2013 ACTA UNIVERSITATIS CAROLINAE PAG. 91–108
PHILOLOGICA 2 / ROMANISTICA PRAGENSIA XIX

ANALISI E CLASSIFICAZIONE AUTOMATICA DEI VERBI ITALIANI: UNO STUDIO SUL CORPUS "LA REPUBBLICA"

DIANA PEPPOLONI
Università per gli Stranieri di Perugia

AN ANALYSIS AND AUTOMATIC CLASSIFICATION OF ITALIAN VERBS: A STUDY BASED ON THE "LA REPUBBLICA" CORPUS

This article is concerned with experiments on the automatic induction of Italian semantic verb classes using k-Means, a standard clustering technique, for the purpose of verifying the plausibility of finding a direct connection between the meaning-bearing components of a verb and its syntactic behaviour. A theoretical foundation has been established in extensive works on semantic verb classes such as Levin (1993) for English and Schulte im Walde (2002, 2003, 2004, 2006) for German: each verb class contains verbs which are similar in their meaning and in their syntactic properties.
Basing our work on this hypothesis, we have conducted a study of the "La Repubblica" corpus, one of the leading corpora freely available for the Italian language, to subsequently obtain an automatic classification of a sample of Italian verbs.
Using probability distributions over verb subcategorisation frames, we obtained an intuitively plausible clustering of 200 verbs into 40, 24, and 10 classes. The automatic clustering was evaluated against independently motivated, hand-constructed semantic verb classes. A series of post-hoc cluster analysis explored the influence of specific frames and frame groups on the coherence of the verb classes, and supported the validity of the syntactic-semantic hypothesis.

Keywords: syntactic-semantic hypothesis – clustering – automatic classification –subcategorization frames – written Italian corpus
Parole chiave: ipotesi sintattico-semantica – clustering – classificazione automatica – frames di sottocategorizzazione – corpus dell'italiano scritto

1. Introduzione

La classificazione del verbo costituisce uno dei problemi più dibattuti nella costruzione di un modello lessicologico che intenda rispondere a domande come: cos'è che permette ad un determinato verbo di comparire in certi tipi di costruzioni, ad esempio transitiva o intransitiva, mentre impedisce che ciò accada nel caso di altri verbi? Quale informazione semantica e argomentale è espressa in un'entrata verbale? Secondo quali modalità essa condiziona o addirittura determina il comportamento sintattico? E viceversa, qual è il ruo-

lo del contesto nella definizione delle proprietà semantiche dei verbi? Come è opportuno procedere per isolare le variabili composizionali del significato (Jezek (2003))?

Per identificare classi di parole, si è fatto solitamente riferimento, in modo alternativo, a due versanti: quello sintattico da un lato, e quello semantico dall'altro. Ad esempio, considerati i verbi *aggiustare*, *arrivare* e *partire* si distinguono sintatticamente una classe transitiva, in cui inserire il verbo *aggiustare*, ed una intransitiva, in cui collocare *arrivare* e *partire*; mentre dati i verbi *correre*, *camminare* e *sostare* è possibile distinguere da un punto di vista semantico-aspettuale una classe di moto per *camminare* e *correre*, ed una di verbi stativi per *sostare*.

Seguendo la tesi per cui gli elementi del significato determinano la realizzazione sintattica dei verbi, sarà allora anche vero che verbi con significato analogo mostreranno le stesse caratteristiche sintattiche. A livello descrittivo questa metodologia ha prodotto ottimi risultati, ma sul piano esplicativo ha rivelato forti incongruenze, dovute alla presenza di numerose parole omogenee dal punto di vista semantico, ma non altrettanto per ciò che riguarda il comportamento sintattico e viceversa.

In base a quanto detto finora, emerge chiaramente come la classificazione verbale sia un'operazione estremamente complessa da realizzare, per tutta una serie di ragioni.

Il presente articolo, che prende spunto dalla mia tesi di dottorato in cui è possibile trovare tutto il materiale utilizzato e gli esperimenti svolti e commentati per esteso, illustra la realizzazione di una classificazione automatica per un campione significativo di verbi dell'italiano. Lo scopo non è meramente quello di confrontarne la validità e l'attendibilità confrontandola con una classificazione manuale precedentemente sviluppata degli stessi, quanto quello di verificare se sia possibile individuare una qualche correlazione tra: il comportamento sintattico dei verbi e le loro proprietà del significato.

Nel primo paragrafo illustriamo l'ipotesi teorica che fa da sfondo agli esperimenti condotti, vale a dire la cosiddetta ipotesi sintattico-semantica, che sviluppa il concetto per cui comportamento sintattico di un verbo, specie per ciò che riguarda l'espressione e l'interpretazione dei suoi argomenti, è largamente determinato dal suo significato (Korhonen (2002)); perciò tale comportamento può essere utilizzato per individuare aspetti linguisticamente rilevanti del significato del verbo stesso. Resta da valutare, ed è quello che si cercherà di fare nel corso dell'articolo, fino a che punto le proprietà di selezione sintattica dei verbi siano riconducibili a particolari dimensioni del loro significato.

Successivamente viene proposta una panoramica d'insieme su una serie di esperimenti sviluppati per il tedesco da Schulte im Walde, che interpretano questa corrispondenza biplanare tra sintassi e semantica in chiave computazionale: usando cioè applicazioni informatiche per ottenere descrizioni automatiche del significato dei verbi considerati.

Si prosegue con la descrizione della vera e propria indagine condotta sui verbi italiani. A partire dalla selezione di un campione di 200 verbi italiani ad alta frequenza, estratti dal corpus di *La Repubblica*, è stata elaborata manualmente una lista ontologica di 40 classi verbali, sul modello di quelle proposte da Levin (1993), (2005) e Schulte im Walde (2002), (2003), (2004), (2006), nei rispettivi lavori. Successivamente si è passati all'analisi del corpus di Repubblica, utilizzando un *parser* sintattico a dipendenze, così da ottenere automaticamente la distribuzione statistica degli schemi di sottocategorizzazione sintattica dei verbi esaminati, nonché delle loro preferenze di selezione. Una volta in possesso delle distribuzioni di frequenza dei verbi rispetto ai frames di sottocategorizzazione e ai fillers nominali, i dati sono

stati sottoposti ad un processo di classificazione automatica, tramite l'uso di algoritmi di clustering. È a questo punto opportuno dedicare un po' di spazio alla descrizione del corpus utilizzato nel corso della nostra analisi. Questo strumento è una raccolta di tutti gli articoli del quotidiano omonimo, compresi tra il 1985 e il 2000, per un totale di circa 326 milioni di *tokens* annotati morfosintatticamente e lemmatizzati con metodi semi-automatici (Baroni et al. (2004)). Sebbene proveniente da un'unica fonte e appartenente esclusivamente all'ambito giornalistico, perciò sbilanciato, il materiale prodotto è stato ritenuto comunque una base adeguata e promettente per lo sviluppo della presente ricerca. Coprendo un arco di 16 anni, tale *corpus*, liberamente consultabile[1], fornisce uno strumento per lo studio sia diacronico che sincronico dell'italiano contemporaneo, nonché per ulteriori studi contrastivi, visto che sono già disponibili *corpora* di testi giornalistici per molte altre lingue.

