



***Dottorato di Ricerca in Linguistica Generale,***

*Storica, Applicata, Computazionale e delle Lingue Moderne*

L-LIN/01 o altro settore S.D.

*Tesi di dottorato:*

# *Metodi computazionali per esplorare l'interfaccia tra sintassi e semantica: il caso dei verbi italiani*

*Candidato:*

Dott.ssa Diana Peppoloni

*Tutori*:

Prof. Alessandro Lenci

Prof. Mariella Bertuccelli

*Presidente:* Prof. Giovanna Marotta

Triennio 2006-2008

*Indice*

# Capitolo 1
# Il Problema della Rappresentazione e della Classificazione verbale

## 1.1 Studi sulla rappresentazione lessicale del verbo

La Linguistica ha attribuito negli ultimi decenni, un ruolo sempre più di primo piano al lessico; quest'ultimo è entrato a far parte anche dello studio sui fenomeni di interfaccia tra i diversi componenti della grammatica (sintassi, semantica etc.). Si è assistito allora allo sviluppo di teorie che assegnano appunto un ruolo dominante al lessico, tra le quali citiamo ad esempio la *Lexical Functional Grammar* (Bresnan 1982), la *Functional Grammar* (Dik 1978, 1979; Vossen 1995; Olbertz et al. 1998) e la *Role and Reference Grammar* (van Valin e Foley 1984), alla creazione di teorie sulla strutturazione del lessico e sulla rappresentazione dell'informazione lessicale quali il *Generative Lexicon* (Pustejovsky 1995), ma anche alla rivalutazione della posizione del lessico all'interno di teorie saldamente orientate in altra direzione, come il *Government and Binding* (Chomsky 1981) di stampo prettamente sintatticista.

All'interno di questo ambito di studi, la categoria del verbo è quella che ha riscosso più attenzione da parte degli specialisti, probabilmente in virtù della difficoltà incontrata nel formalizzare due diversi tipi di situazioni:

- la relazione tra predicato e argomento/i;
- il rapporto tra informazione semantico-lessicale contenuta nel predicato e realizzazione degli argomenti a livello frasale.

L'*argument* difatti è riguarda al contempo la rappresentazione sintattica così come di quella semantica, ed è proprio la sua duplice natura a renderlo uno degli strumenti privilegiati nell'indagine sui rapporti di correlazione tra piano sintattico e piano semantico. I linguisti si sono pertanto prevalentemente concentrati sull'elaborazione di modelli in grado di spiegare gli aspetti semantici e sintattici coinvolti nel meccanismo della proiezione argomentale, ovvero in quel processo mediante il quale il verbo proietterebbe un quadro argomentale, informativo sulla natura ed il tipo di argomenti e sulla relazione tematica che questi intrattengono con il verbo stesso.

Ovviamente sono state sviluppate molteplici soluzioni, in funzione soprattutto delle modalità di approccio al problema. Ad oggi l'ipotesi più largamente condivisa nelle varie teorie a disposizione, avanza la proposta per cui la rappresentazione lessicale del verbo sembra essere organizzata su più livelli, o strutture (struttura argomentale, struttura logica, struttura eventiva, etc.). Ognuno di questi livelli serve a rappresentare delle informazioni di vario genere, ad esempio categoriale, sottocategoriale, semantica, aspettuale, che poi andranno ad interagire con quelle presenti negli altri. Ciò che più interessa sono proprio le modalità di contatto tra queste strutture, ovvero le zone di interfaccia, tra cui ad esempio abbiamo già menzionato la struttura argomentale. In questo ambito l'approccio più accreditato è quello modulare (Jackendoff (1997)): esso sviluppa una chiave interpretativa che, alla reciproca dipendenza tra i vari livelli, abbina un certo grado di autonomia grazie al quale essi funzionerebbero anche in modalità parzialmente indipendente.

Nonostante la tesi largamente condivisa sull'esistenza di un meccanismo di funzionamento a strutture per rappresentare la conoscenza lessicale, c'è invece ancora una certo grado di controversia su quanti e quali livelli di organizzazione delle informazioni sia effettivamente possibile rintracciare, nonché sul tipo di informazioni che ciascuno di essi rispettivamente contiene. Il livello generalmente più studiato è stato senz'altro quello della struttura argomentale; a tutt'oggi rimangono ancora insoluti degli interrogativi:

– circa il numero e la natura delle informazioni in esso raccolte, ovvero sul numero degli argomenti identificabili, sulla distinzione tra argomenti esterni ed interni, sulla definizione di ruolo tematico, sulle restrizioni imposte nella selezione dei possibili argomenti, sul grado di opzionalità nella realizzazione degli argomenti, etc.;
– sulla natura primitiva o altresì derivata di tale livello: ci si chiede in particolare se la sua struttura non possa dipendere dalle caratteristiche semantiche ed aspettuali del verbo.

Recentemente sono stati proposti anche modelli validi concentrati sulle altre strutture identificate, come ad esempio la *struttura logica*, *eventiva* o ancora la *struttura qualia* (Pustejovsky (1995), van Valin e LaPolla (1997)).

I modelli proposti dai diversi studiosi variano infine in riferimento a quali informazioni si suppone facciano parte del componente lessicale e non di altri componenti, come ad esempio quello sintattico.

Intuitivamente sembra ovvio affermare che un'entrata lessicale contenga anzitutto due ordini di informazioni: la prima di tipo categoriale (ad esempio Verbo, o Nome etc.) e l'altra più specifica di tipo sottocategoriale (ad esempio Transitivo, Intransitivo etc.). Spesso però la realtà linguistica sembra smentire questi postulati; in molte lingue, ad esempio, non è raro che uno stesso elemento lessicale possa assumere, nel contesto sintattico, un diverso valore categoriale (è il caso dell'inglese con parole come *book* che fungono sia da Nome che da Verbo), ma anche sottocategoriale (è il caso di tutti quei verbi che prevedono più di una configurazione sintattica).

Diventa allora difficile appurare quali informazioni siano contenute in queste parole: se tutte quelle che la parola può assumere contestualmente, o piuttosto se si tratta di informazioni esterne alla dimensione lessicale, attivate da variabili composizionali.

Sicuramente una valida teoria lessicologica deve anzitutto essere economica, pertanto le rappresentazioni lessicali sulle quali poggia devono essere organizzate in una forma tale da poter essere agevolmente apprese e sfruttate. Ma seppure nella sua economicità, una buona rappresentazione lessicale deve comunque tenere in considerazione tutta una gamma di fenomeni collegati all'uso effettivo delle parole nei vari contesti linguistici. Tra questi ultimi annoveriamo: il comportamento sintattico delle parole, le interpretazioni che queste ultime attivano nei diversi contesti d'uso, la creatività che i parlanti esibiscono nell'utilizzarle, le modalità per cui le parole si combinano o meno tra loro sull'asse sintagmatico, ed infine le relazioni semantiche e formali che si instaurano tra le parole appartenenti ad un certo lessico.

Andiamo ora ad analizzare nello specifico alcuni modelli per la rappresentazione delle proprietà lessicali dei verbi, in linguistica teorica e computazionale.

### 1.1.1 Modelli per la rappresentazione lessicale del verbo

Le rappresentazioni lessicali del verbo proposte nella letteratura teorica degli ultimi decenni ed utilizzati nella ricerca linguistica, vedono una netta e significativa opposizione tra due differenti tipi di approccio:

1. gli *approcci lessicali* con *focus* interno alla parola, che presentano un asse interpretativo che oscilla tra prospettive più sintattiche ed altre più strettamente semantiche;
2. gli *approcci composizionali*, con *focus* esterno alla parola, ovvero nel contesto sintagmatico o frasale (Jezek (2003)).

I modelli che appartengono al primo dei due gruppi descritti, sono accomunati dal fatto che sono tutti basati sull'idea che numero e tipo di argomenti siano dei primitivi lessicali, ovvero costituiscano una serie di informazioni specificate nelle entrate verbali.

Degli studi sviluppati su questa tendenza, verrà approfondito il modello valenziale ed il relativo concetto di relazione attanziale. Nella letteratura specifica sul fenomeno si è soliti attribuire a Lucien Tesnière (1959) le prime riflessioni sui concetti di *valenza*, *attante* e *relazione attanziale*, che per molti aspetti saranno la base per la teoria della struttura argomentale. Prima di allora si distingueva classicamente tra verbi transitivi ed intransitivi, affiancando a queste due categorie i concetti di complemento e reggenza; venivano però trascurati interamente alcuni aspetti, come ad esempio il fatto che la categoria complemento mostra molte complessità al suo interno; basti pensare che alcuni complementi sono retti dal verbo e sono quindi obbligatori, mentre altri no e risultano pertanto facoltativi.

Tesnière sostituisce al concetto di reggenza quello di valenza, mutuato direttamente dalla chimica, per cui il verbo viene raffigurato come una sorta di atomo dotato di appigli, capace di attrarre un numero variabile di elementi all'interno della frase (precisamente da 0 a 3). La valenza del verbo corrisponde proprio al numero complessivo di tali elementi, detti *attanti* e definiti come:

> les persone ou choses qui participent à un degré quelconque au procès (Tesnière 1959: 105)

Le proprietà valenziali permettono di distinguere quattro tipi di verbi, riassumibili come segue:

V0 arg: nevica es. sta nevicando
V1 arg: morire es. è morta una donna
V2 arg: salutare es. Mauro saluto Gianni
V3 arg: dedicare es. Luca ha dedicato una canzone alla moglie

Nella frase, insieme al *nucleo verbale* e agli *attanti*, è possibile distinguere dei componenti circostanziali, che esprimono

> Les circostances de temps, lieu manière etc. dans lesquelles se déroule le process (Tesnière 1959: 102)

Negli esempi che seguono, le parti in grassetto indicano proprio questi elementi circostanziali:

(1) a. Ho comprato un libro **giovedì scorso**
b. Abbiamo incontrato Gianna **al supermercato**

La differenza più evidente tra la teoria valenziale e la grammatica tradizionale, è che mentre la prima distingue tra elementi retti dal verbo (o attanti) e elementi non retti dallo stesso (o circostanziali), la seconda si basa sulla più netta divisione tra Soggetto e complementi.

Il concetto di valenza verbale occupa un posto di rilievo all'interno degli studi di tradizione europea, difatti la letteratura a riguardo è assai vasta. Vanno ricordati in ambito francese i lavori di Maurice e Gaston Gross (1975, 1986), l'ampia letteratura tedesca nota come *grammatica della dipendenza*, che ha portato alla compilazione di dizionari valenziali come quello per i verbi tedeschi di Helbig e Schenkel (1969), o quello per l'italiano di Blumenthal e Rovere (1998). Sempre per l'italiano va menzionato il modello della frase nucleare proposto da Sabatini e Coletti (1997), e adottato nello sviluppo delle voci verbali nel Dizionario Italiano Sabatini Coletti.

La teoria valenziale solleva un primo e consistente quesito legato alla seguente osservazione: è vero che il quadro attivato dal verbo richiama un certo numero di elementi rispetto ai quali qualcosa è predicato, ma è anche vero che non tutti questi elementi vengono espressi a livello sintattico e non con le stesse modalità.

(2) a. Giulia non ha ancora imparato a guidare
b. Marco ha aspettato molto tempo per parcheggiare
Nei casi di (2) a. e b., sembra possibile postulare la presenza di un elemento, in questo caso *classe di veicoli*, non espresso sintatticamente ma presente a livello logico

Questi esempi dimostrano una situazione abbastanza frequente, quella per cui il *pattern* sintattico di un verbo presenta un numero inferiore di elementi, rispetto a quelli che l'enunciato implica a livello logico. Occorrerà allora distinguere tra:

– *valenza sintattica*, ovvero l'insieme degli attanti o argomenti obbligatori, perché obbligatoriamente espressi, pena l'agrammaticalità della frase;
– *valenza semantica*, corrispondente al numero di elementi implicati a livello logico.

È evidente da quanto detto, che la definizione della valenza semantica di un verbo è assai più complessa di quanto non lo sia l'identificazione della sua valenza sintattica, in quanto costituita da ciò che vediamo concretamente della realizzazione della frase.

Il secondo problema collegato alla teoria valenziale è la distinzione tra quelli che Tesnière definisce attanti da un lato, e elementi circostanziali dall'altro; c'è poi da verificare la distinzione correlata tra nucleo e margini della predicazione ed infine tra argomenti e modificatori. Tale questione è ampiamente trattata non solo negli studi sulla rappresentazione lessicale del verbo, ma anche in molte opere contenenti descrizioni di specifiche teorie di grammatica interessate a definire il nucleo funzionale della predicazione (Dik (1989)). Non è sempre chiaro quale criterio sia possibile adottare per effettuare tale distinzione, poiché essa non è netta ed è spesso contraddistinta da casi intermedi (Smith (2000)); l'identificazione degli elementi circostanziali ad esempio, è complicata dal fatto che alcuni componenti che svolgono lo stesso ruolo nella frase possono, a seconda dei casi, essere obbligatori o meno:

(3) Si veda il caso del locativo *al mare* nei seguenti esempi:
a. Mauro va *al mare* tutte le estati
*Mauro va tutte le estati
b. Abbiamo incontrato Mauro *al mare* lo scorso fine settimana
Abbiamo incontrato Mauro lo scorso fine settimana

In (3) a. il caso locativo ha funzione argomentale, mentre in (3) b. esso ha funzione non argomentale, difatti può essere o meno espresso senza alterare la grammaticalità della frase.

Un terzo problema riguarda il tipo di realizzazione sintattica degli attanti, poiché data una valenza è difficile predire come essa si manifesterà sintatticamente in termini di assegnazione del caso.

(4) È il caso di due verbi come *chiamare* e *telefonare* che, sebbene sinonimi per alcuni tratti del significato, mostrano comportamenti sintattici diversi:
a. Gianni chiama <u>Mauro</u> (obj dir)
b. Gianni telefona <u>a Mauro</u> (obj PP)

Un ultimo punto problematico all'interno della teoria valenziale è costituito dalla capacità di variazione nel numero e nella distribuzione degli attanti tipica dei verbi, ovvero dalla cosiddetta *alternanza sintattica*. Con questo termine si fa riferimento nello specifico alla possibilità dei singoli verbi di presentare molteplici configurazioni sintattiche di cui non è sempre immediato interpretare i reciproci rapporti, così come prevedere le restrizioni che operano su di esse in vario modo.

Gli studi successivi a Tesnière hanno individuato una soluzione ai problemi sollevati dalla sua teoria, nel considerare la valenza in termini prettamente sintattici, cioè di argomenti sintatticamente realizzati. Sviluppato in questo senso, il concetto di valenza riusciva a delineare correttamente la configurazione sintattica sviluppata attorno al verbo, ma non teneva conto dei rapporti tra questa e la dimensione semantica dell'entrata lessicale.

Tra i primi tentativi di affrontare i problemi semantici legati alla proiezione argomentale ricordiamo quello basato sull'idea di *ruolo tematico*; esso si propone di formalizzare i vari tipi di relazioni semantiche che intercorrono tra il verbo ed i suoi argomenti.

Fillmore (1968) sostenne nei suoi lavori l'esistenza di una lista finita di casi profondi, o *deep cases*, che avevano il compito di distinguere tipi diversi, ma universalmente validi, di relazioni tematiche tra Verbo e argomenti ad esso correlati. Di casi profondi avevano già parlato studiosi come Katz (1966) nella sua teoria sulle *semantic relations*, Gruber (1965) e Jackendoff (1972) in quella sulle *thematic relations* e Davidson (1967) in riferimento alle *primitive notions* nella logica degli eventi.

Fillmore partiva dall'osservazione per cui il referente di una stessa posizione sintattica all'interno dell'enunciato, può stabilire con il verbo relazioni diverse.

Nell'esempio seguente, il soggetto delle proposizioni a., b. e c. assume tre diversi rapporti rispetto al verbo, diventando rispettivamente: Agente, Beneficiario e Paziente.

| | | |
|---|---|---|
| (5) | a. Mauro ha rotto il vetro | **Agente** |
| | b. Mauro ha ricevuto un regalo | **Beneficiario** |
| | c. Mauro ha preso un pugno | **Paziente** |

L'intento di Fillmore era quindi quello di etichettare le varie relazioni individuate e di formalizzare il nesso tra il tipo di rapporto verbo/argomento, cioè il caso profondo (Agente, Paziente, Goal etc.), e la sua realizzazione sintattica, ovvero soggetto, oggetto diretto, oggetto indiretto etc.

L'ipotesi di fondo è che i casi profondi correlati ad un dato verbo siano in grado di determinarne le configurazioni sintattiche di cui esso fa parte. A livello lessicale, a ciascun verbo si associa un *case frame*, vale a dire una sorta di finestra che elenca i casi profondi obbligatori ed opzionali nella realizzazione sintattica.

In linea con questi principi e sempre in ambito prevalentemente sintatticista, si è mosso anche Noam Chomsky (1981), che rivede la sua versione della teoria del *Government and Binding* includendovi il meccanismo definito *Theta Criterion*; in questo modo Chomsky postula la proiezione, da parte del verbo, di una griglia tematica, *Theta Grid*, contenente tutte le informazioni circa il tipo di relazione che i vari argomenti intrattengono con il verbo (*Thematic Role*).

Il modello basato sui ruoli tematici viene progressivamente affiancato e poi sostituito da altri tipi di rappresentazioni delle informazioni semantiche proprie del verbo. Questi si distinguono da quelli fin qui trattati, perché nascono in ambito semanticista e non si basano su un concetto semantico di tipo relazionale come quello di ruolo, ma ricorrono ad un processo di scomposizione del significato lessicale. Significativo è in questo senso il lavoro di Jackendoff, che ha proposto un modello in cui viene mantenuto il concetto di ruolo tematico, affiancato però ad una rappresentazione semantica articolata, detta *Lexical Conceptual Structure*. Tale struttura è costruita intorno ad una serie di primitivi semantici funzionali quali *cause*, *change*, *go*, *act* a cui sono abbinati degli argomenti denominati *individui*, *stati*, *eventi*. Prendiamo ad esempio la funzione *change* ed osserviamo che essa implica necessariamente tre argomenti, ovvero un individuo, uno stato iniziale ed uno stato finale; agli argomenti viene poi associato un ruolo tematico: nel caso di *change*, l'individuo coinvolto assumerà il ruolo di *Tema*. L'innovazione rispetto ai modelli precedenti, sta nel fatto che pur mantenendo ancora intatta l'idea di ruolo tematico, esso non costituisce più un primitivo semantico, ma al contrario una derivazione determinata dalle posizioni degli argomenti nella Struttura Concettuale. Le rappresentazioni lessicali proposte da Jackendoff contengono inoltre una griglia di sottocategorizzazione che esplicita le categorie sintattiche collegate agli argomenti espressi (NP, PP etc.) .

La *Conceptual Structure* di Jackendoff punta a rappresentare la semantica del verbo, in modo indipendente, autonomo rispetto al livello sintattico; il suo modello decomposizionale trova le sue basi negli studi di semantica generativa americana (Katz e Fodor (1963), Lakoff (1970)) e, ancora, nei lavori di Gruber (1965).

I lavori successivi a Jackendoff inaugurano la presenza di un nuovo concetto che troverà largo impiego nelle rappresentazioni lessicali a venire: quello di *struttura argomentale*.

Ad oggi è opinione diffusa che la struttura argomentale sia un formalismo efficace nella rappresentazione delle informazioni fornite dal verbo sugli argomenti di cui predica qualcosa. Restano comunque alcune questioni insolute sui suoi argomenti, tra cui quante e quali informazioni essa dovrebbe contenere, in che modo tali informazioni dovrebbero rapportarsi reciprocamente ed infine quale sia la sua natura prevalente, se sintattica o piuttosto semantica. Il termine struttura si deve al fatto che essa non è considerata una semplice lista di argomenti, ma invece un insieme composito di informazioni organizzate secondo principi gerarchici e strutturali appunto.

Molti studiosi hanno cercato di integrare adeguatamente questa nozione all'interno di un modello per la rappresentazione del significato lessicale. Tra questi citiamo ad esempio Rappaport, Laughren e Levin (1987) che presentano un modello costruito su due livelli:

1. la *struttura argomentale*, che contiene le informazioni relative al numero e alla natura sintattica degli argomenti;
2. la *struttura concettuale lessicale*, contenente le informazioni più strettamente semantiche descritte utilizzando un formalismo decomposizionale.

Secondo questo modello gli argomenti vengono realizzati sintatticamente grazie ad una serie di regole che specificano la posizione per ogni ruolo associato al rispettivo argomento. Il concetto di ruolo tematico si riduce ad una nozione puramente strutturale ed equivalente alla distinzione tra *argomento interno* ed *argomento esterno* (esterno al sintagma verbale, di solito si tratta di un sintagma nominale che funge da soggetto grammaticale dell'enunciato).

Anche per Jane Grimshaw (1990) la struttura argomentale di un verbo è una parte dell'entrata lessicale, specificamente quella atta a contenere l'informazione di natura più grammaticale. Essa interagisce però in questo modello, non più con un solo livello, bensì con altri due: quello *semantico-lessicale* e quello *aspettuale*. Centrale in questo modello è l'idea che la struttura argomentale sia internamente organizzata: gli argomenti in essa contenuti saranno gerarchicamente organizzati, in base ad un principio definito di prominenza. L'argomento esterno risulterà essere il più prominente, ed inoltre sarebbe possibile formulare una gerarchia di prominenza anche tra gli argomenti interni se ce ne fosse più di uno. Il grado di prominenza (principio di natura sintattica) deriva sia dalle proprietà tematiche che aspettuali del verbo, proprietà che non appartengono alla struttura argomentale, ma che invece si rintracciano nella struttura Concettuale Lessicale ed in quella eventiva.

Grimshaw propende per un'interpretazione nettamente sintattica della struttura argomentale, tanto è vero che nei suoi lavori scinde gli argomenti grammaticali dai partecipanti semantici all'evento, che non sono specificati dalla struttura argomentale.

Più recentemente Pustejovsky (1995) definisce la struttura argomentale come costituita dall'insieme del numero e del tipo degli argomenti logici di un predicato, che lo studioso assegna a quattro differenti categorie:

1. *veri argomenti*, sono quegli argomenti effettivamente realizzati sintatticamente, sia obbligatori che opzionali; es. "*Mauro* legge un libro"
2. *argomenti default*, sono costituiti dall'insieme degli argomenti logici, non necessariamente realizzati sintatticamente; es. "Mauro parcheggia *la macchina"*
3. *argomenti ombra*, sono parte integrante dell'entrata lessicale e possono essere realizzati sintatticamente solo se aggiungono una specificazione;
   es. "Mauro si pettina *con la spazzola"*
   "Mauro si pettina *con la spazzola che gli ho regalato ieri"*

4. *veri modificatori*, sono quegli argomenti che incidono sulla logica dell'enunciato, ma appartengono all'interpretazione situazionale, e non sono specificati nella semantica lessicale del verbo; es. "Mauro è andato in vacanza in Francia *quest'estate"*.

Pustejovsky, grazie a questa organizzazione, riesce ad inserire nella struttura argomentale anche quegli argomenti che designano i partecipanti logici o semantici che non vengono obbligatoriamente esplicitati dal punto di vista sintattico, e che a volte non possono addirittura proprio essere realizzati a questo livello (Pustejovsky (1995)).

Uno dei problemi teorici sollevati dal modello basato sulla struttura argomentale, ovvero sull'ipotesi che gli argomenti di un verbo siano contenuti in una struttura retta da principi di prominenza sintattica e tematica e che tale struttura sia specificata nel lessico, è il confronto con la possibilità dei vari verbi di realizzare comportamenti sintattici diversi, anche per ciò che concerne gli argomenti obbligatori (esempio 6).

(6) *sbattere* transitivo es. Mauro sbatte la porta
*sbattere* intransitivo es. La porta sbatte

In linguistica questo fenomeno viene chiamato alternanza argomentale (Levin (1993)); gli studiosi sono in difficoltà nel determinare quale sia da considerarsi la struttura argomentale primaria e quali quelle derivate, e ancora in base a quali restrizioni certi verbi ammettono più costruzioni sintattiche mentre altri no.

Levin (1993) sostiene che davanti a molteplici configurazioni di uno stesso verbo, soltanto una sarà la struttura argomentale primitiva, le altre invece deriveranno dall'applicazione di regole lessicali, tipiche di singoli verbi o di una classe verbale. La flessibilità del comportamento verbale mette quindi in difficoltà il modello basato sulla struttura argomentale, tanto che i linguisti hanno cercato di formulare rappresentazioni meno rigide, in grado di dare conto di tutta la gamma delle realizzazioni sintattiche di un verbo nei vari contesti d'uso.

### 1.1.2 Le classi aspettuali

Come già detto in precedenza, esiste un altro filone di studi per la rappresentazione dell'informazione lessicale del verbo, che si concentra sulle proprietà aspettuali di quest'ultimo, ritenendo attraverso questi formalismi di poter far luce anche sulle realizzazioni argomentali.

L'idea alla base di questi lavori è che le proprietà aspettuali di un verbo controllino varie situazioni, tra cui le restrizioni sul tipo di strutture argomentali che un verbo può presentare, il numero degli argomenti richiesti, il ruolo tematico ad essi assegnato.

Vendler (1967) postula la presenza di quattro classi verbali ottenute analizzandone le proprietà temporali interne; tale analisi è stata ottenuta verificando le restrizioni dovute all'utilizzo di modificatori avverbiali di tempo, di alcuni tempi verbali e le implicazioni logiche derivate.

*Tab. 1.1 Le classi verbali di Vendler*

| **Aktionsart** | **Proprietà temporali** |
|---|---|
| *States* | hanno una durata, ma non una struttura interna,<br>poiché nell'arco di tempo in cui sono veri non introducono cambiamenti |
| *Activities* | descrivono un processo vero in ogni fase del suo svolgimento e quindi hanno una durata |
| *Achievements* | presentano una culminazione istantanea, perciò hanno un punto terminale obbligatorio, ma non una durata |
| *Accomplishments* | hanno una durata, un punto terminale obbligatorio, e sono veri solo se è raggiunta la fase finale |

Le caratteristiche azionali di un verbo possono essere descritte utilizzando i tratti semantici di *durativ*ità e *telicità*. La classificazione secondo il comportamento azionale non è l'unica calssificazione semantica possibile per i verbi di una lingua, ma l'*azionalità* è particolarmente interessante per i riflessi che ha sulla sintassi, sul sistema del *Tempo* e su quello dell'*Aspetto*. Infatti per verificare se un verbo possiede uno di questi tratti, si utilizzano dei particolari test sintattici. Ad esempio, gli avverbiali temporali del tipo *a lungo* (es. "Chiara ha mescolato l'impasto *a lungo*") sono compatibili con verbi durativi ma non con verbi puntuali: l'avverbiale "a lungo" diventa quindi un test sintattico di verifica della durativ*ità* di un verbo.

La durativ*ità* distingue verbi che sono percepiti come prolungati nel tempo (*correre*, *preparare*, *cantare*, *amare*) da altri che sono invece percepiti come istantanei (*cadere*, *esplodere*, *morire*). Ovviamente anche un evento non durativo ha una sua durata fisica nel tempo, ma viene trattato come se accadesse in un istante.

(7) durativo: "Chiara *prepara* la torta"

non durativo: "La torta *è caduta* dal tavolo"

I verbi durativi sono incompatibili con avverbiali puntuali, mentre i verbi non durativi sono incompatibili con avverbiali durativi come "per *x* tempo":

(8) "Chiara mescolò l'impasto *per tre ore*"

"La torta è caduta dal tavolo *per tre minuti*"

I verbi telici descrivono eventi che tendono verso un fine, un completamento. Se il verbo è telico, il completamento dell'azione è necessario per dire che l'azione è effettivamente avvenuta.

(9) telico: "Chiara *prepara* la torta"

non telico: "Chiara *canta* in cucina"

L'avverbiale "in *x* tempo" non è compatibile con verbi non telici:

(10) "Chiara ha preparato una torta *in due ore*"

*"Chiara ha cantato *in due ore*"

I verbi che appartengono alle classi di *accomplishment* e *achievement* descrivono azioni che culminano in un punto terminale e sono pertanto telici; la classe delle *activities* descrive invece verbi che non presentano questa caratteristica e si dicono pertanto atelici.

Lo stesso verbo può partecipare a descrizioni degli eventi riconducibili a più di uno dei tipi aspettuali menzionati; questo fenomeno suggerisce che le classi aspettuali siano quindi interrelate tra di loro. Tali correlazioni sono ben illustrate analizzando verbi come *see*, *hear*, *recognize*, *understand* etc., utilizzati in frasi che possono coinvolgere sia la classe degli *states*, che quella degli *achievements* (esempio 11).

(11) a. Mauro capisce il tedesco (STATE)
b. Mauro capì subito il significato della lettera (ACHIEVEMENT)

Le teorie sulla realizzazione argomentale coinvolgono raramente le classi aspettuali nella mappatura della sintassi; più spesso fanno riferimento alla definizione di questo genere di proprietà. Di solito viene postulato che le proprietà che producono relazioni tra classi aspettuali e quelle che figurano nella realizzazione argomentale siano le stesse.

L'idea che le proprietà aspettuali influenzino la realizzazione argomentale risale ad Hopper e Thompson (1980): il loro studio include le nozioni di felicità e puntualità tra i fattori capaci di determinare la transitività di un verbo.

I lavori di Tenny (1987, 1992, 1994) sono però i primi a presentare un approccio aspettuale articolato alla teoria della realizzazione argomentale. Il punto di partenza si colloca nella discussa e controversa *Ipotesi dell'Interfaccia Aspettuale*:

> the universal principles of mapping between thematic structure andsyntactic argument structure are governed by aspectual properties. Constraints on the aspectual properties associated with direct internal arguments, indirect internal arguments, and external arguments in syntactic structure constrain the kinds of event participants that can occupy these positions. Only the aspectual part of thematic structure is visible to the universal linking principles (Tenny, 1994: 2)

Le attuali teorie su base aspettuale per la realizzazione argomentale, focalizzano solitamente sulla relazione tra scelta e espressione morfosintattica dell'oggetto diretto e su nozioni quali *telicità*, *misura* e *tema incrementale*. Quale di queste nozioni sia direttamente legata all'oggetto diretto e come questo viene realizzato, varia da teoria a teoria. Consideriamo i vincoli proposti da Tenny:

> Measuring-Out Constraint on Direct Internal Arguments:
> (i) The direct internal argument of a simple verb is constrained so that it undergoes no necessary internal motion or change which "measures out the event" over time (where "measuring out" entails that the direct argument plays a particular role in defining the event);
> (ii) Direct internal arguments are the only overt arguments which can "measure out the event";
> (iii) There can be no more than one measuring-out for any event described by a verb.
>
> (Tenny, 1994: 11)
>
> The Terminus Constraint on Indirect Internal Arguments:
> (i) An indirect internal argument can only participate in aspectual structure by providing a terminus for the event described by the verb. The terminus causes the event to be limited;
> (ii) If the event has a terminus, it also has a path, either implicit or overt;
> (iii) An event as described by a verb can have only one terminus.
>
> (Tenny, 1994: 68)

L'ipotesi dell'interfaccia aspettuale di Tenny, rende conto piuttosto bene dell'osservazione che gli argomenti che sono oggetti diretti di verbi prototipicamente transitivi, sono quelli in grado di "misurare" un evento.

Anche Levin e Rappaport Hovav (1995) verificano la teoria dell'approccio aspettuale alla realizzazione argomentale. Si parte dall'idea che l'inaccusatività di un verbo genera una forma di alternanza argomentale attraverso quello che Levin e Rappaport Hovav chiamano *verbi a comportamento variabile*, ovvero verbi che mostrano la duplice caratteristica di *inaccusativi* e *non ergativi*. Se si ritiene che l'inaccusatività sia codificata sintatticamente, allora il singolo argomento di questi verbi avrà due diverse realizzazioni: soggetto quando non ergativo, e oggetto diretto quando in accusativo.

I verbi che esprimono il modo del movimento hanno mostrato, ad esempio, di rispondere a questa duplice classificazione; essi compaiono sia con l'ausiliare inaccusativo *essere*, che con quello non ergativo *avere* (Centineo (1986, 1996); Van Valin (1990); Zaenen (1993)). Queste diverse opzioni sono ben illustrate nei seguenti esempi in italiano:

(12) a. Mauro ha corso meglio ieri
b. Mauro è corso a casa mia

La felicità è coinvolta nella spiegazione di questo tipo di realizzazione argomentale multipla. La proposta è che gli usi non ergativi siano atelici, mentre quelli inaccusativi siano al contrario telici. Gli esempi in (13) ci aiutano a sostenere questa tesi; stabilito che i verbi inaccusativi hanno un oggetto sottinteso, questa correlazione supporta la connessione ipotizzata tra realizzazione dell'oggetto e telicità.

I vari studi menzionati finora, pur prendendo spunto da orientamenti diversi, presentano un punto in comune che è il fatto di essere *predicate-oriented*; tutti tentano difatti di ricostruire i comportamenti argomentali partendo dalla semantica o dall'aspetto del verbo, basandosi sull'idea che la struttura argomentale derivi dalla semantica lessicale, o dall'*Aktionsart* verbali. Tale tendenza è però incapace di confrontarsi con situazioni come l'alternanza argomentale e la polisemia lessicale.

Si osservi il caso di un verbo come *affondare*; esso presenta un'alternanza di tipo V TR – INTR ES che comporta sia una variazione nel significato del verbo, che un diverso utilizzo nei contesti in cui questo compare (esempio 13).

(13) a. La nave affonda
b. La storia affonda le sue radici nell'antichità

Si è pensato allora che la semantica non fosse un atomo impermeabile ai processi sintattici, bensì un componente sensibile ai processi composizionali e quindi interagente con la semantica degli altri elementi presenti nella frase. Questo genere di intuizione si è rivelata fondamentale in un modello che tenga conto della semantica e delle proprietà sintattiche dell'entrata lessicale nei contesti in cui essa compare; i modelli più fruttuosi in questo senso, sono quelli che focalizzano sulle modalità di interazione della semantica lessicale con gli elementi composizionali.

Appare comunque evidente come i due approcci descritti, lessicale da un lato e composizionale dall'altro, siano complementari nella formalizzazione delle proprietà semantiche, aspettuali e argomentali del verbo su cui agiscono fattori di entrambi i tipi, seppure con un peso diverso a seconda dei casi.

## 1.2 La classificazione del verbo

La classificazione del verbo costituisce indubbiamente uno dei problemi più intricati 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 ruolo del contesto nella definizione delle proprietà semantiche dei verbi? Come è opportuno procedere per isolare le variabili composizionali del significato (Jezek (2003))?

Finora, riflettendo in linea teorica sulla rappresentazione verbale, sono stati discussi esclusivamente i problemi inerenti le singole entrate lessicali, o al più coppie di verbi. Invece per poter elaborare una teoria lessicologica completa, cioè un modello capace di descrivere correttamente l'organizzazione e la struttura del lessico, occorre:

- ragionare in termini di classi di parole, piuttosto che di singole entrate lessicali;
- stabilire i criteri necessari per individuare sottogruppi di parole all'interno di insiemi più ampi.

Rispetto a quanto detto fin qui, bisognerà allora ragionare in termini di classi verbali, ponendosi domande di portata più vasta, come ad esempio: qual è il fattore che determina il fatto che un certo insieme di verbi compaia in una determinata costruzione (es. transitiva) o anche in più di una, e al contempo esclude che questo stesso meccanismo sia valido anche nel caso di altri gruppi di verbi?

Se si vuole rispondere a questo genere di domande, si dovrà necessariamente affrontare l'intricato problema della classificazione verbale.

Solitamente definiamo una classificazione verbale come uno strumento per attribuire delle parole appartenenti alla stessa classe lessicale (es. N, V, AG etc.) a varie sottoclassi, partendo dalle differenze riscontrate nel loro comportamento sintattico, argomentale e grammaticale (Jezek (2003)).

È in questo modo che si può distinguere ad esempio tra verbi transitivi ed intransitivi, ovvero tra quelli che possono accompagnarsi ad un complemento oggetto e quelli che invece non possono:

(14) *uscire* Mauro esce (INTRANSITIVO)
*Mauro esce la porta (TRANSITIVO)
*riparare* Il falegname ripara la porta (TRANSITIVO)
*Il falegname ripara (INTANSITIVO)

Per classificare le entrate lessicali sono state elencate una serie di proprietà sintattiche, generalmente usate in combinazione tra loro (Tab. 1.2)

*Tab.1.2 Proprietà sintattiche per le operazioni classificatorie*

| numero di argomenti |
|---|
| realizzazione sintattica degli argomenti (Sogg, Ogg dir, Ogg indir) |
| posizione degli argomenti nella frase |
| ordine degli argomenti nella frase |

La ragione per cui parole riconducibili alla stessa parte del discorso mostrano comportamenti sintattici diversi, si pensa sia attribuibile al loro *significato*. Se è vera questa ipotesi e gli elementi del significato intervengono direttamente nel comportamento sintattico delle entrate lessicali, allora verbi con sintassi diversa avranno di conseguenza caratteristiche semantiche diverse. Questa è esattamente l'ipotesi da cui parte anche Levin (1993), (2005), nella realizzazione della sua classificazione verbale per l'inglese.

Nel processo di classificazione verbale entra dunque in gioco un meccanismo di interfaccia biplanare, sintattico e semantico-lessicale; l'obiettivo di una valida classificazione dovrebbe essere proprio quello di cogliere le correlazioni tra questi due livelli. È doveroso comunque precisare che questa non è l'unica strategia per portare avanti una classificazione; in alternativa è possibile procedere:

- distinguendo le classi verbali solo in funzione del loro comportamento sintattico (otterremo così una separazione tra verbi monoargomentali, biargomentali e così via);
- isolando unicamente le proprietà semantiche e aspettuali dei verbi, così da ricavare dei tipi semantici o aspettuali dei verbi (avremo ad esempio una distinzione tra verbi stativi, di moto, con Soggetto Animato etc.).

È evidente però che una classificazione che tenga in considerazione la correlazione tra piano sintattico e semantico, risulterà più esaustiva di una che valuti solo uno dei due livelli considerati.

Tornando alla 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; difatti questo è ciò che emerge solitamente dall'osservazione del comportamento verbale.

Vi sono però dei casi che sembrano smentire la presenza di un legame regolare, costante tra il livello semantico e quello sintattico; ci riferiamo a tutti quei verbi con significato simile, ma diverso comportamento sintattico.

(15) Stasera *chiamo* Mauro
Stasera *telefono a* Mauro
Il verbo *chiamare* realizza l'Oggetto come Obj diretto, mentre *telefonare* lo fa come Obj Preposizionale

La riflessione su tali discrepanze, ha portato all'individuazione di tutta una serie di fattori ritenuti responsabili delle incongruenze nell'interfaccia semantico-sintattica:

- assenza di corrispondenza biunivoca tra piano del significato e quello della forma. Questo vuol dire che se non tutti i componenti del significato sono realizzati sintatticamente, allora solo alcuni elementi del significato influenzeranno la forma, mentre altri no. Inoltre sarà altrettanto corretto affermare che la sintassi mostra dunque solo una parte del contenuto semantico delle parole. Uno dei problemi principali è proprio cercare di capire come individuare il tratto semantico corrispondente all'attivazione di un certo comportamento sintattico;
- indipendenza del comportamento sintattico. Il principio di determinazione sintattica fin qui sostenuto non deve essere considerato valido in assoluto, poiché è chiaro che la semantica non può influenzare *in toto* il comportamento sintattico di un'entrata lessicale. Ci saranno difatti dei fenomeni sintattici determinati da altri piani (ad esempio la pragmatica), ed altri ancora giustificati completamente all'interno del componente sintattico stesso.
- alternanza argomentale e polisemia. Nel primo caso singoli verbi hanno la possibilità di realizzare costruzioni sintattiche diverse degli stessi argomenti o di cambiare gli argomenti espressi in base al contesto sintattico; nel secondo invece una stessa entrata lessicale attiva diversi significati nel contesto:

(16) <u>Alternanza argomentale</u>
Mauro *apre* la porta
La porta *si apre*
<u>Polisemia</u>
*mangiare* un panino
*mangiare* la foglia

Specie per quanto riguarda l'alternanza argomentale, la mancanza di una corrispondenza biunivoca tra piano argomentale e piano semantico, comporta che uno stesso verbo, valutato in base a questa caratteristica, possa appartenere a più classi contemporaneamente, andando così a complicare la realizzazione di un'efficace classificazione verbale.

