto (verbo-frame-frequenza) e i dati di output per esprimere le preferenze di selezione sono stati mantenuti nello stesso formato (verbo-frame-synset-frequenza), sostituendo al filler nominale il synset della classe corrispondente.

Nonostante questo lavoro possa considerarsi come un esperimento autonomo, indipendente da LexIt, qui utilizzato soltanto come risorsa lessicale-semantica per la distribuzione del comportamento dei verbi italiani, in un'ottica più ampia l'obiettivo e l'evoluzione naturale del programma sono quelli di integrarsi con le funzionalità di LexIt.

Un'interrogazione diretta del database di LexIt e la possibilità di arricchirlo con tutte le informazioni fornite dall'analisi delle preferenze di selezione su tutti i livelli della rete semantica, sono da considerarsi i possibili futuri sviluppi.

Bibliografia

- Artale et al. ,*WordNet for Italian and Its Use for Lexical Discrimination*, Atti del Quinto Congresso della Associazione Italiana per l'Intelligenza Artificiale, Roma 1997.

- Artale et al., *Lexical Discrimination with the Italian Version of WordNet*, Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid;

- Baroni M. et al., *Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian*, Proceedings of LREC 2004, 1771–1774, Lisboa, 2004.

- Bentivogli L. et al., *Multiwordnet: developing an aligned multilingual database*, First International Conference on Global WordNet, number 1996, Mysore, India, 2002.

- Bentivogli L., E. Pianta, *Looking for lexical gaps*, Proceedings of the ninth EURALEX International Congress, Stuttgart, 2000.

- Bentivogli L., *Relazioni lessicali e semantiche nella costruzione di un lessico computazionale multilingue: problematiche tecniche e filosofiche*, Bologna 1998.

- Blumenthal P. e Rovere G., *Wörterbuch der italienischen Verben*, Ernest Klettverlag, Stuttgart, 1998

- Bosco C. et al., *Evalita'09 Parsing Task: comparing dependency parsers and treebanks*, Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, Reggio Emilia, 2009.

- Bosque I., *Redes: diccionario combinatorio del espanol contemporáneo*, SM Ediciones, Madrid, 2004.

- Chomsky N., *Aspects of the Theory of Syntax*, The MIT press, Cambridge, MA, 1965.

- Church K. and Hanks P., *Word association norms, mutual information, and lexicography*, Computational linguistics, 16(1):22–29, 1990

- Clark S. e Weir D., *Class-Based Probability Estimation using a Semantic Hierarchy*, Computational Linguistics, 28, 2002.

- Dell'Orletta F. et al., *Maximum Entropy for Italian PoS Tagging*, Intelligenza Artificiale, IV(2):10–11, 2007.

- Evert S., *Corpora and collocations*, Corpus Linguistics. An International Handbook, Mouton de Gruyter, Berlin, 2008.

- Francis W. e Kučera H., *Frequency analysis of English usage*, New York 1982.

- Gamallo P., Agustini A., Lopes G., *Clustering syntactic positions with similar semantic requirements*, Computational Linguistics, 31(1):107–146, 2005.

- Grefenstette G., *Explorations in automatic thesaurus discovery*, Springer, 1994;

- Hamp B. e Feldweg H., *GermaNet: a Lexical Semantic Net for German*, Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, 1997

- Katz J.J. e Fodor J.A., *The structure of a semantic theory*, Language, 39(2):170–210, 1963.

- Kullback S. e Leibler R.A., *On information and sufficiency*, Annals of Mathematical Statistics, 22, 79-86, 1951; T.M. Cover, J.A. Thomas, *Elements of information theory*. New York, Wiley, 1991.

- Lenci A. , *SIMPLE: A general framework for the development of multilingual lexicons*, International Journal of Lexicography, 22(4):489–495, 2000.

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

- Light M. and W. Greiff, *Statistical models for the induction and use of selectional preferences*,Cognitive Science, 87:1–13, 2002.

- Light M. e Greiff W., *Statistical Models for the Induction and Use of Selectional Preferences*, Cognitive Science: 26.269–281

- MacWhinney B. e Snow C., *The child language data exchange system*, Journal of Child Language, 12, 1985.

- Magnini B. e Strapparava C., *Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet*, Atti del XXVIII Congresso della Società di Linguistica Italiana, 415-418, Palermo, 1994;

- Magnini B. e C. Strapparava, *Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet*, Atti del XXVIII Congresso della Società di Linguistica Italiana, 415-418, Palermo, 1994

- Manning D. e Schütze H., *Foundations of Statistical Language Processing*, Cambridge Mass.: MIT Press 1999.

- Miller G. A. et al., *WordNet: An online lexical database*, Int. J. Lexicograph. 3, 4, 235–244, 1990;

- Fellbaum, Christiane, *WordNet and wordnets*, Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670, 2005.

- Pianta E. et al., *MultiWordNet: developing an aligned multilingual database*, Proceedings of the First International Conference on Global WordNet, Mysore, India, 2002.

- Resnik P., *Selection and information: a class-based approach to lexical relationships*, Phd dissertation, University of Pennsylvania, 1993.

- Resnik P., *Selectional constraints: an information-theoretic model and its computational realization*, Cognition 61, 127-159, 1996.

- Resnik P., *Selectional preference and sense disambiguation*, Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How, 52–57, 1997.

- Resnik P., *Semantic classes and syntactic ambiguity*, in Proceedings of the workshop on Human Language Technology - HLT '93, 278–283, Morristown, 1993.

- Ribas F., *On Learning More Appropriate Selectional Restrictions*, Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, Dublin, Ireland 1995;

- Roland D. and Jurafsky D., *How verb subcategorization frequencies are affected by corpus choice*, Proceedings of the 36th annual meeting on Association for Computational Linguistics, 1122–1128, 1998.

- Schulte Im Walde S., *Experiments on the Automatic Induction of German Semantic Verb Classes*, Computational Linguistics, 32(2):159–194, 2006.

- Schulte Im Walde S., *GermaNet Synsets as Selectional Preferences in Semantic Verb Clustering*, Proceedings of the GermaNet Workshop, 2003.

- Vossen P., *Categories and classifications in EuroWordNet*, Proceedings of the First International Conference on Language Resources and Evaluation. Granada, 399-407, 1998; *Special Issue on EuroWordNet*, Computer and the humanities, 2-3, 73-251, 1998.

- Wagner A., *Enriching a Lexical Semantic Net with Selectional Preferences by Means of Statistical Corpus Analysis*, Proceedings of the ECAI Workshop on Ontology Learning, 37–42, Berlin, Germany 2000.