L'ultimo paragrafo è stato dedicato alla comparazione dei dati ottenuti e degli esperimenti svolti. Il confronto tra le due classificazioni, una sintattica, estratta in modo automatico in base agli argomenti manifestati dai verbi e al contesto distribuzionale dei nomi con cui questi si accompagnano, l'altra semantica, sviluppata manualmente e a priori dal ricercatore a partire dai tratti semantici espressi, ha fornito:

– un indice di corrispondenza tra i dati ottenuti;
– indicazioni sui meccanismi di funzionamento del sistema verbale italiano.

Dagli esperimenti condotti emerge che la componente sintattica riesce solo parzialmente, e solo per determinate categorie verbali distribuzionalmente simili da questo punto di vista, a catturare le affinità semantiche che accomunano i verbi appartenenti ad una stessa classe.

Il limite linguistico di questo tipo di esperimenti, risiede nella scelta delle proprietà verbali descrittive ritenute maggiormente pertinenti, siano esse sintattiche (frames di sottocategorizzazione) o semantiche (fillers nominali). Il significato di un verbo comprende infatti, tanto le caratteristiche generali che ne determinano l'appartenenza ad una classe di riferimento, quanto quelle specifiche che lo distinguono dagli altri membri con cui condivide tale appartenenza. Se da un punto di vista teorico la distinzione è chiara, all'atto pratico la scelta delle proprietà verbali dipende dallo schema di definizione delle classi che si vogliono rappresentare, e questo varia necessariamente in funzione del tipo di coerenza semantica catturato dalla classificazione.

In ogni caso la combinazione delle proprietà selezionate all'interno del presente articolo, sembra un punto di partenza perfettibile, ma promettente, per la descrizione dei verbi italiani.

2. L'ipotesi di Levin e il problema della classificazione verbale

La ricerca linguistica ha dimostrato come i verbi appartengano a classi diverse a seconda delle rispettive proprietà semantiche e sintattiche (Levin (1993), Pinker (1989)). Ad esempio, verbi che condividono la componente di significato del movimento, come *camminare* o *correre*, riveleranno affinità anche per ciò che riguarda i processi sintattici di sottocategorizzazione, raggruppandosi così all'interno di una stessa classe linguisticamente coerente.

[1] Il *corpus* è consultabile al sito http://www.sslmit.unibo.it/repubblica.

L'assunto che sta a fondamento di questa tesi è che il comportamento sintattico di un verbo, specie per ciò che riguarda l'espressione e l'interpretazione dei suoi argomenti, è largamente determinato dal suo significato (Korhonen (2002)). Perciò tale comportamento può essere utilizzato per individuare aspetti linguisticamente rilevanti del significato del verbo stesso; questa idea prende il nome di *ipotesi semantico-sintattica*.

La classificazione verbale maggiormente utilizzata per approfondire la riflessione linguistica su questo tema e per successive applicazioni (traduzione automatica, *word sense disambiguation*, acquisizione lessicale etc.) è quella proposta da Levin (1993, 2005); quest'ultima si avvale delle informazioni distribuzionali ottenute osservando il comportamento semantico dei verbi considerati, per ricavarne solo successivamente una classificazione sintattica. In altre parole, Levin si preoccupa di capire come gli elementi rilevanti del significato dei verbi si proiettino e si realizzino poi nei diversi *patterns* sintattici, direttamente osservabili nell'uso linguistico dei parlanti, dei verbi stessi.

Le classi di Levin si basano sulla possibilità del verbo di ricorrere in diverse alternanze diatetiche, ovvero in specifiche coppie di *frames* sintattici.

I parlanti nativi sono in grado di produrre giudizi raffinati sull'occorrenza dei verbi con un insieme di possibili combinazioni di argomenti e aggiunti nelle diverse espressioni sintattiche. Ad esempio i parlanti nativi inglesi sanno a quali alternanze diatetiche, ovvero a quali alternanze nell'espressione degli argomenti a cui possono saltuariamente accompagnarsi anche dei cambiamenti nel significato, i verbi possono partecipare. Un esempio di alternanza diatetica in inglese, di cui i parlanti hanno coscienza, è quella associata ai verbi *spray* e *load*, che possono esprimere i propri argomenti in due modi differenti, dando forma alla cosiddetta alternanza locativa.

(1) a. Sharon sprayed water on the plants.
b. Sharon sprayed the plants with water.
(2) a. The farmer loaded apples into the cart.
b. The farmer loaded the cart with apples.

Ma gli stessi parlanti saranno altrettanto consapevoli del fatto che, verbi come *fill* e *cover*, strettamente correlati ai due sopraelencati, non ammettono che una singola opzione.

(3) a. Monica covered a blanket over the baby.*
b. Monica covered the baby with a blanket.
(4) a. Carla filled lemonade into the pitcher.*
b. Carla filled the pitcher with lemonade.

Inoltre i parlanti concordano nei loro giudizi riguardanti le sottili differenze di significato associate alle espressioni alternanti degli argomenti verbali. La capacità di emettere tali giudizi si estende fino alle combinazioni tra argomenti ed aggiunti; ad esempio i parlanti inglesi sanno che le frasi benefattive vengono solitamente introdotte dalla preposizione *for*, ma possono essere realizzate anche come primo oggetto in una costruzione frasale a doppio oggetto.

(5) a. Martha carved a toy out of wood for the baby.
b. Martha carved the baby a toy out of wood.

Naturalmente i parlanti sanno anche altrettanto bene quando queste alternative non sono ammissibili a livello linguistico.

L'idea alla base del lavoro di Levin è quella per cui se le proprietà sintattiche di un verbo sono deducibili in gran parte a partire dal suo significato, allora dovrebbe essere possibile identificare i principi generali che derivano dal comportamento di un verbo, proprio basandosi sull'analisi del significato che esso trasmette. Pertanto se ammettiamo che ci sia correlazione tra significato del verbo e comportamento sintattico dello stesso, allora alcune delle sue proprietà non saranno più contenute nell'entrata lessicale, ma saranno invece predicibili appunto dal significato.

Se i diversi comportamenti delle classi verbali rispetto alle alternanze diatetiche riscontrate derivano dal significato dei verbi stessi, allora ogni classe verbale le cui alternanze diatetiche corrispondono dovrebbero costituire una classe semanticamente coerente. Una volta che tale classe venga identificata, si possono esaminare i suoi membri per isolare le componenti del significato comuni. Questa tecnica di indagine è importante perché permette di analizzare il significato verbale, che è una componente impalpabile rispetto all'elemento sintattico che viene invece esplicitamente realizzato, senza basarsi sulla sola introspezione.