Perciò, in base a quanto detto finora, emerge chiaramente il fatto che la classificazione verbale si presenta come un'operazione estremamente complessa da realizzare, per tutta una serie di ragioni. Tra queste ricordiamo che si tratta di un problema che interessa più di un livello, quello semantico e quello sintattico, e che occorre capire come si sviluppa e prende forma l'interazione tra di essi; inoltre non tutti gli elementi del significato vengono sempre espressi sintatticamente, ma sono comunque importanti perché determinano le restrizioni sulle possibilità combinatorie delle parole. Infine la capacità di un verbo di realizzare la cosiddetta alternanza argomentale, fa sì che questo possa risultare in più di un gruppo della classificazione elaborata.

Non si possono ignorare tutti questi fenomeni destabilizzanti se si vuole sviluppare un modello che rappresenti fedelmente l'organizzazione e la struttura del lessico di una qualsiasi lingua.

### 1.3 Struttura argomentale e alternanza diatetica: il lavoro di Levin

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 delle affinità anche per ciò che riguarda i processi sintattici di sottocategorizzazione, raggruppandosi così all'interno di una stessa classe di appartenenza linguisticamente coerente.

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 (2002a)). 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 indagata e 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.

I risultati del progetto proposto da Levin dovrebbero aprire la strada allo sviluppo di una teoria sulla conoscenza lessicale; quest'ultima dovrebbe dar conto di entrate lessicali linguisticamente motivate per i verbi, che incorporino una rappresentazione del significato dei verbi stessi, e permettano ai significati dei verbi di essere associati correttamente all'espressione sintattica dei loro argomenti.

Le classi di Levin si basano sulla possibilità del verbo di ricorrere in diverse alternanze diatetiche, ovvero in specifiche coppie di *frames* sintattici. La classificazione così elaborata, copre un numero sostanzioso di alternanze possibili nella lingua inglese; ma in ogni caso non si può definire esaustiva. Essa lavora su circa 3200 verbi raggruppati in 48 classi, alcune delle quali vengono suddivise a loro volta in sottoclassi, fino a raggiungere un numero complessivo di 191. Queste partecipano a 79 diverse alternanze diatetiche, che coinvolgono solo complementi di tipo NP e PP.

Le ricerche si concentrano quasi esclusivamente sui verbi, poiché questi sono in grado di illustrare bene le caratteristiche della conoscenza lessicale; inoltre, in virtù della loro capacità di accorpare più argomenti di vario tipo, coinvolgono insiemi di proprietà estremamente complessi. 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.

(17) a. Sharon sprayed water on the plants.
b. Sharon sprayed the plants with water.

(18) 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.

(19) a.* Monica covered a blanket over the baby.
b. Monica covered the baby with a blanket.

(20) 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.

(21) 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.

Andiamo ad analizzare attraverso un esempio, come Levin abbia sfruttato il concetto di alternanza diatetica nell'elaborazione della sua classificazione. Prendiamo come campione la classe dei *Break Verbs*, ovvero quei verbi che esprimono azioni che implicano un cambiamento di stato nell'integrità materiale di una qualche entità. Tale classe si caratterizza per la sua partecipazione alle alternanze (1-3), ma non (4-6) qui di seguito proposte:

1. causative/incoative alternation
*Tony broke the window* ⟷ *The window broke*

2. middle alternation
*Tony broke the window* ⟷ *The window broke easily*

3. instrument subject alternation

*Tony broke the window* with the hammer ⟷ *The hammer broke the window*

4. *with/against alternation

*Tony broke the cup against the wall* ⟷ **Tony broke the wall with the cup*

5. *Conative alternation

*Tony broke the window* ⟷ *Tony broke at the window*

6. *Body-Part possessor ascension alternation

*Tony broke herself on the arm* ⟷ *Tony broke her arm*

Gli esempi appena proposti sono rappresentativi di un'ampia gamma di fenomeni che suggeriscono, nell'insieme, come la conoscenza del parlante circa le proprietà dei verbi vada oltre la mera consapevolezza dell'espressione dei suoi argomenti; proprio la sua abilità nel formulare giudizi sottili sui verbi e le loro proprietà, rende inverosimile che la conoscenza del parlante si limiti a quanto indicato dall'entrata lessicale. Ma allora cos'è che soggiace alla capacità dei parlanti di formulare tali giudizi?

> what enables a speaker to determine the behavior of a verb is its meaning (Hale e Keyser 1987).

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. Un esempio di quanto fin qui detto ci viene dai *frames di sottocategorizzazione*: queste etichette riguardanti le costruzioni sintattiche verbali che non possono essere derivate da principi generali di grammatica, sono considerate proiezioni delle proprietà lessicali delle parole su queste costruzioni.

Sebbene nessuno possa completamente smentire l'assunto per cui parole dal significato simile mostrano quanto meno una tendenza a realizzare lo stesso comportamento sintattico, l'ipotesi che quest'ultimo sia totalmente determinato dalla semantica del termine è assai controversa.

La chiave per verificare questa ipotesi è l'identificazione di una appropriata rappresentazione del significato del verbo. Comunque determinare le componenti appropriate del significato non è un'operazione facile, poiché aprioristicamente si possono elaborare più tipi di classificazioni verbali basate sulla semantica del verbo stesso.

Per verificare l'ipotesi semantico-sintattica, occorrerà stabilire:

- in che misura il significato di un verbo determina il suo comportamento sintattico?
- in che misura questo comportamento sintattico è predicibile? Quali componenti del significato del verbo figurano nelle generalizzazioni rilevanti?

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 viene identificata, si possono esaminare i suoi membri per isolare le componenti di significato comuni.

Questa tecnica di indagine è importante perché permette di indagare il significato verbale, che è un componente impalpabile rispetto all'elemento sintattico che viene invece immediatamente realizzato, senza basarsi sulla semplice introspezione. Le distinzioni indotte dalle alternanze diatetiche aiutano appunto a fornire degli spunti sul significato e più in generale sull'organizzazione del lessico trattato, che mostra, spesso non a livello superficiale, inattese somiglianze e differenze tra i verbi ad esso appartenenti.

Un esempio eclatante ci viene dai cosiddetti *verbi di movimento*, che spesso costituiscono un'unica grande classe nelle varie classificazioni dei verbi inglesi. Invece uno studio sul comportamento sintattico di questi verbi (Levin e Rappaport Hovav (1992)) ha dimostrato come questa classe non sia affatto un gruppo omogeneo al suo interno. Essa include difatti almeno due sottoclassi:

1. una composta dai verbi che esprimono la *direzione* del movimento (ad esempio *arrive*, *come*, *go*);
2. un'altra per indicare la *modalità* del movimento (ad esempio *jump*, *run*, *trot*, *skip*).

Levin, in virtù di questa complessità organizzativa, classifica i verbi di movimento articolandoli in più sottogruppi; di seguito vengono proposte, a livello esemplificativo, le classi individuate dalla ricercatrice per i verbi di movimento che non implicano l'uso di un veicolo (Levin (1993)):

**VERBS OF MOTION**

- **Verbs of inherently directed motion**

  *membri della classe*:

  advance, arrive, ascend, ?climb, come, ?cross, depart, descend, enter, escape, exit, fall, flee, go, leave, plunge, recede, return, rise, tumble

  *proprietà:*

  a. The convict escaped

  b. Locative Preposition Drop Alternation (alcuni verbi):

     1. The convict escaped from the police
     2. The convict escaped the police

  c. *Causative Alternations:

     1. The convict escaped
     2. The collaborators escaped the convict

  d. *Measure Phrase:

     *The convict escaped three miles

  e. Adjectival Perfect Participle:

     1. an escaped convict (a convict that has escaped)
     2. *an escaped jail

f. Depictive Phrase

The convict escaped exhausted

(con l'interpretazione per cui colui che scappa si sente sfinito)

g. *Resultative Phrase:

The convict escaped exhausted

(con l'interpretazione per cui è lo scappare che sfinisce chi fugge)

*commenti:*

il significato di questi verbi include la specificazione della direzione del movimento, anche in assenza di un complemento direzionale esplicito. Nessuno di questi verbi specifica il modo del movimento. Comunque i membri di questa classe non devono comportarsi analogamente sotto tutti gli aspetti; possono ad esempio distinguersi per ciò che riguarda la meta, la fonte, il percorso del movimento. A seconda del verbo considerato, questi tratti possono essere espressi attraverso una frase preposizionale, un oggetto diretto, o entrambe queste alternative.

- **Leave Verbs**

*membri della classe:*

abandon, desert, leave

*proprietà:*

a. We abandoned the area

b. * We abandoned from the area

c. Adjectival Passive Participle (alcuni verbi):

an abandoned house

*commenti:*

questi verbi non specificano il modo del movimento che descrivono. Indicano semplicemente un movimento che si sposta rispetto al punto in cui è iniziato. L'oggetto diretto di questi verbi rappresenta il luogo che viene lasciato. Tale luogo non può essere espresso attraverso una frase preposizionale.

- **Manner of Motion Verbs**

questi verbi descrivono solitamente dei movimenti che tipicamente, ma non necessariamente, implicano uno spostamento, ma nessuno di essi fornisce un'indicazione sulla direzione intrinseca del movimento. I membri di questa classe differiscono tra loro solo per il mezzo o il modo dello spostamento.

- **Roll Verbs**

*membri della classe:*

bounce, drift, drop, float, glide, move, roll, slide, swing

MOTION AROUND AN AXIS: coil, revolve, rotate, spin, turn, twirl, twist, whirl, wind

*proprietà:*

a. The ball rolled

b. The ball rolled down the hill/over the hill/into the gutter

c. Causative/Inchoative Alternation (per la maggior parte dei verbi):

1. Bill rolled the ball down the hill
2. The ball rolled down the hill

d. Locative Preposition Drop Alternation:

1. The ball rolled down the hill
2. * The ball rolled the hill

e. Resultative Phrase:

1. The drawer rolled open
2. *The cart rolled the rubber off its wheels
3. * The cart rolled its way down the hill

f. Adjectival Passive Participle:

a constantly rolled ball

*commenti:*

questi verbi esprimono un movimento tipico delle entità inanimate; in assenza di una frase preposizionale, nessuno di essi specifica la direzione del movimento.

- **Run Verbs**

*membri della classe:*

amble, backpack, bolt, bounce, bound, bowl, canter, carom, cavort, charge, clamber, climb, clump, coast, crawl, creep, dart, dash, dodder, drift, file, flit, float, fly, frolic, gallop, gambol, glide, goosestep, hasten, hike, hobble, hop, hurry, hurtle, inch, jog, journey, jump, leap, limp, lollop, lope, lumber, lurch, march, meander, mince, mosey, nip, pad, parade, perambulate, plod, prance, promenade, prowl, race, roam, roll, romp, rove, run, rush, sashay, saunter, scamper, scoot, scram, scramble, scud, scurry, scutter, scuttle, shamble, shuffle, skidle, skedaddle, skip, skitter, skulk, sleepwalk, slide, slink, slither, slog, slouch, sneak, somersault, speed, stagger, stomp, stray, streak, stride, stroll, strut, stumble, stump, swagger, sweep, swim, tack, tear, tiptoe, toddle, totter, traipse, tramp, travel, trek, troop, trot, trudge, trundle, vault, waddle, wade, walk, wander, whiz, zigzag, zoom

*proprietà:*

a. The horse jumped over/across/into/out of the strema

b. Induced Action Alternation (alcuni verbi):

1. The horse jumped over the fence

   Tom jumped the horse over the fence
2. The lions jumped over through the hoop

   The lion tamer jumped the lions through the hoop

c. Locative Preposition Drop Alternation (alcuni verbi):

1. The horse jumped over the stream
2. The horse jumped the stream

d. *There*-Insertion:

1. A little white rabbit jumped out of the box
2. There jumped out of the box a little white rabbit

e. Locative Inversion:

1. A little white rabbit jumped out of the box
2. Out of the box jumped a little white rabbit

f. Measure Phrase (alcuni verbi):

We walked five miles

g. Resultative Phrase:

1. We walked ourselves into a state of exhaustion

   * We walked into a state of exhaustion
2. Tom ran the soles off his shoes

h. Adjectival Passive Participle (alcuni verbi):

the jumped/run/galloped horse

(un cavallo che viene fatto saltare/correre/galoppare da qualcuno)

i. *Adjectival Perfect Participle:

*the jumped horse

(se si interpreta come un cavallo che ha appena saltato)

l. *Cognate Object:

*The horse jumped a high jump

m. Zero-related Nominals:

a jump, a run, a walk

*commenti:*

molti di questi verbi descrivono le modalità di movimento di entità animate, sebbene alcuni possano andare bene anche per le entità inanimate. L'elemento direzionale non viene mai implicitamente espresso, a meno che non sia specificato da un'esplicita frase direzionale.

I verbi appartenenti alla stessa classe sono "sintatticamente sinonimi", vale a dire che sono sostituibili nello stesso insieme di *frames* sintattici, anche se non necessariamente negli stessi contesti.

Inoltre non è sempre detto che se i membri di una classe si differenzino solo per una o due proprietà (ma condividono tutte le altre), sia opportuno creare delle sottoclassi.

In sintesi lo studio delle alternanze diatetiche ci aiuta nell'identificazione di componenti del significato linguisticamente rilevanti, che determinano il comportamento sintattico del verbo. Per individuare l'insieme completo delle componenti di significato che

compaiono nella rappresentazione lessicale del significato verbale, l'indagine sulle proprietà sintattiche semanticamente rilevanti e il conseguente raggruppamento dei verbi in classi, necessitano di essere fatti su un numero se possibile sempre più ampio di verbi che compaiono in una vasta gamma di costruzioni.

Levin (1993) suddivide il proprio lavoro in due fasi:

1. Individuazione di una lista di alternanze diatetiche rilevanti per la conoscenza lessicale del parlante inglese;
2. Elaborazione di un consistente numero di classi semanticamente coerenti, derivate osservando principalmente la loro rispondenza rispetto alle alternanze diatetiche.

Le classi verbali proposte nascono così, 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.

Ad esempio i componenti del significato *contatto* e *movimento* sono comuni tanto agli *hit verbs*, quanto ai *cut verbs*, in base a quanto manifestato dalla loro partecipazione all'alternanza diatetica conativa. Il solo componente *contatto* è riconoscibile però anche nei *touch verbs* oltre che nelle altre due classi appena menzionate.

Risulta evidente da quanto detto finora 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. La parte più complessa nel raggiungimento di questo obiettivo è lo sviluppo di una teoria valida sulla realizzazione argomentale, nonché di una correlata a quest'ultima sulla rappresentazione della semantica lessicale. Ciò permetterà di rendere conto in modo adeguato sia delle variazioni, che dei tratti uniformi nella realizzazione delle alternanze argomentali.

### 1.4 Le classi verbali proposte da Elisabetta Jezek

Il lavoro proposto dalla Jezek intende trattare il problema della classificazione del verbo italiano, allo scopo di verificare se e fino a che punto le proprietà semantiche ed aspettuali del verbo determinino i comportamenti argomentali. Inoltre lo studio si prefigge di stabilire quali tratti semantici e aspettuali determinino alcuni tipi di restrizioni per le diverse alternanze argomentali consentite dai singoli verbi, e ancora di isolare nei vari casi la loro natura lessicale o co-composizionale.

L'area di indagine del progetto interessa tutto il lessico verbale della lingua italiana; l'*input* iniziale è costituito da 15 liste di verbi ottenute tramite l'interrogazione elettronica di un *database* dell'italiano. Ogni lista contiene verbi che presentano lo stesso tipo di alternanza argomentale, valutata in base a tre parametri: Transitività, Intransitività, Pronominalità (avremo così verbi solo transitivi come *allevare*, verbi transitivi ed intransitivi ma non pronominali come *scoppiare*, verbi intransitivi e pronominali ma non transitivi come *sedere* etc.).

Lo scopo principale dello studio è quello di verificare se gli eventi espressi da verbi che rientrano nelle stesse liste condividono o meno uno o più tratti semantici o aspettuali, e valutare, in caso di riscontro positivo, la matrice lessicale o composizionale dei tratti presenti (Jezek (2003)).

(22) si consideri il caso del verbo *chiudere*, la cui analisi condotta attraverso il filtro delle alternanze argomentali, consente di notare come le proprietà di sottocategorizzazione del verbo (TR=transitivo, INACC=in accusativo, INERG=inergativo), siano assegnate composizionalmente in base alla natura semantica del Soggetto:

*chiudere*

| | |
|---|---|
| TR | Mauro ha chiuso la portiera |
| INACC | la finestra si è chiusa |
| INERG | il cinema ha chiuso |

* la finestra ha chiuso

* il cinema si è chiuso

Per tutte le aree verbali indagate, che vanno poi a costituire un'unica macroarea di interesse, il principale obiettivo resta sempre quello di ricostruire, per quanto possibile, *mappe* di corrispondenza tra proprietà del significato e comportamenti sintattici, ed arrivare così ad individuare classi di verbi sintoniche rispetto a tali corrispondenze.

Nel portare avanti l'analisi, per classificare i vari casi e dar loro successivamente una valida interpretazione, sono stati adottati dei formalismi distinti sulla base dell'informazione rappresentata (semantica o sintattica).

- *La rappresentazione dell'informazione sintattica*

  Essa distingue tra *categoria* e *sottocategoria grammaticale* e si ispira ai quadri di sottocategorizzazione proposto inizialmente da Chomsky (1965).

(23) *aprire*

| | | |
|---|---|---|
| CAT GRAMM: | V | |
| SUBCAT GRAMM: | TR | [$SN_1$ V $SN_2$] |
| | INTR | [$SN_2$ V] |
| | INTR PRON | [$SN_2$ SI V] |

Questo genere di rappresentazione è di tipo descrittivo e non esplicativo, perché descrive le varie configurazioni sintattiche che un verbo può assumere, ma non formula ipotesi sui rapporti che si instaurano tra di esse.

- *La rappresentazione dell'informazione semantica*

  Per descrivere questo tipo di informazione si è dovuto far ricorso a tre diversi formalismi:

  1) il primo di tipo decomposizionale, si basa sull'idea di primitivo semantico e trova la sua origine nella semantica componenziale (Dowty (1979), van Valin (1990), van Valin e La Polla (1997), Jackendoff (1972)). I primitivi semantici utilizzati si ispirano per lo più a Dowty e si compongono come di seguito:

| | | |
|---|---|---|
| FA | indica | AZIONE |
| CAUSA | indica | CAUSATIVITÀ |
| DIVENTA | indica | AVVENIMENTO |

(24) consideriamo i seguenti esempi di rappresentazione semantica:

| | |
|---|---|
| *sciogliersi* | [DIVENTA y <pred>] |
| *giocare* | [FA x <pred>] |

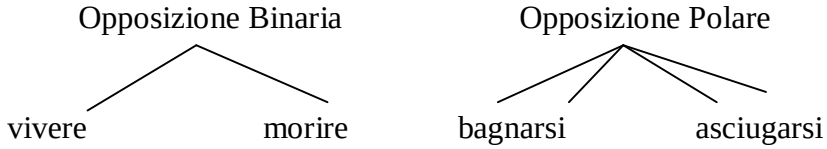
Come in molti altri quadri teorici, gli argomenti associati ad ogni predicato sono statu indicati con le lettere *x, y, z*. La *x* rappresenta sia il primo argomento di una frase transitiva, che l'unico argomento di alcune frasi intransitive (inergative); la *y* rappresenta il secondo argomento di una frase transitiva o l'unico argomento di altre frasi intransitive (inaccusative); infine la *z* costituirà il terzo attante di una frase transitiva.

2) il secondo è composto da una lista di ruoli semantici che descrivono le relazioni che intercorrono tra il verbo ed i suoi argomenti, che si suppone siano in molti casi assegnate a livello contestuale. Sulla base di alcuni studi proposti in questo ambito (Dik (1989), van Valin (1997)), sono stati selezionati i seguenti ruoli:

| | | |
|---|---|---|
| Agent | Ag | Agente |
| Theme | Th | Tema |
| Patient | Pt | Paziente |
| Experiencer | Exp | Esperiente |
| Source | Source | Origine |
| Beneficiary | Ben | Beneficiario |
| Goal | Goal | Destinatario |

3) l'ultimo è un formalismo ad albero che rappresenta le opposizioni semantiche codificate nei predicati (Pustejovsky (2001)); considerando la loro struttura interna, sono state scelti due tipi di opposizione: quella binaria (che prevede unicamente due poli) e quella polare (che prevede degli stadi intermedi).

(25)




Nello spoglio delle 15 liste contenenti verbi con comportamento argomentale analogo, volto a individuare le correlazioni esistenti tra elementi della forma e elementi del significato, e a definire la natura lessicale o composizionale dei tratti rilevanti, ci interessano particolarmente:

- la variazioni nella realizzazione degli argomenti ovvero le alternanze argomentali. Le variazioni prese in considerazione riguardano gli argomenti sintatticamente *obbligatori*, cioè necessari a garantire grammaticalità e compiutezza semantica all'enunciato. L'elemento verrà considerato come obbligatorio in base al contesto d'uso in cui esso compare, ovvero in base al significato che l'elemento lessicale genera in tale contesto;
- le variazioni nella configurazione sintattica, come la selezione dell'ausiliare (avere o essere) e della marca pronominale ([+si] vs [-si]) negli usi intransitivi;
- le variazioni di significato connesse all'alternanza sintattico-argomentale.

In sostanza l'analisi delle liste dei verbi italiani, distinte in base al tipo di alternanze presentate lungo l'asse della Transitività-Intransitività-Pronominalità, ha consentito di individuare interessanti correlazioni tra *tipo di alternanza sintattica* e presenza di specifiche *proprietà semantiche e aspettuali*. In particolare è emerso come alcuni tratti aspettuali, ad esempio la presenza/assenza di telicità, emergano prepotentemente nella sintassi, condizionando il *pattern* di alternanza presentato dai verbi.

I dati ottenuti mostrano però come non tutte le proprietà semantiche e aspettuali delle entrate verbali siano assegnate a livello lessicale, e come alcune di loro siano assegnate a livello composizionale. Difatti, se in alcuni casi è possibile individuare correlazioni sistematiche tra le interpretazioni soggiacenti a due o più realizzazioni sintattiche consentite

dallo stesso elemento lessicale (ad esempio la regola che permette di esprimere o meno il componente causale), non sempre queste possono essere sufficienti per contenere l'intera semantica della stesso verbo, che in combinazione con altri Soggetti e Oggetti può dar vita a nuovi e distinti significati.

(26) *cambiare*
   a. Cambiare macchina
   b. Cambiare idea

In generale l'analisi mostra come il piano composizionale abbia una forte rilevanza nella definizione del senso e porta quindi dati importanti a sostegno di questa stessa ipotesi.

Occorre, in ultima analisi, evidenziare due aspetti caratterizzanti del lavoro della Jezek:

1. in primo luogo il *focus* su un ridotto numero di alternanze. Vengono di fatto considerate solo quelle transitivo-intransitivo-inaccusativo;
2. in secondo luogo il metodo di classificazione, che è sostanzialmente sintattico. La Jezek classifica i verbi sulla base del loro essere transitivi, intransitivi etc, dopodichè cerca di ricondurre tali variazioni di classe a qualche tratto semantico. di fatto quindi le sue non sono delle vere e proprie classi semantiche come quelle formulate da Levin.

# Capitolo 2
# La Rappresentazione dei verbi nei lessici computazionali

## 2.1 Che cos'è un'Ontologia

Il tipo di modello per la rappresentazione semantica che andremo a descrivere in questa sezione, è quello *simbolico*, in cui il contenuto semantico delle parole si concretizza in una sua proiezione su un'ontologia di concetti.

Le ontologie non sono altro che sistemi di simboli selezionati allo scopo di rappresentare un insieme di concetti; in tale linguaggio di descrizione della conoscenza, è essenziale specificare quali siano i concetti atomici rilevanti e quali i modi e le forme in cui questi si strutturano.

La definizione di ontologia più largamente accettata è quella di Tom Gruber, secondo il quale

> an ontology is an explicit specification of a conceptualisation.

In questa sede ci limiteremo a considerare nel dettaglio le ontologie linguistiche, ovvero quei sistemi simbolici che si occupano di rappresentare i concetti, nello specifico i significati, codificati nelle espressioni del linguaggio naturale.

> A set of knowledge terms, including the vocabulary, the semantic interconnections and some simple rules of inference and logic (Hendler 2001).

Schematicamente, quello che si vuole esprimere attraverso questo tipo di ontologie sono le seguenti caratteristiche:

- una conoscenza lessicale, formata da un insieme di parole, intese come stringhe di caratteri;
- una conoscenza semantica, che raccoglie in sé i significati delle parole e le relazioni che intercorrono fra di esse.

Le principali ontologie finora costruite nascono originariamente come lessici computazionali e prendono il nome di EuroWordNet, WordNet, FrameNet etc. Vediamo nel dettaglio le caratteristiche fondanti di alcune tra queste.

### 2.1.1 Il progetto WordNet

Il lavoro proposto da Levin, e quelli prodotti in seguito sulla stessa linea teorica, si fondano su una base *sintagmatica*, ovvero sulla struttura interna delle parole che determina dove esse possono posizionarsi e che cosa possono fare all'interno di un enunciato. In effetti in essi vengono evidenziate le caratteristiche intrinseche dei verbi, come la loro struttura argomentale o la loro capacità di realizzare o meno una certa alternanza diatetica; tali studi non ci dicono però molto sul rapporto che questi verbi intrattengono tra di loro o tra le connessioni che si instaurano tra le varie classi semantiche.

Il progetto WordNet, al contrario, si fonda su un altro tipo di struttura a base *paradigmatica*, ovvero sulle relazioni che le parole intrecciano tra di loro. Infatti, come vedremo in seguito più nel dettaglio, le parole in WordNet sono entità concettuali che assumono un posto specifico all'interno dell'ontologia, in virtù dei rapporti che stabiliscono con gli altri termini della rete (sinonimia, antonimia, toponimia etc.).

WordNet è un lessico computazionale concepito per la lingua inglese basato sulle attuali teorie psicolinguistiche (Harley (2008)); sviluppato a partire dal 1985 presso il Cognitive Science Laboratory dell'Università di Princeton (Fellbaum (1998)), esso è in grado di distinguere i diversi significati di una parola e di produrre una lista di *synonym sets* e di glosse, utili proprio a differenziare i vari sensi.

Nel 1985, agli albori del progetto, psicologi cognitivi e linguisti computazionali, parlavano del significato delle parole in termini di reti e diagrammi composti da:

- nodi rappresentanti i significati;
- e da archi indicanti le relazioni intercorrenti fra i significati.

(1) *Table* e *furniture* corrisponderanno a due nodi nell'ontologia, collegati tra loro da un arco che esprimerà la relazione *A table is a kind of furniture*.

La semantica lessicale di tipo relazionale sfrutta quindi relazioni come quella di iperonimia, per descrivere il significato delle parole, procedimento questo alternativo a quello adottato dalla semantica lessicale composizionale, che basa la sua descrizione sulla scomposizione del significato nei costituenti che lo compongono. Nelle prime fasi del progetto, WordNet nasce proprio come uno strumento atto a testare la validità di questo approccio alternativo al lessico, perciò la lista delle entrate lessicali era ridotta a quelle ritenute più utili su questo fronte.

Un antecedente importante di WordNet sono le reti semantiche, quali quelle elaborate da Quillian, Collins etc. Esse costituiscono una classe di sistemi di rappresentazione, basati sull'idea generale di utilizzare come strumento di rappresentazione un grafo, in cui ad ogni nodo è associata un'entità concettuale di qualche tipo (ad esempio un concetto, il significato di un enunciato, o il significato di un elemento lessicale). Le relazioni, di tipo logico o associativo, fra entità concettuali diverse sono rappresentate mediante gli archi che connettono i nodi. Al di là di questa caratterizzazione generale, i vari tipi di rete semantica sono molto diversi fra loro. I tipi di nodi e di archi che è possibile utilizzare, la loro interpretazione, le regole sintattiche che consentono di comporli in una rete, ed i meccanismi di inferenza che sono definiti sulle reti variano notevolmente nei molteplici sistemi di rappresentazione che sono stati via via elaborati. Le reti semantiche furono sviluppate originariamente in base a motivazioni di carattere psicologico, e spesso, soprattutto nel corso degli anni settanta, utilizzate dagli oppositori della logica in contrapposizione ai classici formalismi logico matematici. Rispetto a tali formalismi, le reti semantiche infatti avrebbero dovuto consentire di costruire basi di conoscenza con una struttura associativa più simile a quella ipotizzabile per la memoria umana (ad esempio, utilizzando gli archi della rete per rappresentare non soltanto relazioni di tipo puramente logico, ma anche la "distanza concettuale" fra due rappresentazioni). Inoltre, la maggior parte dei sistemi a rete semantica (e più ancora i sistemi a *frame*, che delle reti semantiche sono parenti prossimi) prevedevano la possibilità di rappresentare concetti per mezzo di tratti prototipici anziché esclusivamente mediante condizioni necessarie e/o sufficienti. Questi fattori, e inoltre il fatto che in origine le reti semantiche venissero utilizzate soprattutto come formalismi per la rappresentazione del significato in programmi per la comprensione del linguaggio naturale e come modelli psicologici per la rappresentazione di concetti lessicali hanno fatto sì che *frame* e reti semantiche fossero visti come la proposta sviluppata in Intelligenza Artificiale per rispondere al problema della rappresentazione del significato lessicale.

I lavori di Quillian sulla memoria associativa (Quillian 1967, 1968) costituiscono il punto di partenza universalmente riconosciuto delle ricerche sulla rappresentazione della conoscenza mediante reti semantiche in IA. Gli interessi di Quillian erano di tipo eminentemente psicologico. Il suo scopo era di fornire un modello dell'organizzazione della memoria semantica di un essere umano, in modo da rappresentare il significato lessicale di termini del linguaggio naturale, e di eseguire vari tipi di inferenza a partire da tali rappresentazioni. Ciò avrebbe dovuto consentire la simulazione di alcune capacità linguistiche umane, quali il confronto del significato di due parole, o la "comprensione" di un testo in linguaggio naturale (dove per comprensione si intende la costruzione automatica di una rappresentazione del significato a partire dal testo assunto in *input*). L'intento era di rappresentare esclusivamente la componente "oggettiva", non emotiva o puramente soggettiva, del significato. I dati di partenza per la costruzione della base di conoscenza erano forniti dalle definizioni di un dizionario. Nonostante i suoi interessi specificamente rivolti al

linguaggio naturale, Quillian ipotizzava che il tipo di rappresentazione da lui proposto avesse una validità che andasse oltre l'ambito linguistico, e assumeva che la struttura della memoria semantica fosse la stessa della memoria generale. Nel modello di Quillian esistono due generi principali di nodi: *nodi tipo* (*type node*) e *nodi esemplare* (*token node*, o, semplicemente, *token*). Il significato lessicale di ogni voce è rappresentato mediante un *type node*. Ad ogni *type node* corrisponde nella rete un *piano* (*plane*), vale a dire una struttura che rappresenta la descrizione del significato corrispondente. Il piano corrispondente ad un *type node* circoscrive la porzione di rete che rappresenta la definizione corrispondente. Come in un dizionario ogni voce è definita utilizzando altre voci definite altrove nel dizionario stesso, così, per definire una voce lessicale in un piano, si fa riferimento alla rappresentazione di altre voci definite nella rete. Ciò avviene mediante nodi *token*, i quali consentono di far riferimento dall'interno di un piano ad altre definizioni presenti nella rete. Ogni *token* è collegato da un arco di tipo opportuno al *type node* che ne esprime la definizione. Quindi, mentre ad ogni significato di una voce lessicale corrisponde nella rete al più un solo *type node*, non esiste un limite al numero di *token* di una voce presenti nel modello: essi sono tanti quante le volte che la voce viene utilizzata in altre definizioni.

Le relazioni fra i vari nodi *token* che concorrono a una definizione all'interno di un piano sono espresse mediante archi che appartengono ad un insieme predefinito di tipi:

1. archi che rappresentano una relazione di sottoclasse;
2. *modification pointer*: sono archi che consentono di utilizzare un *token* per modificare avverbialmente o come aggettivo un secondo *token*;
3. archi che esprimono relazioni di *disgiunzione* fra nodi; essi possono essere utilizzati per rappresentare i molteplici significati di una parola oppure per rappresentare espressioni che denotano un insieme ottenuto per disgiunzione;
4. archi che esprimono relazioni di *congiunzione*;
5. archi che consentono di rappresentare relazioni arbitrarie fra due *token* utilizzando un terzo *token* per esprimere il tipo di relazione che tra essi sussiste; essi vengono raffigurati con frecce doppie.

Il formalismo di Quillian presenta altre caratteristiche, alcune delle quali verranno largamente riprese e sviluppate nelle reti semantiche successive. E' possibile associare ai nodi *token* vari tipi di *specification tag*, vale a dire valori numerici che servono ad indicare, ad esempio, il numero di istanze che deve avere un *token*, o la sua rilevanza nella definizione di un concetto. Secondo Quillian, espressioni come *a, six, much, very, probably, not, perhaps* dovrebbero essere rappresentate non come nodi distinti, ma come *specification tag* associate ai *token* di un piano.

Alla definizione di una parola possono essere associate *clue words*, che indicano con quali altri concetti un concetto deve essere messo in relazione nei casi tipici. Ad esempio, la definizione del verbo *to comb* ha come oggetto la *clue word hair*, per indicare che di solito l'oggetto di *to comb* è *hair*. Le *clue words* sono quindi le antesignane dei *default value* dei sistemi di rappresentazione successivi, e in particolare dei *frame*s, sono cioè valori tipici che si assumono come veri in mancanza di informazioni contrarie più specifiche.

Abbiamo visto che ogni piano contiene la definizione di un concetto lessicale. Tale definizione non è tuttavia completa, in quanto i vari *token* che vi compaiono rimandano a loro volta ad altri piani, cioè ad altre definizioni rappresentate nel modello. La parte della definizione compresa in un singolo piano viene detta da Quillian la *definizione immediata* di una parola. Alla definizione immediata si contrappone il *concetto completo* di una parola (*full word concept*), che comprende tutti i nodi *type* e *token* che contribuiscono, in maniera diretta o indiretta, alla definizione. Il *full word concept* di una parola si ottiene partendo dal nodo *type* della parola stessa (detto nodo "patriarca"), si percorrono quindi tutti gli archi del piano che le corrisponde e, per ogni *token* raggiunto, si visitano tutti i piani dei relativi nodi *type*, ripetendo ricorsivamente l'operazione e visitando, ogni volta che si incontra un *token*, il piano

del *type* corrispondente. Un insieme non strutturato di nodi *type* e *token* non costituisce un'adeguata rappresentazione di un *full concept*, poiché mancano completamente le informazioni sulle relazioni fra i nodi. E' quindi necessario, man mano che la ricerca procede, mantenere traccia anche dei vari archi che sono stati percorsi.

> Nel modello della memoria il concetto completo di una parola è definito come l'insieme di tutti i nodi che possono essere raggiunti tramite un processo esaustivo di percorrimento avente origine nel corrispondente *type node* patriarca, assieme alla somma totale delle relazioni fra questi nodi specificate da archi fra *token* e *token* di uno stesso piano" (Quillian 1968: 238).

Si noti che questo processo di ricerca può partire da qualsiasi nodo *type* della rete, poiché a ogni nodo *type* sono associati dei *token* che rimandano a loro volta ad altri *type*. Non ci sono infatti nodi primitivi nelle reti di Quillian, ed ogni voce è definita sempre nei termini di altre definizioni nella rete (così come in un dizionario ogni voce richiama sempre qualche altra voce).

Lo scopo dichiarato del modello di Quillian era quello di simulare la comprensione di enunciati espressi nel linguaggio naturale. Fornendogli in *input* un enunciato sconosciuto, il sistema avrebbe dovuto essere in grado di ricavare una rappresentazione del suo significato sulla base delle definizioni presenti nella rete semantica. Le prestazioni effettive del modello sono tuttavia molto più limitate: di fatto, esso è in grado soltanto di confrontare il significato di due voci lessicali. Sottoponendogli due parole la cui definizione è rappresentata nella rete, il programma individua le più importanti somiglianze e differenze fra i loro significati. Dopo di che, una seconda componente del programma genera un enunciato, scritto in un sottoinsieme semplificato dell'inglese, che esprime tali somiglianze e differenze.

Il meccanismo inferenziale utilizzato per confrontare due parole si basa su una forma di *attivazione diffusa* (*spread activation*) dei nodi della rete. Il confronto avviene individuando i punti in cui i *full concepts* delle due parole si intersecano. Partendo dai nodi patriarca delle due definizioni da confrontare, il programma procede visitando man mano i nodi vicini (siano essi *type* o *token*) ai quali può accedere percorrendo i vari tipi di archi associativi. I nodi via via visitati vengono "attivati", etichettandoli per mezzo di una *activation tag*. In questo modo, attorno a ciascun patriarca si crea una zona di nodi attivati che si espande lentamente in tutte le direzioni. Si noti infatti che il processo di attivazione procede alternativamente su ciascuno dei due concetti da confrontare, in modo da simulare un'elaborazione eseguita parallelo. Ogni *activation tag* comprende il nome del nodo patriarca da cui è partito il processo di attivazione. Questo consente di evitare circoli viziosi e di identificare le intersezioni. Ogni volta che un nodo viene raggiunto da un processo di attivazione, il programma controlla se quel nodo era già stato attivato a partire dallo stesso patriarca. In tal caso quel percorso viene abbandonato, in quanto già visitato precedentemente. Se invece quel nodo presenta la *activation tag* col nome di un altro patriarca, allora è stato identificato un punto in cui i due *full concept* si intersecano. Infine, se il nodo non presenta alcuna *activation tag*, allora il nodo viene etichettato, e il processo di attivazione continua. Ogni *activation tag* comprende anche il nome del nodo visitato immediatamente prima durante il processo di attivazione. Questo consente, una volta individuata un'intersezione fra le sfere di attivazione di due concetti, di ricostruire i percorsi (*path*) che dai patriarchi conducono al nodo d'intersezione. Tali percorsi sono necessari al sistema per generare le proprie risposte. I *path* che conducono dai nodi patriarca al nodo intersezione vengono utilizzati per generare le risposte del sistema. Ogni risposta viene costruita sulla base dei nodi del percorso e dei tipi di archi che li collegano (utilizzando talvolta anche le informazioni contenute nelle immediate vicinanze dei nodi percorsi).

Tornando al progetto WordNet, risulta subito evidente come in esso non troviamo solo verbi, ma una serie di parole appartenenti a più parti del discorso (nomi, verbi, aggettivi,

avverbi), ed inoltre in esso non viene trattato solo il senso primario di un termine, bensì l'insieme dei molteplici significati che lo caratterizzano.

L'idea iniziale era proprio quella di fornire agli utenti una risorsa *on line* più spendibile rispetto ai dizionari tradizionali disponibili in rete, che consentivano soltanto una semplice ricerca alfabetica.

La differenza più lampante tra questi due strumenti è che, se il dizionario organizza le parole disponendole alfabeticamente per facilitarne la consultazione, WordNet invece memorizza le informazioni in base al loro significato e alla categoria sintattica di appartenenza (nomi, verbi, avverbi, aggettivi), collegandole tra di loro tramite vari tipi di relazioni (sinonimia, antonimia, troponimia etc.).

WordNet divide il significato di una parola in due concetti:

1. quello di *Word Form*, ovvero la forma scritta del termine;
2. quello di *Word Meaning*, vale a dire il concetto espresso da tale termine.

Quindi il punto di inizio della classificazione delle parole secondo WordNet, sono le relazioni che si stabiliscono fra un lemma ed il suo significato.

La base di questa teoria risiede nella cosiddetta Matrice Lessicale (Figura 2.1): nelle righe vengono elencati i significati delle parole e nelle colonne i lemmi. Ad esempio per la colonna relativa alla Word Form *function*, i possibili Word Meaning, corrispondenti alle righe della Matrice Lessicale, potrebbero essere: *mathematical relation, subroutine, religious cerimony*, quando *function* appartiene alla categoria sintattica dei nomi, oppure *operate* e *officiate*, nel caso sia utilizzata come verbo.

*Figura 2.1 La Matrice Lessicale di WordNet*

| Word Meanings | Word Forms | | | | |
|---|---|---|---|---|---|
| | $F_1$ | $F_2$ | $F_3$ - - - - - - | | $F_n$ |
| $M_1$ | $E_{1,1}$ | $E_{1,2}$ | | | |
| $M_2$ | | $E_{2,2}$ | | | |
| $M_3$ | | | $E_{3,3}$ | | |
| ⋮ | | | | ⋱ | |
| $M_m$ | | | | | $E_{m,n}$ |

La presenza di un valore non nullo di Eij all'interno della matrice implica che la forma Fj viene espressa dal significato Mi. Se abbiamo più valori non nulli sulla stessa riga ci troviamo davanti a dei sinonimi: ovvero forme diverse con lo stesso significato; più valori non nulli sulla stessa colonna esprimono invece un concetto di polisemia: vale a dire una stessa forma Fj con più significati.

WordNet è quindi organizzato su *relazioni semantiche* che coinvolgono i rapporti fra significati (rappresentati nei cosiddetti *synsets*) e su *relazioni lessicali* che stabiliscono i rapporti tra i singoli lemmi (ovvero tra le forme delle parole).

La costruzione della base di dati si è confrontata con l'esistenza di due teorie: la *teoria costruttiva* e la *teoria differenziale*.

Secondo la *teoria costruttiva* un'accurata costruzione di un concetto deve essere supportata da un numero sufficiente di informazioni. Tali informazioni devono consentire di caratterizzarlo in modo da distinguerlo da altri possibili concetti lessicali e di fornirne una corretta definizione.

La teoria differenziale, molto meno rigida, prevede che la rappresentazione di un concetto possa essere espressa solo con elementi che permettano di distinguerlo da altri.

Per essere più chiari possiamo ricorrere ad un esempio: la parola *pianta*.

Essa può avere i seguenti significati:

- *nome generico che indica qualsiasi vegetale fornito di organi specializzati;*

- *proiezione orizzontale di un oggetto;*
- *parte inferiore del piede;*

Nella teoria costruttiva per differenziare i due significati dobbiamo fornire un numero sufficiente di informazioni che ci consentano di distinguerli.

In quella differenziale basta fornire una lista di forme che lo possano esprimere.

Il significato M può essere quindi espresso con una lista di forme (F1, F2, …); in questo modo abbiamo per ogni significato, una lista di forme fra di loro in relazione di sinonimia. Questo insieme viene indicato appunto come *Synonym Set*, o meglio conosciuto nella sua forma contratta di *Synset*.

Ritornando al nostro esempio, per distinguere i due significati sarebbe stato sufficiente citarne due sinonimi: *vegetale* per il primo, e *mappa* per il secondo. Nel caso in cui non esista un sinonimo appropriato a differenziare quel significato da altri, si fa ricorso ad una glossa ovvero una breve spiegazione del significato. Per il terzo significato del nostro esempio si potrebbe utilizzare la glossa: *parte inferiore del piede*.

WordNet si basa su delle relazioni semantiche fra concetti, fra le quali la sinonimia gioca senz'altro un ruolo fondamentale, ma non è l'unica ad essere utilizzata per la costruzione dell'ontologia.

Le relazioni che verranno descritte saranno quelle fondamentali, anche se non le uniche, implementate da WordNet:

· *Sinonimia*
· *Antonimia*
· *Iponimia/Iperonimia*
· *Meronimia/Olonimia*

**Sinonimia**

Una prima definizione della relazione di Sinonimia, peraltro molto rigida, fu elaborata inizialmente dal filosofo Leibniz e recita come segue:

definizione 1: due concetti sono fra loro sinonimi se la sostituzione di uno con l'altro non cambia il valore di verità della frase nella quale viene effettuata la sostituzione.

In base a questa definizione, due parole fra di loro legate da una relazione di sinonimia sono piuttosto rare da individuare in una qualsiasi proposizione data. Si utilizza quindi una definizione più debole, non più legata alla frase, ma al contesto a cui si fa riferimento:

definizione 2: due concetti sono fra loro sinonimi in un contesto linguistico C se la sostituzione di un concetto con l'altro nel contesto C non ne altera il valore di verità (Miller et al. (1990)).

Così la sostituzione della parola *pianta* con *mappa* non ne altera il significato in topografia, ma in altri contesti la stessa sostituzione potrebbe risultare del tutto inappropriata.

Oltre a questa definizione in termini di vero/falso, esiste un'altra scuola filosofica di pensiero secondo la quale i sinonimi possono essere pensati anche secondo un concetto di *similarità*. Così una relazione di similarità semantica è sufficiente a rendere due concetti fra loro sinonimi.

Ritornando sempre al nostro esempio: *albero*, *arbusto* e *pianta* secondo le definizioni date in precedenza non possono essere fra loro sostituite, in quanto esistono anche piante che non sono necessariamente alberi o arbusti, ma secondo una definizione di similarità essi appartengono allo stesso *synset*, in quanto consentono la distinzione rispetto ad altri significati.

**Antonimia**

Una relazione alla quale non sempre è facile dare una definizione, ma che compare assai di frequente è l'antonimia.

L'antonimo di una parola *x* viene definito quasi sempre in modo assoluto come *not-x*. Così *ricco* e *povero* sono fra loro antonimi, anche se essere non ricchi non implica necessariamente essere poveri: infatti, si può essere al contempo né ricchi, né poveri.

**Iponimia/Iperonimia**

Le relazioni di iponimia e iperonimia sono relazioni fra significati che rispettano la definizione seguente:

Definizione 3: un concetto rappresentato dal *synset* {x1, x2, x3,…} viene detto iponimo del concetto rappresentato dal *synset* {y1, y2,…} se si può accettare una frase costruita come: *Un x è un (un tipo di ) y*.

L'iponimia è transitiva ed antisimmetrica; essa genera una struttura semantica gerarchica secondo la quale gli iponimi denotano una sottoclasse della classe o categoria denotata dall'iperonimo (es. *cane* e *animale*).

**Meronimia/Olonimia**

Questa è una relazione semantica, che esprime il concetto di *parte di.*

Definizione 4: {x1, x2,…} è un meronimo di un concetto rappresentato da {y1, y2,…} se si possono accettare frasi come *x è parte di y*.

La relazione di meronimia è transitiva (con le riserve che verranno espresse più avanti) e antisimmetrica e può anch'essa essere usata per costruire relazioni gerarchiche.

Un esempio dell'applicazione di questo sistema relazionale può essere dato da: *becco* ed *ala* che sono meronimi di *uccello*, mentre *canarino* sarà invece un iponimo dello stesso termine *uccello*. Per il sistema gerarchico, *becco* e *ala* sono essi stessi anche meronimi di *canarino*.

L'applicazione della transitività invece ci porta a dire che se *dita* è meronimo di *mano* e *mano* è meronimo di *arto*, allora *dita* è meronimo di *arto*, ovvero le *dita* sono parte della *mano* e quindi anche dell'*arto*. Questo non è sempre vero è dipende dal tipo di relazione *parte di* che si instaura fra le parti.

Esistono pertanto varie tipologie di relazioni *parte di*, per l'esattezza ne sono state individuate sei:

1. componente/oggetto;
2. elemento/insieme;
3. porzione/intero;
4. materiale/oggetto;
5. azione/attività;
6. località/area.

Un ultimo aggiunto è stato inserito più tardi nel tempo e si tratta del rapporto:

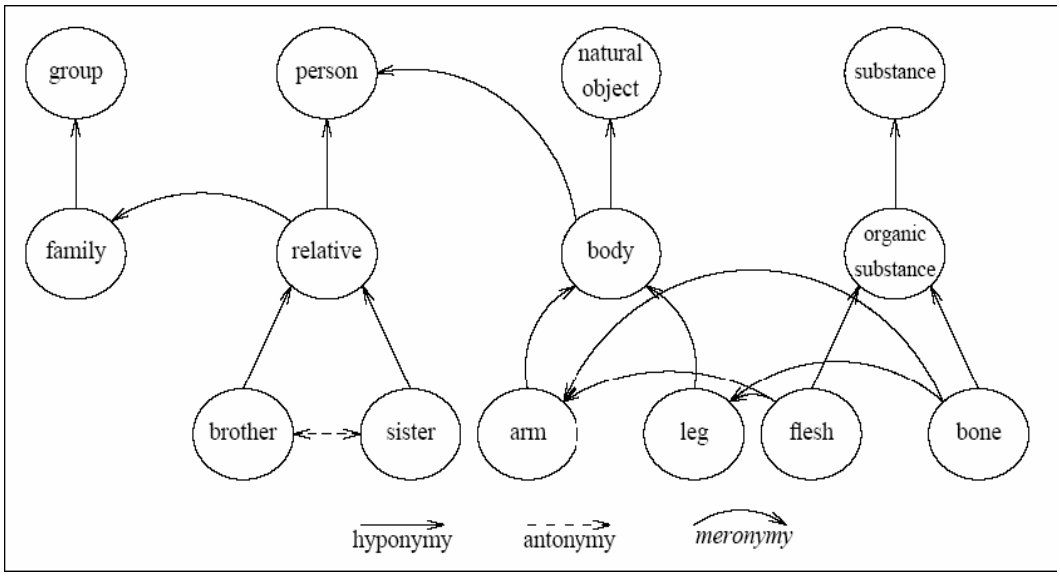
7. fase/processo.

Considerate nel loro insieme, le parole sono state ripartite in quattro reti semantiche, una per ogni classe: nomi, verbi, aggettivi e avverbi; tale suddivisione implica l'assenza in WordNet di qualsiasi tipo di informazione sulle proprietà sintagmatiche delle parole stesse (Fellbaum (1998)).

Ciascuna delle quattro reti semantiche poggia su relazioni semantiche differenti, poiché differenti sono i costituenti che ne compongono il nucleo:

- i modificatori, vale a dire aggettivi e avverbi, sono disposti in coppie di antonimi, come *buono* vs *cattivo*;
- i nomi rispondono principalmente alla relazione gerarchica di iperonimia;
- ed infine per i verbi la relazione più diffusa è quella di troponimia, espressa dalla formula *To V1 is to V2 in some way* (Fellbaum (1998), Widdows (2004)).

*Figura 2.2 Un esempio delle relazioni di meronimia, antonimia ed iponimia rappresentate in WordNet*




L'idea di utilizzare dei *synonym sets* (*synsets*) per rappresentare i concetti lessicali, deriva dalla volontà e dalla necessità di rappresentare quanto più correttamente possibile, le forme che le parole possono assumere ed i rispettivi significati ad esse associabili. Ogni *synset* conterrà tutte quelle parole che esprimono uno stesso concetto, perciò un utente, nel richiamare un determinato concetto attraverso una parola atta ad esprimerlo, potrà rintracciare anche tutte le altre parole che lo lessicalizzano all'interno dell'ontologia. I *synsets* sono organizzati gerarchicamente, in modo che, risalendo la rete semantica a ritroso, si arrivi ai *root nodes* dell'albero linguistico, detti anche *unique beginners* poiché non dipendono da nessun'altro nodo in una scala di generalità dei concetti espressi, e sono pertanto le radici di tutti gli altri rami (o archi) dell'albero che da essi dipendono (Figura 2.3)

*Figura 2.3 Lista degli unique beginners di WordNet per i nomi*

| | |
|---|---|
| {act, action, activity} | {natural object} |
| {animal, fauna} | {natural phenomenon} |
| {artifact} | {person, human being} |
| {attribute, property} | {plant, flora} |
| {body, corpus} | {possession} |
| {cognition, knowledge} | {process} |
| {communication} | {quantity, amount} |
| {event, happening} | {relation} |
| {feeling, emotion} | {shape} |
| {food} | {state, condition} |
| {group, collection} | {substance} |
| {location,place} | {time} |
| {motive} | |

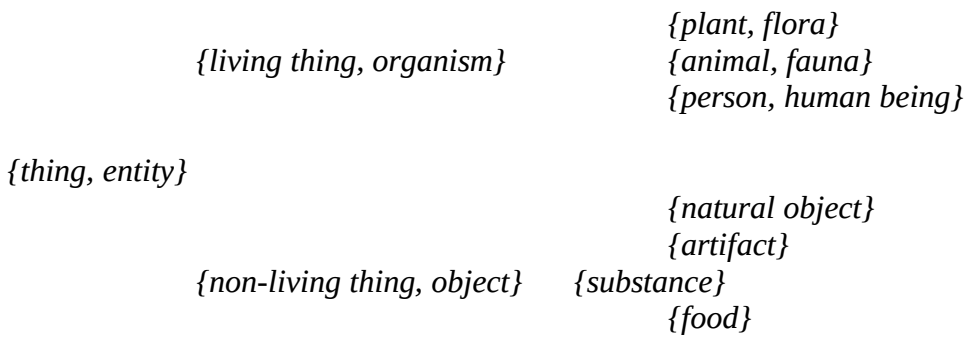
La relazione di iperonimia/iponimia è quella su cui gli studiosi basano la teoria sul sistema di memoria semantica. Esso è di tipo gerarchico e può essere schematizzato come un albero (nel senso grafico del termine).

L'albero viene costruito seguendo la catena di termini in relazione di iponimia. La struttura creata è una sequenza di livelli che va da molti termini specifici al livello più basso a pochi termini generici a livello più alto.

Con questo sistema si ovvia al problema della ridondanza, soprattutto per basi di dati che contengono molti termini: il termine al livello *n* ha tutte le proprietà del termine al livello *n-1* ad esso collegato e ne aggiunge delle altre, e così via scendendo o risalendo nella scala gerarchica. Basta quindi memorizzare solo le informazioni che caratterizzano l'oggetto stesso, mentre si possono tralasciare quelle che già sono memorizzate per il relativo iperonimo.

Secondo la teoria dell'organizzazione gerarchica, la costruzione dell'albero dovrebbe partire da un'unica gerarchia. Se così fosse, il livello più generico, la radice, sarebbe semanticamente pieno. In linea di principio si potrebbe mettere come radice un termine astratto come {entità} e mettere {oggetto, cosa} e {idea} come suoi immediati iponimi; in pratica questo porta a pochissimo contenuto semantico.

*Figura 2.4 Relazioni fra concetti primitivi*



Negli ultimi decenni, l'attenzione di psicologi e linguisti si è però concentrata più che sull'organizzazione dei nomi, sulle caratteristiche tipiche dei verbi. Obiettivo comune è diventato allora:

- caratterizzare la struttura del lessico verbale;
- rappresentare quest'ultimo come una parte fondante della conoscenza linguistica dei parlanti.

WordNet funge dunque da esperimento per valutare se un certo modello del lessico si adatta bene anche al gruppo dei verbi o meno; difatti al suo interno rintracciamo molte informazioni implicite sulle classi verbali contenute e sulle loro proprietà semantiche e sintattiche.

Allo scopo di organizzare il lessico verbale come una rete relazionale, si è scelto, per due ragioni, di suddividere il lessico verbale in *campi semantici*; in primo luogo questa ripartizione fornisce un'organizzazione iniziale semanticamente fondata delle migliaia di verbi polisemici tipici del lessico inglese. In seconda istanza, i ricercatori che si sono occupati dell'individuazione dei suddetti campi semantici, hanno rilevato che le parole che sono legate tra loro da relazioni semantiche e lessicali solitamente appartengono allo stesso campo semantico. Risulta dunque evidente che un'analisi di tipo relazionale del lessico implica necessariamente anche un'analisi di quest'ultimo in riferimento ai campi semantici di appartenenza.

Alcuni campi semantici oggetto di un'attenzione particolare da parte dei linguisti, come quelli contenenti i termini riferiti a piante o colori, si sono rivelati organizzati su relazioni di iponimia. In base a quanto detto, *red* appartiene al campo semantico dei colori perché è un tipo di colore; similmente *sprint* e *run* possono essere assegnati al campo semantico dei *motion verbs*, perché *to sprint* significa *to run* in qualche modo (e *to run* significa a sua volta *to move* in qualche modo).

Dato che la maggior parte degli studi condotti si sono prevalentemente concentrati sui nomi, non sono ancora state codificate relazioni semantiche e lessicali altrettanto esatte per i verbi. Si è però giunti alla conclusione che, qualsiasi relazione si scelga per collegare tra loro i concetti verbali, sembra essere comunque ragionevole l'intuizione per cui tali relazioni connetteranno primariamente tutti quei verbi appartenenti allo stesso campo semantico.

Quindi la divisione dei verbi in campi semantici, effettuata inizialmente su base puramente intuitiva, rivela successivamente delle relazioni fondanti che organizzano i verbi ed i concetti verbali.

Una prima ripartizione viene operata tra:

- i verbi che indicano *azioni* ed *eventi*;
- i verbi che denotano *stati*.

La maggior parte dei verbi appartengono al primo gruppo menzionato e vengono ripartiti nei seguenti 14 campi ancora più specifici*: verbi di movimento, di percezione, di contatto, di comunicazione, di competizione, di cambiamento, di cognizione, di consunzione, di creazione, di emozione, di possesso, di parti e funzioni corporee* ed infine *verbi riferiti a comportamenti ed interazioni sociali*.

Tale classificazione prende spunto:

- in parte da alcune classi verbali considerate come semanticamente correlate, discusse da Miller e Johnson-Laird (1976);
- e in parte da un'idea di classificazione semantica che sembrava appropriata perché capace di sistemare virtualmente tutti i verbi trattati.

I predicati che sono evidenti elaborazioni del concetto di *be*, inclusi *resemble*, *belong* e *soffice*, non possono rientrare nelle 14 classi sopra menzionate; questi verbi stativi costituiscono un insieme semanticamente eterogeneo e quindi sono l'unico gruppo che non costituisce un campo semantico. Al suo interno ritroviamo sia gli ausiliari che i verbi di

controllo come *want, fail, prevent* e *succede*, così come i verbi aspettuali quali ad esempio *begin*.

In conclusione i 15 gruppi considerati sono sembrati adeguati per collocare tutti i *synsets* verbali che si sono aggiunti nel corso degli anni (già WordNet 1.5 conteneva più di 11500 *synsets* verbali). Va comunque sottolineato che il confine tra i vari campi verbali resta estremamente vago ed indeterminato; ad esempio molti verbi non possono essere classificati univocamente come verbi di cognizione o di comunicazione (è il caso di *wonder, speculate, confirm* e *judge*, per citarne solo alcuni). Analogamente, un verbo come *whistle* in *The bullet whistled past him*, può essere classificato sia come un verbo di emissione del suono, che come un verbo di movimento (Atkins e Levin (1991)). In WordNet si è cercato di trattare questi verbi come parole polisemiche, rintracciabili all'interno di più campi semantici. In ogni caso in WordNet non sembra essere particolarmente rilevante a quale gruppo specifico un determinato verbo appartenga, poiché all'interno dell'ontologia il suo significato viene rappresentato principalmente dalle relazioni che esso intrattiene con altri verbi ed altri *synsets*.

La divisione del lessico verbale in campi semantici non fornisce solo un valido appiglio per organizzare un vasto insieme di dati, ma è anche motivata dall'assenza di una singola radice verbale o *unique beginner*, che funga da testa dell'intero corpo dei verbi.

Lyons (1977), avendo notato questa mancanza, propone un insieme di radici verbali che includono: *act, move, get, become, be, make*.

Pulman (1983) suggerisce invece solo *do* e *be*, evidenziando in tal modo la distinzione tra verbi di azione e verbi stativi, già presente nelle categorie concettuali maggiori riprese da Jackendoff (1983) *event* e *state*. Sebbene questi primitivi semantici sembrano essere buoni candidati come radici verbali, si è visto che per WordNet l'adozione dei soli *be* e *do* come *unique beginners* non era completamente appropriata.

Tanto per cominciare, anche questi concetti semanticamente non elaborati sono polisemici; WordNet distingue ben 12 sensi per *do* e altrettanti per *be*. Chiaramente alcuni di questi significati non li qualificano come *unique beginners* (*do* in *do my hair* o *do my room in blue* esprime evidentemente concetti semanticamente molto specifici ed elaborati), ma ci sono comunque troppi significati di base che rendono impossibile isolarne uno tra tutti come senso predominante da cui discendono tutti gli altri.

Secondariamente, si ritiene che le particolari relazioni semantiche stabilite per realizzare la rete dell'ontologia, rendono inopportuno collegare verbi astratti come *do*, al livello successivo dei subordinati come *communicate* e *move*. Ad esempio, laddove sembra esserci una connessione gerarchica appropriata tra verbi come *communicate* e *chat*, lo stesso legame risulta essere molto meno fondato tra i verbi *do* e *communicate*; difatti questi due concetti sembrano essere molto più distanti l'uno dall'altro di quanto non lo siano *communicate* e *chat*.

Similmente *move* e *run* possono essere correlati in modo soddisfacente, ma *do* e il suo immediato subordinato, ovvero lo stesso *move*, sembrano essere semanticamente distinti e lontani.

In terzo luogo non sembra esserci un'evidenza psicolinguistica in base alla quale le persone collegano nella loro mente verbi come *do* ad altri come *move*, mentre tale evidenza sussiste nelle coppie associative come *move* e *run* (Chaffin, Fellbaum e Jenei, 1994).

L'adozione di *unique beginners* come quelli proposti da Lyons (1977) e Pulman (1983), sembra essere inadeguata rispetto all'aspirazione di WordNet di riflettere l'intera organizzazione lessicale dei parlanti. Per tale motivo sono stati fissati più *unique beginners* per ognuno dei 14 campi semantici. È infatti frequente che all'interno di un campo semantico non tutti i verbi presenti siano raggruppabili sotto un solo *unique beginner*; alcuni membri possono essere rappresentati in modo soddisfacente soltanto tramite alberi indipendenti. I verbi di movimento ad esempio hanno due *top nodes* omofoni, che latori di due diversi

concetti: *move1* e *move2* che esprimono rispettivamente un movimento traslatorio e un movimento senza spostamento.

Come i nomi e gli aggettivi, anche i verbi in WordNet sono raggruppati insieme come gruppi di sinonimi (*synonym sets*). Comunque se si adotta la definizione di sinonimi, quali parole sostituibili reciprocamente negli stessi contesti linguistici, si potranno trovare allora solo pochi casi di sinonimia verbale (come *shut* e *close*) nel lessico inglese. A volte emergono ad esempio delle sottili differenze di significato tra sinonimi apparenti, grazie all'analisi delle rispettive preferenze di selezione. È il caso di *rise* e *fall* che possono selezionare come loro argomento un'entità astratta come *temperature* o *prices*, mentre i loro sinonimi più stretti, *ascend* e *descend*, non possono fare altrettanto. In generale si è evitato di sistemare in uno *synset* verbi che differiscono significativamente rispetto alle preferenze di selezione correlate.

Se la relazione più frequente individuata per la categoria ontologica dei nomi era quella di iponimia, per i verbi si parla invece di *troponimia* (Fellbaum e Miller, 1990); questa può essere espressa grazie alla formula *To $V_1$ is to $V_2$ in some particolar manner*. La relazione di troponimia può essere considerata un tipo particolare di inclusione; difatti ogni troponimo $V_1$ di un verbo più generale $V_2$, include anche evidentemente lo stesso $V_2$. Ad esempio per considerare *march* come un troponimo di *walk*, l'atto del *marching* deve implicare necessariamente quello del *walking*. Inoltre le attività cui si riferiscono un troponimo ed il suo immediato superordinato sono sempre coestensive a livello temporale. Difatti si deve camminare ad ogni istante mentre si sta marciando.

Le tassonomie verbali costruite sulla relazione di troponimia tendono ad essere più superficiali di quelle nominali, non superando in effetti i quattro livelli di connessione nella scala ontologica. Più si scende nella gerarchia verbale, più la varietà di nomi che un verbo può realizzare come suoi argomenti, crolla sensibilmente; questo avviene perché in modo inverso cresce la specificità del significato del verbo considerato.

Per quanto riguarda l'informazione associata ad ogni verbo in WordNet possiamo dire che ne viene specificata la struttura predicato-argomento: sono attribuiti ruoli tematici ai nomi che fungono da argomenti del verbo e vengono specificate le proprietà semantiche riguardanti le classi di nomi che possono costituire l'argomento per un verbo. Infine, per ogni verbo del *synset*, vengono descritte una o più strutture (*frames*) che specificano le caratteristiche di sottocategorizzazione dei verbi (indicando le frasi in cui possono comparire).

Come esempio di quanto finora descritto, riportiamo di seguito una delle classi verbali elaborate per WordNet, nello specifico quella dei *motion verbs*. Direzione, modo e scopo del movimento sono le linee guida di entrambe le classificazioni; comunque da un confronto con la classe proposta da Levin, emerge una riduzione nella granularità della classificazione, che risulta più semplificata in WordNet rispetto a quella elaborata manualmente.

**Verbs of motion**

- inherently directed motion (arrive, go…)
- leave verbs
- manner of motion
  - roll verbs (bounce, float, move…)
  - run verbs (bounce, float, jump)
- manner of motion using a vehicle
  - vehicle name verbs (bike…)
  - verbs not associated with vehicle names (fly…)
- waltz verbs (boogie, polka…)
- chase verbs (follow, pursue…)
- accompaign verbs

### 2.1.2 Il *database* lessicale multilingue EuroWordNet

Obiettivo del progetto, finanziato dalla Comunità Europea, iniziato nel Marzo 1996 e terminato nel 1999, era la costruzione di un database lessicale multilingue coerente ed affidabile, e allo stesso tempo in grado di conservare le diversità e le ricchezze delle diverse lingue. Il modello scelto fu il *Merge Model* secondo il quale il metodo da seguire è quello di tenere separate le strutture semantiche dei diversi linguaggi.

Secondo questa logica, il primo passo è quello di sviluppare un database per ogni lingua: questa operazione può essere fatta autonomamente. Il passo successivo prevede lo sviluppo di una parte interlinguistica che metta in relazione i *synsets* dei diversi WordNet con quelli del WordNet di Princeton che più si avvicinano nel significato.

Le lingue coinvolte in questo progetto sono in tutto otto: olandese, inglese, italiano, tedesco, spagnolo, presenti fin dall'inizio del progetto, e francese, ceco, estone, introdotte in un secondo momento.

Ogni WordNet specifico della lingua considerata è strutturato secondo le stesse linee guida del WordNet originario: cioè i sinonimi sono raggruppati in *synsets*, i quali a loro volta sono legati fra di loro da relazioni semantiche. In aggiunta, ogni significato viene associato ad un synset di Princeton WordNet tramite una relazione di equivalenza, in modo da creare così un *database* multilingue.

La previsione all'inizio del lavoro era quella di inserire nel *database* approssimativamente 25000 *synsets* per ogni lingua. Il vocabolario doveva contenere tutte le forme della parola di base per ogni lingua, e in più dei sottovocabolari per domini specifici in modo da illustrare la possibilità di integrare tale terminologia nel lessico generico. Ogni WordNet specifico viene mantenuto in maniera del tutto indipendente in un *database* lessicale centrale, mentre le relazioni di equivalenza fra significati in lingue diverse vengono mantenute attraverso le relazioni di equivalenza con i *synsets* di WordNet.

Vediamo ora come può essere schematizzata l'implementazione di EuroWordNet.

Si costruisce una *Top-Ontology* comune, in base alle gerarchie appena citate e ai nodi più frequentemente coinvolti nelle relazioni.

Si delinea successivamente la presenza di due strutture costitutive fondamentali:

1. un insieme di moduli specifici, uno per ogni linguaggio;
2. un modulo indipendente dal linguaggio.

Le relazioni di equivalenza fra *synsets* di lingue diverse vengono esplicitate attraverso una struttura chiamata *Inter-Lingual-Index* (ILI).

Ogni *synset* in un WordNet monolingue avrà almeno una relazione di equivalenza con un *record* della struttura ILI. Secondo lo schema, i *synsets* presenti in WordNet appartenenti a lingue diverse e collegati allo stesso *record* ILI, dovrebbero essere fra loro equivalenti (in senso lessicale).

La struttura iniziale dell'*Inter-Lingual-Index* contiene una lista di *synsets* equivalente a quelli esistenti nel WordNet centrale; essa è comunque destinata ad incrementare il proprio numero di *records*, con l'introduzione di concetti specifici di altre lingue.

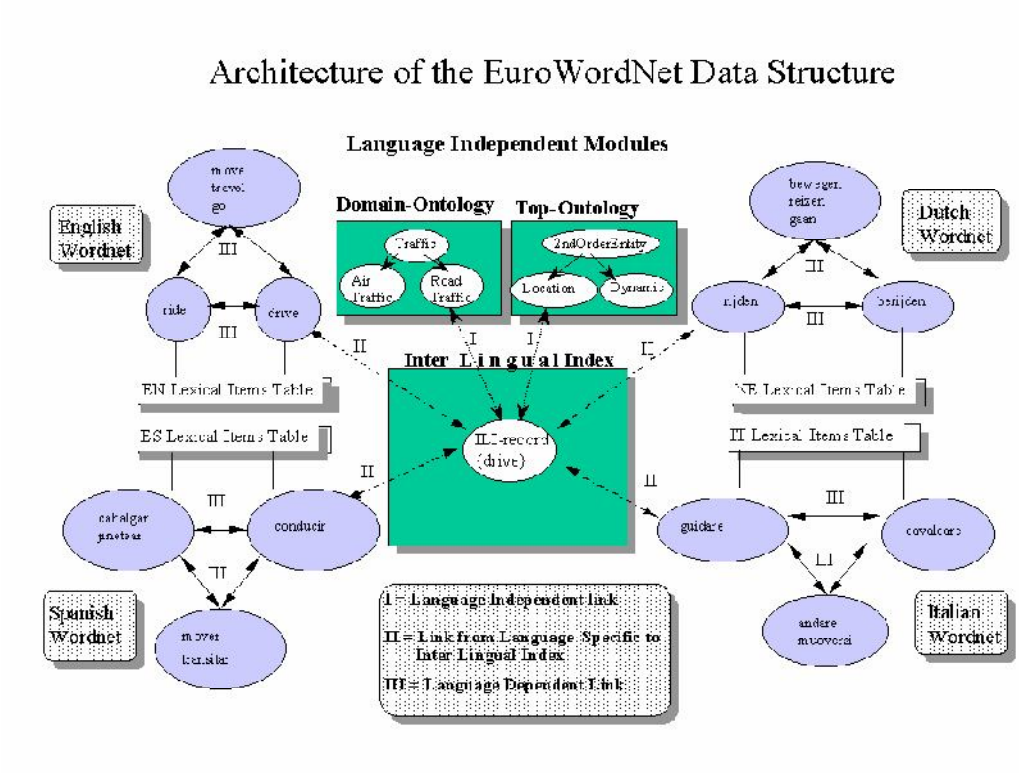
Nella Figura 2.5 si può vedere come *synsets* di linguaggi diversi sono connessi allo stesso *record* della struttura ILI; nello specifico, è presente un esempio relativo alla parola *drive*. Inoltre, sempre la stessa figura ci dà una rappresentazione schematica dei differenti moduli e delle relazioni che intercorrono fra di essi.

Nel centro si trovano i moduli indipendenti dalla lingua (in verde), attorno ai quali troviamo quelli di ogni specifico idioma. Non esiste alcuna relazione interna fra *record* della struttura ILI, essa serve solamente come elemento di connessione.

I vantaggi dell'architettura che abbiamo appena illustrato sono i seguenti:

- le relazioni multilingui non devono essere considerate una per una, ma vengono ridotte a relazioni fra significati;
- estensioni future del *database* (come ad esempio l'introduzione di una nuova lingua) non rimettono in discussione tutto il *database*; avvengono invece proprio attraverso l'utilizzo dell'ILI, inteso come insieme di concetti a cui far riferimento per l'introduzione di un nuovo WordNet;
- per aumentare l'efficienza della ricerca interlinguistica, e quindi dell'interconnessione fra WordNet diversi, è sufficiente agire su un'unica struttura.

*Figura 2.5 Schema di EuroWordNet per la parola "drive"*




Entrando più nel dettaglio e riferendoci sempre allo schema dell'architettura di EuroWordNet si può notare che esistono due ulteriori strutture indipendenti dal linguaggio, rappresentanti altrettante diverse ontologie, a cui possono essere collegati i *record* della struttura ILI:

- la *Top-Concept Ontology* (TCO o TO), che è una rappresentazione strutturata secondo tre livelli (ordini) dei concetti indipendenti dalle lingue. Ad esempio: *Object*, *Location*, *Dynamic*, *Static*;
- la *Domain Ontology* (DO), una gerarchia che divide i significati per campo semantico, cioè per *argomento*. Ad esempio: *Traffic*, *Road-Traffic*, *Air-Traffic*.

I *records* di entrambe le strutture possono essere collegati ai significati specifici di ogni linguaggio attraverso le relazioni di equivalenza rese esplicite dai riferimenti presenti nella struttura ILI. Così, come si può vedere nella Figura 2.5, i concetti *Location* e *Dynamic* presenti nel secondo ordine della *Top Concept Ontology* sono direttamente collegati alla parola *drive* della ILI. Attraverso le relazioni di equivalenza essi sono indirettamente collegati ai concetti relativi ad altre lingue, nel caso della lingua italiana: *andare*, *cavalcare*, *guidare*, *muovere*.

Lo scopo principale della TCO è di fornire una struttura comune per i concetti più importanti di tutti i WordNet collegati. Essa consiste di 63 gruppi semantici di base, che classificano un insieme di 1310 concetti fondamentali comuni a tutti le lingue, individuati secondo i seguenti criteri:

- numero delle relazioni ad essi associate;
- posizione nella gerarchia tassonomica;
- frequenza in un *corpus*.

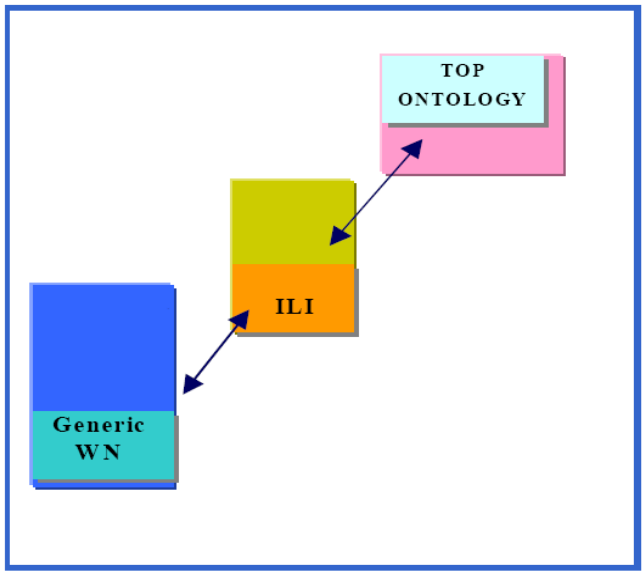
### 2.1.3 La costruzione di un *database* per l'italiano. Due progetti a confronto: ItalWordNet e Italian WordNet

Il *database* ItalWordNet (IWN) è formato dai seguenti componenti:

- una generica rete WordNet, costruita come estensione di quella sviluppata in EuroWordNet, contenente 46.000 lemmi, corrispondenti a 49.000 *synsets* e a 65.000 sensi delle parole;
- un *Inter-Lingual-Index* che è una versione presa da WN1.5, contenente tutti i *synsets* rintracciabili in WN1.5 ma non le relazioni che intercorrono tra di essi. Questo modulo viene usato anche in EuroWordNet per collegare i wordnets delle diverse lingue codificate. Anche in IWN i *synsets* dell'italiano sono collegati a questo indice interlingue, in modo da rendere tale risorsa utilizzabile anche per applicazioni multilingue;
- una *Top Ontology* (TO), ovvero una gerarchia composta da 60 concetti indipendenti dalla lingua d'uso, che riflettono quelle che vengono considerate distinzioni semantiche fondamentali (Roventini et al. (2003)).

Tutti questi componenti ed i loro reciproci legami, ovvero l'architettura globale di ItalWordNet, vengono presentati nella Figura 2.6.

*Figura 2.6 Architettura generale del database ItalWordNet*




La nozione di base attorno alla quale si costruisce IWN, così come per WordNet ed EuroWordNet, è quella di *synset*, ovvero un insieme di parole sinonime appartenenti alla stessa parte del discorso, che risultano intercambiabili in almeno un contesto.

I *synsets* sono collegati fra loro principalmente da relazioni di iponimia, vale a dire di tipo IS-A, ma all'interno di IWN sono codificati anche altri tipi di relazioni, in parte ereditate dall'architettura di EuroWordNet, per descrivere al meglio i rapporti che intercorrono tra i *synsets*. In particolare è stato arricchito l'insieme di relazioni proposto in WordNet, con tutte quelle relazioni applicabili a *synsets* appartenenti a diverse parti del discorso. In WordNet

ogni parte del discorso forma una rete separata con proprie relazioni interne, pertanto concetti affini vengono comunque separati solo in virtù della loro appartenenza a diverse parti del discorso. Ad esempio non c'è nessuna relazione che colleghi il sostantivo *collection* al verbo *to collect*, sebbene essi descrivano un processo identico.

Per evitare questa ripartizione in parti del discorso, in EuroWordNet viene proposta una distinzione alternativa che si rivolge all'ordine semantico di appartenenza delle entità a cui i significati delle parole fanno riferimento (Lyons (1977)): le entità di 1° ordine vengono realizzate dai nomi concreti, quelle di 2° ordine dai verbi, dagli aggettivi e dai nomi indicanti proprietà, processi, stati o eventi, ed infine quelle di 3° ordine dai nomi astratti che indicano proposizioni indipendenti dal tempo e dallo spazio. Sulla base di questa distinzione sia in EuroWordNet che in ItalWordNet, vengono applicate alle diverse parti del discorso, le varie relazioni codificate in modo trasversale.

Questo approccio sembra essere più corretto del precedente, non solo da un punto di vista strettamente teorico, dato dal fatto che quest'ultima distinzione è fatta su base unicamente semantica, ma anche perché IWN ha inoltre ereditato da EuroWordNet la distinzione tra *relazioni interne* ad una lingua e *relazioni di equivalenza*; le prime collegano tra loro *synsets* appartenenti ad una lingua specifica, mentre le altre uniscono i *synsets* dell'italiano con l'*Inter-Lingual-Index* al fine di garantire successive applicazioni multilingue.

Sempre per l'italiano, è stato sviluppato parallelamente un progetto che prende il nome di Italian WordNet, basato sull'assunto che la maggior parte delle relazioni concettuali valide per l'inglese (circa 72000 relazioni *Is a* e 5600 relazioni *part of*) possano essere condivise anche dall'italiano (Artale, Magnini, Strapparava (1997)).

Per ciò che concerne la classificazione dei verbi in Italian WordNet, sono stati compiuti numerosi sforzi per aggiungere all'interno del progetto le restrizioni di selezione.