Le classi verbali proposte originano perché un insieme di verbi con uno o più componenti di significato condivisi mostra un comportamento sintattico affine. Alcuni componenti del significato attengono trasversalmente a più classi, così come molte proprietà sintattiche sono comuni a numerose classi verbali. Risulta evidente quindi, che l'assunto teorico fondamentale del lavoro di Levin sia in definitiva la nozione di componente del significato, e non quella di classe verbale, che ne è piuttosto una diretta conseguenza. Pertanto sarà l'identificazione di tali elementi del significato la vera sfida su cui lavorare in futuro.

3. Modelli computazionali: il lavoro di Schulte im Walde

Sabine Schulte im Walde riprende il quadro teorico proposto da Levin e lo applica a tecniche di indagine computazionali; la sua classificazione dei verbi tedeschi si prefigge due obiettivi principali:

1. utilizzare empiricamente ed investigare la relazione prestabilita tra significato del verbo e comportamento sintattico dello stesso;
2. indagare i necessari parametri tecnici sottostanti qualsiasi analisi basata su tecniche di *clustering*.

Schulte im Walde sostiene l'ipotesi per cui esisterebbe una stretta connessione tra il significato lessicale di un verbo ed il suo comportamento, e sfrutta questa relazione tra significato e comportamento per indurre una classificazione automatica sviluppata sulla base dei tratti sintattici descrittivi di un verbo (evidentemente più facili da ottenere che non quelli semantici), verificando successivamente se tale classificazione distribuzionale finirà col coincidere con quella semantica. Ci si aspetta pertanto che i verbi che appartengono alla stessa classe semantica, sovrappongano il proprio comportamento riguardo alle alternanze (Schulte im Walde (2004), (2006)), ovvero a quelle costruzioni alternative a livello di interfaccia semantico-sintattica, capaci di esprimere gli stessi concetti, o simili, per uno stesso verbo.

Per modellare il comportamento del verbo rispetto all'alternanza verbale con strumenti automatici, è stato elaborato il modello statistico di una grammatica per il tedesco, che fornisce informazione lessicale empirica specializzata, ma non unicamente ristretta, sui *frames* di sottocategorizzazione dei verbi (Schulte im Walde (2002), (2003)). Tale modello grammaticale descrive i verbi considerati a tre livelli diversi dell'interfaccia sintattico-semantica:

– le strutture sintattiche (rilevanti nel catturare le funzioni degli argomenti);
– le preposizioni (determinanti nel distinguere, ad esempio, tra locazione e destinazione);
– le preferenze di selezione (assegnazione dei tipi semantici) degli argomenti.

Ogni livello incrementa l'informazione prodotta da quello precedente; come si può rilevare, tale lavoro di rifinitura dell'informazione ottenuta comincia da una pura definizione sintattica e aggiunge via via informazione semantica. Alla descrizione sintattico-semantica affinata dei verbi tedeschi ottenuta come appena descritto, viene applicato un algoritmo di *clustering k-Means* (Forgy (1965)), in modo da indurre automaticamente una classificazione semantica dei verbi.

L'algoritmo *k-Means* permette di suddividere gruppi di oggetti in *K* partizioni sulla base dei loro attributi; si assume che gli attributi degli oggetti possano essere rappresentati come vettori e che quindi formino uno *spazio vettoriale*[2]. Di seguito vengono proposti degli esempi di *clusters* ricavati dall'applicazione della metodologia appena descritta; per ciascuno di essi, i verbi che appartengono alla stessa classe *gold standard* vengono presentati disposti su una riga, accompagnati dall'etichetta relativa alla classe di appartenenza:

(a) nieseln regnen schneien – *Weather*
(b) dämmern – *Weather*
(c) beginnen enden – *Aspect*
bestehen$_2$ existieren – *Existence*
liegen sitzen stehen – *Position*
laufen – *Manner of Motion: Locomotion*

Andiamo ora a vedere invece come Schulte im Walde ha elaborato la propria classificazione manuale. Sono stati selezionati 168 verbi tedeschi poi raggruppati manualmente all'interno di 43 classi semantiche, per valutare, attraverso la comparazione, la validità tà nell'esecuzione delle tecniche di *clustering* utilizzate. Tale classificazione manuale si basa sulla pura intuizione semantica del parlante, e prende spunto da lessici pre-esistenti

[2] In matematica, lo spazio vettoriale (chiamato anche spazio lineare) è una struttura algebrica di grande importanza. Si tratta di una generalizzazione dell'insieme formato da tutti i vettori del piano cartesiano ordinario o dello spazio tridimensionale dotato di un'origine. Si dice vettore una qualsiasi grandezza rappresentabile con un segmento orientato di retta o uno ad esso equipollente; esso è definibile quindi attraverso tre parametri:

1. modulo o intensità	= lunghezza del segmento
2. direzione	= retta su cui giace il segmento o una ad essa parallela
3. verso	= orientazione del segmento

compilati a priori, senza tralasciare quelle voci fortemente ambigue che potrebbero mettere fuori strada il meccanismo di *clustering*. Difatti lo scopo prefissato non è quello di ottenere un *clustering* perfetto dei 168 verbi esaminati, bensì quello di investigare sia il potenziale che i limiti della metodologia utilizzata da questo tipo di *clustering*. La classificazione è completata da una descrizione dettagliata di ciascuna classe, strettamente correlata alla *scenes-and-frames semantics* di Fillmore (1977), (1982), utilizzata in ambito computazionale nel progetto *FrameNet* (Johnson et al. (2002)). La definizione della classe in base alla semantica dei *frames* contiene una descrizione in prosa della scena, in cui si rintracciano: il partecipante principale al *frame*, i ruoli modificatori e le varianti del *frame* che descrivono la scena.

La comparazione tra la classificazione manuale e quella automatica, prodotta con l'ausilio di tecniche di *clustering*, ha permesso di chiarire alcuni punti su quale sia effettivamente la natura del rapporto tra significato e comportamento verbali:

- già una descrizione puramente sintattica del verbo permette di classificare con successo quei verbi che concordano con le descrizioni dei rispettivi *frames* sintattici (ad esempio i verbi appartenenti alla classe *support*). Il *clustering* fallisce invece laddove incontriamo verbi semanticamente simili, ma che differiscono nel comportamento sintattico, ed anche per tutti quei verbi che mostrano un comportamento sintattico simile, ma nessuna similarità semantica;
- rifinire l'informazione sintattica del verbo attraverso l'introduzione delle preposizioni, è estremamente efficace in una lingua come il tedesco, per la buona riuscita del *clustering*. Il miglioramento sottolinea il fatto che verbi con significato simile, mostrano anche affinità nell'esprimere specifici complementi preposizionali o comunque modificazioni di tipo più generale (ad esempio le preposizioni direzionali per i verbi di movimento);
- i risultati del *clustering* vengono ulteriormente migliorati dalla definizione delle preferenze di selezione, ma il miglioramento non è così soddisfacente come quello che deriva dall'introduzione delle preposizioni.