In primo luogo sono stati estratti i vari significati del verbo italiano da un dizionario cartaceo, per poi confrontarne l'utilizzo concreto all'interno di un *corpus* di testi generici per l'italiano. Ogni senso del verbo è stato poi abbinato ad uno o più *synsets* del WordNet inglese; questa fase è stata realizzata manualmente con il supporto di un'interfaccia grafica, che include quattro strumenti di lavoro:

1. un dizionario bilingue con più di 30000 lemmi;
2. un grafo che permette la visualizzazione dell'abbinamento con il WordNet inglese;
3. il WordNet bilingue, che funziona esattamente come la versione inglese, ma con la possibilità aggiuntiva di accedere alla rete semantica dell'italiano;
4. infine, le schede di lavoro permettono l'inserimento, la modifica e la verifica dei dati per ciascun *synset*.

Il risultato di questa fase è l'estensione del WordNet inglese con i *synsets* dell'italiano, come mostra la Figura 2.7

*Figura 2.7 Corrispondenza tra i synsets dell'italiano e dell'inglese per il verbo "scrivere" (write)*

| Synset Label | Italian Synset | English Synset |
|---|---|---|
| Write | {scrivere redigere comporre} | {write compose pen indite} |
| Write-Music | {scrivere comporre} | {compose write write_music} |
| Write-Communicate | {scrivere comunicare_per_iscritto} | {write communicate_by_writing} |
| Write-Publish | {scrivere pubblicare} | {publish write} |
| Write-Send | {inviare mandare scrivere spedire} | {mail write post send} |

Il passo successivo è definire quali siano i *frames* di sottocategorizzazione per ogni verbo; ciò include sia informazioni sintattiche (posizione degli argomenti, preposizioni degli

oggetti indiretti etc.), che informazioni semantiche (ruoli tematici, restrizioni selettive etc.). L'informazione sintattica è associata ai singoli verbi, mentre l'informazione semantica è relativa a tutto il *synset* (i partecipanti semantici, ad esempio, sono condivisi da tutti i verbi appartenenti al *synset*).

Le restrizioni di selezione sono state costruite ricorrendo alla gerarchia dei *synsets* dei nomi; per la definizione di queste componenti sono state considerate due diverse possibilità:

1. restrizioni di selezione ottenute dai *frames* forniti da WordNet;
2. restrizioni di selezione ottenute dall'intera gerarchia dei nomi di WordNet.

Riguardo alla prima ipotesi, WordNet descrive tutti i verbi inglesi avvalendosi di un insieme di 35 diversi *frames* sintattici, che a loro volta implicano soltanto due restrizioni: *something* e *somebody*.

(2) i *frames* forniti per il verbo *write* nel *synset* {Publish, Write} presentati sottoforma di modelli, in cui i puntini possono essere sostituiti con il tema del verbo stesso:
Somebody……s
Somebody……s Something

Il problema che nasce dall'utilizzo di queste due restrizioni è che sono completamente avulse dai *synsets* dei nomi, perciò devono essere confrontati sistematicamente con i *synsets* appropriati nella gerarchia dei nomi. Il concetto *Somebody* include non solo il *synset Person*, ma anche tutti i *synsets* che denotano gruppi di persone che possono reggere il ruolo tematico di Agente. Possiamo definire *Somebody* usando la seguente combinazione booleiana di *synsets*:

Somebody: Person v People v People-Multitude v
(Social-Group v ¬ (Society v Subculture v
Political – System v Moiety v clan))
Something: è definito come il complemento di *Somebody*.

Nella seconda ipotesi le restrizioni di selezione sono estrapolate dall'intera gerarchia dei nomi; come esempio si veda la Figura 2.8, che illustra i sensi per il verbo *scrivere* (*write*) rintracciati in Italian WordNet.

Per ogni senso viene riportato un nome convenzionale, che definisce il *synset* in modo non ambiguo, e le posizioni argomentali ammesse per quel senso, con l'indicazione delle restrizioni di selezione. La combinazione appropriata di *synsets* per una posizione argomentale deve essere abbastanza generale da preservare tutte le letture umane, e abbastanza circoscritta da distinguere tra i vari sensi sia del verbo che del nome.

Individuare restrizioni selettive adeguate si rivela un processo difficile e lungo; esso richiede infatti una ricerca approfondita nella gerarchia dei nomi di WordNet. Per raggiungere un compromesso valido tra potere discriminante e livello di precisione, è stato adottato un procedimento empirico fatto di passaggi successivi di raffinazione. All'inizio sono state considerate restrizioni di selezione generali, che sono state successivamente convalidate da risultati sperimentali. Questo processo iterativo termina con complesse selezioni di restrizione per i verbi considerati.

La tassonomia verbale di Italian WordNet si basa sulla relazione di troponimia, definita dalla co-occorrenza:

- sia dell'implicazione lessicale (*entailment*);
- sia della co-estensione temporale tra le coppie di verbi.

Ogni volta che si verifica una relazione di troponimia tra due verbi, si realizza anche una relazione *Is a* tra le restrizioni di selezione corrispondenti.

*Figura 2.8 Entrate Lessicali per Scrivere (Write)*

| WordNet Synset | Subject | Object | Indirect-Object |
|---|---|---|---|
| Write | Somebody | Written-Material∨<br>Symbolic-Repres ∨ Saying∨<br>Correspondence ∨ Sentence∨<br>Message ∨ Message-Content∨<br>Code ∨ Symbol ∨ Date∨<br>Language-Unit ∨ Property∨<br>Address-Speech ∨ Print-Media | −− |
| Write-Music | Person | Music | −− |
| Write-Communicate | Somebody | (Written-Material ∧ ¬Section)∨<br>Symbolic-Repres ∨ Saying∨<br>Sentence ∨ Name∨<br>Message ∨ Message-Content∨<br>Code ∨ Date ∨ Property | Somebody |
| Write-Publish | Somebody | Written-Material∨<br>(Print-Media ∧ ¬Section) | Print-Media∨<br>Publishing-House |
| Write-Send | Somebody | Correspondence ∨ Message∨<br>Letter-Missive | Somebody |

## 2.2 Il lessico SIMPLE-CLIPS

I lessici computazionali hanno come obiettivo comune quello di fornire una rappresentazione esplicita del significato delle parole, perché questo possa essere utilizzato direttamente dalle varie applicazioni computazionali.

Essi aggiungono alla rappresentazione del significato di una parola le informazioni necessarie per stabilire delle connessioni tra parole di lingue diverse.

Il progetto SIMPLE (*Semantic Information for Multipurpose Plurilingual LExica*), ha portato alla definizione di un'architettura per lo sviluppo di lessici computazionali semantici e alla costruzione di lessici computazionali per 12 lingue europee[1]. I lessici SIMPLE offrono una rappresentazione articolata e multidimensionale del contenuto semantico dei termini lessicali. Il modello di rappresentazione semantica di SIMPLE è usato anche per la realizzazione di CLIPS, che include 55.000 entrate lessicali con informazione fonologica, morfologica, sintattica e semantica (Ruimy et al. (2003)).

Il modello SIMPLE costituisce un'architettura per lo sviluppo di lessici computazionali nel quale il contenuto semantico è rappresentato da una combinazione di diversi tipi di entità formali con i quali si cerca di catturare la multidimensionalità del significato di una parola (Lenci et al. (2000), Lenci e Calzolari (2004)).

Il *background* teorico-linguistico del *database* SIMPLE-CLIPS, è costituito dal *Lessico Generativo* di Pustejovsky (1995); quest'ultimo definisce la semantica di un'entrata lessicale tramite la cosiddetta *qualia structure*, una rappresentazione ricca e strutturata della forza relazionale di un'entrata lessicale.

L'importanza di questa struttura è che essa permette di superare il patrimonio unidimensionale catturato tramite le relazioni iperonimiche *standard*, consentendo invece l'espressione di aspetti ortogonali del senso di una parola.

Generalmente le entrate lessicali sono organizzate in base a relazioni tassonomiche, considerato che molti dei sensi di una parola sono caratterizzabili interamente nei termini di relazioni gerarchiche con altre unità lessicali.

Nonostante ciò, un sostanzioso insieme di sensi delle parole mostrano una rete assai più complessa di dimensioni lessicali ortogonali; queste, evidentemente, non possono essere comprese e rappresentate esaustivamente nei termini di una mera relazione iperonimica. La *qualia structure* permette che la multidimensionalità del significato possa essere codificata

[1] Catalano, Danese, Finlandese, Francese, Greco, Inglese, Italiano, Olandese, Portoghese, Spagnolo, Svedese, Tedesco

attraverso quattro *ruoli qualia*, che esprimono aspetti essenziali e complessi del significato di una parola:

1. il *formal role*, che identifica un'entità in mezzo ad altre entità e ne indica inoltre la posizione all'interno dell'ontologia dei tipi;
2. il *constitutive role*, che rivela la composizione dell'entità analizzata e quella dei suoi elementi costitutivi;
3. l'*agentive role*, che fornisce le informazioni necessarie sull'origine e la direzione dell'entità considerata;
4. il *telic role*, che specifica la funzione dell'entità.

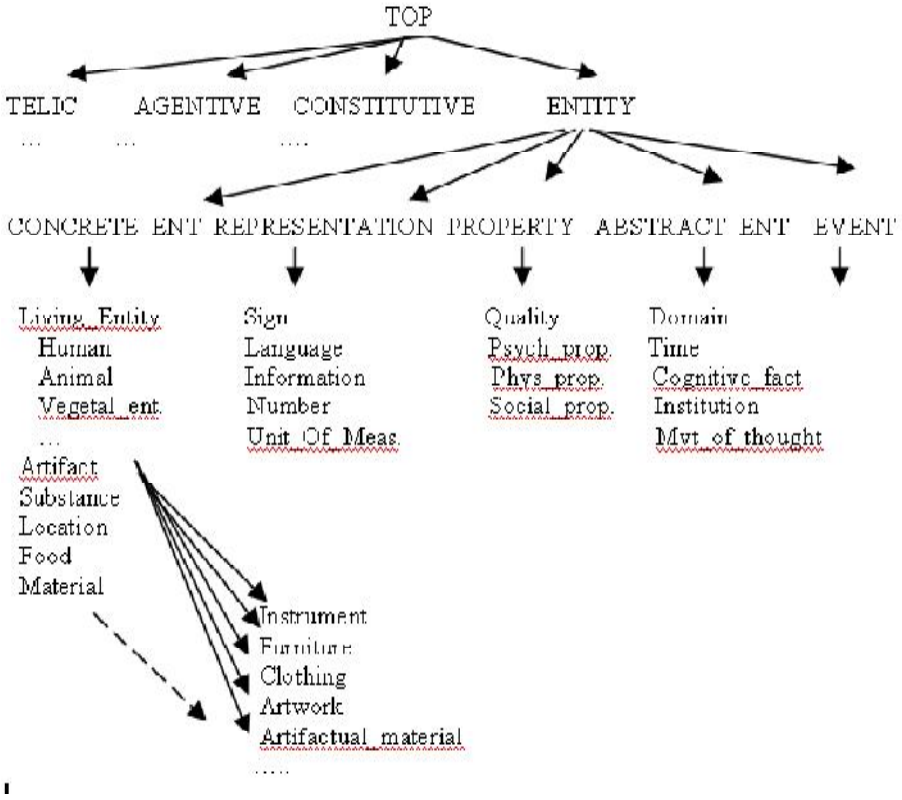
In SIMPLE-CLIPS le entrate lessicali sono costituite dai vari sensi di una parola e vengono dette *unità semantiche* (SemU). Ogni SemU viene assegnata ai quattro *ruoli qualia*; questo procedimento di assegnazione risulta estremamente chiaro nell'esempio fornito da Ruimy et al. (2003) riguardante la parola *fotografo*. Per questa unità semantica vengono codificati diversi tipi di relazioni teliche, con lo scopo di preservare nella sua completezza l'informazione per cui essere un fotografo può essere sia una professione (is_the_activity_of) che un hobby (is_the_ability_of).

Nella *Extended Qualia Structure* la rilevanza di una certa relazione rispetto alle altre, è evidenziata dal diverso peso attribuito ad ognuno dei suoi attuali usi nella definizione di un tipo. Il peso serve ad indicare se la relazione è *type defining* o meno, ovvero se codifica un'informazione che caratterizza intrinsecamente un tipo semantico o se invece convoglia un'informazione opzionale, generalmente sulla conoscenza del mondo esterno.

L'ontologia SIMPLE-CLIPS distingue inoltre tra una *Core Ontology* ed una *Recommended Ontology*; la prima si compone dell'intera gerarchia dei tipi semantici generali o superiori, ovvero quelli che incontrano il più ampio consenso tra le lingue codificate e forniscono tutte le informazioni essenziali per descrivere i sensi di una parola. La seconda include invece i tipi semantici di livello inferiore, vale a dire quelli più specifici, che forniscono ovviamente l'informazione più granulare sul significato di una parola (Ruimy et al. (2003)).

In SIMPLE il processo di codifica all'interno dell'ontologia è stato guidato dai cosiddetti *templates*, ovvero strutture schematiche che permettono ad un tipo semantico di essere raggruppato in uno specifico *cluster* informativo, considerato cruciale per la sua definizione. L'uso dei *templates* è interessante perché garantisce uniformità e consistenza alla codifica.

*Figura 2.9 L'ontologia SIMPLE-CLIPS*




Il lessico italiano SIMPLE-CLIPS è formato da entrate semantiche dei verbi, nomi e aggettivi, descritti ricorrendo ad un'ampia gamma di informazioni che andremo a vedere di seguito più nel dettaglio.

- *Informazione sull'assegnazione del tipo*
  Assegnare un tipo semantico ad un'entrata lessicale, implica una valutazione dell'eredità portata dall'informazione sulla gerarchia del tipo. Questa ci indica la posizione del tipo, e quindi della SemU che rappresenta quel certo tipo, all'interno dell'intera gerarchia dei tipi.
- *Dominio*
  I domini, ripartiti in classi (un insieme di 350 ambiti), forniscono l'informazione sull'argomento dei testi in cui la SemU ricorre più frequentemente.
- *Struttura Qualia*
  Formal role
  Fornisce un'ampia caratterizzazione di un'entità rispetto alle altre, espresse dalla relazione iperonimica *is a* per i nomi e le entità che connotano eventi. La relazione *is a* fornisce un'informazione più granulare rispetto al solo tipo semantico e consente un'ulteriore sottotipizzazione delle entrate che condividono lo stesso *template*, ad esempio la relazione *is a* per rettile, felino, pachiderma, permetterà di differenziare e sottoclassificare le entrate codificate all'interno del tipo EARTH_ANIMAL.
  Come regola generale viene assegnato l'iperonimo più vicino, evitando il più possibile relazioni *is a* circolari.
  Per gli aggettivi, secondo il procedimento adottato anche in WordNet ed in contrasto con la codifica di nomi e verbi, il *formal role* non è espresso tramite una relazione di iperonimia, bensì di antonimia.

Constitutive role

Esprime la composizione interna dell'entità, attraverso un insieme di relazioni del tipo: *is_a_member_of, is_a_part_of, resulting_state, has_a_property*.

Il qualia costitutivo è particolarmente utile nella definizione degli aggettivi; in questo ambito i componenti del significato, che sono essenziali per cogliere la natura aggettivale, sono espressi in termini di tratti (es. *movement, space, substance* etc.).

Agentive role

Fornisce l'informazione sull'origine dell'entità considerata; alcune delle relazioni agentive tipicamente usate nel lessico italiano sono: *created_by,result_of, caused_by, agentive_cause, agentive_experience,* etc.

Telic role

Serve a specificare la funzione dell'entità, lo scopo per cui essa esiste oppure è stata creata; il ruolo telico ed agentivo non vengono mai usati per rappresentare gli aggettivi, poiché sono considerati più adatti a rappresentare la semantica dei sostantivi. Alcune tra le relazioni teliche più utilizzate sono: *used_as, used_for, object_of_the_activity, in direct_telic,* etc.

- *Polisemia regolare*

Nel lessico SIMPLE-CLIPS, i sensi dei nomi che sono semanticamente correlati sono descritti in base ad un insieme di 20 classi di alternanza di senso.

Questo tipo di ambiguità di senso genera tipi complessi, rappresentati in SIMPLE-CLIPS collegando tra loro le diverse *SemU* di un'entrata lessicale, che vengono a far parte di una classe polisemica regolare. Tale legame è espresso in ogni *template* ritenuto rilevante, attraverso il nome della coppia dei tipi semantici a cui i sensi alternativi appartengono.

- *Sinonimia*

È una relazione assegnata a quelle SemU codificate nei *top templates* per cui le relazioni tassonomiche risultavano ingiustificate; la relazione di sinonimia è inoltre utile per la codifica degli aggettivi, specie quelli altamente polisemici.

- *Informazioni derivazionali*

I legami trasversali tra categorie diverse di parole come la derivazione, sono marcati attraverso dei collegamenti che uniscono l'entità derivata alla sua base di riferimento. È stato elaborato un insieme di relazioni che distinguono tra i vari tipi di derivazione possibili (es. aggettivi deverbali e denominali, nomi deverbali, verbi denominali etc.).

- *Tratti semantici*

Il ricorso all'individuazione di tratti semantici, permette di raggruppare SemU codificate in tipi semantici diversi, ma che condividono almeno un componente del significato, in *clusters* semanticamente coerenti (es. *plus_collective, plus_edible,* etc.)

- *Struttura argomentale*

Uno degli aspetti più interessanti in SIMPLE-CLIPS, è la possibilità di codificare predicati lessicali per ogni SemU predicativa; ad ogni argomento è assegnato un ruolo semantico (selezionato in una lista predefinita di ruoli) e l'informazione inerente la sua caratterizzazione semantica. Quest'ultima non viene interpretata come una rigorosa restrizione, ma piuttosto come una preferenza combinatoria in situazioni prototipiche.

Intendiamo ora analizzare nello specifico il trattamento dei verbi all'interno del progetto SIMPLE-CLIPS, sia livello sintattico che semantico.

Considerando il comportamento sintattico di un'unità morfologica verbale, vengono create entrate sintattiche diverse nei seguenti casi:

- *diversità di arità e/o funzioni dei complementi*
- il numero di complementi non opzionali è diverso nelle strutture in esame;
- a parità di numero di complementi, la funzione sintattica di uno di loro è diversa rispetto alle altre strutture a confronto;

- numero di complementi e funzioni divergono;

(3) *disporre i libri negli scaffali* *disporre di due auto*

| **P0** | **P1** | **P2** | **P0** | **P1** |
|---|---|---|---|---|
| **subject** | **object** | **adjunct** | **subject** | **oblique** |

- *opzionalità di un complemento*

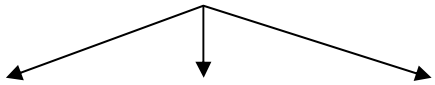
un complemento risulta opzionale in uno dei significati del verbo mentre è obbligatorio nell'altro, com'è spesso possibile osservare nei sensi figurati.

(4) *evadere (dal carcere)* *evadere dalla realtà*
*fuggire (da una casa in fiamme)* *fuggire dalle cattive compagnie*
*attraversare (la strada)* *attraversare un momento difficile*

- *alternanza di realizzazione di un complemento*

dato un verbo, in uno dei suoi significati una determinata posizione del *frame* può essere occupata alternativamente da diverse categorie sintagmatiche (5.b) mentre un altro significato non consente che un'unica realizzazione sintattica della stessa posizione (5.a).

(5) a. Luca evita Maria
b. Luca ha evitato



che Maria si ferisse una sciagura di dover partire

- *possibilità di nominalizzazione*

sensi diversi di verbi omografi o polisemici che condividono la stessa struttura sintattica, vengono codificati in entrate diverse se si comportano diversamente rispetto alla possibilità di nominalizzare.

(6) a. rialzare i pressi il rialzo dei prezzi
b. rialzare la testa * il rialzo della testa

I criteri semantici sono perciò presi in considerazione solo se hanno un riflesso a livello sintattico.

Alcuni verbi hanno complementi interni tutti opzionali che possono essere quindi tutti realizzati, o solo in parte, o tutti assenti.

(7) a. Durante la riunione Luca ha parlato
b. Durante la riunione Luca ha parlato del progetto
c. Durante la riunione Luca ha parlato con i colleghi
d. Durante la riunione Luca ha parlato del progetto con i colleghi

In questo caso un'unica entrata, con una marca di opzionalità su ogni complemento, rende conto di tutte le possibili strutture:

```
<Construction
        id="i-ppconopt-ppdiorinfdiopt"
        syntlabel="Clause"
        selfinsertion="1">
            <InstantiatedPositionC
                range="0"
                optional="YESO"
                positionc="Psubj">
            <InstantiatedPositionC
                range="1"
                optional="YESO"
                positionc="Poblppcon">
            <InstantiatedPositionC
                range="2"
                optional="YESO"
                                positionc="Poblppdiorinfdi"></Construction>
```

In altri casi non si tratta di complementi opzionali, bensì di uso intransitivo e transitivo:

(8) a. Il bambino parla
b. Luca parla inglese

In tal caso, due diverse entrate devono essere create per catturare l'informazione riguardante la diversa sottocategoria del verbo.

Altri verbi ancora presentano delle restrizioni di compresenza dei complementi:

- l'assenza di una posizione esclude la presenza dell'altra; ad esempio nella frase seguente, l'assenza del complemento oggetto inibisce la presenza dell'oggetto indiretto:

(9) a. Luca rifuta
b. Luca rifiuta un bacio
c. Luca rifiuta un bacio a Maria
d. * Luca rifiuta a Maria

Poiché nel modello teorico di SIMPLE-CLIPS, ovvero LE PAROLE, l'opzionalità condizionata è un criterio di creazione di una nuova struttura sintattica, avremo due strutture sintattiche per il verbo rifiutare:

- con un oggetto opzionale (9.a);
- con un oggetto obbligatorio e complemento indiretto opzionale (9.b, 9.c)

a)
```
<Construction
        id="topt8infdiCsC"
        syntlabel="Clause"
        selfinsertion="1">
            <InstantiatedPositionC
                range="0"
                optional="YESO"
                positionc="PsubjCsC">
            <InstantiatedPositionC
                range="1"
                optional="YESO"
                positionc="Pobj8infdiCsC">
            <SyntFeatureClosed
                featurename="CONTROLT"
                value="SUBJECTCONTROL"/>
    </Construction>
```

b)
```
<Construction
        id="t-indopt"
        syntlabel="Clause"
        selfinsertion="1">
            <InstantiatedPositionC
                range="0"
                optional="YESO"
                positionc="Psubj">
            <InstantiatedPositionC
                range="1"
                optional="NOO"
                positionc="Pobj">
            <InstantiatedPositionC
                range="2"
                optional="YESO"
                positionc="Pind"></Construction>
```

Nel quadro del progetto LE PAROLE, in cui sono stati codificati a livello sintattico 3000 verbi, è stato descritto un gran numero di comportamenti sintattici verbali. Nel progetto CLIPS, la codifica dei verbi si avvale di queste descrizioni creando nuove strutture sintattiche, secondo il modello di quelle esistenti, alla luce delle necessità emerse dalla codifica di nuove unità lessicali. Sono stati individuati, descritti e utilizzati nella codifica 725 comportamenti sintattici verbali (comprendenti sia le proprietà sintattiche del verbo che la descrizione del suo contesto sintattico) corrispondenti a 671 diversi quadri di sottocategorizzazione (descrizione del solo contesto sintattico).

L'informazione concernente i verbi viene codificata a livello di realizzazione di posizione, tenendo conto dei seguenti criteri:

- categoria grammaticale;
- funzione sintattica;
- tratti morfosintattici;
- tratti lessicali;
- informazione relativa al controllo.

Nel lessico SIMPLE-CLIPS vengono trattate sintatticamente varie strutture verbali: verbi transitivi, intransitivi, intransitivi pronominali, transitivi pronominali, riflessivi apparenti o impropri, riflessivi propri, reciproci, modali, costruzioni a sollevamento, costruzioni con complementi predicativi e costruzioni impersonali. Sono inoltre prese in considerazione strutture argomentali monovalenti, bivalenti, trivalenti e tetravalenti.

Dal punto di vista semantico, in SIMPLE sono stati valutati due possibili tipi di approccio al predicato:

1. il primo consiste nella definizione di predicati primitivi astratti, che hanno il vantaggio di essere validi a livello multilingue; ad esempio i verbi di sentimento USem*amare*, USem*sentire*, sono collegati ad un PredFEEL;
2. la seconda strategia prevede la definizione di *predicati lessicali*, secondo cui *l'espressione del predicato coincide con il nome di una USem*: ad esempio i verbi di sentimento USem*amare*, USem*sentire*, ricevono i Pred*amare* e Pred*sentire*.

In SIMPLE, nonché in CLIPS, si è optato per il secondo tipo di approccio, cioè per il criterio del predicato lessicale, che è senz'altro più sensibile alla lingua e facilita, tra le altre cose, l'esplicitazione della corrispondenza tra i livelli sintattico e semantico.

La rappresentazione predicativa comprende le seguenti informazioni:

- *struttura argomentale* del predicato, nel senso del numero degli argomenti;
- *assegnazione* del predicato;
- *tipo di legame* che l'entrata intrattiene con il suo predicato;
- descrizione della struttura argomentale: *ruoli semantici* e *restrizione di selezione degli argomenti*;
- esplicitazione della *corrispondenza tra posizioni del quadro sintattico e argomenti del frame semantico*.

Una volta descritto e rappresentato nel lessico, il predicato instaura una fitta rete di corrispondenze che stabiliscono delle connessioni tra gli argomenti semantici individuati e le posizioni da questi assunte nella costruzione sintattica, e permettono inoltre di legare i partecipanti dello scenario espresso dal predicato alla concreta realizzazione della lingua. La corrispondenza tra argomenti e posizioni, che possono essere fino a quattro elementi, è di diversi tipi:

- *isomorphic*, quando la valenza è la stessa per i due livelli di descrizione (*monovalent*: un argomento, una posizione sintattica, *bivalent*: due argomenti, due posizioni sintattiche) e c'è perfetta corrispondenza per cui tutte le posizioni sintattiche mappano gli argomenti semantici nello stesso ordine di posizione;
- *crossed*, quando la valenza è la stessa, ma l'ordine delle posizioni e degli argomenti è diverso. Questo tipo di legame è assegnato ai nomi deverbali quando, per esempio,

nella costruzione nominale il primo argomento (Arg0) corrisponde alla seconda posizione del predicato (Pos1) e il secondo argomento (Arg1), il paziente, corrisponde alla posizione zero (Pos0).

(10) ARG0 *costruire* **ARG1**
la costruzione **dell'edificio** (Pos0) da parte della ditta (Pos1)

- *augmented*, si verifica quando un predicato possiede più argomenti rispetto alle posizioni del quadro sintattico, ed esistono quindi argomenti che non sono mappati sulle posizioni sintattiche: è il caso ad esempio di un argomento che è incluso nel significato di una parola, ma che non fa parte del suo quadro sintattico, ovvero il classico argomento-ombra.

(11) a. Sciare **con gli sci**
b. Inscatolare **in una scatola**

- *reduced*, quando le posizioni del quadro sintattico sono più numerose degli argomenti del predicato, e conseguentemente un complemento sintattico non figura come argomento nel *frame* semantico.

(12) *Il motore non carbura bene*
la frase in sintassi ha due posizioni, poiché l'*adjunct* occupa una posizione, i-adjadvp-xa; Pos0 Pos1. Ciò corrisponde semanticamente ad un predicato Pred*carburare*-Arg0, nel quale l'*adjunct* della sintassi non compare come argomento.

Riportiamo di seguito l'elenco della categoria verbale dei verbi di movimento, inserita nel progetto SIMPLE-CLIPS:

**MOVE**

- ballare
- camminare
- contorcere
- danza
- danzare
- dondolare
- dondolio
- frullare
- movimento
- muovere
- nuotare
- oscillare
- precedere
- salire
- seguire
- strisciare

**CAUSED MOTION**

- calciare
- dondolare

**CAUSE ACT**

- esplodere
- muovere

**CHANGE OF LOCATION**

- muovere
- partenza
- partire
- salire
- spostare
- viaggiare

**CAUSE CHANGE LOCATION**

- muovere
- spostamento
- spostare

È evidente, dall'analisi della categoria proposta, come il fattore discriminante nella classificazione sia la causalità o meno del movimento o del cambiamento di stato; ciò è dovuto al forte peso riconosciuto in SIMPLE-CLIPS all'alternanza causativo-incoativo, nella classificazione dei verbi.

Inoltre, come in genere avviene, la classificazione di Levin, in quanto manuale, risulta essere estremamente più granulare e ridondante rispetto a quella proposta nel lessico sopra menzionato.

Il lessico SIMPLE-CLIPS oltre ad avere il merito di tentare di superare, con la sua complessa architettura, i limiti di quelle ontologie che spesso appiattiscono la ricchezza concettuale delle entrate lessicali sulla sola dimensione tassonomica, ha anche un valore aggiunto nella ricusabilità dei dati ottenuti, non solo in ambito monolingue, ma anche in un panorama plurilingue, visto che il modello è stato concepito in modo da stabilire le basi per un successivo collegamento fra i lessici creati per le varie lingue europee.

# Capitolo 3
# Approcci computazionali alla classificazione verbale

## 3.1 La famiglia dei *Word Space Models*

Come già accennato, all'interno dei modelli teorici rappresentativi del lessico è possibile rintracciare una seconda famiglia che è quella dei cosiddetti *word space models* o modelli semantici distribuzionali. Tali modelli non si basano su una rappresentazione e classificazione dei dati ottenuta *a priori* con l'intervento del ricercatore, bensì sull'utilizzo di risorse linguistiche computazionali, elaborate allo scopo di costruire classificazioni automatiche dei dati disponibili.

A sostegno di questi modelli si trova la concezione del lessico inteso come uno spazio semantico di parole e quindi il parallelismo tra le proprietà del significato e quelle dello spazio inteso in senso geometrico.

Sviluppando l'assunto per cui il significato è uno spazio di parole, il lessico viene concepito di conseguenza come uno spazio metrico i cui elementi, ovvero le parole, sono separati da distanze più o meno ampie, in virtù del loro grado di similarità.

> Vector similarity is the only information present in WordSpace: semantically related words are close, unrelated words are distant (Schütze, 1993: 896)

I modelli basati sui vettori rappresentano il significato delle parole utilizzando come unici dati le statistiche distribuzionali; poiché si ritiene che sia il contesto che circonda una certa parola a fornire le informazioni fondamentali sul suo significato (Harris (1968)), la similarità semantica sarà misurabile attraverso le distribuzioni statistiche di co-occorrenza delle parole stesse nei testi. Il principio che soggiace a questa tipologia di modelli spaziali prende il nome di *ipotesi distribuzionale* e prevede che due parole sono tanto più semanticamente simili, quanto più tendono a ricorrere in contesti linguistici simili (Miller e Charles (1991)).

> You shall know a word by the company it keeps (Firth, 1957: 11)

Nei *word space models* dunque, i significati non vengono organizzati secondo lo schema delle definizioni dei sensi in un dizionario, ma invece secondo rappresentazioni contestuali, che incarnano l'astrazione dell'informazione nell'insieme dei contesti linguistici naturali in cui una certa parola ricorre (Charles (2000)). L'ipotesi distribuzionale ha guadagnato terreno grazie alla disponibilità di *corpora* testuali di grandi dimensioni e di tecniche statistiche sempre più raffinate e adatte all'estrazione di schemi distribuzionali dei lessemi. In tal modo essa trova la sua attuazione concreta in modelli computazionali per la costruzione di spazi semantico-lessicali. Lo scopo di questi modelli è quello di riuscire a quantificare la similarità semantica tra le parole, ovvero di valutare la misura in cui si sovrappongono i contesti linguistici in cui esse ricorrono.

All'interno di questo schema generale possiamo applicare differenti variabili, algoritmiche e rappresentazionali, che saranno scelte alternativamente nella costruzione degli spazi distribuzionali, proprio in virtù delle diverse finalità teoriche o applicative cui ciascun modello ambisce (Lenci (2009)).

Ciò che accomuna queste diverse realizzazioni dell'ipotesi distribuzionale, è il principio per cui misurare la similarità semantica tra due parole corrisponda a definire la misura in cui i contesti linguistici in cui esse ricorrono si sovrappongono.

Il concetto di spazio semantico viene affiancato per analogia a quello di spazio geometrico: in quest'ultimo a ciascun punto dello spazio si associa un vettore composto da *n* numeri che rappresentano le sue coordinate rispetto ad *n* assi cartesiani, ovvero le dimensioni dello spazio; allo stesso modo, il contenuto di una parola sarà rappresentato dalla sua posizione all'interno di uno spazio definito da un sistema di coordinate, definito dai contesti linguistici in cui la parola stessa può ricorrere (Lenci (2009)).

A livello formale, lo spazio semantico è determinato dalle quattro coordinate <*T*, *B*, *M*, *S*> (Lowe (2001), Padò e Lapata (2007)):

- *T* è l'insieme delle parole *target* che formano gli elementi che compongono lo spazio e di cui questo ci dà una rappresentazione semantica;
- *B* è la base che definisce le dimensioni stesse dello spazio e contiene i contesti linguistici rispetto ai quali viene valutata la similarità distribuzionale delle parole *target*;
- *M* è una matrice di co-occorrenza, che contiene la rappresentazione vettoriale di ogni parola presente in *T*. Ognuna di queste corrisponde ad una riga della matrice *M*, le cui colonne presentano invece gli elementi di *B*. Il valore di una cella della matrice indica allora la frequenza di co-occorrenza della parola considerata in un dato contesto di riferimento; si consideri nella Tabella 3.1 il caso della parola *presidente* che ricorre 7 volte nel contesto di *repubblica*, nel quale non figurano invece mai né torta né panino ad esempio (Lenci (2009)).
-

*Tabella 3.1 Matrice di co-occorrenza tra le parole*

| | **dire** | **mangiare** | **aprire** | **pensare** | **repubblica** | **gustoso** |
|---|---|---|---|---|---|---|
| **ministro** | 6 | 2 | 5 | 4 | 1 | 0 |
| **presidente** | 10 | 3 | 2 | 3 | 7 | 0 |
| **torta** | 0 | 4 | 2 | 0 | 0 | 3 |
| **panino** | 0 | 7 | 0 | 0 | 0 | 1 |

- *S* è la metrica che misura la distanza tra i punti all'interno dello spazio semantico. Per determinare la posizione di due parole occorre, infatti, comparare i loro vettori rispetto a tutte le dimensioni che li costituiscono. Maggiore è il numero di dimensioni nelle quali i due vettori presentano valori simili, maggiore sarà la loro vicinanza nello spazio e quindi anche la loro similarità semantica.

  Una delle misure più comuni di vicinanza spaziale tra due vettori normalizzati è il coseno dell'angolo che essi formano: se due vettori sono geometricamente allineati sulla stessa linea nella stessa direzione, l'angolo tra di loro è 0°, il coseno è 1 e la similarità è massima. Al contrario se i due vettori sono indipendenti o ortogonali, il loro angolo è prossimo a 90°, il coseno equivale a 0 e si è in assenza di similarità. Altri tipi di misurazione della distanza tra i vettori sono la *metrica euclidea* e la *Manhattan distance*.

Il ricorso al parametro della vicinanza spaziale per definire la similarità semantica, non è né accidentale, né tanto meno arbitrario. Sembra essere invece un sistema estremamente intuitivo e naturale per la descrizione del concetto di similarità. Lakoff e Johnson nei loro lavori (Lakoff e Johnson (1980), (1999)) hanno proprio evidenziato come le metafore costituiscano sia il materiale grezzo delle idee astratte, che i nostri strumenti di base per iniziare a ragionare su fenomeni complessi, ad esempio sul linguaggio o il significato linguistico. Secondo i due studiosi le metafore che andiamo a formulare in tal senso, compongono un nucleo piuttosto limitato ed essenziale, direttamente legato alla nostra esistenza fisica nel mondo. Le relazioni spaziali sono assolutamente salienti in questo ambito: collocazione, direzione, vicinanza sono tutte proprietà imprescindibili della nostra esistenza

corporea. Tra le metafore fondamentali costruite su questi elementi, è imprescindibile quella definibile come *similarity-is-proximity*:

> two things that are deemed to be similar in some sense are conceptualized as being close to or near each other, while dissimilar things are conceptualized as being far apart or distant from each other […] This also applies to meanings: it is intuitive, if not evitable, to use the similarity-is-proximity metaphor when talking about similarities of meaning. Words with similar meanings are conceptualized as being far apart (Sahlgren, 2006: 19)

Naturalmente la realizzazione della metafora *similarity-is-proximity* presuppone l'intervento di un'ulteriore metafora geometrica, ovvero quella per cui *entities-are-locations*. Difatti perché due cose possano essere descritte come vicine, esse devono necessariamente possedere la dimensione della spazialità, cioè devono occupare collocazioni diverse nello spazio concettuale. Quando ci riferiamo ai significati apostrofandoli come reciprocamente vicini o lontani, inevitabilmente li definiamo come entità all'interno di uno spazio semantico, nel quale la loro distanza diventa misurabile. I *word space models* catturano le proprietà semantiche delle parole in uno spazio multidimensionale, tramite vettori costruiti su grandi *corpora*, nonché osservando i modelli distribuzionali di co-occorrenza tra parole vicine.

Se si lavora con i modelli semantici distribuzionali, non ha senso concettualizzare una singola parola o una singola collocazione, poiché ciò non ci aiuterebbe a migliorare la nostra conoscenza della stessa. È solo quando lo spazio si popola di altre parole che la concettualizzazione acquista valore; se assumiamo che il significato è una relazione tra parole, allora è chiaro che in un sistema relazionale di questo tipo, non si può parlare dei vari sensi di una parola isolata. Il significato ed anche il vettore corrispondente, che lo rappresenta geometricamente, non ha alcun valore di per sé, ma serve invece a stabilirne la posizione nello spazio rispetto ai contesti linguistici in cui ricorre.

In un simile modello rappresentativo il significato viene perciò considerato come:

> a relation among words. In such a relational system, one cannot talk about the meaning of a word in isolation; words have meaning only in virtue of their relations to other words – meaning is a property of the system as a whole (Kintsch, 2007:91)

Il significato nasce solo dalle configurazioni di punti nello spazio, distribuiti proporzionalmente al loro grado di similarità. Conseguenza di quanto appena detto è che le dimensioni del vettore non sono interpretabili direttamente e singolarmente, poiché

> The distribution of an element will be understood as the sum of all its environments (Harris, 1970: 775)

L'identificazione del significato come spazio di parole, genera un modello di rappresentazione semantica sostanzialmente diverso da quello tipico della tradizione linguistica, fondato piuttosto su strutture simboliche quali *frames*, tratti semantici e così via.

Mentre i modelli ontologici assegnano al contesto una funzione prettamente *discriminativa*, poiché qui esso agisce come elemento disambiguante, che permette di selezionare nel repertorio dei sensi di una parola un certo significato ritenuto appropriato in una specifica situazione d'uso. I *word space models* invece capovolgono questa prospettiva e attribuiscono al contesto una funzione *costitutiva* del significato, per cui il contenuto informativo di una parola trae fondamento dai contesti linguistici cui esso partecipa e su cui si plasma (Charles (2000)).

Altre differenze fondanti rispetto ad ontologie come ad esempio WordNet sono le seguenti:

- gli elementi dello spazio nei *word space models* sono parole e non entità concettuali o sensi come nelle reti semantiche;
- nei *word space models* il contenuto semantico di un lessema nasce solo dai suoi rapporti di similarità distribuzionale tradotti in distanze nello spazio;

- le reti semantiche sono intrinsecamente discrete a livello strutturale e le relazioni che legano i nodi in esse contenute sono di tipo qualitativo; invece i *word space models* presentano una struttura continua e puramente quantitativa (quantificano la distanza dei punti).