Da dove deriva l'impredicibilità della codifica e degli effetti delle proprietà verbali, specie rispetto alle preferenze di selezione? Per rispondere occorre riprendere la distinzione tra proprietà comuni ad una classe verbale, e proprietà specifiche dei singoli verbi ad essa appartenenti. Difatti non tutte le proprietà di tutti i verbi appartenenti ad una certa classe sono simili, tanto che si potrebbe rifinire la descrizione dei tratti indefinitamente. Il significato dei verbi comprende sia le proprietà generali che fanno sì che un certo verbo appartenga ad una determinata classe, sia quelle specifiche che lo distinguono dagli altri membri della stessa classe. Finché definiamo i verbi in base alle caratteristiche comuni che mostrano, il *clustering* funziona correttamente; nel momento in cui introduciamo proprietà specifiche, l'effetto benefico delle prime viene annullato.

4. Le fasi del progetto di classificazione

Gli esperimenti presentati in questo articolo, considerando gli assunti teorici fin qui illustrati, operano un confronto tra un'iniziale classificazione semantica effettuata *a priori*, indipendentemente da misurazioni di tipo distribuzionale, ed un'altra di tipo sintat-

tico ricavata automaticamente a partire da un *corpus* di riferimento. Le eventuali corrispondenze e divergenze tra le due, permetteranno di valutare empiricamente, sulla base di dati statistici, l'influenza delle correlazioni tra comportamento sintattico e proprietà semantiche dei verbi considerati. In questo contesto allora, il *clustering* e la classificazione semantica *a priori* sono essenzialmente da considerarsi come strumenti inseriti all'interno di un'indagine esplorativa più complessa, il cui fine non è riducibile alla valutazione della loro più o meno ampia efficacia. La suddetta analisi mira altresì ad individuare quali proprietà semantiche condividono i verbi che in italiano presentano proprietà sintattiche distribuzionali simili, e in che misura tali proprietà comuni trovano una corrispondenza nelle classi semantiche naturali.

Per definire la classificazione semantica, si è deciso di partire dalla formulazione di una lista ontologica di classi verbali, sul modello di quelle individuate da Levin e Schulte im Walde, che contenessero in totale duecento verbi prototipici, estratti dal *corpus* di La Repubblica.

Di seguito (Tabella 4.1), si propone un estratto esemplificativo di alcune delle classi semantiche individuate.

Tab. 4.1 Esempio delle classi semantiche elaborate a partire dal corpus di Repubblica

ID	Numero Classe	Classe	Verbo
1	1	Aspect	iniziare
1	1	Aspect	cominciare
1	1	Aspect	continuare
1	1	Aspect	finire
1	1	Aspect	terminare
1	1	Aspect	smettere
2	2	Propositional Attitude	sapere
2	2	Propositional Attitude	pensare
2	2	Propositional Attitude	credere
2	2	Propositional Attitude	dubitare
2	2	Propositional Attitude	ritenere
2	2	Propositional Attitude	considerare

Per realizzare la suddetta classificazione semantica, è stato scelto un insieme di 200 verbi, ripartiti in 40 classi semantiche in base alla similarità del loro significato lessicale e concettuale; infine ad ogni classe è stata assegnata una nomenclatura corrispondente ad una certa categoria concettuale. I nomi delle classi verbali italiane sono forniti in inglese e corredati dall'indicazione numerica della classe ed eventuale sottoclasse di appartenenza. La ripartizione in classi semantiche non è altro che un metodo di catalogazione che permette di operare generalizzazioni sui verbi e sulle loro proprietà semantiche, catturandone una larga parte del significato, senza entrare nell'ambito dei dettagli idiosincratici propri di ciascun verbo considerato. La classificazione si basa primaria-

mente sull'intuizione semantica del parlante, piuttosto che sui comportamenti sintattici dei verbi considerati; difatti troviamo dei verbi che pur appartenendo alla stessa classe semantica, mostrano caratteristiche sintattiche differenti:

(6)

<u>*Propositional Attitude*</u>	
sapere (+ obj dir)	Paolo sa la canzone a memoria
dubitare (+ comp_di)	Paolo dubita delle sue capacità
<u>*Communication*</u>	
dire (+obj dir)	Paolo dice il proprio parere
conversare (+comp_con, +comp_di)	Paolo conversa con gli amici
<u>*Basis*</u>	
vertere (+comp_su)	La discussione verte sulla situazione politica
concernere (+obj dir)	Il dibattito concerne la situazione politica

Inoltre, alcuni dei verbi considerati sono ambigui, quindi possono appartenere a più di una delle classi sviluppate (esempio 7); in questo caso è stato necessario che il ricercatore compia inizialmente una scelta, sulla base delle proprietà semantiche condivise dai verbi di una certa classe ad un livello generale.

(7)

piangere: appartiene alla classe semantica *Facial expression*, ma potrebbe essere inserito anche nel gruppo denominato *Emotion*;
spiegare: si trova nella classe definita *Description*, presenta però affinità semantiche anche con i verbi che fanno capo alla classe *Communication*;
consumare e divorare: sono stati inseriti nella classe *Consumption*, ma rientrano semanticamente anche nel concetto di *Elimination*.

Gli identificatori che distinguono le varie classi semantiche, fanno riferimento a due diversi livelli semantici:

1. alcuni appartengono ad un livello più generico, ad esempio *Inference*;
2. altri invece scendono più in profondità, suddividendo i verbi in sottogruppi molto più granulari, ad esempio *Motion (manner), Motion (directed), Motion (cross)*.

La classificazione semantica elaborata in questa sede, si propone di coprire trasversalmente quanti più domini semantici possibili: si va da processi puramente mentali e cognitivi (*Inference*, *Speculation*), ad altri collegati ad esempio a processi corporei (*Facial Expression*) o a fenomeni atmosferici (*Weather*).

Principalmente, l'utilità della classificazione semantica, risiede nel fatto che essa permette di esprimere generalizzazioni sui singoli verbi, in base alle proprietà semantiche da questi mostrate e condivise; dunque essa rappresenta uno strumento pratico per catturare un'ampia fetta della conoscenza verbale, senza dover scendere nella definizione dei tratti distintivi di ciascun verbo. Così, verbi come *finire* ed *iniziare* si trovano nella stessa classe, che prende il nome di *Aspect*, poiché condividono il tratto semantico prevalente dell'aspettualità. Inoltre la classificazione semantica manuale rappresenta una sorta di *gold standard*, che funziona da parametro di valutazione dell'affidabilità

e delle prestazioni dei successivi esperimenti di *clustering* condotti sullo stesso gruppo di verbi.

Dopo aver selezionato i verbi all'interno del *corpus* di Repubblica ed averli ripartiti in classi per formare una prima classificazione semantica su base manuale, si è passati all'analisi del *corpus* stesso, per ottenere in modo automatico la distribuzione statistica dei *patterns* (schemi) di sottocategorizzazione. I dati del *corpus* sono stati elaborati attraverso un *parser* sintattico a dipendenze, addestrato presso l'Istituto di Linguistica Computazionale del CNR di Pisa. Un *parser* è un programma che si occupa di assegnare una descrizione sintattica ad una certa frase. Nel caso specifico del *parser* sintattico a dipendenze, la struttura sintattica è rappresentata attraverso relazioni binarie di dipendenza tra termini lessicali (Tesnière (1959)). Il concetto di dipendenza coinvolge un elemento che funge da testa ed uno che si comporta invece come dipendente; centrali sono dunque i criteri (sia semantici che sintattici) utili a stabilire a quale di queste due categorie appartengono gli elementi coinvolti nella relazione.

Il *parser* sviluppato all'interno di questo lavoro, non si basa su una grammatica formulata *a priori* dal linguista, ma sul risultato di una fase di apprendimento automatico a partire da un *corpus* addestrato per l'italiano, composto da 79.654 parole (4.162 frasi) annotate con informazioni morfosintattiche e dipendenze grammaticali. Il sistema di *parsing* utilizzato, prende il nome di *Maltparser* (Nivre et al. (2006)). A partire da un *corpus* di addestramento esso costruisce un modello probabilistico per l'assegnazione delle operazioni di *parsing*. Ad ogni passo della computazione il sistema sceglie l'operazione di *parsing* più probabile data la parola in *input*, i suoi tratti morfosintattici, il contesto e le relazioni di dipendenza già individuate.

Le dipendenze sintattiche individuate dal *parser*, sono state poi utilizzate per estrarre i *patterns* di sottocategorizzazione dei verbi trattati; principalmente si tratta, oltre ai due *frames* più frequenti ovvero *transitivo* ed *intransitivo*, di *frames* preposizionali, come ad esempio *comp_a*, *comp_di*, *comp_in*, di *frames* preposizionali doppi come *comp_a#comp_con*, ed infine di strutture frasali introdotte sempre da preposizioni, come nel caso di *inf_a*. I risultati ricavati in *output* sono stati elaborati tramite uno *script* effettuato in *Perl*. I dati così trattati contengono informazioni sui verbi selezionati, corredate da varie indicazioni sulla loro frequenza nel *corpus*, dai lemmi con cui sono attestati nello stesso, dai rapporti di dipendenza che li legano, e da altre informazioni addizionali, come la scelta degli ausiliari cui si accompagnano.

Una volta in possesso delle distribuzioni di frequenza dei verbi rispetto ai *frames* di sottocategorizzazione, si possono sottoporre i dati al processo di classificazione automatica, operazione effettuata tramite un algoritmo di *clustering* capace di raggruppare una lista di elementi in classi, in base al loro grado di similarità all'interno dello spazio vettoriale in cui l'algoritmo stesso li colloca. Perciò vettori-parola simili saranno inseriti in uno stesso gruppo, mentre quelli dissimili andranno a collocarsi logicamente in gruppi diversi.

Tutte le tecniche di *clustering* si basano sul concetto di distanza tra due elementi, detti vettori. Se la distanza risulta essere un concetto fondamentale, l'appartenenza o meno ad uno stesso *cluster* dipenderà strettamente da quanto l'elemento esaminato è appunto distante dall'insieme stesso; più questo sarà vicino e più sarà considerato simile, mentre il caso inverso sarà indice di dissimilarità (Manning e Schütze (1999)).

Una volta che si dispone di un insieme di vettori-parola e che si è misurata la distanza tra di essi, si possono raggruppare tali vettori in *clusters* che risultano simili in base alle rappresentazioni sintattiche che contengono (Lenci e Calzolari (2004)).

Il *clustering* ci permette, all'interno di uno spazio vettoriale, di individuare:
- gli elementi verbali in esso presenti, raggruppati in categorie;
- le varie classi semantiche attorno a cui si costruisce una certa regione dello spazio vettoriale (Widdows (2004)).

Non assegnando preventivamente alcuna nomenclatura ai dati, il *clustering* originerà una classificazione automatica e non controllata *a priori*, poiché i risultati dipendono esclusivamente dalle divisioni naturali assunte dai dati (Manning e Schütze (1999)); il ricercatore condurrà poi la sua analisi, basandola sul confronto tra tali *clusters* ed una eventuale classificazione da lui precedentemente stabilita.

Gli esperimenti di *clustering* sono stati realizzati tramite un programma liberamente disponibile in rete che prende il nome di *Cluto*. Cluto è un *package* statistico per la creazione di *clusters* e l'analisi dei vari gruppi di dati ottenuti. Esso fornisce tre diverse classi di algoritmi di *clustering*, che operano sia direttamente sullo spazio delle caratteristiche dell'oggetto, sia sullo spazio di similarità dell'oggetto (Karypis (2003)).

5. Analisi degli esperimenti svolti

Gli esperimenti di classificazione svolti sui dati a nostra disposizione, sono stati condotti utilizzando l'algoritmo predefinito di *clustering* di Cluto, il *repeated bisection*, per il quale si rende necessario, come per il *k-means*, specificare il numero *k* di *clusters* di partenza[3]. Nei nostri esperimenti a tale numero corrisponde quello delle classi verbali nelle quali saranno ripartiti gli elementi del *clustering*, ovvero i 200 verbi selezionati *a priori*. Sono state effettuate tre diverse tipologie di ripartizioni, variando il numero di classi semantiche e dunque di *clusters*:

1. la prima prevede 40 classi in uscita;
2. la seconda 24;
3. infine la terza soltanto 10.

Queste tre diverse classificazioni intrattengono tra di loro rapporti di tipo gerarchico: vale a dire che le 40 classi sono sottoclassi delle 24, che a loro volta sono sottoclassi delle 10. La suddivisione in 40 classi risulta molto più precisa e granulare tanto che le classi vengono anche ripartite in sottogruppi, mentre quelle in 24 e 10 sono senz'altro più generali. Di seguito riportiamo degli esempi di quanto appena detto:

(8) La classe *Transfer of Possession*, che nella ripartizione in 24 classi compare come un unico gruppo di verbi, viene ulteriormente scomposta nella divisione in 40 classi, dando origine ai seguenti sottogruppi: *Transfer of Possession (obtaining)*, *Transfer of Possession (giving-gift)*, *Transfer of Possession (giving-supply)*.

[3] Per eventuali dettagli sul funzionamento dell'algoritmo utilizzato in Cluto, si rimanda al manuale del *software* contenuto in Karypis (2003).

Nella divisione in 10 classi troviamo delle macro-categorie, come nel caso di *Cognition*, che racchiudono molte delle classi che nelle altre due ripartizioni compaiono separatamente: *Communication, Perception, Propositional Attitude, Moaning, Emotion*.