Abbiamo visto come i modelli semantici distribuzionali utilizzino la metafora geometrica del significato come base rappresentativa. Ma un *word space model* non è solo una rappresentazione spaziale dei significati, esso è anche un modo per costruire tale spazio. Ciò che distingue infatti questo tipo di modelli da altri modelli geometrici del significato, è proprio che lo spazio stesso è costruito senza alcun intervento umano, e senza il ricorso ad alcuna conoscenza aprioristica o a vincoli sulle affinità del significato. La similarità tra le parole è estratta automaticamente dai dati linguistici, tramite l'osservazione empirica del loro effettivo uso nel linguaggio che li ospita. I dati che popolano i *word space models* sono derivati da statistiche sulle proprietà distributive di co-occorrenza delle parole. Come già detto l'idea è proprio quella di sistemare parole con proprietà distributive simili in regioni simili dello spazio semantico, cosicché il grado di vicinanza reciproca sarà il riflesso del grado di similarità linguistica distributiva.

È inevitabile sottolineare il ruolo cruciale dell'ipotesi distribuzionale per connettere lo spazio algebrico, nel quale le distanze tra parole sono determinate da una metrica che dipende dalle statistiche di co-occorrenza delle stesse, al contenuto semantico dei termini rappresentati. Lo spazio vettoriale si limita a registrare le posizioni delle parole nei contesti linguistici che le ospitano; il coseno poi misurerà la similarità degli schemi distribuzionali delle parole, per cui due parole vicine nello spazio vettoriale sono soltanto due parole che presentano distribuzioni statistiche di co-occorrenza simili nei contesti linguistici. È però l'ipotesi distribuzionale che realizza l'accostamento tra la similarità nelle distribuzioni e la similarità corrispondente nel significato, interpretando lo spazio geometrico come spazio semantico, cioè come uno spazio che definisce il contenuto semantico delle parole che lo popolano.

Nella Tabella 3.2 riportiamo un'importante caratteristica della competenza lessicale fornita dai *word space models*, vale a dire quella sui giudizi di similarità semantica tra parole. La tabella presenta infatti le distanze che intercorrono tra alcuni nomi della lingua italiana presi dal *corpus* di *La Repubblica*, calcolate grazie alla misurazione dei rispettivi coseni (Lenci (2009)).

*Tabella 3.2 Coseni tra parole in uno spazio distribuzionale*

| | cane | animale | sentimento | odio | auto | aereo | presidente |
|---|---|---|---|---|---|---|---|
| **animale** | **0.53** | | | | | | |
| **sentimento** | 0.04 | 0.14 | | | | | |
| **odio** | -0.01 | 0.05 | **0.52** | | | | |
| **auto** | 0.22 | 0.10 | -0.04 | -0.07 | | | |
| **aereo** | 0.15 | 0.03 | 0.03 | -0.03 | **0.25** | | |
| **presidente** | -0.03 | -0.01 | -0.02 | 0.04 | -0.005 | 0.03 | |
| **ministro** | 0.04 | 0.07 | -0.06 | -0.03 | 0.002 | 0.02 | **0.16** |

Ad un maggiore valore del coseno corrisponde una minore distanza delle parole all'interno dello spazio distribuzionale; in effetti nella tabella le parole più simili da un punto di vista semantico, come *auto* e *aereo*, presentano anche un coseno più elevato (valori in grassetto). Sembra esserci dunque una corrispondenza tangibile tra l'idea del significato come spazio di parole e le intuizioni semantiche dei parlanti di una lingua; conseguentemente la

similarità semantica di due parole può essere correttamente misurata attraverso la loro proiezione in uno spazio distribuzionale, quale quello finora descritto.

I modelli che descriveremo nelle sezioni seguenti non sono altro che applicazioni dei *word space models* a *tasks* specifici, quale quello della categorizzazione verbale. Difatti le distribuzioni di co-occorrenza sono un valido sistema per identificare le proprietà semantiche salienti dei verbi; tali distribuzioni possono poi essere formalmente rappresentate come spazi vettoriali distribuzionali. Ad ogni verbo è associato un vettore che ne riassume varie caratteristiche distribuzionali: ad esempio quante volte il verbo compare con un certo *frame* come in Schulte im Walde, o con un certo tratto precedentemente selezionato come in Merlo & Stevenson e Joanis.

## 3.2 Esempi di classificazioni automatiche per i verbi inglesi

In ogni lingua i verbi possono essere raggruppati insieme, all'interno di classi semantiche che condividono componenti comuni del significato dei singoli membri. È stato dimostrato a livello linguistico come i verbi appartenenti a tali classi condividano anche caratteristiche sintattiche, dal momento che la semantica di un verbo ne determina, anche se solo parzialmente, il comportamento sintattico (Pinker (1989), Levin (1993)).

In linguistica computazionale, il legame semantico-sintattico è stato sfruttato per costruire classificazioni semantiche automatiche dei verbi, utili nella costruzione di lessici disponibili per un ampio numero di applicazioni, quali la traduzione automatica (Dorr (1997)), la generazione del linguaggio naturale (Stede (1999)) e l'*information retrieval* (Klavans e Kan (1998)).

Recentemente sono state molte le ricerche che hanno lavorato su tale principio per automatizzare, almeno in parte, il processo di classificazione verbale; gli studi prodotti si fondano sull'individuazione di tratti specifici per quantificare o descrivere le caratteristiche sintattiche dei verbi, postulando che l'informazione sull'uso sintattico di un certo verbo può fornire indicazioni sulla semantica dello stesso.

### 3.2.1 Il lavoro di Merlo e Stevenson

Come risorsa alternativa alle classificazioni manuali, come quella proposta da Levin (1993) per i verbi inglesi, vengono applicati dei metodi computazionali, quali il *clustering*, per inferire una classificazione verbale a partire da un *corpus* di dati. In rapporto alla tipologia di classi verbali da ricavare, gli approcci automatici modificano la scelta degli algoritmi di *clustering* utilizzati per sviluppare la classificazione. Un altro parametro fondamentale e al contempo estremamente variabile nell'induzione automatica delle classi semantiche verbali, è la scelta delle caratteristiche valutate rilevanti per ciascun verbo.

Dato che il *target* della classificazione determina la similarità, o la dissimilarità tra i verbi, la selezione di alcune caratteristiche come tratti imprescindibili nella definizione di un verbo, a scapito di altre, sarà un criterio primario nell'orientamento della similarità di interesse.

Un interessante esperimento di classificazione è stato condotto nello studio di Merlo e Stevenson (2001), il cui intento è appunto di classificare automaticamente 60 verbi inglesi sulla base della loro struttura argomentale, cioè della loro capacità di assegnare ruoli tematici ai partecipanti dell'enunciato. Il metodo di indagine è *corpus-based* ed utilizza tecniche di analisi statistiche.

Le due ricercatrici si sono avvicinate al problema della classificazione ritenendolo un attendibile strumento per descrivere l'organizzazione lessicale, in grado di catturare le proprietà generali dei verbi, prescindendo invece da quelle idiosincratiche (Palmer (2000)).

I verbi considerati vengono ripartiti nelle tre classi atte in inglese a contenere verbi opzionalmente intransitivi:

1. *unergative;*
2. *unaccusative;*
3. *object-drop.*

La distinzione nelle tre classi di appartenenza è motivata da proprietà linguistiche direttamente collegate alla struttura argomentale dei verbi; esse catturano, in effetti, le distinzioni nell'assegnazione dei ruoli tematici ai diversi *frames*.

*Tabella 3.3 Lista dei verbi usati negli esperimenti*

**Table 5**
Verbs used in the experiments.

| Class Name | Description | Selected Verbs |
|---|---|---|
| Unergative | manner of motion | *jumped, rushed, marched, leaped, floated, raced, hurried, wandered, vaulted, paraded* (group 1); *galloped, glided, hiked, hopped, jogged, scooted, scurried, skipped, tiptoed, trotted* (group 2). |
| Unaccusative | change of state | *opened, exploded, flooded, dissolved, cracked, hardened, boiled, melted, fractured, solidified* (group 1); *collapsed, cooled, folded, widened, changed, cleared, divided, simmered, stabilized* (group 2). |
| Object-Drop | unexpressed object alternation | *played, painted, kicked, carved, reaped, washed, danced, yelled, typed, knitted* (group 1); *borrowed, inherited, organized, rented, sketched, cleaned, packed, studied, swallowed, called* (group 2). |

Va sottolineato che le varie classi presentano gli stessi *frames* di sottocategorizzazione, pertanto un'indagine condotta unicamente su questo parametro sarebbe poco significativa. Per contro, ogni classe attesta un'unica possibilità di assegnamento tematico, la quale permette di inserire i verbi in modo specifico all'interno della relativa classe di riferimento. Gli indicatori statistici utilizzati per attribuire a ciascun verbo opzionalmente transitivo la propria classe di appartenenza, individuano l'informazione necessaria attraverso la sola alternanza transitivo *vs* intransitivo. Saranno questi indicatori a costituire l'*input* per un algoritmo, capace di produrre una classificazione automatica per il sistema delle nostre tre classi verbali.

*Tabella 3.4 Sommario delle assegnazioni dei ruoli tematici per ogni classe*

| | transitive | | intransitive |
|---|---|---|---|
| Classes | Subject | Object | Subject |
| Unergative | Agent (of causation) | Agent | Agent |
| Unaccusative | Agent (of causation) | Theme | Theme |
| Object-drop | | Theme | Agent |

Effettivamente tutte e tre le classi considerate, partecipano all'alternanza diatetica che correla una forma transitiva ed un'altra intransitiva dello stesso verbo;

(1) Unergative: a) The horse raced past the barn
b) The jockey raced the horse past the barn

Unaccusative: a) The butter melted in the pan
b) The cook melted the butter in the pan

Object-drop: a) The boy played
b) The boy played soccer

ma ognuna presenta un particolare tipo di alternanza diatetica, determinato dalle relazioni semantiche degli argomenti col verbo stesso.

Sebbene la granularità della classificazione proposta differisca da quella presentata da Levin nei suoi lavori, Merlo e Stevenson concordano con lei nell'ipotesi di fondo per cui le proprietà semantiche di un verbo si riflettono nel suo comportamento sintattico. Il comportamento su cui focalizza Levin è quello dell'alternanza diatetica; anche i verbi considerati da Merlo e Stevenson sottostanno a questo principio e presentano una diversa proiezione con i suoli semantici per ciò che concerne la realizzazione dell'alternanza transitivo/intransitivo. La partecipazione o meno di un verbo ad una data alternanza diatetica è un fattore chiave nel progetto di Levin per classificare il verbo stesso; in Merlo e Stevenson invece, come in altri lavori computazionali, l'idea è stata ampliata mostrando come le statistiche sulle alternanze di un verbo sono effettivamente in grado di catturare l'informazione sulla sua classe di appartenenza (Lapata (1999), McCarthy (2000), Lapata e Brew (1999)).

Merlo e Stevenson verificano quindi l'assunto per cui le distinzioni evidenziate si riflettono a livello statistico nei *corpora* indagati, tramite una metodologia computazionale sperimentale (Merlo e Stevenson (2001)), in grado di dar conto delle tre ipotesi seguenti:

1. le caratteristiche lessicali catturano le differenze a livello di struttura argomentale delle diverse classi; la capacità di determinare quale sia l'insieme delle caratteristiche rilevante nel distinguere le classi verbali considerate, è la chiave per la riuscita di qualsiasi classificazione automatica (Merlo e Stevenson (2001)). Le proprietà semantiche rilevanti delle classi verbali (*causativity, transitivity, animacy of subject* etc.) sono approssimativamente stimabili, enumerando i tratti sintattici che le realizzano nella frase;
2. i tratti linguisticamente distintivi mostrano differenze distribuzionali per le tre classi, evidentemente correlate all'esperienza linguistica, poiché riscontrabili a livello testuale, nel *corpus* di riferimento;
3. le distribuzioni statistiche delle caratteristiche esaminate contribuiscono a determinare una classificazione verbale automatica.

L'utilizzo di metodi statistici *corpus-based* per ottenere il livello di classificazione desiderato, poggia sull'ipotesi per cui: le differenze tra le varie classi verbali per ciò che riguarda la loro struttura argomentale, si riflettono statisticamente sull'uso dei verbi che le compongono, e tali statistiche possono essere estratte automaticamente da un ampio *corpus* annotato.

I risultati degli esperimenti condotti hanno raggiunto un'accuratezza del 69,8%; pertanto le tecniche di apprendimento basate sull'analisi dei *corpora*, a partire da informazioni correlate agli usi linguistici, costituiscono un buon punto di partenza per lo studio dei comportamenti linguistici umani.

### 3.2.2 Gli studi condotti da McCarthy e Korhonen

Una parte importante del lavoro di produzione di una classificazione verbale automatica, ispirata comunque a quella realizzata manualmente da Levin (1993), è quella

della verifica della partecipazione o meno di un dato verbo ad una specifica alternanza diatetica.

McCarthy e Korhonen (1998), propongono proprio un metodo di identificazione automatica della partecipazione verbale ad un'alternanza diatetica. Il progetto si basa sempre sull'assunto per cui ai cambiamenti sintattici si accompagnano dei cambiamenti anche nel significato del verbo, e sull'importanza rivestita dalle alternanze nella classificazione del verbo stesso, per cui un certo ruolo semantico ricopre diversi ruoli grammaticali in realizzazioni alternanti (*Role Switching Alternation*).

Secondo McCarthy e Korhonen, le alternanze riducono drasticamente il fenomeno linguistico della ridondanza nel lessico, dato che i *frames* di sottocategorizzazione non devono essere enumerati per ogni verbo se si ricorre all'utilizzo di un *marker* che specifichi a quali verbi l'alternanza si applichi. Inoltre, attraverso le alternanze, vengono generalizzati aspetti del comportamento inerenti interi gruppi di verbi, poiché di solito i membri del gruppo sono semanticamente correlati.

A diverse alternanze diatetiche corrisponde una diversa enfasi e sfumatura del significato, anche se rispetto ad uno stesso contenuto di base; questi sottili cambiamenti di significato risultano assai importanti nei processi di *Natural Language Generation*.

McCarthy e Korhonen partono dall'estrazione, con l'ausilio di un *parser*, di *frames* di sottocategorizzazione da *corpora* di riferimento, lavorando però, a differenza di Merlo e Stevenson, su un quantitativo di dati molto più ampio, che rende certo più difficoltosa l'individuazione dei tratti salienti del verbo. Una volta individuati, questi stessi *frames* vengono ripartiti in 161 classi e sottoposti all'analisi di un filtro, che elimina quelli che ricorrono nei vari verbi con una frequenza minore a quella attesa. Il lessico ottenuto dai risultanti *frames* di sottocategorizzazione estratti dal *parser*, conterrà la lista dei verbi contenuti nel *corpus* considerato, e i relativi *frames* ad esso abbinati (McCarthy (2000)). La frequenza con cui i *frames* ottenuti sono attestati nel *corpus* di riferimento, è un parametro altamente significativo per lo studio condotto, poiché ci fornisce una stima esatta della produttività delle varie alternanze. L'acquisizione automatica dei *frames* di sottocategorizzazione elimina le onerose revisioni tipiche di un approccio manuale e consente di valutare come il comportamento verbale possa variare in base ai sottolinguaggi, ai settori e al momento storico di riferimento.

Oltre ai *frames* di sottocategorizzazione sono state estratte le preferenze di selezione dei verbi considerati; queste sono state rappresentate attraverso dei modelli computazionali detti *Assocation Tree Cut Models* (Abe e Li (1998)), ovvero insiemi di classi semantiche che tagliano trasversalmente la gerarchia degli iperonimi di WordNet (Miller et al. (1993)), coprendo tutti i nodi disgiuntamente. I punteggi associativi dei verbi rispetto ai nomi rintracciati nell'ontologia, vengono calcolati per le varie classi a partire dalla frequenza dei nomi che ricorrono con un verbo *target* ed indipendentemente dal verbo stesso. Il punteggio ottenuto indicherà il grado di preferenza tra la classe (c) e il verbo (v) in una specifica *slot* (soggetto, oggetto o frasePP).

Nella Figura 3.1, vediamo un esempio di ATCM, che opera sulla *slot Oggetto* del verbo *build*; per un altro verbo sarebbe stato necessario un livello differente di taglio; ad esempio *eat* avrebbe richiesto un taglio a livello dell'iponimo *food* della categoria ontologica *object*.

Trovare il migliore insieme rappresentativo di classi è quindi la chiave per produrre un buon modello per le preferenze di selezione; per raggiungere questo obiettivo Abe e Li usano il principio della Minima Lunghezza Descrittiva (MDL), il quale afferma che il modello migliore minimizza la somma di *i* (il numero di *bits* necessari per codificare il modello) e di *ii* (ovvero il numero di *bits* che occorrono per codificare i dati all'intenro del modello). Grazie a questo sistema si ottiene un compromesso proficuo tra un modello semplice, ed un altro che invece descrive i dati in modo efficiente.

*Figura 3.1 ATCM per la slot Oggetto del verbo build*

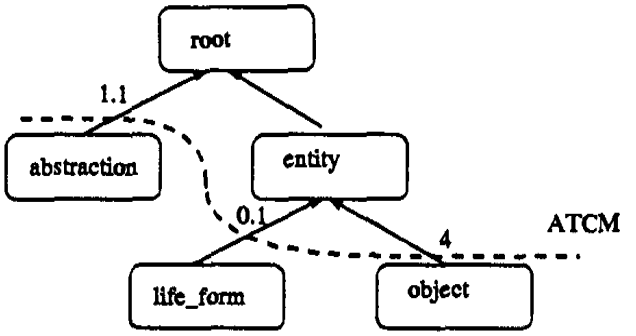



Figure 1: ATCM for *build* Object slot

P
er i verbi che partecipano ad un'alternanza ci si aspetta che i dati presenti nelle *slots* alternanti dei rispettivi *frames* di sottocategorizzazione siano piuttosto omogenei; ciò dipenderà dalla misura in cui l'alternanza è applicabile al significato predominante del verbo e alla maggioranza dei restanti significati degli argomenti cui esso si accompagna.

L'ipotesi di McCarthy e Korhonen è che se l'alternanza è ragionevolmente produttiva (realizzazione sintattica) e ricorre nella maggior parte delle frasi in cui un certo verbo compare, allora le preferenze di selezione relative alle *slots* interessate dall'alternanza stessa dovrebbero essere simili tra loro (componente semantica).

Proprio in virtù dell'importanza in questo tipo di esperimenti sulle alternanze diatetiche e del confronto dei dati estratti automaticamente con una classificazione sviluppata in modo manuale, Korhonen insieme a Briscoe (2004) ha lavorato all'estensione delle classi semantiche di Levin, elaborando:

- 57 nuove classi;
- 106 nuove alternanze diatetiche (realizzando uno strumento potenziato di supporto per l'individuazione delle nuove classi in modo semi-automatico).

L'utilità di disporre di nuove classi risiede nel fatto che una classificazione verbale più ampia, avrà una copertura più estensiva del lessico verbale inglese, specie per ciò che concerne l'acquisizione dei *frames* di sottocategorizzazione.

Le classificazioni disponibili fino a quel momento, anche quelle più quotate (Pinker (1989), Levin (1993)), si limitano tutte solo ad alcuni tipi di classi e vi includono solo pochi esemplari di verbi. Ad esempio, la tassonomia elaborata da Levin (1993) si riferisce principalmente a quei verbi che assumono come complementi nomi e frasiPP; inoltre fornisce un campione ridotto di membri per ciascuna classe proposta.

Già Dang et al. (1998) avevano arricchito la tassonomia di Levin corredandola di classi interstizie, ovvero di insiemi di classi speciali che racchiudevano tutti quei verbi che nel lavoro di Levin appartenevano a più di una classe in virtù del loro essere regolarmente polisemici.

La classificazione originaria prodotta da Levin è stata creata:

- selezionando manualmente un insieme di alternanze diatetiche dalle risorse linguistiche disponibili;
- classificando i verbi in rapporto alla loro appartenenza alle suddette alternanze;

- raggruppando i verbi in base alla loro partecipazione ad insiemi di alternanze.

Korhonen e Briscoe sviluppano invece un approccio diverso rispetto a quello appena descritto, procedendo a:

- assemblare un insieme di alternanze per i verbi non trattati da Levin (cfr. Tabella 3.5);
- selezionare dalle risorse linguistiche disponibili un insieme di classi semantiche per i verbi individuati nella fase precedente;
- valutare se l'insieme dei verbi di ogni classe candidata possono essere correlati tra loro con l'ausilio delle alternanze diatetiche, ed eventualmente creare nuove classi.

*Tabella 3.5 Esempi di alcune nuove alternanze introdotte*

| Category | Example Alternations | Alternating SCFs |
|---|---|---|
| Equi | *I advised Mary to go* ↔ *I advised Mary* | 53 ↔ 24 |
| | *He helped her bake the cake* ↔ *He helped bake the cake* | 33 ↔ 142 |
| Raising | *Julie strikes me as foolish* ↔ *Julie strikes me as a fool* | 143 ↔ 29 |
| | *He appeared to her to be ill* ↔ *It appeared to her that he was ill* | 99 ↔ 12 |
| Category switches | *He failed in attempting to climb* ↔ *He failed in the climb* | 63 ↔ 87 |
| | *I promised Mary to go* ↔ *I promised Mary that I will go* | 54 ↔ 52 |
| PP deletion | *Phil explained to him how to do it* ↔ *Phil explained how to do it* | 90 ↔ 17 |
| | *He contracted with him for the man to go* ↔ *He contracted for the man to go* | 88 ↔ 15 |
| P/C deletion | *I prefer for her to do it* ↔ *I prefer her to do it* | 15 ↔ 53 |
| | *They asked about what to do* ↔ *They asked what to do* | 73 ↔ 116 |

Ogni classe semantica individuata è stata valutata come segue:

1. sono stati estratti tutti i *frames* di sottocategorizzazione dei verbi membri;
2. grazie ad un confronto con la tassonomia di Levin e con la lista delle 106 nuove alternanze diatetiche precedentemente ottenute, sono state estrapolate tutte le alternanze in cui tali *frames* di sottocategorizzazione comparivano;
3. laddove siano state rintracciate una o più alternanze e sia stato identificato un numero minimo di due verbi membri, è stata creata una nuova classe verbale.

Le fasi 1 e 2 sono state realizzate attraverso procedure automatiche, mentre la terza è stata prodotta manualmente; si parla pertanto di una procedura di tipo semi-automatico.

L'ipotesi in sintesi è che: le alternanze giudicate inizialmente rilevanti ci aiutino ad estrarre *frames* di sottocategorizzazione aggiuntivi, che a loro volta ci indirizzano verso l'identificazione di nuove alternanze aggiuntive.

I *frames* di sottocategorizzazione e le alternanze così ottenuti, costituiranno la descrizione sintattico-semantica di ogni nuova classe. Per tutte quelle classi che presentano un numero insufficiente di membri al loro interno, vengono cercati nuovi elementi tramite WordNet (Miller (1990)). Sebbene WordNet classifichi i verbi su base puramente semantica, le regolarità sintattiche studiate da Levin si riflettono, in una certa misura, nella correlazione semantica espressa dalla particolare struttura dell'ontologia stessa (Fellbaum (1999)). A tale proposito, Dorr e Jones (1996) e Dorr (1997) hanno dimostrato nei loro studi come verbi rappresentati come sinonimi in WordNet, esibiscono un comportamento sintattico simile a quello che caratterizza il sistema di classificazione di Levin.

Pertanto vengono così rintracciati nuovi verbi tramite l'analisi dei sinonimi, dei troponimi, degli iperonimi e dei termini coordinati e/o antonimi dei verbi membri già individuati e disposti, come spiegato precedentemente, nelle nuove classi verbali risultanti (cfr. Tabella 3.6).

Un simile procedimento sottende come suo fondamento la già menzionata ipotesi semantico-sintattica: ovvero il fatto che verbi con significato affine, rivelino comportamenti sintattici altrettanto simili.

*Tabella 3.6 Nuove classi verbali*

| Class | | Example Verbs |
|---|---|---|
| 1. | URGE | *ask, persuade* |
| 2. | FORCE | *manipulate, pressure* |
| 3. | ORDER | *command, require* |
| 4. | WANT | *need, want* |
| 5. | TRY | *attempt, try* |
| 6. | WISH | *hope, expect* |
| 7. | ENFORCE | *impose, risk* |
| 8. | ALLOW | *allow, permit* |
| 9. | ADMIT | *include, welcome* |
| 10. | CONSUME | *spend, waste* |
| 11. | PAY | *pay, spend* |
| 12. | FORBID | *prohibit, ban* |
| 13. | REFRAIN | *abstain, refrain* |
| 14. | RELY | *bet, count* |
| 15. | CONVERT | *convert, switch* |
| 16. | SHIFT | *resort, return* |
| 17. | ALLOW | *allow, permit* |
| 18. | HELP | *aid, assist* |
| 19. | COOPERATE | *collaborate, work* |
| 20. | SUCCEED | *fail, manage* |
| 21. | NEGLECT | *omit, fail* |
| 22. | LIMIT | *restrict, restrain* |
| 23. | APPROVE | *accept, object* |
| 24. | ENQUIRE | *ask, consult* |
| 25. | CONFESS | *acknowledge, reveal* |
| 26. | INDICATE | *demonstrate, imply* |
| 27. | DEDICATE | *devote, commit* |
| 28. | FREE | *cure, relieve* |
| 29. | SUSPECT | *accuse, condemn* |
| 30. | WITHDRAW | *retreat, retire* |
| 31. | COPE | *handle, deal* |
| 32. | DISCOVER | *hear, learn* |
| 33. | MIX | *pair, mix* |
| 34. | CORRELATE | *coincide, alternate* |
| 35. | CONSIDER | *imagine, remember* |
| 36. | SEE | *notice, feel* |
| 37. | LOVE | *like, hate* |
| 38. | FOCUS | *focus, concentrate* |
| 39. | CARE | *mind, worry* |
| 40. | DISCUSS | *debate, argue* |
| 41. | BATTLE | *fight, communicate* |
| 42. | SETTLE | *agree, contract* |
| 43. | SHOW | *demonstrate, quote* |
| 44. | ALLOW | *allow, permit* |
| 45. | EXPLAIN | *write, read* |
| 46. | LECTURE | *comment, remark* |
| 47. | SUGGEST | *propose, recommend* |
| 48. | OCCUR | *happen, occur* |
| 49. | MATTER | *count, weight* |
| 50. | AVOID | *miss, boycott* |
| 51. | HESITATE | *loiter, hesitate* |
| 52. | BEGIN | *continue, resume* |
| 53. | STOP | *terminate, finish* |
| 54. | NEGLECT | *overlook, neglect* |
| 55. | CHARGE | *commit, charge* |
| 56. | REACH | *arrive, hit* |
| 57. | ADOPT | *assume, adopt* |

### 3.2.3 L'esperimento condotto da Joanis

In questo lavoro, Joanis (2003) parte dall'espansione dell'approccio statistico condotto da Merlo e Stevenson (2001), per definire uno spazio generale di caratteristiche utile nella classificazione dei verbi inglesi. I tratti individuati da Joanis sono appunto degli indicatori statistici dell'uso generale del verbo, ad esempio, dell'uso specifico delle *slots* sintattiche, e della partecipazione a diverse alternanze tra quelle definite nella prima parte del lavoro di Levin (1993). Nello sviluppo dello spazio contenente tali caratteristiche, sono stati inclusi degli indicatori che sono potenzialmente utili per la classificazione verbale, ma che sono anche facilmente estraibili dal *corpus* di riferimento.

Joanis si è poi occupato di stimare il valore dei tratti individuati, a partire dal conteggio delle frequenze dell'uso specifico dei vari verbi all'interno di un ampio *corpus* di riferimento per l'inglese, ovvero il *British National Corpus*. A differenza del *Wall Street Journal Corpus*, utilizzato in alcuni dei precedenti lavori correlati (Merlo e Stevenson (2001)), il *BNC* è un *corpus* generale; ciò vuol dire che se nel *WSJ* ci si può attendere che prevalga il significato finanziario di un qualsiasi verbo polisemico, nel *BNC* al contrario non ci si aspetta il dominio preponderante di uno specifico settore di utilizzo della lingua. Si è perciò deciso di utilizzare il BNC invece del WSJ, al fine di ottenere delle stime sulle caratteristiche individuate per ogni verbo, che fossero maggiormente rappresentative dell'uso generale dell'inglese. È stato utilizzato un *parser* parziale per identificare le costruzioni sintattiche necessarie per calcolare il valore delle caratteristiche rintracciate (Abney (1991), (1997)). Questo programma detto SCOL, permette di estrarre soggetti e oggetti diretti con un ragionevole grado di accuratezza; consente inoltre di identificare anche sintagmi preposizionali potenzialmente associati al verbo candidato ed oggetti indiretti, anche se con un grado di attendibilità minore.

Per ciascun verbo, è stato ricavato un vettore composto dalla misura di tali caratteristiche che sono:

- da un lato indicatori statistici dell'uso del verbo in inglese;
- dall'altro dei membri all'interno delle classi verbali semantiche.

*Tabella 3.7 Categorie delle caratteristiche individuate e numero delle singole caratteristiche*

| § in Text | Feature Category | #Features |
|---|---|---|
| 2.1 | Syntactic slots | 76 |
| | Slot overlap | 40 |
| | "Empty" words | 4 |
| 2.2 | Passive | 2 |
| | POS of the verb | 6 |
| | Aux, modal, Adv | 13 |
| | Derived forms | 3 |
| 2.3 | Animacy of NPs | 76 |

Dal momento che si ricorre ad un ampio spazio di caratteristiche (il nucleo dello spazio include 234 caratteristiche), è evidente che non tutte hanno la stessa utilità all'interno dello specifico scopo classificatorio che si prefigge il presente progetto. Sono stati allora sperimentati alcuni approcci computazionali per la selezione automatica delle caratteristiche stesse; il sistema prescelto prende il nome di C5.0. Questo modello automatico permette di focalizzare su due questioni fondamentali:

1. se le caratteristiche individuate funzionano bene;
2. se sia possibile identificare quali caratteristiche siano le più rilevanti e significative all'interno del progetto.

Uno degli scopi principali del progetto è quello di delineare uno spazio di caratteristiche che possa essere applicato in modo generale per ottenere un'ampia gamma di distinzioni tra le classi verbali.

In conclusione, sono stati condotti dieci esperimenti classificatori che coinvolgono dalle due alle tredici classi, tramite cui viene verificata in modo sperimentale l'utilità delle caratteristiche individuate e delle tecniche di selezione di queste ultime. I risultati derivanti mostrano in sostanza che i sistemi di classificazione costruiti sul nucleo dello spazio delle caratteristiche individuate come spiegato sopra, riducono il tasso di errore del 40% o più rispetto ad una base casuale condotta sui dieci esperimenti menzionati.

### 3.3 Il caso dei verbi tedeschi: il progetto di Schulte im Walde

La ricerca descritta finora riguardava comunque esperimenti svolti solo sull'inglese, ma lo stesso quadro teorico e gli stessi procedimenti, sono stati applicati con successo anche ad altre lingue come il tedesco, da ricercatori come, tra i più accreditati, Schulte im Walde. 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 ogni analisi basata su tecniche di *clustering*.

Anche Schulte im Walde nei suoi lavori sostiene l'ipotesi per cui esisterebbe una stretta connessione tra il significato lessicale di un verbo ed il suo comportamento: in una certa misura il significato lessicale determina il comportamento sintattico di un dato verbo, specie nella scelta degli argomenti cui si accompagna (Pinker (1989), Levin (1993)).

Possiamo utilizzare questa relazione significato/comportamento per indurre una classificazione formulata 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.

Tale intuizione ha costituito poi il fondamento per le operazioni successive di acquisizione automatica delle classi semantiche verbali del tedesco. 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.

Nell'esempio (4) vengono descritte le alternanze più comuni per la classe dei *Verbi di Movimento con un Veicolo*:

(4) a. *Der Wagen fährt in die Innenstadt*
La macchina guida verso il centro della città
b. *Die Frau fährt nach Haus*
La signora guida verso casa
c. *Der Filius fährt einen blauen Ferrari*
Il figlio guida una Ferrari blu

d. *Der Junge fährt seinen Vater zum Zug*
Il ragazzo guida il padre al treno

I partecipanti nella struttura concettuale espressa dal verbo sono: un veicolo, un conducente, una persona condotta ed una direzione. In *a.* il veicolo è espresso come soggetto della costruzione transitiva del verbo, con una frase preposizionale ad indicare la direzione. In *b.* il conducente è realizzato come soggetto della costruzione transitiva, ed una frase preposizionale indica la direzione. In *c.* il conducente diventa il soggetto della frase transitiva, accompagnato da un sostantivo all'accusativo che rappresenta il veicolo. In *d.* il conducente è espresso con funzione di soggetto all'interno di una costruzione di transitiva, con un sostantivo all'accusativo che indica la persona condotta, e una frase preposizionale che indica invece la direzione.

Anche se un certo partecipante non viene sempre realizzato all'interno dell'alternanza, il suo contributo è definito implicitamente dal verbo; ad esempio in *a.* il conducente non viene dichiarato apertamente, ma il parlante sa che c'è comunque un conducente, così come in *b.* e *d.* non compare alcun veicolo, sebbene noi sappiamo che c'è comunque un veicolo (Schulte im Walde (2004)).

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 intuire tale lavoro di rifinitura dell'informazione ottenuta, comincia da una pura definizione sintattica e aggiunge via via informazione semantica. Una classificazione indotta su base puramente sintattica è destinata a fallire se si ha a che fare ad esempio con verbi semanticamente simili, ma con comportamenti sintattici differenti; l'aggiunta di informazione riguardante le preposizioni che introducono obbligatoriamente, ma anche opzionalmente, gli argomenti realizzati dai verbi si è rivelata assai utile al buon funzionamento del *clustering* semantico. Infine la definizione delle preferenze di selezione e quindi delle restrizioni nell'assegnazione dei tipi semantici agli argomenti verbali, migliorano ulteriormente i risultati ottenuti dai precedenti livelli di analisi. Sulla base della descrizione sintattico-semantica più elaborata dei verbi tedeschi, viene applicato un algoritmo di *clustering k-Means* (Forgy (1965)), in modo da indurre 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]. L'obiettivo che l'algoritmo si pone è

[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

di minimizzare la varianza totale *inter-cluster*. Ogni *cluster* viene identificato mediante un centroide o punto medio. L'algoritmo segue una procedura iterativa. Inizialmente crea *K* partizioni e assegna ad ogni partizione i dati da *clusterizzare*, o casualmente o usando alcune informazioni euristiche; quindi calcola il centroide di ogni gruppo. Costruisce poi una nuova partizione associando ogni dato al *cluster* il cui centroide è più vicino ad esso; infine vengono ricalcolati i centroidi per i nuovi *clusters* e così via, finché l'algoritmo non converge. L'algoritmo ha acquistato notorietà dato che converge molto velocemente. Infatti, si è osservato che generalmente il numero di iterazioni sono minori del numero di punti. In termini di prestazioni l'algoritmo non garantisce il raggiungimento dell'ottimo globale; la qualità della soluzione finale dipende largamente dal *set* di *clusters* iniziale e può, in pratica, ottenere una soluzione ben peggiore dell'ottimo globale. Dato che l'algoritmo è estremamente veloce, è possibile applicarlo più volte ricorsivamente e fra le soluzioni prodotte scegliere quella più soddisfacente. Un altro svantaggio dell'algoritmo è che esso richiede di scegliere il numero di *clusters (k)* da trovare: se i dati non sono opportunamente ripartiti si ottengono risultati discutibili. 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*
(d) kriechen rennen – *Manner of Motion: Locomotion*
eilen – *Manner of Motion: Rush*
gleiten – *Manner of Motion: Flotation*
starren – *Facial Expression*
(e) klettern wandern – *Manner of Motion: Locomotion*
fahren fliegen segeln – *Manner of Motion: Vehicle*
fließen – *Manner of Motion: Flotation*
(f) festlegen – *Constitution*
bilden – *Production*
erhöhen senken steigern vergrößern verkleinern – *Quantum Change*
(g) töten – *Elimination*
unterrichten – *Teaching*
(h) geben – *Transfer of Possession (Giving): Gift*

Il fatto che ci sono dei verbi che vengono clusterizzati semanticamente sulla base delle loro proprietà empiriche *corpus-based* e *knowledge-based* (ad esempio i *weather verbs* del *cluster* (a)), indica:

- una relazione tra le componenti del significato dei verbi ed il loro comportamento;
- che l'algoritmo di *clustering* è capace di trarre vantaggio dalle descrizioni linguistiche e di astrarre dal rumore delle distribuzioni.

I verbi a bassa frequenza hanno costituito un problema negli esperimenti di *clustering*; le loro distribuzioni sono più rumorose dei verbi a più alta frequenza, perciò tipicamente andranno a costituire dei *clusters rumorosi*. Inoltre l'ambiguità dei verbi non può essere modellata dall'algoritmo *hard clustering* di tipo *k-Means*; i verbi ambigui vengono difatti assegnati

- ad uno dei *clusters* corretti;

- ad un *cluster* i cui componenti hanno distribuzioni simili a quelle dei verbi ambigui considerati;
- ad un *cluster* composto da un singolo elemento.

L'interpretazione dei risultati del *clustering* evidenzia delle proprietà del significato verbale che potrebbero essere trascurate nella classificazione manuale, indotta su base puramente semantica e quindi intuitiva: ad esempio potrebbero emergere delle classi separate nella classificazione manuale, ma clusterizzate insieme dall'algoritmo. Ad esempio le classi *Perception* e *Observation* che, come vedremo, nella classificazione manuale risultano separate, compaiono unite nel *clustering*; esse sono caratterizzate dal fatto che, in entrambi i casi, i verbi contenuti esprimono un'osservazione, con il riferimento aggiuntivo ad un'abilità fisica, come quella di ascoltare, nei verbi di percezione.

Perciò potremmo concludere che la classificazione manuale elaborata è più granulare e dettagliata di quella automatica, con particolari difficoltà evidenziate proprio dalle sottoclassi.

Andiamo ora proprio a vedere come si compone questa classificazione manuale. La Schulte im Walde ha selezionato 168 verbi tedeschi e li ha raggruppati manualmente in 43 classi semantiche, per valutare, attraverso la comparazione, la validità e l'attendibilità nell'esecuzione delle tecniche di *clustering* utilizzate. Una simile classificazione manuale si basa primariamente sulla pura intuizione semantica del parlante, e prende spunto da lessici pre-esistenti compilati a priori, senza tralasciare quelle voci fortemente ambigue che potrebbero mettere fuori strada il meccanismo di *clustering*. Nonostante sarebbe stato più agevole operare una classificazione su base sintattica (evidenza empirica), considerare solo ampie classi di verbi ad alta frequenza ed evitare complicazioni nel raggruppamento ignorando fenomeni come l'ambiguità semantica, ciò non sarebbe stato conforme allo scopo prefissato inizialmente, che 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*. Di seguito vengono elencate le classi semantiche elaborate da Schulte im Walde per il tedesco:

1. *Aspect*: anfangen, aufhören, beenden, beginnen, enden
2. *Propositional Attitude*: ahnen, denken, glauben, vermuten, wissen
3. *Desire*
    (a) *Wish*: erhoffen, wollen, wünschen
    (b) *Need*: bedürfen, benötigen, brauchen
4. *Transfer of Possession (Obtaining)*: bekommen, erhalten, erlangen, kriegen
5. *Transfer of Possession (Giving)*
    (a) *Gift*: geben, leihen, schenken, spenden, stiften, vermachen, überschreiben
    (b) *Supply*: bringen, liefern, schicken, vermitteln$_1$, zustellen
6. *Manner of Motion*
    (a) *Locomotion*: gehen, klettern, kriechen, laufen, rennen, schleichen, wandern
    (b) *Rotation*: drehen, rotieren
    (c) *Rush*: eilen, hasten
    (d) *Vehicle*: fahren, fliegen, rudern, segeln
    (e) *Flotation*: fließen, gleiten, treiben
7. *Emotion*
    (a) *Origin*: ärgern, freuen
    (b) *Expression*: heulen$_1$, lachen$_1$, weinen
    (c) *Objection*: ängstigen, ekeln, fürchten, scheuen
8. *Facial Expression*: gähnen, grinsen, lachen$_2$, lächeln, starren
9. *Perception*: empfinden, erfahren$_1$, fühlen, hören, riechen, sehen, wahrnehmen
10. *Manner of Articulation*: flüstern, rufen, schreien

11. *Moaning*: heulen$_2$, jammern, klagen, lamentieren
12. *Communication*: kommunizieren, korrespondieren, reden, sprechen, verhandeln
13. *Statement*
    (a) *Announcement*: ankündigen, bekanntgeben, eröffnen, verkünden
    (b) *Constitution*: anordnen, bestimmen, festlegen
    (c) *Promise*: versichern, versprechen, zusagen
14. *Observation*: bemerken, erkennen, erfahren$_2$, feststellen, realisieren, registrieren
15. *Description*: beschreiben, charakterisieren, darstellen$_1$, interpretieren
16. *Presentation*: darstellen$_2$, demonstrieren, präsentieren, veranschaulichen, vorführen
17. *Speculation*: grübeln, nachdenken, phantasieren, spekulieren
18. *Insistence*: beharren, bestehen$_1$, insistieren, pochen
19. *Teaching*: beibringen, lehren, unterrichten, vermitteln$_2$
20. *Position*
    (a) *Bring into Position*: legen, setzen, stellen
    (b) *Be in Position*: liegen, sitzen, stehen
21. *Production*: bilden, erzeugen, herstellen, hervorbringen, produzieren
22. *Renovation*: dekorieren, erneuern, renovieren, reparieren
23. *Support*: dienen, folgen$_1$, helfen, unterstützen
24. *Quantum Change*: erhöhen, erniedrigen, senken, steigern, vergrößern, verkleinern
25. *Opening*: öffnen, schließen$_1$
26. *Existence*: bestehen$_2$, existieren, leben
27. *Consumption*: essen, konsumieren, lesen, saufen, trinken
28. *Elimination*: eliminieren, entfernen, exekutieren, töten, vernichten
29. *Basis*: basieren, beruhen, gründen, stützen
30. *Inference*: folgern, schließen$_2$
31. *Result*: ergeben, erwachsen, folgen$_2$, resultieren
32. *Weather*: blitzen, donnern, dämmern, nieseln, regnen, schneien

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* (Baker et al. (2002), Johnson et al. (2002)). La definizione della classe in base alla semantica dei *frames* contiene una descrizione in prosa della scena, il partecipante principale al *frame*, i ruoli modificatori e le varianti del *frame* che descrivono la scena.

La Schulte im Walde ha prodotto un elenco delle principali varianti dei *frames* così come venivano rintracciate nel *corpus*, marcate dai ruoli partecipanti e da almeno una frase come esempio per ogni verbo, utilizzando i rispettivi *frames* (Figura 3.8).

*Tabella 3.8 Descrizione della classe verbale degli Aspects Verbs*

***Aspect Verbs***: *anfangen, aufhören, beenden, beginnen, enden*

Scene: [$_E$ An event] begins or ends, either internally caused or externally caused by [$_I$ an initiator]. The event may be specified with respect to [$_T$ tense], [$_L$ location], [$_X$ an experiencer], or [$_R$ a result].

Frame Roles: I(nitiator), E(vent)

Modification Roles: T(emporal), L(ocal), (e)X(periencer), R(esult)

| Frame | Participating Verbs & Corpus Examples |
|---|---|
| $\mathbf{n}_E$ | + anfangen, aufhören, beginnen / $+_{adv}$ enden / ¬ beenden |
| | Nun aber muß [$_E$ der Dialog] **anfangen**.<br>... bevor [$_E$ der Golfkrieg] **angefangen** hatte ...<br>... damit [$_E$ die Kämpfe] **aufhören**.<br>Erst muß [$_E$ das Morden] **aufhören**.<br>[$_E$ Der Gottesdienst] **beginnt**.<br>[$_E$ Das Schuljahr] **beginnt** [$_T$ im Februar].<br>[$_X$ Für die Flüchtlinge] **beginnt** nun [$_E$ ein Wettlauf gegen die Zeit].<br>[$_E$ Sein Zwischenspiel] bei der Wehrmacht **endete** ... [$_R$ glimpflich].<br>[$_E$ Die Ferien] **enden** [$_R$ mit einem großen Fest].<br>[$_E$ Druckkunst] ... **endet** [$_R$ beim guten Buch].<br>[$_E$ Die Partie] **endete** [$_R$ 0:1].<br>[$_L$ An einem Baum] **endete** in Höchst [$_E$ die Flucht] ...<br>[$_E$ Der Informationstag] ... **endet** [$_T$ um 14 Uhr]. |
| $\mathbf{n}_I$ | + anfangen, aufhören / ¬ beenden, beginnen, enden |
| | [$_I$ Die Hauptstadt] muß **anfangen**.<br>... daß [$_I$ er] [$_T$ pünktlich] **anfing**.<br>Jetzt können [$_I$ wir] nicht einfach **aufhören**.<br>Vielleicht sollte [$_I$ ich] **aufhören** und noch studieren. |
| $\mathbf{n}_I$<br>$\mathbf{a}_E$ | + anfangen, beenden, beginnen / ¬ aufhören, enden |
| | Nachdem [$_I$ wir] [$_E$ die Sache] **angefangen** haben,<br>... [$_I$ er] versucht, [$_E$ ein neues Leben] **anzufangen**.<br>[$_I$ Die Polizei] **beendete** [$_E$ die Gewalttätigkeiten].<br>[$_T$ Nach dem Abi] **beginnt** [$_I$ Jens] [$_L$ in Frankfurt] [$_E$ seine Lehre] ... |
| $\mathbf{n}_I$<br>$\mathbf{a}_E$<br>$[P]$ | + anfangen, beenden, beginnen / ¬ aufhören, enden |
| | Wenn [$_E$ die Arbeiten] [$_T$ vor dem Bescheid] **angefangen** werden ...<br>Während [$_X$ für Senna] [$_E$ das Rennen] **beendet** war ...<br>... ehe [$_E$ eine militärische Aktion] **begonnen** wird ... |
| $\mathbf{n}_I$<br>$\mathbf{i}_E$ | + anfangen, aufhören, beginnen / ¬ beenden, enden |
| | [$_I$ Ich] habe nämlich [$_E$ zu malen] **angefangen**.<br>[$_I$ Ich] habe **angefangen**, [$_E$ Hemden zu schneidern].<br>[$_I$ Die Bahn] will [$_T$ 1994] **anfangen** [$_E$ zu bauen].<br>... daß [$_I$ der Alkoholiker] **aufhört** [$_E$ zu trinken].<br>... daß [$_I$ die Säuglinge] einfach **aufhören** [$_E$ zu atmen].<br>In dieser Stimmung **begannen** [$_I$ Männer] [$_E$ Tango zu tanzen] ...<br>[$_I$ Tausende von Pinguinen] **beginnen** [$_E$ dort zu brüten]. |
| $\mathbf{n}_I$<br>$\mathbf{p}_E$ : *mit* | + anfangen, aufhören, beginnen / ¬ beenden, enden |
| | Erst als [$_I$ der versammelte Hofstaat] [$_E$ mit Klatschen] **anfing**.<br>Aber [$_I$ wir] müssen endlich [$_E$ damit] **anfangen**.<br>[$_I$ Der Athlet] ... kann ... [$_E$ mit seinem Sport] **aufhören**.<br>... müßten noch [$_I$ viel mehr Frauen] [$_E$ mit ihrer Arbeit] **aufhören** ...<br>Schließlich zog [$_I$ er] einen Trennstrich, **begann** [$_E$ mit dem Entzug] ...<br>[$_I$ Man] **beginne** [$_E$ mit eher katharsischen Werken]. |
| $\mathbf{n}_I$<br>$\mathbf{p}_E$ : *mit*<br>$[P]$ | +anfangen, aufhören, beginnen / ¬ beenden, enden |
| | Und [$_E$ mit den Umbauarbeiten] könnte **angefangen** werden.<br>[$_E$ Mit diesem ungerechten Krieg] muß sofort **aufgehört** werden.<br>[$_T$ Vorher] dürfe [$_E$ mit der Auflösung] nicht **begonnen** werden.<br>... daß [$_E$ mit dem Umbau] ... **begonnen** werden kann. |

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 (ad esempio *unterstützen* non appartiene alla classe *support*, poiché richiede un oggetto espresso all'accusativo invece che al dativo). Il *clustering* fallisce inoltre per tutti quei verbi che mostrano un comportamento sintattico simile, ma nessuna similarità semantica (ad esempio molti dei verbi considerati sottocategorizzano un oggetto all'accusativo, pertanto vengono erroneamente clusterizzati insieme);
- rifinire l'informazione sintattica del verbo attraverso l'introduzione delle preposizioni, è estremamente utile per la buona riuscita del *clustering*, sia laddove queste sono obbligatorie che dove invece sono opzionali. Il miglioramento sottolinea il fatto che verbi con significato simile, mostrano anche affinità nell'esprimere specifici complementi preposizionali (ad esempio *glauben/denken an*$_{Akk}$) 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. Da un punto di vista pratico è chiara la distinzione tra i due tipi di proprietà descritti, ma a livello teorico non è possibile identificare una scelta univoca perfetta per codificare le caratteristiche verbali.

Seppure l'analisi condotta tramite tecniche di *clustering* necessiti di essere corretta e completata manualmente, rappresenta comunque un'utile base per lo sviluppo di risorse semantiche per il lessico.

Il lavoro condotto da Schulte im Walde e quello invece proposto da Stevenson & Merlo e da Joanis rappresentano due modi diversi di approcciare computazionalmente il problema della classificazione verbale.

Nel caso di Stevenson & Merlo e di Joanis, quest'ultima è modellata come un *task* di categorizzazione; si parte perciò da una serie di classi o categorie predefinite, quali quelle proposte da Levin, e si valuta la possibilità di riuscire a categorizzare (o classificare) automaticamente nuovi verbi. Ciò implica una conoscenza *a priori* delle classi, e un'indagine sulle nostre capacità di utilizzo corretto delle stesse nei compiti di classificazione. A livello computazionale tutto questo si riflette nell'uso di algoritmi supervisionati, vale a dire che necessitano proprio di dati pre-classificati per apprendere e classificare a loro volta.

Schulte im Walde invece modella la classificazione come un *task* di *clustering*; in questo caso l'ineteresse non è rivolto ad imparare ad usare una classificazione pre-esistente, quanto a scoprire che tipi di classi emergono dai dati. Da qui deriva l'uso di un metodo computazionale invece non-supervisionato, quale il *clustering* appunto, che non richiede dati pre-classificati, ma cerca di individuare le ripartizioni migliori all'interno di un gruppo di dati fornito, sulla base dei rapporti di similarità che tra questi inetrcorrono. La classificazione *a priori* di cui si avvale Schulte im Walde, viene usata pertanto solo per valutare i *clusters* in *output* e non per imparare a classificare.

# Capitolo 4
# Esperimenti di classificazioni computazionali distribuzionali sui verbi italiani

## 4.1 Il livello di analisi sintattico

### 4.1.1 La realizzazione di una classificazione semantica

Le classi verbali possono essere definite in modo indipendente tanto su base sintattica quanto semantica.

Per ciò che concerne il piano semantico, già a partire dall'antichità sono stati proposti dei criteri per distinguere tipi diversi di parole in base al loro significato o ad alcuni aspetti del loro significato. Ad esempio, già Aristotele nella *Poetica* distingueva tra *ónoma*, parola il cui significato non è dotato di temporalità, e *rhēma*, parola il cui significato è dotato di una dimensione temporale.

Un criterio semantico fondamentale per molti linguisti moderni, ad esempio J. Lyons (1977), anche se in parte già presente nelle riflessioni antiche, è quello che distingue tra:

- parole che si riferiscono a delle entità (persone, animali, piante, luoghi, oggetti fisici e immaginari, concreti e astratti, più o meno delimitati), come ad esempio *bicchiere, cane, acqua, traffico*;
- parole che attribuiscono una proprietà alle entità, come *risplendere* nella frase *il sole risplende*, o che descrivono la relazione esistente tra due o più entità, come *illuminare* nella frase *il sole illumina la stanza*;
- parole che esprimono le qualità delle entità, come *rosso* nel sintagma *un ombrello rosso*.

Esistono straordinarie convergenze tra le proprietà formali delle parole (quindi le proprietà morfologiche e sintattiche) e il loro significato, seppur ad un livello piuttosto astratto. Ad esempio, vi è una tendenza preponderante delle parole che indicano persone, luoghi, cose concrete o astratte ad essere nomi, delle parole che descrivono proprietà di cose o relazioni tra cose ad essere verbi, ed infine delle parole che indicano qualità di cose ad essere aggettivi. Questa correlazione già evidenziata nell'antichità, è stata successivamente riproposta all'attenzione della comunità scientifica da numerosi linguisti, come E. Sapir (1921) e J. Lyons (1977), i quali hanno cercato di offrirne una descrizione accurata, con l'ausilio degli strumenti della linguistica moderna. Si tratta di un fatto tutt'altro che scontato: l'esistenza di convergenze tra le categorie ontologiche (come le cose e i fatti del mondo) e le categorie linguistiche (ad esempio i nomi e i verbi) è un dato interessantissimo, che sottolinea lo stretto rapporto che intercorre tra la categorizzazione concettuale e quella linguistica.

I lavori fin qui considerati, a partire dagli studi pionieristici della Levin, fino ad arrivare ai progetti più direttamente connessi con le potenzialità di automazione fornite dalla linguistica computazionale, poggiano tutti su una comune ipotesi di fondo: l'esistenza di classi verbali identificabili, parallelamente, tanto in virtù delle proprietà distribuzionali (piano sintattico), quanto dei tratti semantici che mostrano (piano semantico).

La classificazione semantica, attualmente, può essere effettuata soltanto manualmente dal linguista, che si basa fondamentalmente sulla propria esperienza e conoscenza del linguaggio. Per il comportamento sintattico si può procedere in modo diverso poiché esso, manifestandosi formalmente, è oggettivamente misurabile. Sarà allora interessante verificare punti di contatto e divergenze, derivanti dal confronto tra una prima classificazione effettuata a mano, ed un'altra basata invece su una distribuzione statistica dei tratti sintattici.

Numerosi, come abbiamo visto nel capitolo 3, sono i progetti sviluppati sulla base di queste considerazioni, ma nessuno si è ancora cimentato nello studio del sistema verbale dell'italiano.

Pertanto gli esperimenti presentati in questo lavoro, considerando gli assunti teorici fin qui illustrati, si propongono di operare un confronto tra un'iniziale classificazione semantica effettuata *a priori* indipendentemente da misurazioni di tipo distribuzionale, ed un'altra di tipo sintattico ricavata automaticamente a partire da un *corpus* di riferimento. Tale raffronto verrà effettuato tramite un algoritmo di *clustering*, ed avrà come obiettivo principale quello di

realizzare e valutare uno studio *corpus-based* della struttura argomentale, a cavallo tra sintassi e semantica, di un insieme di verbi italiani usando metodi di estrazione semi-automatica delle proprietà sintattiche dei verbi (*frames* di sottocategorizzazione). In questo contesto infatti, il *clustering* e la classificazione semantica *a priori* sono soltanto degli strumenti inseriti all'interno di un'indagine esplorativa più ampia. Quest'ultima mira ad individuare quali proprietà semantiche condividono verbi che in italiano presentano proprietà sintattiche distribuzionali simili, e in che misura tali proprietà comuni trovano una corrispondenza in *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 Repubblica. 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, il materiale prodotto è stato considerato comunque una base promettente per lo sviluppo di un *corpus* di dimensioni notevoli, e di facile costruzione. Coprendo un arco di 16 anni, tale *corpus*, liberamente consultabile[3], fornisce uno strumento per lo studio diacronico oltre che sincronico dell'italiano contemporaneo, nonché per studi contrastivi, visto che sono già disponibili *corpora* di testi giornalistici per molte altre lingue.

Nella Tabella 4.1 consultabile in appendice, viene proposta la classificazione semantica operata su una selezione di verbi, scelti in quanto ritenuti maggiormente adatti allo scopo; essa ricalca quelle precedentemente elaborate da Levin e Schulte im Walde, pur tenendo conto delle peculiarità rilevanti del sistema verbale italiano, che non consentono un'esatta corrispondenza[4].

Un insieme di 200 verbi sono stati ripartiti in 40 classi semantiche in base alla similarità del loro significato lessicale e concettuale, successivamente 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 primariamente sull'intuizione semantica e non sui comportamenti sintattici; difatti è possibile trovare dei verbi che appartengono alla stessa classe, ma che mostrano caratteristiche sintattiche differenti:

(1) <u>*Propositional Attitude*</u>

| | |
|---|---|
| *sapere* (+ obj dir) | Mauro sa la canzone a memoria |
| *dubitare* (+ comp_di) | Mauro dubita delle sue capacità |

<u>*Communication*</u>

| | |
|---|---|
| *dire* (+obj dir) | Mauro dice il proprio parere |
| *conversare* (+comp_con, +comp_di) | Mauro conversa con gli amici |

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

[4] Verbi come *realisieren*, che corrisponde all'italiano *realizzare*, nella classificazione di Schulte im Walde appartengono alla classe definita *Observation* (con l'accezione di realizzare un'idea, un pensiero etc.), mentre in italiano possiamo collocarlo all'interno di quella classe che indica la *Produzione*.

*Basis*

| | |
|---|---|
| *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 2); in questo caso ho dovuto compiere inizialmente una scelta, basandomi sulle proprietà semantiche condivise dai verbi di una certa classe ad un livello generale.

(2) *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 suddividono 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 quante più aree possibili dei domini semantici: 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*).

L'utilità principale 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.

### 4.1.2 Tecniche di estrazione automatica dei *frames* di sottocategorizzazione: il *parser* sintattico a dipendenze

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 proceduto all'analisi del *corpus* stesso, per ottenere in modo automatico la distribuzione statistica dei *patterns* (schemi) di sottocategorizzazione.

Si è deciso di elaborare i dati del *corpus* 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; tale idea è espressa esaustivamente nei capitoli introduttivi dell'opera di Tesnière (1959):

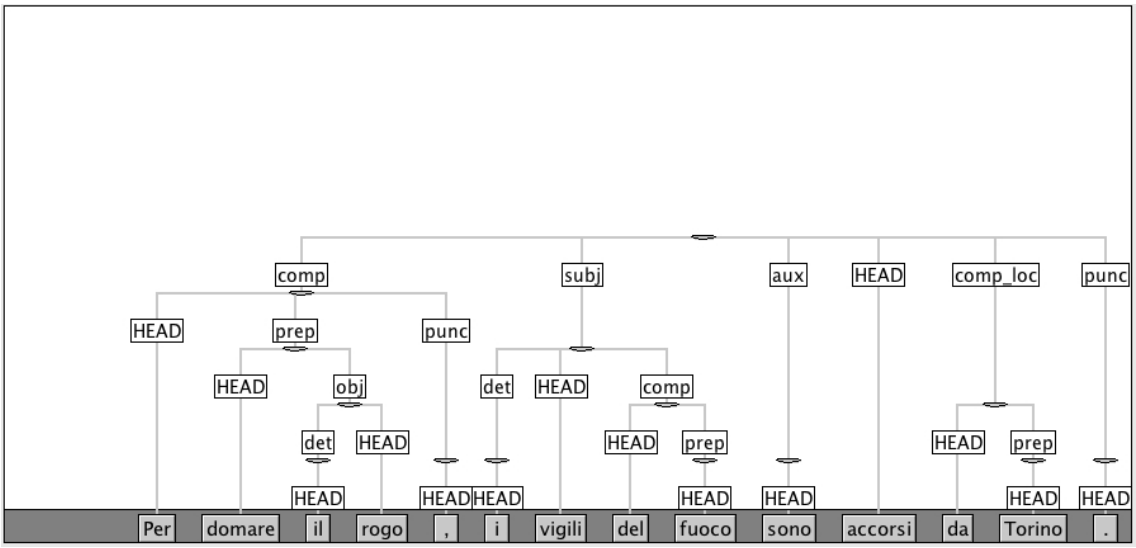
> La phrase est un *ensemble oragnisé* dont les éléments constituants sont les *mots*. [1.2] Tout mot qui fait partie d'une phrase cesse par lui même d'être isolé comme dans le dictionnaire. Entre lui et ses voisins, l'esprit aperçoit des connexions, dont l'ensemble forme la charpente de la phrase. [1.3] Les connexions

> structurales établissent entre les mots des rapports de *dépendance*. Chaque connexion unit en principe un terme *supérieur* et un terme *inférieur*. [2.1] Le terme supérieur reçoit le nom de *régissant*. Le terme inférieur reçoit le nom de *subordonné* (Tesnière 1959 : 11-13).

Il concetto di dipendenza, come si evince dalla citazione, coinvolge un elemento che funge da testa ed uno che si comporta invece come dipendente; centrali saranno dunque i criteri (sia semantici che sintattici) utili a stabilire a quale di queste due categorie appartengono gli elementi coinvolti nella relazione (Figura 4.1).

Se è vero che le relazioni di dipendenza identificate dai *parser* hanno due caratteristiche fondamentali, sono cioè binarie (in quanto coinvolgono una testa ed un suo dipendente) ed asimmetriche (poiché se la testa domina il suo dipendente, non è altrettanto vero l'inverso; quindi la testa intrattiene un certo tipo di relazione gerarchica con il suo dipendente, ma non viceversa), sarà possibile allora rappresentarle tramite un grafo direzionato in cui gli elementi lessicali rappresentano i nodi, mentre le relazioni tra testa e subordinati costituiranno gli archi del grafo stesso.

*Figura 4.1 Esempio di parsing a dipendenze*




Le restrizioni che il grafo deve rispettare sono essenzialmente tre:

1. connessione, ogni nodo deve essere legato almeno ad un altro nodo all'interno del grafo;
2. non-ciclicità, il grafo non deve contenere delle relazioni cicliche:
   ad esempio *Un "figlio" ha un solo "padre"*;
3. singola-testa, ogni nodo deve dipendere da una ed una sola testa.

I programmi di *parsing* si dividono principalmente in due gruppi: quelli che assegnano una struttura sulla base di una grammatica fornita inizialmente in *input*, e quelli che invece apprendono automaticamente le regole grammaticali su cui si basa il linguaggio analizzato (*parser* statistici).

Il *parser* sviluppato all'interno di questa tesi, appartiene al secondo gruppo e non si basa dunque 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 *parser* utilizzato, prende il nome di *Maltparser*. 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. Un *corpus* testuale annotato con codici linguistici, normalmente a livello sintattico e morfosintattico, prende il nome di *Treebank*; le annotazioni presenti descrivono la struttura delle frasi attraverso alberi di rappresentazione sintattica, individuano le unità linguistiche che le compongono e associano a ciascuna unità un codice, o identificatore, che indica la relativa categoria sintattica di appartenenza. Nella maggior parte dei casi è previsto un solo livello di annotazione, anche se esistono sistemi più complessi che possono presentarne in numero maggiore. In questo modo, ad esempio, le annotazioni riguardanti le relazioni di dipendenza superficiali, possono essere separate da un livello di rappresentazione della struttura profonda della frase basato su relazioni di tipo semantico, oppure da un livello di rappresentazione delle relazioni funzionali tra i sintagmi.

Il funzionamento del *Maltparser* dipende da tre tipi di restrizioni:

1. da algoritmi di *parser* deterministici utili per costruire i grafi delle dipendenze;
2. da modelli in grado di ricostruire la storia delle azioni condotte dal *parser*, così da poter predire quelle future;
3. da un algoritmo di apprendimento in grado di discernere la storia delle azioni del *parser* e di rappresentarle (Nivre, Hall e Nilsson (2006)).

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, grazie all'ausilio ed alla supervisione del Prof. Alessandro Lenci, 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.

(3)
arrivare freq=105499
0# 26889 0.25 essereAux=0.99 avereAux=0.00 pass=0.00
comp_a# 17892 0.17 essereAux=0.99 avereAux=0.00 pass=0.00
comp_in# 6412 0.06 essereAux=0.99 avereAux=0.00 pass=0.00
comp_da# 4514 0.04 essereAux=0.99 avereAux=0.00 pass=0.00
inf_a# 3815 0.04 essereAux=0.99 avereAux=0.00 pass=0.00
si#comp_a# 2475 0.02 essereAux=0.99 avereAux=0.00 pass=0.01
comp_ad# 1776 0.02 essereAux=0.99 avereAux=0.00 pass=0.00
comp_con# 1730 0.02 essereAux=0.99 avereAux=0.00 pass=0.00
comp_su# 1360 0.01 essereAux=0.99 avereAux=0.00 pass=0.00
pred# 1293 0.01 essereAux=0.99 avereAux=0.00 pass=0.00
comp_a#comp_con# 896 0.01 essereAux=0.99 avereAux=0.00 pass=0.00
comp_a#comp_in# 838 0.01 essereAux=0.99 avereAux=0.00 pass=0.00

L'esempio 3, presenta un estratto dello *script* elaborato tramite Perl; per comprenderne il contenuto occorre illustrare alcune delle formule convenzionali in esso contenute. Partiamo dalla prima riga, in cui compare il nome del verbo preso in esame insieme alla relativa frequenza d'uso nel *corpus*; successivamente si incontra il simbolo 0, che sta ad indicare che si tratta di un verbo intransitivo (nel senso di 0 valente, ovvero senza dipendenze nella frase), seguito dalla specificazione degli ausiliari con cui il verbo *arrivare* si accompagna, e dall'indice della frequenza con cui ciò si verifica. Come evidenzia lo *script*, il

verbo *arrivare* si combina nella totalità dei casi con l'ausiliare *essere,* mentre non si trova mai con *avere*.

(4) Mauro è arrivato in città
*Mauro ha arrivato in città
La soluzione con l'ausiliare *avere* è linguisticamente inaccettabile

Nelle righe successive dello *script*, vengono elencati i complementi e le strutture frasali specifici del verbo di riferimento, corredati dalle rispettive preposizioni, dagli ausiliari e dagli indici di frequenza, relativi alla loro ricorrenza nel *corpus* con quel verbo.

(5) *comp_da*
Mauro arriva da Firenze
*comp_a#comp_con*
Mauro arriva a Roma con l'aereo
*inf_a*
Mauro arriva a mangiare anche tre pizze

Infine occorre precisare che lo *script* non tiene conto della presenza del soggetto, poiché quest'ultimo, se posposto rispetto al verbo, genera confusione nel *parser* che tende ad identificarlo piuttosto come un oggetto diretto.

### 4.1.3 Le operazioni di *clustering*

Una volta in possesso delle distribuzioni di frequenza dei verbi rispetto ai *frames* di sottocategorizzazione, si è pronti a sottoporre tali dati al processo di classificazione automatica, operazione possibile 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; la bontà delle analisi ottenute tramite questi algoritmi dipende essenzialmente da quanto è significativa la metrica utilizzata per definirne questo parametro e da come sono composti al loro interno i vettori, ovvero da come scegliamo le loro dimensioni.

Se la distanza risulta essere dunque un concetto fondamentale, l'appartenenza o meno ad uno stesso *cluster* dipenderà strettamente da quanto l'elemento esaminato è distante dall'insieme stesso; più questo sarà vicino e più sarà considerato simile, mentre il caso inverso sarà indice di dissimilarità (Manning, Schütze (1999)).

All'interno dell'insieme degli algoritmi di *clustering* distinguiamo due grandi famiglie, che prendono il nome di: *hard clustering* e *soft clustering*. Nel caso della famiglia *hard*, ciascun vettore appartiene ad un solo *cluster* con grado di appartenenza unitario, invece per la famiglia *soft*, ciascun vettore può appartenere a diversi *clusters* con differente grado di similarità.

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.

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

Possiamo perciò definire le operazioni di *clustering* come una forma di apprendimento: inizialmente gli elementi vengono raggruppati in *clusters*, su cui poi si

possono fare delle generalizzazioni, a partire dalle nostre conoscenze su alcuni membri dei *clusters* stessi. Non assegnando preventivamente delle nomenclature ai dati, dal *clustering* deriverà una classificazione automatica e non controllata *a priori*, poiché i risultati dipendono esclusivamente dalle divisioni naturali assunte dai dati (Manning, Schütze (1999)); il ricercatore condurrà poi la sua analisi, basandola sul confronto tra tali *clusters* ed una eventuale classificazione da lui precedentemente stabilita.

Nella presente tesi, a ciascun verbo viene assegnata una rappresentazione vettoriale; le dimensioni del vettore dipendono dalla distribuzione di frequenza di ogni verbo rispetto ai suoi *frames* di sottocategorizzazione all'interno del *corpus*. Tramite l'applicazione dell'algoritmo di *clustering* alle suddette rappresentazioni vettoriali, è possibile ricostruire lo spazio di similarità semantica tra i verbi, ovvero il loro grado di distanza, e quindi ottenere classi di verbi semanticamente simili sul piano delle loro distribuzioni con diversi *frames* di sottocategorizzazione (Lenci, Calzolari (2004)).

I nostri 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. Cluto 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)).

Questi algoritmi sono basati su paradigmi partizionali e agglomerativi. Cluto fornisce un insieme di sette diversi criteri funzionali che possono essere usati per guidare sia gli algoritmi di *clustering* agglomerativi che quelli partizionali; molti di questi criteri funzionali hanno dimostrato di produrre soluzioni di *clustering* di alta qualità per insiemi di dati con un alto numero di dimensioni semantiche.

Cluto offre anche strumenti per analizzare i *clusters* individuati e comprendere le relazioni che intercorrono tra gli oggetti attribuiti ad ogni singolo *cluster* e fra i diversi *clusters*; inoltre offre anche delle applicazioni per visualizzare graficamente le soluzioni dei *clusters* ottenuti.

Cluto può anche identificare le caratteristiche prototipiche di ogni *cluster*, ovvero quei tratti che meglio lo descrivono e lo distinguono dagli altri. Questo insieme di caratteristiche può essere usato per incrementare la conoscenza dell'insieme degli elementi assegnati ad ogni singolo *cluster* e produrre una breve sintesi sul suo contenuto. Infine Cluto fornisce delle modalità di visualizzazione che possono essere usate per cogliere le relazioni tra *clusters*, elementi e caratteristiche degli stessi.

Gli algoritmi di Cluto sono stati ottimizzati per operare su ampi insiemi di dati, sia in termini di numero di oggetti che di dimensioni considerati; ciò è vero in particolare per gli algoritmi partizionali, che possono, per definizione, *clusterizzare* velocemente insiemi di dati contenenti decine di migliaia di elementi e svariate migliaia di dimensioni.

## 4.2 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, ovvero il *repeated bisection*, per il quale si rende necessario, come per il *k-means*, specificare il numero *k* di *clusters* di partenza[5].

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*.

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

Sono state effettuate tre diverse tipologie di esperimenti, 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, inserite in appendice nelle Tabelle 4.1, 4.2 e 4.3, 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:

(6) 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 Possesion (obtaining), Transfer of Possession (giving-gift), Transfer of Possesion (giving-supply).*

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 con Cluto, 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 e sono consultabili in appendice nelle Tabelle 4.4, 4.5 e 4.6.

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. In effetti 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 divisione iniziale era assolutamente troppo generica per poter operare qualunque tipo di riflessione su di essa (Figura 4.2); grazie invece all'introduzione del logaritmo della frequenza (Figura 4.3), la classificazione risulta estremamente più raffinata, granulare e conseguentemente più efficace al fine di comprendere quanto i componenti sintattici siano in grado di riflettere il significato dei verbi e quali tra questi sono più significativi in questo processo.

Nella matrice proposta nella Figura 4.2 le colonne rappresentano i *frames* di sottocategorizzazione scelti (in questa sede ne vengono proposti solo alcuni a livello esemplificativo), mentre le righe corrispondono ai verbi trattati e ripartiti graficamente in *clusters*. Le caselle di colore rosso più o meno intenso, riflettono il grado di ricorrenza di un

certo tratto per un determinato verbo; i dati risultano estremamente sparsi, come evidenziano i grandi spazi bianchi nella matrice e le colonne più significative sono prevalentemente quelle corrispondenti a *ogg_dir#* e *0#*, ovvero ai *frames* che rappresentano rispettivamente la transitività o l'intransitività dei vari verbi. Come già accennato, questa indicazione è troppo generica e non permette di rilevare quali siano gli altri tratti prototipici che i verbi presentano sia singolarmente che in comune con gli altri membri del *cluster* cui appartengono.

*Figura 4.2 Esempio della matrice costruita calcolando la frequenza dei frames di sottocategorizzazione*

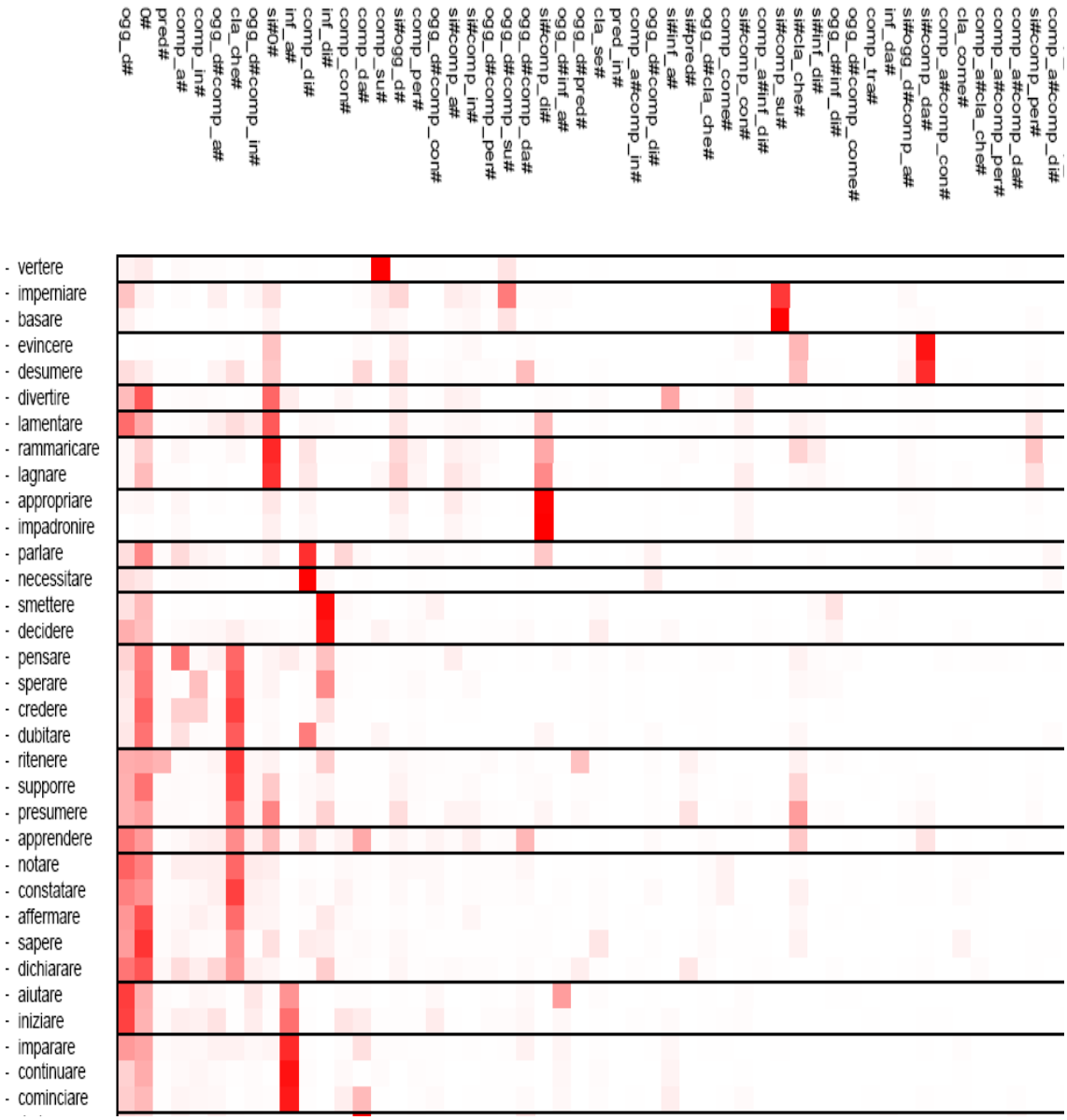
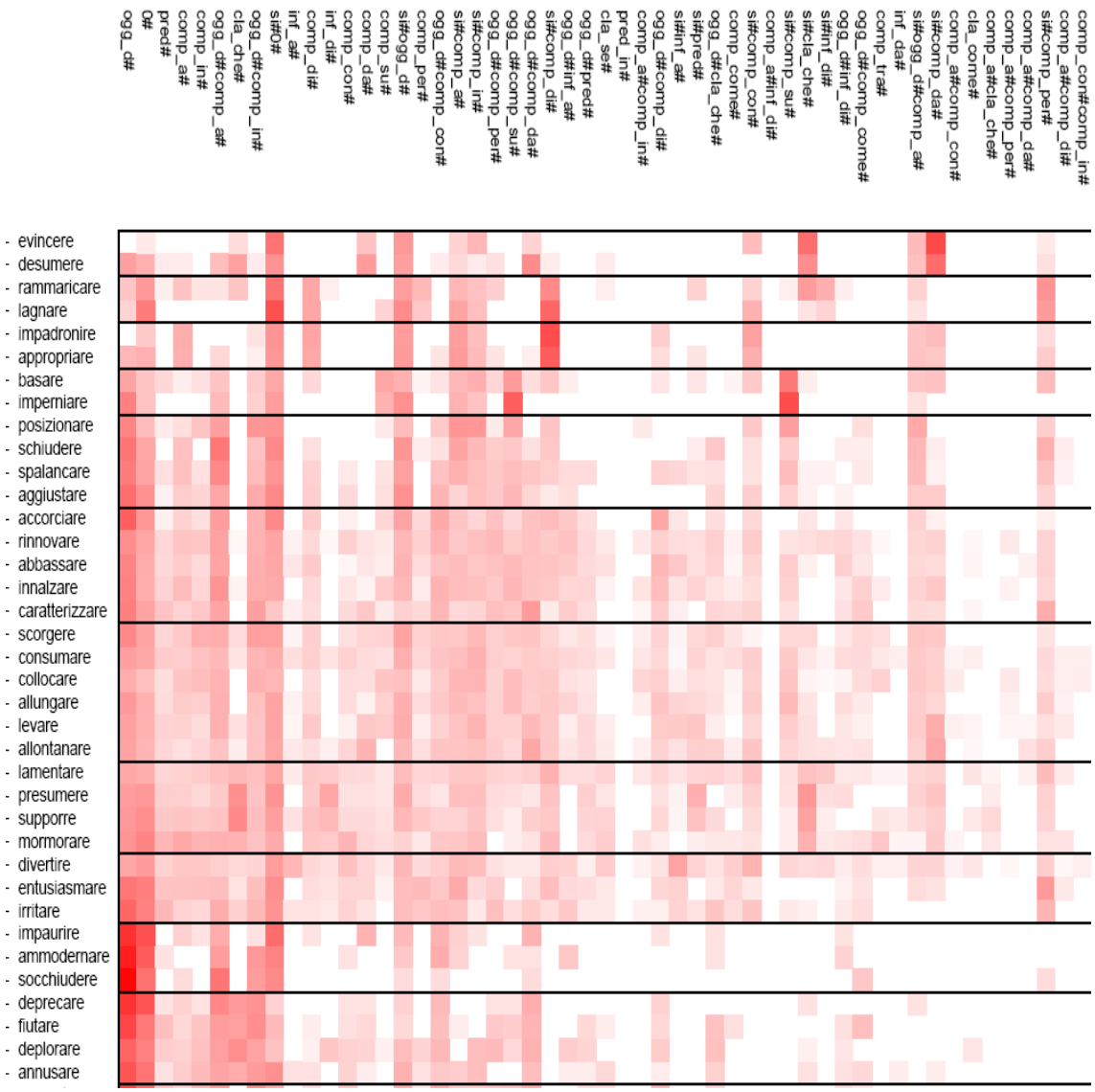


*Figura 4.3 Esempio di matrice costruita in base al logaritmo della frequenza dei frames di sottocategorizzazione*




Nella Figura 4.3 è subito evidente come la colorazione della matrice sia molto più uniforme e distribuita, priva degli ampi spazi bianchi che invece caratterizzano quella in Figura 4.2. Questo vuol dire che nelle procedure di *clustering* viene considerata significativa una più vasta gamma di *frames* di sottocategorizzazione, fatto questo che permette una valutazione più esaustiva e completa sulle strutture sintattiche rilevanti per la rappresentazione del significato verbale. Si consideri ad esempio il verbo *evincere*, per il quale la matrice evidenzia *frames* come *si#cla_che#* (es. *da ciò si evince che l'imputato è innocente*), o ancora *si#comp_da#* (es. *l'innocenza dell'imputato si evince dalle prove addotte*), ovvero specifiche costruzioni sintattiche importanti per valutare il perché dell'appartenenza di quel dato verbo ad un certo *cluster* e quindi per coglierne indicazioni sulla sua semantica, anche tramite il confronto con gli altri membri del *cluster* stesso (in questo caso il verbo *desumere*). Il riferimento ai tratti sintattici tipici di ciascun *cluster*, ovvero ai tratti ricorrenti e distintivi dei verbi che a tale *cluster* appartengono, è possibile grazie ad un'applicazione messa a disposizione da Cluto che fornisce tutta una serie di strumenti di analisi per la verifica dei dati ottenuti. In effetti il *software* elabora dei *reports*,

che contengono tra l'altro proprio indicazioni sulle proprietà distintive e caratteristiche per ogni *cluster* (esempio 7).