Un altro parametro variabile nello sviluppo degli esperimenti svolti, è stato quello della selezione del numero di *frames* di sottocategorizzazione utilizzati, rappresentativi delle strutture sintattiche dei verbi considerati. Tutte e tre le tipologie di classificazione automatica sopra elencate, hanno operato su un gruppo iniziale di 105 *frames*, poi su un altro più ridotto di 50 ed infine sull'ultimo di soli 25. I *frames* di sottocategorizzazione sono stati selezionati in base alla loro frequenza globale all'interno del *corpus* parsato di Repubblica.

Per ciò che riguarda l'utilizzo dei suddetti *frames*, si è scelto inoltre di non considerare l'indice della frequenza relativa con cui un determinato *frame* ricorre con un certo verbo, quanto piuttosto il logaritmo di tale frequenza, allo scopo di evitare l'influsso negativo dei verbi ad alta frequenza ed ottenere così dati più significativi per l'analisi linguistica. In effetti nel primo caso Cluto era in grado di accorpare solo macro-gruppi di verbi, che corrispondevano sostanzialmente alla grande divisione tra transitivi ed intransitivi, poiché questi sono in assoluto i *frames* più frequenti. Questo perché i *frames* di sottocategorizzazione seguono una *distribuzione zipfiana* rispetto ai verbi cui si riferiscono: quindi avremo un insieme estremamente ridotto di *frames* molto frequenti (essenzialmente *transitivo* ed *intransitivo*), ed invece un gruppo molto numeroso di *frames* poco frequenti. Se si ricorre semplicemente alla frequenza come indice statistico, i primi finiscono per dominare, mentre l'uso del logaritmo permette di livellare il divario tra le varie frequenze. La suddivisione iniziale era assolutamente troppo generica per poter operare qualunque tipo di riflessione su di essa; grazie invece all'introduzione del logaritmo della frequenza, la classificazione è risultata significativamente più raffinata, granulare e quindi più utile per comprendere quanto i componenti sintattici siano in grado di riflettere il significato dei verbi e quali tra questi sono più significativi in questo processo.

Il variare del numero delle classi considerate e dei *frames* sintattici utilizzati, ci permette di valutare quanto un'analisi computazionale basata esclusivamente su parametri sintattici è o meno in grado di catturare il significato dei verbi e, conseguentemente, di raggrupparli in *clusters* semanticamente coerenti, ovvero di verificare il grado di validità e riscontro dell'ipotesi sintattico-semantica.

Successivamente, oltre ai *frames* di sottocategorizzazione, rappresentativi delle strutture argomentali dei verbi esaminati, è stato inserito negli esperimenti anche un altro livello di analisi, quello dei *fillers* nominali, ovvero dei tipi semantici partecipanti alle diverse costruzioni argomentali di ciascun verbo. Laddove sintassi e preposizioni non sembrano sufficienti a catturare nella sua interezza la complessità della struttura argomentale dei verbi, questo nuovo parametro permette di ottenere risultati sicuramente più apprezzabili ed efficaci. L'ipotesi di fondo è che i nomi siano associati ad un tipo semantico, che caratterizza la loro dimensione concettuale; i nomi portano informazioni sugli eventi e sulle situazioni in cui sono tipicamente coinvolti. I verbi utilizzano tali informazioni, in base alle loro restrizioni e preferenze di selezione. Scopo di questo tipo di indagine è quello di esaminare le interazioni tra i tipi semantici e i processi

combinatori, attraverso l'analisi della distribuzione delle parole nei vari contesti d'uso. Il contesto d'uso è formato dai nomi con cui i verbi si accompagnano, in una certa posizione argomentale.

Ogni verbo presenta delle preferenze di selezione nel realizzare i propri argomenti, ovvero restrizioni semantiche operate da una certa parola all'interno dell'ambiente sintagmatico in cui si colloca (Brockmann e Lapata (2003)).

(9) un verbo come *mangiare* selezionerà tipicamente:
– entità animate nel ruolo di *soggetto*;
– entità commestibili in quello di *oggetto*.

Le preferenze di selezione di un verbo sono individuabili con maggiore evidenza laddove tali restrizioni vengono violate, piuttosto che nei casi in cui sono al contrario assecondate.

(10) La *montagna* mangia *sincerità*

In questo enunciato, sia la restrizione semantica valevole per il subj, che quella riferita all'obj dir sono state contravvenute, difatti il significato di cui esso è portatore non può essere accettato dai parlanti.

Se analizziamo, ad esempio, la tipologia di nomi che più spesso si accompagnano al verbo *mangiare*, vedremo che tipicamente l'argomento *obj dir* sarà realizzato da parole come *cibo*, *pasto*, *cena*, *pranzo* piuttosto che da altre come *fiume*, *montagna* o *luna*, che risultano pertanto improprie in quel contesto. Il significato lessicale di un verbo può essere pertanto rappresentato a livello contestuale, ovvero tenendo conto degli incontri ripetuti che esso intrattiene con determinate parole in vari contesti d'uso.

A questo punto, si è trattato di dare una rappresentazione semantica delle proprietà di selezione dei *frames* dei verbi considerati. Queste dipendono essenzialmente dai nomi che possono comparire in un certo *frame*; pertanto ogni verbo è stato rappresentato come un vettore, le cui dimensioni descrivono la distribuzione statistica dei possibili argomenti nominali. Dunque due verbi saranno tanto più simili, quanto più tenderanno ad avere argomenti nominali simili. Questa è una prima approssimazione che si è cercato di dare, della nozione di preferenze di selezione. Il modello proposto nel presente studio, è stato costruito usando una matrice (Figura 5.1) che ha come righe i verbi considerati, e come colonne non più i *frames* sintattici di sottocategorizzazione, bensì i *fillers* nominali dei verbi stessi. Si tratta, più precisamente, dei 3000 nomi più frequenti che ricorrono come *fillers* dei *frames* estratti dal *corpus* di Repubblica. I valori della matrice esprimono la forza di associazione tra il verbo ed un dato nome, calcolata attraverso la misura associativa detta Simple Log Likelihood.

Di fatto, i nomi inseriti nella matrice, ci forniscono le preferenze di selezione dei verbi selezionati; ovviamente non si tratta dello stesso tipo di preferenze di selezione ottenute da Schulte im Walde, che utilizza le classi ontologiche di WordNet, bensì di preferenze di selezione intese in termini di *lexical sets* selezionati dai verbi stessi. Quello che ne deriva è, dunque, un vero e proprio spazio semantico o *word space*, dove ogni riga è costituita da una singola parola detta anche parola *target*, ed ogni colonna rappresenta un certo