(7) Il *report* fa riferimento al *clustering* delle 40 classi verbali, sulla base dei 105 *frames* di sottocategorizzazione selezionati; analizziamo i tratti descrittivi e distintivi del *cluster* numero 1, cui appartengono i verbi *evincere* e *desumere*. Le caratteristiche segnalate sono le seguenti:

```
Descriptive: si#comp_da# 22.2%, si#cla_che# 13.6%, si#0# 12.6%,
             si#ogg_d#comp_da# 10.6%, si#ogg_d#  7.9%
Discriminating: si#comp_da# 15.9%, si#cla_che#  9.9%,
                si#ogg_d#comp_da# 8.1%, comp_in#  4.0%, ogg_d#  3.8%
```

In effetti i verbi che compongono questo *cluster*, tipicamente costruiscono le loro strutture argomentali in base alle caratteristiche sopra indicate:

`si#comp_da#`

a. *Si evince dal testo l'intenzione dell'autore*
b. *Si desume dal tono delle sue parole la sua vera intenzione*

`si#cla_che#`

a. *Si evince che l'evolversi della situazione è imminente*
b. *Si desume che l'evolversi della situazione è imminente*

Le variabili nel numero delle classi considerate e dei *frames* utilizzati, ci permettono appunto 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 il grado di validità e riscontro dell'ipotesi sintattico-semantica.

Sempre a tale scopo sono stati condotti anche esperimenti con 34 *frames* che escludono la presenza delle preposizioni, per verificare il peso l'intervento di questi elementi nella costruzione sintattica e semantica dei verbi selezionati per l'italiano. Alcuni tra i più significativi sono ad esempio: ogg_dir#comp#, comp#comp#, si#ogg_dir#comp#, che evidentemente lasciano sottospecificate le preposizioni che realizzano gli argomenti indicati. L'elenco completo di questa categoria di *frames* è consultabile in appendice, nella Tabella 4.7.

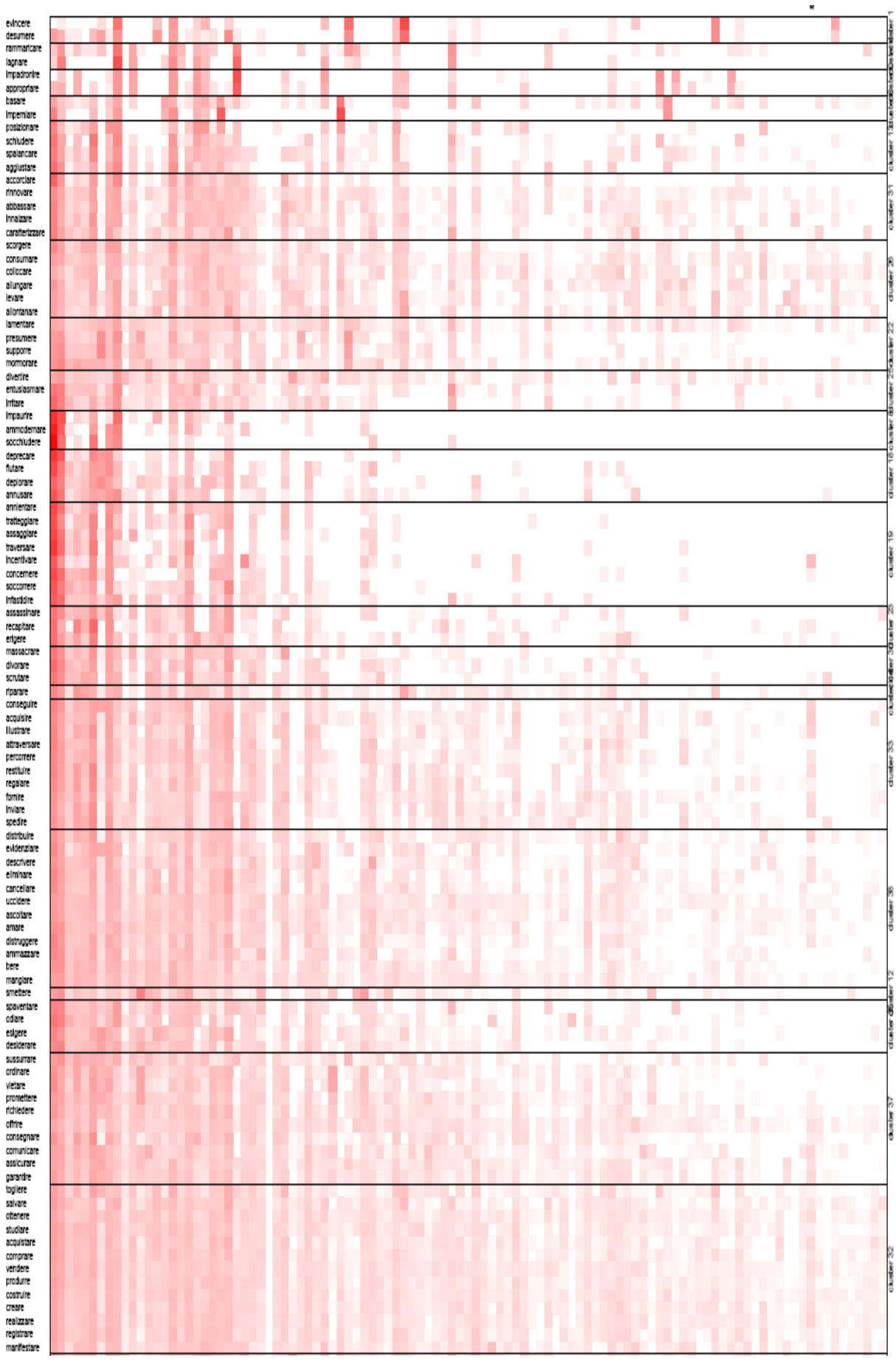
In questa fase non si è tenuto conto dei *fillers* argomentali di questi *patterns*; per tale motivo il *clustering* tiene conto solo delle diverse distribuzioni di frequenza di un verbo rispetto ad un certo *pattern* sintattico, senza considerare i diversi argomenti che tale *pattern* può avere.

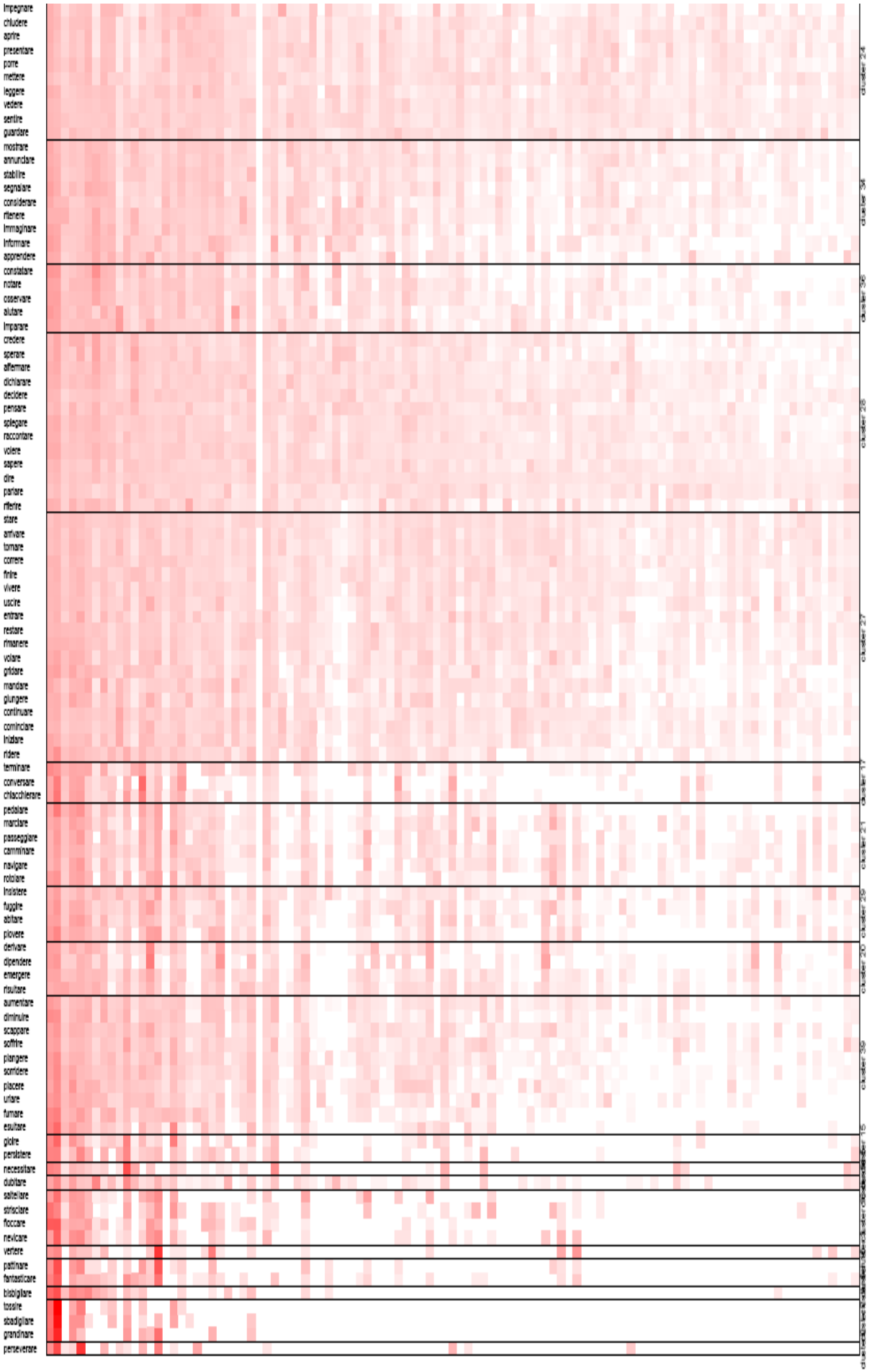
Di seguito vengono discussi i risultati delle diverse classificazioni, prodotti attraverso la variazione dei parametri considerati. Iniziamo con l'analizzare la classificazione ottenuta mantenendo costante il numero di *frames* di sottocategorizzazione a 105 e alternando invece il numero di classi in *output* (40, 24 e 10).

### 4.2.1 Esperimenti di classificazione automatica su 40 classi verbali

Prima di iniziare ad analizzare i singoli *clusters* propri di questa specifica classificazione, proponiamo nella Figura 4.4 un esempio di come vengono presentati in Cluto i risultati ottenuti.

*Figura 4.4 Il clustering delle 40 classi verbali*




impegnare
chiudere
aprire
presentare
porre
mettere
leggere
vedere
sentire
guardare
mostrare
annunciare
stabilire
segnalare
considerare
ritenere
immaginare
informare
apprendere
constatare
notare
osservare
aiutare
imparare
credere
sperare
affermare
dichiarare
decidere
pensare
spiegare
raccontare
volere
sapere
dire
parlare
riferire
stare
arrivare
tornare
correre
finire
vivere
uscire
entrare
restare
rimanere
volare
gridare
mandare
giungere
continuare
cominciare
iniziare
ridere
terminare
conversare
chiacchierare
pedalare
marciare
passeggiare
camminare
navigare
rotolare
insistere
fuggire
abitare
piovere
derivare
dipendere
emergere
risultare
aumentare
diminuire
scappare
soffrire
piangere
sorridere
piacere
urlare
fumare
esultare
gioire
persistere
necessitare
dubitare
saltellare
strisciare
fioccare
nevicare
vertere
pattinare
fantasticare
bisbigliare
tossire
sbadigliare
grandinare
perseverare
cluster 24
cluster 34
cluster 36
cluster 28
cluster 27
cluster 17
cluster 21
cluster 29
cluster 20
cluster 30
cluster 15

Nella classificazione in 40 *clusters*, sono numerose le classi verbali che tendono in larga parte a soddisfare la corrispondenza semantico-sintattica (esempio 8), ovvero che fanno corrispondere ad una coerenza nei tratti sintattici condivisi dagli elementi che compongono il *cluster*, anche una coerenza semantica comune.

(8) CLUSTER 0
impadronire(si)
appropriare(si)

Questi due verbi, che nella classificazione semantica elaborata *a priori* appartengono entrambi alla classe definita *Transfer of Possesion*, vengono raggruppati insieme anche nelle operazioni automatiche di *clustering*, poiché condividono tipicamente il tratto sintattico *#si#comp_di*:

a. Il giocatore si impadronisce del pallone
b. Il truffatore si appropria di una somma di denaro

CLUSTER 8
pedalare
marciare
passeggiare
camminare
navigare

Anche questi verbi mostrano una reciproca coerenza semantica, difatti appartengono tutti alla categoria semantica dei *Verbs of Motion*, sebbene al loro interno siano riconoscibili dei sottogruppi. Ad esempio *camminare* viene classificato come *Motion (manner)*, mentre *pedalare* come *Motion (manner-vehicle)*. I tratti sintattici condivisi da questi elementi riguardano soprattutto l'uso delle preposizioni e vengono individuati da Cluto come segue:

- 0# (ovvero intransitività) *Mauro passeggia*
- #comp_in *Mauro naviga in mare aperto*
- #comp_a *Mauro cammina a passo svelto*

CLUSTER 25
divertire
irritare
entusiasmare

A questo *cluster* appartengono verbi della classe semantica *Emotion*; Cluto li raggruppa insieme in virtù della loro coerenza sintattica, visto che presentano tutti e tre *si#0#* come tratto sintattico prevalente:

a. *Mauro si diverte*
b. *Mauro si irrita*
c. *Mauro si entusiasma*

Ci sono altri casi in cui l'individuazione del tratto sintattico tipico di un certo gruppo di verbi genera confusione nel programma di *clustering*, spesso perché esso è troppo generico e non consente un riconoscimento inequivocabile delle possibili sfumature di significato dei verbi che lo adottano (esempio 9). Questo è particolarmente evidente nei verbi altamente o esclusivamente intransitivi, che formano dei *clusters* estremamente confusi e disomogenei al loro interno.

(9)

| CLUSTER 7 | CLUSTER 16 |
|---|---|
| tossire | saltellare |
| sbadigliare | strisciare |
| grandinare | fioccare |
| | nevicare |

In questi *clusters* troviamo classificati insieme, verbi appartenenti a categorie semantiche molto diverse tra loro:
- *Facial expression* (tossire, sbadigliare);
- *Weather* (grandinare, fioccare, nevicare);
- *Motion (manner)* (saltellare, strisciare).

Il motivo di tale confusione semantica è che questi sono tutti verbi fortemente intransitivi e che utilizzano anche le stesse preposizioni (su, in, a), perciò Cluto non riesce attraverso il solo comportamento sintattico a differenziarli e ripartirli in modo coerente. A questo proposito, occorre specificare due limiti intrinseci al modo in cui sono stati rappresentati i dati:
- il fatto che non vengono distinti i diversi sensi delle preposizioni (ad esempio *a* locativo e *a* dativo);
- il fatto che le varie preposizioni locative (ad esempio *a*, *in*) sono trattate come occorrenze del tutto separate.

In questo esempio, l'informazione sintattica, ovvero il comportamento dei verbi, è insufficiente per catturarne il significato e costruire classi semantiche accettabili. Difatti il tratto dell'intransitività è estremamente generico e la specificazione delle preposizioni non è in grado di raffinarlo; in italiano l'uso delle preposizioni nei costrutti verbali non è spesso così significativo come in altre lingue come il tedesco (cfr. lavori della Schulte im Walde), dove queste si accompagnano anche a casi specifici (dativo, accusativo, genitivo) e la loro combinazione determina mutazioni evidenti del significato dei verbi cui si accompagnano.

In altri casi invece, la mancanza di corrispondenza tra la classificazione automatica e quella semantica, dipende da un'ambiguità di fondo di quest'ultima dovuta al suo essere stata costruita *a priori* da un linguista (esempio 10). La ripartizione in classi semantiche dipende dalle scelte personali operate da quest'ultimo, dalle proprie competenze sulla lingua, dallo scopo dell'esperimento ed è quindi suscettibile a modificazioni e correzioni e, sicuramente, non è l'unica classificazione possibile per quel gruppo di verbi.

(10)

| CLUSTER 13 |
|---|
| posizionare |
| schiudere |
| spalancare |

Nella classificazione semantica proposta nella Tabella 4.1 questi verbi appartengono a classi semantiche diverse:
- *bring into position* (posizionare);
- *opening* (schiudere, spalancare).

In realtà si deve considerare il fatto che tutti implicano anche un'idea di spostamento o comunque di cambiamento di posizione, inoltre essi condividono tipicamente il tratto sintattico *si#0#*, pertanto il programma di *clustering* tende a raggrupparli insieme.

È innegabile, e comunque significativo per una valutazione dei dati, che Cluto produca dei *clusters* semanticamente incoerenti ed immotivati, insieme ad altri poco significativi per

una successiva analisi linguistica poiché contengono un solo elemento al loro interno, o al contrario molto popolati, ma al contempo troppo generici:

(11)

| CLUSTER 23 | CLUSTER 2 |
|---|---|
| assassinare | perseverare |
| recapitare | |
| erigere | |
| **CLUSTER 9** | **CLUSTER 12** |
| bisbigliare | smettere |

| CLUSTER 27 | | |
|---|---|---|
| stare | finire | giungere |
| arrivare | uscire | continuare |
| tornare | entrare | cominciare |
| restare | volare | ridere |
| correre | gridare | iniziare |
| rimanere | mandare | |

È senz'altro interessante evidenziare alcuni rapporti di similarità, che il programma di *clustering* riesce comunque a catturare nella classificazione automatica; si pensi, ad esempio, a verbi come *bere* e *mangiare* che appartengono allo stesso *cluster*, ma non si trovano insieme agli altri verbi della stessa classe semantica *Consumption* (ad es. *fumare* e *consumare*), presumibilmente poiché indicano processi affini, ma dissimili rispetto a questi ultimi. Questa sfumatura di significato fa parte, in effetti, del bagaglio di competenze del parlante dell'italiano ed è perciò linguisticamente motivata. Quanto appena detto, può venire considerato come un limite della classificazione manuale; è infatti possibile, che l'aver raggruppato insieme i quattro verbi sopra menzionati in un'unica classe, che abbiamo denominato *Consumption*, sia stata un'operazione troppo generica ed arbitraria, poiché tende ad offuscare differenze importanti che li distinguono.

Cluto assegna a questo insieme di verbi una serie di caratteristiche descrittive che li accomunano e che ci permettono di capire perché il programma li accorpa insieme, nonostante le discrepanze semantiche; tali tratti sono:

- comp_in (*rimanere in gara*, *stare in casa*, *continuare in segreto*, etc.);
- comp_a (*gridare a squarciagola*, *correre a perdifiato*, *cominciare a studiare*, *entrare a casa*, etc).

Il *cluster* 27 è chiaramente dominato dai *frames* locativi introdotti dalla preposizione *a*; esso risulta piuttosto omogeneo al suo interno, se non fosse per la presenza di verbi come *gridare*, *ridere*, *cominciare* e *iniziare*. Questi potrebbero essere stati inclusi, perché il programma di *clustering* non riesce a distinguere l'uso locativo della preposizione *a*, dalle altre funzioni che essa può rivestire. Per questo stesso motivo, in altri casi, non è stato possibile distinguere alcuni verbi di movimento da quelli di posizione.

Finora si è partiti dall'analisi dei *clusters*, per andare a verificarne la coerenza interna, ovvero la misura in cui i verbi di un *cluster* appartengono alle stesse classi semantiche.

A questo punto, occorre sviluppare anche la riflessione opposta: vale a dire data una certa classe semantica costruita *a priori*, andare a confrontare come sono stati ripartiti i verbi in essa contenuti all'interno dei *clusters* di arrivo. Se, ad esempio, consideriamo una classe composta da cinque verbi, di cui tre si trovano in un *cluster* e due in un altro, questo fenomeno ci rivelerà informazioni interessanti sul fatto che verbi che appartengono *ontologicamente* alla stessa classe, distribuzionalmente mostrano comportamenti diversi (esempio 12).

(12) *Impadronirsi* ed *appropriarsi* si riferiscono entrambi alla classe *Transfer of Possession (obtaining)*, e formano inoltre insieme il *cluster 0*. Gli altri verbi appartenenti alla stessa classe semantica, non si trovano però raggruppati in uno stesso *cluster*; ciò avviene perché, evidentemente, questa non è una classe semantica molto omogenea dal punto di vista distribuzionale. Difatti:

- *impadronirsi* ed *appropriarsi* costruiscono argomenti del tipo *comp_di*
  a. Mauro si appropria delle mie idee;
  b. Mauro si impadronisce delle mie idee.
- *acquisire* e *conseguire*, che insieme si trovano nel *cluster 33*, costruiscono argomenti del tipo *ogg_dir+comp_a*
  c. Mauro acquisisce la proprietà della casa a pieno diritto;
  d. Mauro consegue la laurea a pieni voti.
- *acquistare*, *comprare* e *ottenere*, presentano un tratto distintivo molto generico, ovvero quello della transitività, che li colloca nel *cluster 32* insieme ad altri verbi semanticamente molto diversi fra loro, come *salvare*, *studiare*, *produrre* etc.
  e. Mauro acquista una giacca;
  f. Mauro compra una sciarpa;
  g. Mauro ottiene una promozione;
  h. Mauro salva un bagnante;
  i. Mauro studia filosofia;
  l. Mauro produce bottoni.

Ancora, i verbi appartenenti alla classe semantica *Motion (manner)*, si trovano raggruppati in tre *clusters* differenti, poiché distribuzionalmente presentano *pattern* di sottocategorizzazione diversi. Avremo allora che:

- *camminare*, *marciare*, *passeggiare* e *rotolare*, appartengono al *cluster 21*, che costruisce argomenti del tipo *0#+comp_in*;
- *saltellare* e *strisciare*, sono collocati nel *cluster 16* insieme a verbi semanticamente molto diversi da loro, quali *fioccare* e *nevicare*, proprio perché tutti presentano lo stesso tratto descrittivo prevalente, ovvero *0#+comp_su*
  a. Mauro saltella su un piede;
  b. Il serpente striscia sulla pancia;
  c. La neve fiocca sui tetti;
  d. Nevica sulla città.
- *correre* appartiene al *cluster 27*, che presenta il tratto estremamente generico dell'intransitività, accogliendo pertanto al suo interno verbi dal significato vario, quali ad esempio: *ridere*, *gridare*, *restare*. Per poter raffinare ulteriormente questo tipo di *clusters* occorre aggiungere dell'informazione semantica, poiché la sintassi da sola non è evidentemente sufficiente.

Ai fini della nostra analisi linguistica, è molto utile anche valutare quali sono le classi semantiche meglio rappresentate nei risultati delle operazioni di *clustering*; in altre parole quali sono le classi formulate *a priori*, che meglio delle altre trovano un corrispettivo distribuzionale omogeneo. Nell'esempio 13, possiamo vedere come i *Consumption Verbs* e i *Desire (need) Verbs*, non siano soddisfacenti da questo punto di vista:

(13) - *bere* e *mangiare* si trovano insieme nel *cluster 38*;
- *consumare* appartiene al *cluster 26*;
- *divorare* fa parte del *cluster 30*;
- *fumare* si colloca nel *cluster 39*.

La ragione di questa discontinuità risiede nel fatto che questi sono verbi semplicemente transitivi, capaci di distinguersi da altri transitivi magari per le preferenze di selezione semantica con cui si accompagnano, ma non tanto per i *patterns* di sottocategorizzazione che presentano. Perciò la procedura di *clustering* improntata su base prettamente sintattica, non sarà in questo caso capace di cogliere queste distinzioni di stampo puramente semantico.

La stessa cosa accade per la classe semantica dei *Desire (need) Verbs*, composta dai verbi *esigere*, *necessitare* e *richiedere*, che appartengono tutti a *clusters* sintattici diversi:

- *esigere* si trova nel *cluster 35*, che presenta come tratto argomentale *si#0#*
  a. Si esige che la situazione cambi;
- *necessitare* costituisce un *cluster* a se stante, il numero 5, caratterizzato dal tratto descrittivo *comp_di*
  b. Mauro necessita di un paio di scarpe nuove;
- *richiedere* appartiene al *cluster 37*, che presenta come tratto tipico *ogg_dir+comp_a*
  c. Il docente richiede agli studenti più impegno

Al *cluster 37* appartengono verbi dal significato molto vario, ma uniti dallo stesso *pattern* di sottocategorizzazione:

- *sussurrare* es. Mauro sussurra una frase ad un amico;
- *ordinare* es. Il generale ordina l'attenti ai suoi sottoposti;
- *vietare* es. Il preside vieta l'assemblea agli studenti;
- *promettere* es. La mamma promette un regalo al bambino;
- *offrire* es. Mauro offre la cena agli amici;
- *consegnare* es. Il postino consegna una lettera al signor Rossi;
- *comunicare* es. Il giornalista comunica una notizia ai telespettatori;
- *assicurare* es. L'esercito assicura il suo appoggio al Presidente;
- *garantire* es. La polizia garantisce la sicurezza ai cittadini.

Anche in questo caso, l'indicazione sintattica fornita è troppo generica, da sola, per catturare le sfumature di significato proprie dei verbi *clusterizzati*.

Al contrario, ci sono invece delle classi di verbi, come nel caso di *Aspect* o *Communication*, che si dimostrano molto più omogenee dal punto di vista distribuzionale e che, quindi, trovano una maggiore corrispondenza semantico-sintattica nei rispettivi *clusters* di appartenenza (esempio 14).

(14) Classe semantica: *Aspect*
Verbi contenuti:

*cominciare*, *continuare*, *finire*, *iniziare* } *cluster 27*

*smettere* → *cluster 12*

*terminare* → *cluster 17*

Come dimostra la ripartizione effettuata da Cluto, la maggior parte dei verbi appartenenti alla categoria *Aspect*, formano un unico *cluster*, poiché questi sono accomunati dal tratto sintattico della transitività. *Smettere* e *terminare* si trovano invece in *clusters* diversi, perché diverso è il tratto sintattico che li descrive, ovvero *inf_di*.

Anche la classe *Communication* è molto coerente dal punto di vista distribuzionale; difatti i verbi in essa contenuti formano due soli *clusters*:


### 4.2.2 Esperimenti di classificazione automatica su 24 classi verbali

Mantenendo inalterato il numero dei *frames* di sottocategorizzazione utilizzati, ma riducendo il numero delle classi a 24, è possibile notare un aumento della confusione semantica all'interno dei singoli *clusters*; prevedibilmente ciò si deve al fatto che molti dei verbi che nella ripartizione in 40 classi formavano da soli dei *clusters* separati, sono stati ora accorpati insieme da Cluto proprio in virtù della diminuzione del numero delle classi disponibili in cui inserire i verbi stessi. Inoltre si deve considerare il fatto che le dimensioni del vettore contengono esclusivamente la distribuzione del verbo rispetto a *patterns* di sottocategorizzazione, che non tengono conto in nessuna misura dei tratti semantici degli argomenti.

Di seguito vengono riportati degli esempi di quanto appena detto:

(15) CLUSTER 11
dubitare
necessitare
gioire
persistere

Seppure questi verbi appartengono a classi semantiche distinte, Cluto li raggruppa insieme in virtù dei loro comuni tratti sintattici e preposizionali:
- *0#* *Il maltempo persiste*
- comp_di# *Mauro gioisce della vittoria*

Anche in questo caso il comportamento sintattico dei verbi non è sufficiente a catturarne automaticamente il significato.

CLUSTER 4
impaurire
ammodernare
socchiudere

Lo stesso dicasi per quest'ultimo *cluster*, al cui interno si trovano verbi semanticamente disomogenei tra loro, ma che mostrano una certa coerenza nei tratti sintattici con cui tipicamente costruiscono la rispettiva struttura argomentale; vale a dire:
- *ogg_d#* *Mauro ammoderna l'appartamento*
- *si#0#* *Mauro si impaurisce*

Sebbene in numero minore, è possibile comunque individuare anche in questa classificazione dei *clusters* semanticamente convincenti:

(16) CLUSTER 3
imperniare
basare

Tali verbi condividono la categoria semantica *Basis*, nonché il seguente tratto sintattico, che ne permettono il raggruppamento anche a livello automatico:

- *si#comp_su#* *La relazione si basa sugli esperimenti condotti*
  *La sua vita si impernia sul lavoro*

Gli altri due verbi che appartengono alla stessa categoria semantica, ovvero *vertere* e *concernere*, si trovano in clusters diversi, rispettivamente l'8 e il 10, proprio perché mostrano altri tratti sintattici distintivi:

- *comp_su#* *es. La relazione verte sugli esperimenti condotti*
- *ogg_dir#* *es. La relazione concerne gli esperimenti condotti*

Anche per questa classificazione è dunque possibile condurre l'analisi opposta a quella fin qui presentata: ovvero valutare quanto le classi semantiche *a priori* di partenza trovino una certa corrispondenza nei *clusters* sintattici di arrivo.

Anche in questo caso occorre tener conto del grado di omogeneità distribuzionale dei verbi che compongono le varie classi, ed in più del fatto che queste ultime sono molto meno granulari rispetto alla classificazione in 40 classi. Ciò risulta evidente, ad esempio, per i *Motion Verbs* che non presentano più la distinzione in: *Manner*, *Cross*, *Manner-Vehicle* e *Directed*, e formano pertanto un'unica grande classe. Vediamo nell'esempio 17, come questa classe viene raggruppata dalla procedura di *clustering*:

(17) trattandosi di una classe molto vasta, i verbi in essa contenuti presenteranno *pattern* di sottocategorizzazione simili soltanto a gruppi, formando pertanto altrettanti *clusters* sintattici diversi:

- una larga parte dei verbi considerati presentano il tratto descrittivo dell'intransitività, e vengono perciò raggruppati insieme nel *cluster 13*; è il caso ad esempio di:
  *arrivare* es. Il treno arriva in orario
  *entrare* es. La sposa entra in chiesa
  *giungere* es. Il plotone giunge a destinazione
  *tornare* es. Il soldato torna a casa
  *uscire* es. Gli studenti escono da scuola
  *volare* es. L'aereo vola su Parigi
- molti dei restanti verbi di movimento costruiscono argomenti del tipo *comp_in*, e si trovano nel *cluster 14*; tra questi citiamo:
  *marciare* es. I soldati marciano in fila indiana
  *navigare* es. Il panfilo naviga in mare aperto
  *abitare* es. Mauro abita in Francia
- gli altri verbi appartenenti alla classe semantica denominata *Motion*, si collocano singolarmente in altrettanti *clusters* sparsi, proprio a causa dei diversi *frames* di sottocategorizzazione che li caratterizzano e li dividono, sintatticamente dagli altri. È il caso di verbi come:
  *attraversare* (*cluster 18*) *+ogg_dir+comp_a* es. Il bambino attraversa la strada a piedi
  *pattinare* (*cluster 8*) *+comp_su* es. Il bambino pattina sul ghiaccio
  *scappare* (*cluster 22*) *+comp_a* es. Il ladro scappa a gambe levate

Nonostante la minore granularità di questo secondo tipo di classificazione, alcune delle classi semantiche proposte *a priori*, trovano comunque una corrispondenza soddisfacente nei *clusters* prodotti da Cluto. Nell'esempio 18 ne elenchiamo alcune tra le più significative, accompagnate dall'indicazione dei rispettivi tratti sintattici descrittivi tipici, che ne determinano il raggruppamento:

(18) Classe semantica: *Quantum of Change*
Verbi contenuti:

*abbassare*, *accorciare*, *allungare*, *innalzare* } *cluster 16* (+*ogg_dir*)

*aumentare*, *diminuire* } *cluster 22* (+*0#*)

Classe semantica: *Result*
Verbi contenuti:

*risultare*, *emergere*, *derivare* } *cluster 9* (+*comp_da*)

Classe semantica: *Inference*
Verbi contenuti:

*desumere*, *evincere* } *cluster 1* (+*si#comp_da*)

*dipendere* ⟶ *cluster 9* (*comp_da*)

### 4.2.3 Esperimenti di classificazione automatica su 10 classi verbali

Il terzo tipo di esperimenti condotti, prevede una classificazione con solo 10 *clusters* in *output*, che nonostante il considerevolmente ridotto numero di classi presenta molta più coerenza semantica di quanto non faccia quella intermedia. Questo aspetto può essere determinato da una maggiore naturalezza, propria della classificazione a priori cui si fa riferimento.

Esaminiamo alcuni dei *clusters* prodotti da Cluto in quest'ultimo raggruppamento:

(19)

| CLUSTER 0 | CLUSTER 1 |
|---|---|
| desumere | appropriare |
| evincere | impadronire |
| | lagnare |
| | rammaricare |

Il primo dei *clusters* proposti presenta due verbi che appartengono entrambi alla classe semantica denominata *Inference*, ed è perciò semanticamente giustificato; l'altro mostra comunque una certa coerenza, ma presenta due sottogruppi:
- il primo si riferisce alla classe semantica detta *Transfer of Possession* ed include i verbi *appropriare* ed *impadronire*;
- il secondo si riferisce invece alla classe semantica etichettata come *Basis* e comprende i verbi *lagnare* e *rammaricare*.

Entrambi questi sottogruppi costruiscono la propria struttura sintattica con dei tratti sintattici abbastanza specifici da permettere a Cluto di produrre una classificazione raffinata e granulare per gli stessi. Questi tratti sono:

- per il *cluster 0*:

| | |
|---|---|
| *si#comp_da#* | *Come si evince dal racconto dei testimoni* |
| *si#cla_che#* | *Si desume che lo svolgimento della prova è irregolare* |

- per il *cluster 1*:
  *si#comp_di#* *Mauro si appropria della mia idea*
  *si#0#* *Mauro si rammarica*

Ci sono poi altri *clusters*, sempre all'interno di questo tipo di suddivisione in 10 classi, che si presentano come estremamente variegati al loro interno, ma che contengono comunque dei nutriti sottoinsiemi semanticamente coerenti. Dato il numero delle classi estremamente limitato, queste tendono a diventare sempre più miste al loro interno; è il caso, ad esempio, del *cluster 5*:

(20)
finire, iniziare, cominciare, continuare } *Aspect*

vivere, stare, rimanere, restare } *Be in Position*
mettere, porre } *Bring into Position*
} *Position*

dire, parlare, raccontare } *Communication*

credere, pensare, sapere } *Propositional Attitude*

volare, uscire, entrare, arrivare, tornare, giungere } *Motion*

Infine citiamo anche per quest'ultima classificazione, il caso di *clusters* completamente immotivati a livello semantico, come quello del *cluster 3*:

(21) CLUSTER 3

| | |
|---|---|
| perseverare | fantasticare |
| tossire | vertere |
| sbadigliare | nevicare |
| grandinare | fioccare |
| pattinare | strisciare |
| saltellare | |

Come abbiamo fatto per le precedenti due classificazioni, condurremo anche in questo caso un'analisi inversa sugli esperimenti condotti; nonostante le classi semantiche di partenza

siano estremamente più ampie, rispetto agli altri due casi, cercheremo comunque di valutarne la corrispondenza e la compattezza sintattica e distribuzionale.

Di seguito vengono proposte le classi semantiche che meglio si riflettono nei *clusters* di arrivo, proprio in virtù dei *patterns* di sottocategorizzazione simili, che caratterizzano i verbi in esse contenute (esempio 22).

(22) Classe semantica: *Aspect*
Verbi contenuti:

*cominciare*, *iniziare*, *finire*, *continuare* } *cluster 5* (*+comp_a*)

*smettere*, *terminare* } *cluster 9* (*+comp_in*)

L'omogeneità distribuzionale dei verbi contenuti in questa categoria semantica, fa si che essi siano raggruppati in pochi *clusters*, semanticamente coerenti al loro interno.

Lo stesso vale per i verbi appartenenti alla classe semantica *Support*, che condividono pattern di sottocategorizzazione simili tra loro:
Categoria semantica: *Support*
Verbi contenuti:

*aiutare*, *salvare* } *cluster 6* (*ogg_dir+comp_a*)

*incentivare*, *soccorrere* } *cluster 4* (*ogg_dir*)

Viceversa l'esempio 23, contiene il caso di verbi appartenenti ad una stessa classe semantica di partenza, ma dal comportamento sintattico dissimile; il programma di *clustering* li ha pertanto raggruppati in altrettanti *clusters* differenti. Ancora una volta si rende necessaria l'aggiunta di informazione semantica negli esperimenti condotti, se si vogliono cogliere le differenze e le affinità di significato tipiche di alcuni gruppi di verbi, che altrimenti andrebbero perse con il solo utilizzo della sintassi.

(23) Classe semantica: *Argument*
Verbi contenuti:

*derivare*, *dipendere*, *emergere*, *risultare* } *cluster 9* (*+comp_in*)

*desumere*, *evincere* } *cluster 0* (*+si#comp_da*)

*basare* (*+ogg_dir*), *imperniare* (*+comp_su*) } *cluster 2* (+si#0#)

*concernere* → *cluster4*

*vertere* → *cluster3*

### 4.2.4 Variazione numerica dei *frames* di sottocategorizzazione nei vari tipi di classificazione

Verifichiamo ora gli effetti prodotti sulla classificazione automatica dalla riduzione del numero dei *frames* di sottocategorizzazione utilizzati, che passa da 105 a 50. Questo mutamento comporta, nella ripartizione in 40 classi, una vistosa sgranatura a livello quantitativo dei componenti dei singoli *clusters*, che arrivano a contenere poco più di tre o quattro elementi ciascuno nella maggior parte dei casi.

Ad esempio, i verbi *evincere* e *desumere*, appartenenti alla classe semantica *Inference*, che nelle classificazioni finora analizzate si presentavano sempre insieme all'interno dello stesso *cluster*, in questo caso ne costituiscono due separati, nello specifico il numero 0 e il numero 3.