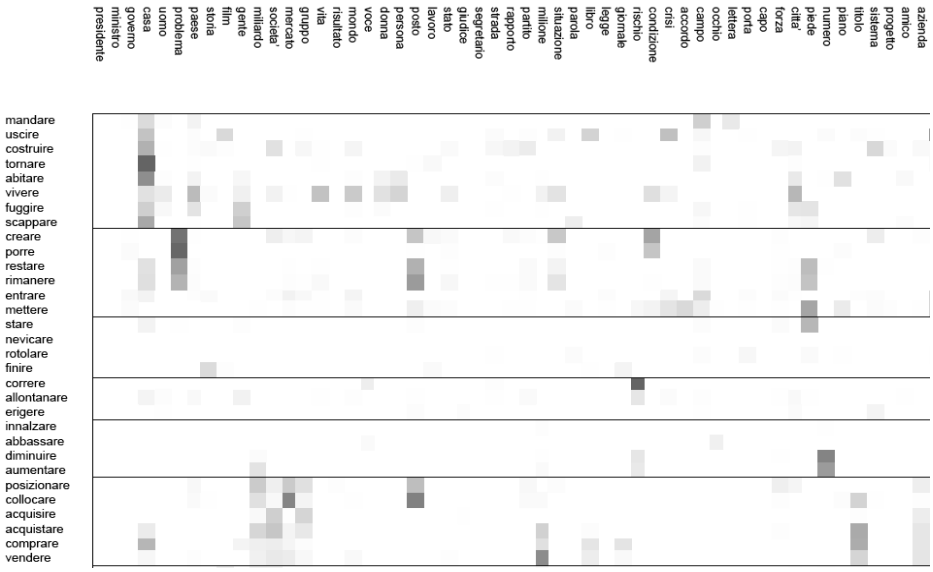

Figura 5.1 Estratto della matrice costruita con i fillers nominali

contesto linguistico. La similarità tra le due parole viene calcolata tramite una misura di associazione e rappresentata nella matrice dal colore rosso, più o meno intenso, del quadrato che simboleggia la loro intersezione.

Da un confronto tra i *clusters* ottenuti sulla base dei *frames* di sottocategorizzazione, e quelli che risultano dall'utilizzo dei *fillers* nominali, emerge come per alcune classi verbali la similarità risulti assai più evidente a livello semantico piuttosto che sintattico. Prendiamo il caso della classe semantica dei *Desire(need) verbs*, contenente i verbi *esigere*, *necessitare* e *richiedere*; negli esperimenti condotti con i *frames* di sottocategorizzazione, questa classe viene mal rappresentata dal programma di *clustering*, tanto che i verbi in essa contenuti sono collocati in tre diversi *clusters*. Cluto non è pertanto in grado di coglierne l'affinità semantica considerando esclusivamente le loro proprietà sintattiche, poiché essi presentano tratti argomentali assai diversi. Così non è, invece, nel caso delle preferenze di selezione, che ci permettono di associare correttamente questi verbi, proprio come farebbero intuitivamente i parlanti dell'italiano, in virtù dei nomi con cui essi prevalentemente compaiono a livello distribuzionale.

(11)
esigere
necessitare } *cluster* 10
richiedere

I nomi con cui più frequentemente figurano sono: *intervento*, *investimento*, *rispetto*, *cura*, *impegno*; questa associazione distribuzionale trova riscontro correttamente in frasi come:

a. Lo studio *esige/necessita/richiede* impegno
b. Il paziente *esige/necessita/richiede* cure efficaci
c. La situazione *esige/necessita/richiede* l'intervento dello Stato

Un altro esempio di classe semantica che trova un corrispettivo distribuzionale più soddisfacente nelle preferenze di selezione, piuttosto che nei *frames* di sottocategorizzazione, è quello dei *Manner of Articulation verbs*. Questa classe contiene i seguenti verbi:

bisbigliare
sussurrare
mormorare } *gridare*
urlare

Nelle operazioni di *clustering* effettuate considerando i *frames* di sottocategorizzazione ad essi associati, i risultati non sembrano soddisfacenti, difatti, anche in questo caso, ogni verbo viene collocato separatamente in un *cluster* differente.

Al contrario, il parametro delle preferenze di selezione consente non solo di rispecchiare più fedelmente la classificazione elaborata *a priori*, ma addirittura di raffinarla ulteriormente, tramite l'individuazione di sottoinsiemi di verbi semanticamente affini tra loro. Vediamo dunque come Cluto ha ripartito i suddetti verbi e quali sono le preferenze di selezione a cui essi si accompagnano più frequentemente:

bisbigliare
sussurrare } *cluster* 0
mormorare
preferenze di selezione: *orecchio, parola, voce, preghiera*
gridare
} *cluster* 8
urlare
preferenze di selezione: *slogan, scandalo, miracolo, rabbia, gente*

I due *clusters* presentano delle evidenti sfumature di significato:

nel *cluster* 0 troviamo tutti quei verbi che implicano un abbassamento del tono di voce, una modulazione e un controllo del messaggio espresso. Per questo motivo tra i sostantivi associati più frequentemente troviamo parole come *preghiera*;

a. Il fedele ha *bisbigliato/sussurrato/mormorato* una preghiera

nel *cluster* 8, invece, figurano i verbi che comportano un innalzamento del tono di voce, un'esasperazione del modo di articolazione nel veicolare il messaggio. Difatti i nomi a cui si accompagnano evocano ben altre situazioni e stati d'animo nei parlanti.

b. Paolo ha *gridato/urlato* di rabbia

Similmente a quanto appena detto per i *Manner of Articulation verbs*, anche i *Motion verbs*, sebbene più numerosi e più sparsi all'interno della classificazione, vengono raggruppati in virtù del tipo di movimento che esprimono. Ad esempio, nel *cluster* 14 troviamo tutti verbi che indicano il modo del movimento (*Motion (manner)*), mentre nel *cluster* 7, tutti quelli che sottendono l'idea di attraversamento (*Motion (cross)*).

(12)
marciare
pedalare
saltellare
camminare } *cluster* 14

preferenze di selezione: ritmo, salita, strada, passo
a. Paolo *marcia/pedala/cammina* in salita
percorrere
attraversare } *cluster* 7
traversare

preferenze di selezione: oceano, fiume, frontiera, fase (senso figurato)
b. Il battello *percorre/attraversa/traversa* il fiume
c. L'adolescente *percorre/attraversa/traversa* una fase travagliata della vita

Infine citiamo il caso di classi semantiche che trovano un corrispettivo esatto nella classificazione automatica effettuata utilizzando come parametro le preferenze di selezione; tra queste citiamo i *Moaning verbs* che, proprio come nella classificazione semantica sviluppata *a priori*, costituiscono un gruppo compatto, inserendosi in un unico *cluster* (esempio 13).
(13)
deplorare
deprecare
lagnare(si)
lamentare(si)
rammaricare(si) } *cluster* 31

preferenze di selezione: *assenza, mancanza, violenza, comportamento, decisione*
a. L'insegnante *deplora* l'assenza dello studente
L'insegnante *depreca* l'assenza dello studente
L'insegnante *si lagna* dell'assenza dello studente
L'insegnante *si lamenta* per l'assenza dello studente
L'insegnante *si rammarica* dell'assenza dello studente

Come evidenziato dall'esempio (24) a., i verbi contenuti in questa classe semantica mostrano una forte affinità di significato, tanto che si associano alle stesse preferenze di selezione; al contrario, essi presentano tratti argomentali differenti, infatti nella classificazione basata sui *frames* di sottocategorizzazione, non vengono accorpati insieme, ma appartengono a *clusters* diversi.

Ovviamente occorre menzionare anche quei *clusters* che, pur appoggiandosi alle preferenze di selezione, risultano comunque imprecisi e confusi; tra questi citiamo il *cluster* 39, al cui interno troviamo verbi come: *ridere, fumare, smettere, sapere*, che appartengono a classi semantiche disparate (*Facial Expression*, *Consumption*, *Aspect* e *Propositional Attitude*).

6. Conclusioni

Gli studi condotti negli ultimi decenni, hanno dimostrato l'utilità dei metodi linguistico-computazionali nell'estrazione di informazioni importanti per vari aspetti del lessico dai *corpora* disponibili. In questo studio si è cercato di cogliere, in particolar modo, le modalità combinatorie delle parole nei loro contesti d'uso linguistico, per studiarne poi di conseguenza le proprietà sintattico-semantiche. Gli strumenti utilizzati sono di due tipi: da un lato l'annotazione dei testi, dall'altro l'analisi statistica delle loro componenti.

L'ipotesi linguistica che soggiace al metodo di analisi del presente studio, è che verbi semanticamente simili di solito presentano simili proprietà di sottocategorizzazione. Il problema è valutare fino a che punto tali proprietà di selezione sintattica dei verbi siano riconducibili a particolari dimensioni del loro significato, cioè fino a che punto le proprietà semantiche di un verbo dipendano dai suoi comportamenti sintattici. Dagli esperimenti condotti si è verificato che la componente sintattica riesce solo parzialmente, e solo per certe categorie di verbi distribuzionalmente simili da questo punto di vista, a catturare le affinità semantiche che accomunano i verbi appartenenti ad una stessa classe. Il passaggio dalla sintassi alla semantica è possibile, poiché il significato lessicale di un verbo può essere inteso come una rappresentazione contestuale; in altre parole: dagli incontri ripetuti di un predicato con un certo termine in vari contesti d'uso, deriva la costruzione di una sua rappresentazione contestuale (*ipotesi distribuzionale*). I parlanti di una lingua sanno riconoscere quanto due parole siano semanticamente affini, e tale intuizione può dipendere da un uso simile di quelle stesse parole, ovvero dalla loro presenza in contesti linguistici simili. In conclusione, la combinazione delle proprietà selezionate all'interno del lavoro negli esperimenti proposti, sembra costituire un punto di partenza promettente per la descrizione dei verbi italiani. Ovviamente è necessario un lavoro di analisi e revisione manuali, condotto successivamente sui dati risultanti.

Le direzioni possibili per le ricerche future sono di vario tipo; in primo luogo, si potrebbe estendere l'elaborazione manuale di classi semantiche dei verbi italiani, al fine di includere una gamma più completa di classi verbali. Si potrebbe poi pensare all'utilizzo di un algoritmo di *clustering* di tipo *soft*, invece che *hard* come è stato fatto nell'ambito di questo studio, poiché esso ha la capacità di assegnare i verbi a più *clusters*, introducendo così un metodo di indagine per il fenomeno dell'ambiguità verbale.

Un altro punto fondamentale da approfondire nelle ricerche successive, è quello dello sviluppo di tecniche più raffinate per esprimere la semantica dei *frames* (ruoli semantici, preferenze di selezione etc.). L'uso di vettori che rappresentano *fillers* nominali, è solo un'iniziale tentativo di approssimazione, che merita e richiede ulteriori approfondimenti.

BIBLIOGRAFIA

Baroni, M. et al. (2004): Introducing the la Repubblica corpus: a large annotated, TEI (XML) compliant corpus of newspaper Italian. *Proceedings of LREC*, 2004, Lisbon: ELDA, pp. 1771–1774.

Brockmann, C. – Lapata, M. (2003): Evaluating and combining approaches to selectional preference acquisition. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest: Hungary, pp. 27–34.

Fillmore, C. J. (1977): Scenes and Frame Semantics, Linguistic Structures Processing. In Zampolli, A. (ed.), *Fundamental Studies in Computer Science, No. 59*. North Holland Publishing, pp. 55–88.
Fillmore, C. J. (1982): Frame Semantics. In *Linguistics in the Morning Calm.* Seoul: Hanshin Publishing Co., pp. 111–137.
Forgy, E. W. (1965): Cluster analysis of multivariate data: efficiency vs interpretability of classifications. In *Biometrics*, num. 21, pp. 768–769.
Jezek, E. (2003): *Classi di verbi tra semantica e sintassi.* Pisa: ETS edizioni.
Johnson, C. R. et al. (2002): *FrameNet : Theory and Practice*. Technical Report-02009. Berkeley, CA: International Computer Science Institute.
Karypis, G. (2003): *CLUTO 2.1.1. A Clustering Toolkit.* Technical Report. Department of Computer Science: University of Minnesota.
Korhonen, A. (2002): *Subcategorisation acquisition*. PhD Thesis published as Technical Report UCAM-CL-TR-530. Computer Laboratory: University of Cambridge.
Lenci, A. – Calzolari, N. (2004): Linguistica computazionale. Strumenti e risorse per il trattamento automatico della lingua. *Mondo Digitale,* vol. 3, num. 2, pp. 56–69.
Levin, B. (1993): *English Verb Classes and Alternations: a Preliminary Investigation.* Chicago: IL, University of Chicago Press.
Levin, B. – Rappaport Hovav, M. (2005): *Argument Realization.* Cambridge University Press.
Manning, C. – Schütze, H. (1999): *Foundations of Statistical Natural Language Processing*. MIT Press.
Nivre, J. – Hall, J. – Nilsson, J. (2006): MaltParser: a Data-Driven Parser-Generator for Dependency Parsing. *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC2006)*. Genoa: Italy, pp. 2216–2219.
Pinker, S. (1989): *Learnability and Cognition: the acquisition of Argument Structure.* Cambridge, MA: MIT Press.
Schulte im Walde, S. (2002): A Subcategorization Lexicon for German Verbs induced from a Lexicalized PCFG. *Proceedings of the 3rd Conference on Language Resources and Evaluation,* Las Palmas de Gran Canaria: Spain.
Schulte im Walde, S. (2003): Experiments on the Choice of Features for Learning Verb Classes. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics,* Budapest: Hungary.
Schulte im Walde, S. (2004): Induction of Semantic Classes for German Verbs. In Langer, S. – Schnorbusch, D. (eds.), *Semantik im Lexicon*. Tübingen: Gunter Narr Verlag.
Schulte im Walde, S. (2006): Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics,* 32(2), pp. 159–194.
Tesnière, L. (1959): *Eléménts de syntaxe structurale*. Klincksieck: Paris.
Widdows, D. (2004): *Geometry and Meaning.* Stanford California: CSLI Books.

Diana Peppoloni
Università per Stranieri di Perugia
Piazza Fortebraccio, 4 – 06123 Perugia – Italia
dianapeppoloni@libero.it