In altri casi i *clusters*, sebbene più esigui rispetto agli esperimenti condotti su 105 *frames* di sottocategorizzazione, mostrano un'aumentata coerenza semantica al loro interno:

(24)

| CLUSTER 4 | | CLUSTER 6 | |
|---|---|---|---|
| lagnare | *Moaning* | sbadigliare | *Facial Expression* |
| rammaricare | | tossire | |

| CLUSTER 16 | |
|---|---|
| conversare | *Communication* |
| chiacchierare | |

| CLUSTER 26 | | CLUSTER 30 | |
|---|---|---|---|
| rotolare | *Motion* | distribuire | *Transfer of Possession* |
| navigare | | spedire | |
| passeggiare | | inviare | |
| camminare | | fornire | |
| marciare | | regalare | |
| pedalare | | restituire | |

Altri *clusters* invece, specie quelli contenenti verbi prevalentemente intransitivi, non risultano affatto migliorati dalla riduzione dei *frames* considerati, in quanto comunque troppo generici per fornire indicazioni utili al programma di *clustering* per una ripartizione in classi semanticamente motivate. Eccone di seguito alcuni esempi tra i più significativi:

(25)

| CLUSTER 11 | CLUSTER 13 |
|---|---|
| fantasticare | nevicare |
| grandinare | saltellare |
| pattinare | |

Alla diminuzione del numero *frames* di sottocategorizzazione corrisponde, nella classificazione in 24 *clusters*, un'accresciuta incoerenza semantica. I *clusters* risultano per lo più semanticamente poco convincenti e motivati; inoltre si passa da gruppi di verbi numericamente molto esigui e spesso scollegati tra loro, ad altri invece assai corposi e necessariamente quindi composti da verbi anche molto diversi fra loro nel significato, seppur affini sintatticamente (esempio 26). Il diradarsi dei *frames* utilizzati, li rende più generici rispetto a quelli degli altri esperimenti condotti; ciò tende ad amplificare i problemi nel rintracciare un corrispettivo semantico-sintattico preciso, all'interno della classificazione automatica.

(26)

| CLUSTER 12 | CLUSTER 11 | |
|---|---|---|
| bisbigliare | saltellare | fioccare |
| gioire | strisciare | nevicare |
| necessitare | persistere | |

CLUSTER 21

| | | |
|---|---|---|
| iniziare | comprare | salvare |
| mandare | vendere | registrare |
| consegnare | produrre | realizzare |
| riparare | creare | leggere |
| togliere | mettere | considerare |
| bere | costruire | informare |
| mangiare | porre | lamentare |
| studiare | presentare | divertire |
| uccidere | aprire | impegnare |
| ascoltare | chiudere | |
| ottenere | manifestare | |
| acquistare | consumare | |

Anche nella divisione in 10 classi, la riduzione dei *frames* di sottocategorizzazione ha prodotto un accorpamento dei verbi in *clusters* molto consistenti, ma comunque più coerenti al loro interno, in rapporto alla classificazione intermedia in 24 classi. Si prenda ad esempio il *cluster 5*:

(27) esso contiene i seguenti sottogruppi di verbi:

| | | | |
|---|---|---|---|
| posizionare<br>basare<br>imperniare | *Basis* | abbassare<br>accorciare<br>allungare<br>innalzare | *Quantum of Change* |
| schiudere<br>spalancare | *Opening* | | |
| collocare<br>levare<br>posizionare | *Bring into Position* | aggiustare<br>rinnovare | *Renovation* |

Come già accennato, a differenza di quanto accade ad esempio nei lavori condotti da Schulte im Walde per il tedesco, la struttura linguistica dell'italiano non assegna alle preposizioni un ruolo altrettanto significativo nel catturare l'interfaccia semantico-sintattica dei verbi. Ciò è tanto più evidente negli esperimenti condotti considerando soltanto 34 *frames* di sottocategorizzazione non preposizionali, in cui i *clusters* ottenuti non risultano sensibilmente peggiorati rispetto alle precedenti classificazioni.

Nell'esempio 28 verifichiamo appunto il caso di alcuni gruppi di verbi, che non vengono minimamente inficiati dall'assenza delle preposizioni, proprio perché queste non compaiono nella costruzione dei loro tratti sintattici preponderanti.

(28)

| CLUSTER 7 | CLUSTER 0 | CLUSTER 22 | |
|---|---|---|---|
| basare | appropriare | camminare | rotolare |
| posizionare | impadronire | passeggiare | |
| imperniare | | navigare | |

Nelle classificazioni a 24 e 10 classi, il cambiamento più significativo che si ottiene inserendo i *frames* non preposizionali, è una granularità molto ridotta dei *clusters* (esempio 29), probabilmente perché questi si basano ora su costrutti sintattici essenziali, non specificati dall'intervento di alcuna preposizione appunto.

(29) CLUSTER 0 (10 classi)

| | |
|---|---|
| evincere | appropriare |
| desumere | basare |
| rammaricare | posizionare |
| lagnare | imperniare |
| impadronire | |

Un miglioramento tangibile nella classificazione automatica dei verbi, si ottiene invece utilizzando solo i *frames* di sottocategorizzazione a più alta frequenza, in particolare quelli con un indice > di 0.01. Questo indice si riferisce alla frequenza relativa del *frame*, data la frequenza del verbo; ad esempio, 0.01 vuol dire che quel *frame* copre l'1% di tutte le occorrenze del verbo cui si associa.

I risultati sono molto più precisi per tutte e tre le tipologie di *clustering* effettuate; con questa procedura infatti si eliminano i *frames* meno utilizzati sintatticamente da un verbo e quindi conseguentemente le costruzioni meno frequenti e che spesso veicolano significati secondari o espressioni idiomatiche del verbo stesso.

Nell'esempio 30, vengono riportati alcuni dei *clusters* migliori per le tre classificazioni effettuate.

(30) La classificazione in 40 *clusters* risulta molto migliorata dal ricorso ai soli *frames* più frequenti, come dimostrano i seguenti gruppi di verbi:

| CLUSTER 7 | | CLUSTER 21 | |
|---|---|---|---|
| continuare | *Aspect* | ridere | *Facial Expression* |
| cominciare | | sorridere | |
| iniziare | | tossire | |
| | | sbadigliare | |
| | | piangere | |
| | | gioire | |

| CLUSTER 30 | | |
|---|---|---|
| esultare | saltellare | *Motion* |
| correre | strisciare | |
| navigare | rotolare | |
| passeggiare | nevicare | |
| camminare | | |

Anche i *clusters* della classificazione intermedia, quella composta da 24 classi, risultano piuttosto migliorati dall'uso dei *frames* ad alta frequenza:

| CLUSTER 7 | | CLUSTER 11 | | |
|---|---|---|---|---|
| spaventare | *Emotion* | rotolare | marciare | *Motion* |
| impaurire | | strisciare | pedalare | |
| infastidire | | saltellare | navigare | |
| amare | | camminare | correre | |
| irritare | | passeggiare | volare | |
| odiare | | | | |
| socchiudere | | | | |
| ammodernare | | | | |

Il *cluster 11* è più vasto in realtà, ma è comunque popolato a maggioranza da verbi che veicolano il significato del movimento.

Gli esperimenti con i *frames* ad alta frequenza, danno risultati poco precisi solo nel terzo tipo di classificazione, quella composta da 10 *clusters*, perché le classi contengono molti verbi e risultano perciò troppo miste e semanticamente disomogee.

Alcune tra queste, pur presentando verbi che appartengono a classi semantiche diverse, mostrano comunque un certo grado di affinità nei significati che veicolano:

CLUSTER 2

apprendere, desumere, evincere — *Learning*; *Inference*

ritenere, pensare — *Propositional Attitude*

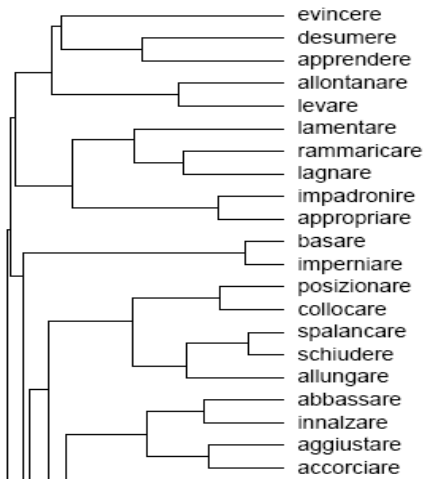
presumere, supporre, immaginare — *Speculation*

Tutti i significati convogliati rimandano a processi cognitivi legati all'acquisizione di nuove informazioni.

Oltre alla funzione di *clusterizzazione* dei dati, Cluto fornisce anche altri strumenti di indagine per approfondire l'analisi dei risultati ottenuti. Uno di questi consiste nell'elaborazione di un grafo ad albero, che costruisce delle connessioni tra i verbi *inter-* ed *intra-clusters* al fine di permettere la visualizzazione delle connessioni che intercorrono tra gli elementi dei *clusters* stessi. I verbi sono legati da archi che simboleggiano il loro legame di maggiore o minore prossimità; anche all'interno di uno stesso *cluster* troveremo dunque verbi classificati come più vicini e quindi più simili, rispetto agli altri elementi dello stesso. In tal modo si creerà, insomma, una sorta di scala gerarchica di similarità o di prossimità che unisce i verbi trattati nella classificazione.

La Figura 4.5 mette in evidenza come anche all'interno di uno stesso *cluster* ci siano verbi legati da un significato più strettamente connesso rispetto ad altri. Si prenda ad esempio il verbo *allontanare*, che nel grafo viene unito immediatamente a *levare*, ed ancora i verbi *lamentare*, *rammaricare* e *lagnare* uniti da uno stesso ramo; solo successivamente questi due archi si ricongiungeranno nella gerarchia costruita dal grafo. Questo vuol dire che Cluto giudica più vicini, e quindi più simili tra loro, *allontanare* e *levare*, di quanto non lo siano invece *allontanare* e *lagnare*.

*Figura 4.5 Il grafo ad albero elaborato da Cluto*



Cluto offre anche la possibilità di tenere conto di altri due parametri di valutazione, nell'analisi dei *clusters* ottenuti. Questi due parametri sono:

- l'*entropia* (indica il modo in cui i concetti delle diverse classi sono distribuiti in ogni *cluster*);
- la *purezza* (rappresenta il grado in cui un *cluster* contiene concetti derivanti da una sola classe di appartenenza).

(31) Si prenda ad esempio il *cluster 0* dell'esperimento condotto su 40 classi e con 105 *frames* di sottocategorizzazione; Cluto ne misura entropia e purezza come segue:
- l'entropia corrisponde a 0.000;
- la purezza invece si attesta ad 1.000.

I valori in effetti sono rispondenti, poiché entrambi i verbi che formano per intero il *cluster*, ovvero *impadronire* ed *appropriare*, appartengono alla medesima classe, quella etichettata come *Transfer of Possesion*.

Entropia e purezza vengono chiamate in Cluto, misure esterne di qualità, poiché servono a valutare la soluzione del *clustering* rispetto ad un *gold standard*. Entrambi i parametri sono compresi nei valori che vanno da 0 a 1: se l'entropia è pari a 0 il *cluster* è perfetto, poiché contiene concetti presi da una singola classe; per la purezza invece è l'inverso, più il valore tende verso l'1, più il *cluster* sarà accurato (esempio 31).

Evidentemente i risultati sono assolutamente perfettibili e anzi, le incongruenze presenti nella classificazione automatica, ci spingono a proseguire nella nostra ricerca, inserendo un ulteriore livello di analisi che vada a rifinire l'informazione in *output*: ci stiamo riferendo nello specifico all'introduzione delle preferenze di selezione relative ad ogni verbo e quindi ai tipi semantici partecipanti alle diverse costruzioni argomentali.

### 4.2.5 Utilità delle preferenze di selezione

Laddove in alcuni casi, come è stato in precedenza riscontrato, sintassi e preposizioni non sono sufficienti a catturare nella sua interezza la complessità della struttura argomentale dei verbi esaminati, occorrerà introdurre un nuovo parametro per cercare di ridurre ulteriormente le incongruenze residue; la scelta è ricaduta sulle preferenze di selezione.

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)).

(32) 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.

(33) 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

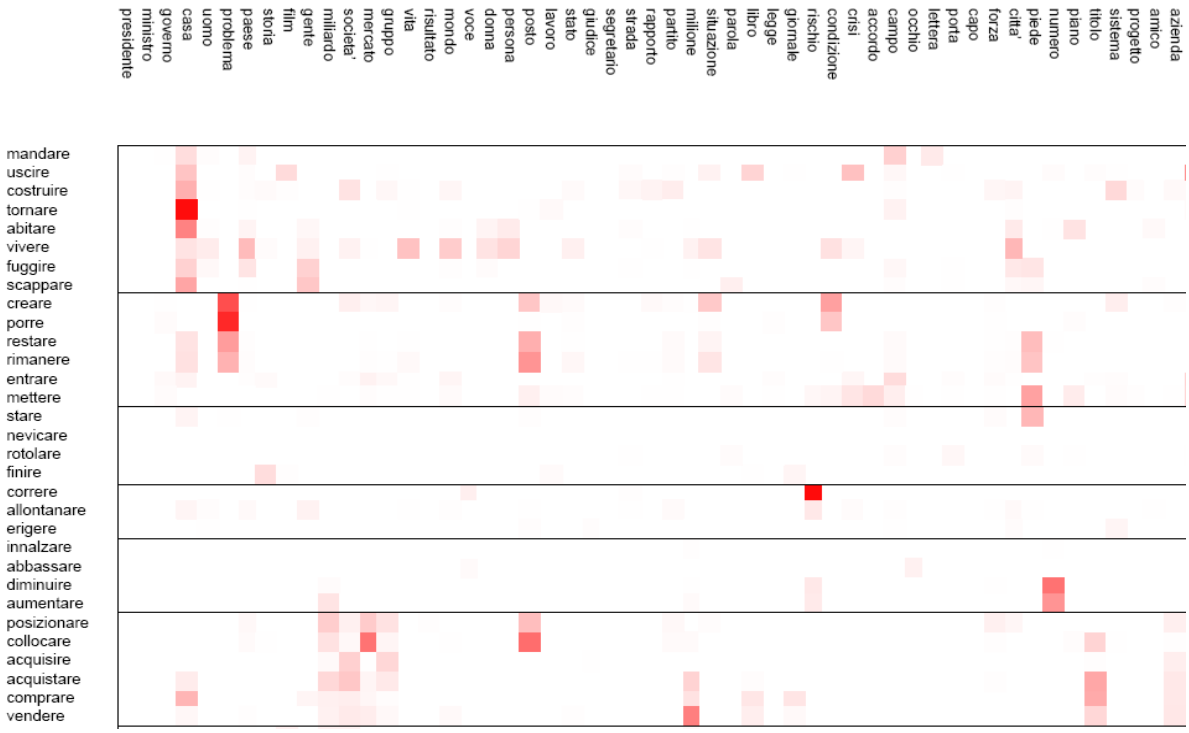
pertanto improprie in quel contesto. Il significato lessicale di un verbo può essere pertanto rappresentato a livello contestuale, ovvero tenendo conto degli incontri ripetuti che esso intrattiene con determinate parole in vari contesti d'uso.

A questo punto, si è trattato di dare una rappresentazione semantica delle proprietà di selezione dei *frames* dei verbi considerati. Tipicamente questa operazione viene effettuata assegnando ai *frames* dei *tipi semantici*, come CIBO, ANIMATO, LUOGO, tratti da ontologie come WordNet.

In questo lavoro, si è scelto invece di percorrere un'altra strada. In fin dei conti le preferenze di selezione di un predicato, dipendono dai nomi che possono comparire in un certo *frame*; pertanto abbiamo rappresentato ogni verbo 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 nella presente tesi, è stato costruito usando una matrice (Figura 4.6) che ha come righe i verbi considerati, e come colonne non più i *frames* sintattici di sottocategorizzazione (come illustrato negli esempi forniti nei paragrafi precedenti), 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.

*Figura 4.6 Estratto della matrice costruita con i fillers nominali*




I valori della matrice esprimono la forza di associazione tra il verbo ed un dato nome, misurata attraverso la Simple Log Likelihood.

Le misure di associazione (Mutual Information, Log Likelihood etc.) possono essere usate per i *fillers* più significativi di un dato verbo. Si è scelto di utilizzare la Log Likelihood poiché, a differenza della Mutual Information che tende a favorire i *fillers* maggiormente idiosincratici per ciascun verbo (ovvero quelli meno frequenti e con più argomenti), essa permette di ridurre questo effetto, bilanciando la specificità del *filler* con la sua frequenza.

La formula della Simple Log Likelihood è la seguente:

$$\text{Simple LL (v, nome)} = 2^{*}\left(O * \log\frac{O}{E}\right) - (O - E)$$

O è la frequenza osseravata della coppia verbo-nome del *corpus*

$$O = fq\ (v\ nome)$$

E è la frequenza attestata della coppia verbo-nome, in caso di mancanza di associazione

$$E = \frac{fq\ (v) * fq\ (nome)}{N}$$

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 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; l'analisi degli esperimenti svolti su 40 classi verbali, servirà proprio a motivare il perché di tale preferenza.

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 (come già evidenziato negli esempi presentati nel paragrafo 4.2.1). 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.

(34) esigere<br>necessitare<br>richiedere } *cluster* 10

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* impegno<br>Lo studio *necessita* impegno<br>Lo studio *richiede* impegno
b. Il paziente *esige* cure efficaci<br>Il paziente *necessita* cure efficaci<br>Il paziente *richiede* cure efficaci

c. La situazione *esige* l'intervento dello Stato
La situazione *necessita* l'intervento dello Stato
La situazione *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, ci 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*
*mormorare*
} *cluster* 0

preferenze di selezione: *orecchio, parola, voce, preghiera*

*gridare*
*urlare*
} *cluster* 8

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* una preghiera
     Il fedele ha *sussurrato* una preghiera
     Il fedele ha *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* di rabbia
     Paolo ha *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)*).

(35) *marciare*, *pedalare*, *saltellare*, *camminare* } *cluster* 14

preferenze di selezione: ritmo, salita, strada, passo

a. Paolo *marcia* in salita
   Paolo *pedala* in salita
   Paolo *cammina* in salita

*percorrere*, *attraversare*, *traversare* } *cluster* 7

preferenze di selezione: oceano, fiume, frontiera, fase (senso figurato)

b. Il battello *percorre* il fiume
   Il battello *attraversa* il fiume
   Il battello *traversa* il fiume

c. L'adolescente *percorre* una fase travagliata della vita
   L'adolescente *attraversa* una frase travagliata della vita
   L'adolescente *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 36).

(36) *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 (36) 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*).

# 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.

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.

Particolare attenzione è stata data all'analisi delle strutture argomentali dei verbi, che sono state scomposte per individuare al loro interno gli elementi minimi rilevanti:

- alternanze argomentali;
- *frames* sintattici;
- preferenze di selezione;
- ruoli semantici.

Nella tesi presentata, sono stati citati alcuni dei lavori più significativi in questo ambito disciplinare, come quello di Korhonen (2002) che si è occupata dell'estrazione automatica dei *frames* di sottocategorizzazione dai *corpora*, o ancora quello di Schulte im Walde (2006) sull'induzione di classi verbali su base distribuzionale, ovvero sull'indagine di proprietà semantiche condivise da classi di predicati che si comportano similmente a livello distribuzionale, sull'analisi delle alternanze argomentali dei verbi selezionati, e sullo studio della validità dell'ipotesi sintattico-semantica.

Gli esperimenti proposti nella tesi utilizzano, come base di partenza, i dati estratti dal *corpus* di Repubblica (in cui sono contenuti circa 326 milioni di *tokens*); su di essi è stata poi condotta un'operazione prettamente linguistica, ovvero l'analisi sintattica a dipendenze, detta anche *parser*, tramite il programma Malt Parser. Quest'ultimo rappresenta la struttura sintattica di una frase, attraverso relazioni binarie asimmetriche di dipendenza tra termini lessicali.

Successivamente si è passati all'estrazione automatica dei *frames* sintattici dal suddetto *corpus* ed alla verifica della loro significatività statistica rispetto ad ogni singolo verbo. Per compiere questo passaggio, dalle dipendenze ottenute dal *parser* ai *frames* di sottocategorizzazione, è stato necessario costruire una matrice basata sul logaritmo della frequenza dei diversi *frames* per i vari verbi.

L'ipotesi linguistica che soggiace a questa metodologia di analisi, è che verbi semanticamente simili di solito presentano simili proprietà di sottocategorizzazione. Il problema che soggiace all'interno della tesi è valutare fino a che punto tali proprietà di selezione sintattica dei verbi sono riconducibili a particolari dimensioni del loro significato.

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.

Grazie alle statistiche distribuzionali ricavate dai *frames* di sottocategorizzazione estratti dal *corpus*, è possibile indurre in modo automatico, classi di verbi distribuzionalmente simili (Schulte im Walde (2006)). I verbi vengono così accorpati in gruppi, detti anche *clusters*, in virtù del fatto che tendono a ricorrere con gli stessi *frames* sintattici. Ogni verbo viene pertanto rappresentato da un vettore di numeri, le cui dimensioni contengono la misura di associazione tra il verbo stesso e i vari *frames* a cui esso si associa.

Il programma di *clustering* utilizzato, nello specifico Cluto, organizza i verbi in insiemi sulla base della similarità vettoriale; i *clusters* rappresentano allora delle classi di verbi simili dal punto di vista delle proprietà di sottocategorizzazione mostrate. I verbi che appartengono allo stesso *cluster*, mostrano vari gradi di somiglianza anche a livello semantico.

Questo 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.

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. Questo fenomeno è comprovato dal fatto che spesso la nostra conoscenza sul significato di una parola, deriva proprio dall'osservazione che facciamo del suo uso linguistico.

Sono stati sviluppati, a tal proposito, una grande quantità di *corpora* testuali dai quali è possibile estrarre informazioni sui contesti linguistici più prototipici di una parola. In essi ogni parola viene considerata come un vettore distribuzionale, in cui le dimensioni del vettore stesso registrano l'associazione statistica tra la parola stessa ed il contesto linguistico che la ospita, ovvero le proprietà combinatorie di quella parola nei testi.

Secondo l'*ipotesi distribuzionale*, la similarità vettoriale è interpretabile appunto come similarità semantica tra le parole.

Nella tesi, il contesto linguistico dei 200 verbi italiani selezionati è stato ottenuto ricavando le preferenze di selezione di questi ultimi, ovvero i nomi a cui essi più spesso si accompagnano. In questa fase, è stato escluso quindi il ricorso al comportamento sintattico del verbo, valutando invece solo i suoi tratti semantici significativi. I risultati del *clustering* ne hanno beneficiato, mostrando un sensibile miglioramento; difatti, da un confronto tra i *clusters* così ottenuti con le 40 classi di una classificazione semantica manuale sviluppata *a priori* con funzione di *gold standard*, emergono numerose e significative corrispondenze e sovrapposizioni.

Il limite linguistico di questo tipo di esperimenti, sta nella scelta delle proprietà verbali descrittive ritenute maggiormente pertinenti, siano esse sintattiche (*frames* di sottocategorizzazione) o semantiche (preferenze di selezione). Il significato dei verbi comprende infatti, sia le caratteristiche generali per le rispettive classi, sia quelle specifiche che distinguono ogni verbo dagli altri membri appartenenti alla stessa classe. Da un punto di vista teorico la distinzione è chiara, ma a livello pratico la scelta delle proprietà verbali dipende dallo schema di definizione delle classi verbali, e questo varia necessariamente in funzione del tipo di coerenza semantica catturato dalla classe.

Finora molti studi si sono concentrati sulla sola sinonimia, ma soprattutto negli esperimenti su larga scala sono emerse altre relazioni semantiche significative all'interno delle varie classi verbali.

In ogni caso la combinazione delle proprietà selezionate all'interno della tesi negli esperimenti proposti, sembra costituire un punto di partenza promettente per la descrizione dei verbi italiani. Ovviamente per aggiungere completezza alle informazioni ottenute automaticamente dalle operazioni di *clustering*, è stato 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, la definizione manuale di classi semantiche per i verbi italiani potrebbe essere ulteriormente estesa, al fine di includere un numero maggiore ed una più vasta gamma di classi verbali. L'estensione della classificazione manuale potrebbe essere utile come *gold standard* per futuri esperimenti di *clustering*, ma anche come risorsa per applicazioni di *Natural Language Processing*. Si potrebbe poi pensare all'utilizzo di un algoritmo di *clustering* di tipo *soft*, invece che *hard* come è stato fatto nell'ambito di questa tesi, poiché esso ha la capacità di assegnare i verbi a più *clusters*, introducendo così un metodo di indagine per il fenomeno dell'ambiguità verbale. Rispetto alle tecniche di *hard clustering*, questo tipo di algoritmi sarebbero utili per individuare nuove componenti del significato verbale e nuove classi semanticamente correlate.

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.

In ogni caso restano ancora dei quesiti linguistici insoluti, come ad esempio in che modo la competenza semantica di un verbo dipende dal fatto che lo osserviamo in determinati contesti linguistici e non in altri? Ovvero, in che misura le distribuzioni d'uso dei verbi sono assorbite dai parlanti nelle loro rappresentazioni semantiche? E ancora, in che termini le proprietà semantiche di intere classi di espressioni linguistiche sono correlate alle loro proprietà sintattiche?

# Bibliografia

Abe, N., Li, H. (1998). 'Generalizing Case Frames Using a Thesaurus and the MDL Principle' in *Computational Linguistics 24(2)*, 217-244

Abney, S. (1991). 'Parsing by chunks' in Berwick, R., Abney, S. e Tenny, C. (eds.), *Principle-Based Parsing*, Boston, MA, Kluwer Academic Publishers

Abney, S. (1997). 'The SCOL manual version 0.1b' disponibile su *http://www.research.att.com/~abney/*

Artale, A., Magnini, B., Strapparava, C. (1997). 'WordNet for Italian and its Use for Lexical Discrimination' in *Proceedingsof the 5th Congress of the Italian Association for Artificial Intelligence* (AI*IA '97)Rome

Atkins, B.T.S. e Levin, B. (1991). 'Admitting Impediments' in Zernik, U. (ed.), *Exploiting On-line Resources to Build a Lexicon*, Hillsdale, NJ, Lawrence Erlbaum Associates, 233-262

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

Blumenthal, P. e Rovere, G. (1998). *Wörterbuch der italienischen Verben*, Stuttgart-Düsseldorf-Leipzig, Klett

Bresnan, J. (1982). *The Mental Representation of Grammatical Relations*, Cambridge, MA:MIT Press

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

Brown, P.F. et al. (1992). 'Class-based n-gram models of natural language' in *Computational Linguistics, 18(4)*, 467-479

Centineo, G. (1986). 'A Lexical Theory of Auxiliary Selection in Italian' in *Davis Working Papers in Linguistics 1*, University of California, Davis, CA, 1-35

Centineo, G. (1996). 'A Lexical Theory of Auxiliary Selection in Italian' in *Probus 8*: 223-271

Chaffin, R., Fellbaum, C., Jenei, J. (1994). *On the Organization of Verbs in the Mental Lexicon*, Trenton, NJ, The College of New Jersey

Charles, W. (2000). 'Contextual Correlates of Meaning' in *Applied Psicholinguistics, 21*, 505-524

Chomsky, N. (1965). *Aspects of the Theory of Syntax*, Cambridge, MA, the MIT Press
(1981). *Lectures on Government and Binding*, Dordrecht, Foris

Clark, S. e Weir, D. (2002). 'Class-based probability estimation using a semantic hierarchy' in *Computational Linguistics, 28(2)*, 187-206

Dang, T.H. et al. (1998). 'Investigating Regular Senses Extensions based on Intersective Levin Classes' in *Proceedings of COLING/ACL*, Montreal, Canada, 293-299

Davidson, D. (1967). 'Truth and Meaning' in *Synthese, vol.17*, 304-323

Dik, S.C. (1978). *Stepwise Lexical Decomposition*, Lisse, Peter de Ridder
(1979). *Functional Grammar*, Dordrecht, Foris
(1989). *The Theory of Functional Grammar. The Structure of the Clause*, a cura di Kees Hengeveld, Berlin, Mouton de Gruyter

Dorr, B.J. (1997). 'Large-Scale Dictionary Construction for Foreign Languages Tutoring and Interlingual Machine Translation' in *Journal of Machine Translation, 12:4*, 271-325

Dorr, B.J. e Jones, D. (1996). 'Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues' in *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, Copenhagen, Denmark, 322-327

Dowty, D.R. (1979). *Word Meaning and Montague Grammar*, Dordrecht, Reidel

Fellbaum, C. (1998). *WordNet: an Electronic Lexical Database*, MIT Press
(1999). 'The organization of Verbs and Verb Concepts in a semantic Net' in Saint-Dizier, P. (ed.), *Predicative Forms in Natural Language and in Lexical knowledge Bases*, Netherlands, Kluwer Academic Publishers, 93-110

Fellbaum, C. e Miller, G. (1990). 'Folk psychology or semantic entailment? A reply to Rips and Conrad' in *The Psychological Review, 97*, 565-570

Fillmore, C.J. (1968). 'The Case for Case' in Bach, E. e Harms, R.T. (eds.), *Universals in Linguistic Theory*, New York, Holt, Rinehart and Winston, 1-88
(1977). 'Scenes and Frame Semantics, Linguistic Structures Processing' in Zampolli, A. (ed.), *Fundamental Studies in Computer Science, No. 59*, North Holland Publishing, 55-88
(1982). 'Frame Semantics' in *Linguistics in the Morning Calm*, Seoul: Hanshin Publishing Co., 111-137

Firth, J.R. (1957). *Papers in Linguistics 1934-1951*, London: Oxford University Press

Forgy, E.W. (1965). 'Cluster analysis of multivariate data: efficiency vs interpretability of classifications' in *Biometrics 21*, 768-769

Grimshaw, J. (1990). *Argument Structure*, Cambridge, MA, MIT Press

Gross, G. (1986). 'Syntaxe des Noms' in *Langue française69*, Paris, 128

Gross, M. (1975). *Méthodes en Syntaxe*, Paris, Hermann

Gruber, J. (1965). *Studies in lexical relations*, Doctoral Dissertation, Cambridge, MA: MIT Press

Hale, K. e Keyser, S. (1987). 'A view from the middle' in *Lexicon Project Working Papers 10*, 1-36

Harley, T. (2008). *The Psychology of Language: From data to theory*, Hove: Psychology Press

Harris, Z. (1968). *Mathematical Structures of Language*, Interscience Publishers
(1970). 'Distributional structure' in *Papers in structural and transformational Linguistics*, 775-794

Helbig, G. e Schenkel, W. (1969). *Wörterbuch zur Valenz und Distribution deutscher Verben*, Leipzig, VEB Bibliographisches Institut

Hendler, J. (2001). 'Agents and the Semantic Web' in *IEEE Intelligent Systems 16(2)*, 30-37

Hopper, P.J. e Thompson, S.A. (1980). 'Transitivity in Grammar and Discourse' in *Language 56, 4*, 251-299

Jackendoff, R. (1972). *Semantics and Cognition*, Cambridge, MA: MIT Press
(1997). *The architecture of language Faculty*, Cambridge, MA: MIT Press

Jezek, E. (2003). *Classi di verbi tra semantica e sintassi*, Pisa, ETS edizioni

Joanis, E. (2003). *Automatic Verb Classification Using a General Feature Space*, Master of Science Graduate, Department of Computer Science, University of Toronto

Johnson, C.R. et al. (2002). *FrameNet : Theory and Practice*, Technical Report-02009, Berkeley, CA, International Computer Science Institute

Katz, J.J. (1966). *The Phylosophy of Language*, New York, Harper and Row

Katz, J.J. e Fodor, J. (1963). 'The structure of a semantic Theory' in *Language,39*, 178-210

Karypis, G. (2003). *CLUTO 2.1.1. A Clustering Toolkit Technical Report*, Department of Computer Science, University of Minnesota

Kintsch, W. (2007). 'Meaning in Context' in Landauer, T.K., McNamara, D., Dennis, S. e Kintsch, W. (eds.), *Handbook of Latent Semantic Analysis*, Mahwah, NJ: ERlbaum, 89-105

Klavans, J. e Kan, M.Y. (1998). 'Role of verbs in document analysis' in *COLING ACL*, 680-686

Korhonen, A. (2002). *Subcategorisation acquisition*, PhD Thesis published as Technical Report UCAM-CL-TR-530. Computer Laboratory, University of Cambridge

Korhonen, A. e Briscoe, T. (2004). 'Extended Lexical-Semantic Classification of English Verbs' in *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*, Boston, MA

Lakoff, G. (1970). 'Global Rules' in *Language,46*, 627-639

Lakoff, G. e Johnson, M. (1980). *Metaphors we live by*, Chicago: University of Chicago Press
(1999). *Philosophy in the flesh: the embodied mind and its challenge to western thought*, New York: Basic Books

Lapata, M. (1999). 'Acquiring lexical generalizations from corpora: A case study for diathesis alternations' in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, MD, 397-404

Lapata, M. e Brew, C. (1999). 'Using subcategorization to resolve verb class ambiguity' in *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, 266-274

Lapata, M. et al. (2001). 'Evaluating smoothing algorithms against plausibility judgments' in *Proceedings of the 39th Annual Meeting of the Association for Computational Languistics*, Toulouse, France, 346-353

Lenci, A. (in print). 'Spazi di parole: metafore e rappresentazioni semantiche' in *Paradigmi*, 2009

Lenci, A. et al. (2000). 'SIMPLE: A General Framework for the Development of Multilingual Lexicons' in *International Journal of Lexicography, XIII(4)*, Oxford University Press, 249-263

Lenci, A. e Calzolari, N. (2004). 'Linguistica computazionale. Strumenti e risorse per il trattamento automatico della lingua' in *Mondo Digitale, vol.3, num.2*, 56-69

Levin, B. (1993). *English Verb Classes and Alternations: a Preliminary Investigation*, Chicago, IL, University of Chicago Press

Levin, B. e Rappaport Hovav, M. (1992). 'The lexical semantics of Verbs of motion: the perspective from unaccusativity' in Roca, I.M. (ed.), *Thematic Structure: Its Role in Grammar*, Berlin, Foris, 247-269

(1995). *Unaccusativity: at the syntax-lexical interface*, Cambridge, MA: MIT Press

(2005). *Argument Realization*, Cambridge University Press

Li, H. e Abe, N. (1998). 'Generalizing case frames using a thesaurus and the MDL principle' in *Computational Linguistics, 24(2)*, 217-244

Lowe, W. (2001). 'Towards a Theory of Semantic Space' in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society, Philadelphia, PA*, 576-581

Lyons, J. (1977). *Semantics*, Cambridge, Cambridge University Press

Manning, C., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press

McCarthy, D. (2000). 'Using Semantic Preferences to identify Verbal Participation in Role Switching Alternations' in *Proceeddings of the first Conference of the North American Chapter of the Association for Computational Linguistics, (NAACL)*, Seattle, WA

McCarthy, D. e Korhonen, A. (1998). 'Detecting verbal participation in diathesis alternation' in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguists, vol.2*, Montreal, 1493-1495

Merlo, P. e Stevenson, S. (2001). 'Automatic Verb Classification based on Statistical Distribution of Argument Structure' in *Computational Linguistics, 27:3*, 373-408

Miller, G. et al. (1990). 'WordNet: an on-line lexical Database' in *International Journal of Lexicography, 3(4)*, 235-*244*

(1993). 'Introduction to WordNet: an on-line Lexical Database. Description of WordNet' disponibile su *http://clarity.princeton.edu:80/~wn/*

Miller, G. e Charles, W. (1991). 'Contextual Correlates of Semantic Similarity' in *Language and Cognitive Processes, 6(1)*, 1-28

Miller, G. e Johnson-Laird, P.N. (1976). *Language and Perception*, Cambridge, MA, Harvard University

Nivre, J., Hall, J. e Nilsson, J. (2006). 'MaltParser: a Data-Driven Parser-Generator for Dependency Parsing' in *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC2006)*, Genoa, Italy, 2216-2219

Olbertz, H. et al. (eds.) (1998). *The Structure of the Lexicon in Functional Grammar*, Amsterdam, Benjamins

Padò, S.G. e Lapata, M. (2007). 'Dependency-Based Construction of Semantic Space Models' in *Computational Linguistics, XIII/2*, pp.161-199

Palmer, M. (2000). 'Consistent Criteria for sense distinctions' in *Special Issue of Computers and the Humanities, SENSEVAL98:Evaluating Word Sense Disambiguation Systems, 34(1-2)*, 217-222

Pereira, F. et al. (1993). 'Distributional clustering of English words' in *Proceedings of 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH, 187-190

Pinker, S. (1989). *Learnability and Cognition: the acquisition of Argument Structure*, Cambridge, MA: MIT Press

Pulman, S.G. (1983). *Word Meaning and Belief*, London: Croom Helm Ltd, and New Jersey: Ablex Publishing Corp.

Pustejovsky, J. (1995). *The Generative Lexicon*, Cambridge, MA, The MIT Press

(2001). 'Type Structure and Logic of Concepts' in Bouillon, P. e Busa, F. (eds.), *The Language of Word Meaning*, Cambridge, Cambridge University Press, 109-145

Quillian, M.R. (1967). 'Word Concepts: a Theory and Simulation of Some Basic Semantic Capabilities', in *Behavioural Science 12*, 410-430

(1968). 'Semantic Memory' in Minsky, M. (ed.), *Semantic Information Processing*, Cambridge, MA: MIT Press, 227-270

Rappaport Hovav, M., Laughren, M. e Levin, B. (1987). 'Levels of Lexical Representation' in *Lexicon Project Working Papers 20*, Cambridge, MA, Center for Cognitive Science, MIT

Resnik, P.S. (1993). 'Selection and Information: A Class-based Approach to Lexical Relationships', Ph.D. thesis, University of Pennsylvania, Philadelphia

Roventini, A. et al. (2003). 'ItalWordNet: Building a Large Semantic database for the Automatic treatment of Italian' in *Computational Linguistics*, in *Pisa, Special Issue, XVIII-XIX, IEPI, Tomo II*, Pisa-Roma, 745-791

Ruimy, N. et al. (2003). 'A Computational Semantic Lexicon of Italian: SIMPLE' in Zampolli, A., Calzolari, N. e Cignoni, L. (eds.), *Computational Linguistics*, in *Pisa, Special Issue, XVIII-XIX, IEPI, Tomo II*, Pisa-Roma, 821-864

Sabatini, F. e Coletti, V. (1997). *Dizionario della Lingua Italiana*, Firenze, Giunti Editore

Sahlgren, M. (2006). *The Word-Space Model: using distributional analysis to represent syntagmatuc and paradigmatic relations between words in high-dimensional vector spaces*, PhD Dissertation, Department of Linguistics, Stockholm University

Sapir, E. (1921). *Language: an Introduction to the Study of Speech*, New York: Hartcourt, Brace

Schulte im Walde, S. (2002). 'A Subcategorization Lexicon for German Verbs induced from a Lexicalized PCFG' in *Proceedings of the 3rd Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain

(2003). 'Experiments on the Choice of Features for Learning Verb Classes' in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary

(2004). 'Induction of Semantic Classes for German Verbs' in Langer, S. e Schnorbusch, D. (eds.), *Semantik im Lexicon*, Tübingen, Gunter Narr Verlag

(2006). 'Experiments on the Automatic Induction of German Semantic Verb Classes' in *Computational Linguistics, 32(2)*, 159-194

Schütze, H. (1993). 'Word Space' in Hanson, S.J., Cowan, J.D. e Giles, C.L. (eds.), *Advances in Neural Information Processing System 5*, San Mateo, CA, Morgan Kaufmann Publishers, 895-902

Smith, D.E. (2000). *Grammar*, Brandon, Manitoba, Brandon University Press

Stede, M. (1999). *Lexical Semantics and Knowledge Representation in multilingual Text Generation*, Boston/Dordrecht/London: Kluwer Academic Publishers

Tenny, C.L. (1987). *Grammaticalizing Aspect and Affectedness*, Doctoral Dissertation, MIT

(1992). 'The Aspectual Interface Hypothesis' in Sag, I.A. e Szabolesi, A. (eds.), *Lexical Matters*, Stanford, California: CSLI Publications, 1-27

(1994). *Aspectual roles and the syntax-semantics interface*, Dordrecht, Kluwer

Tesnière, L. (1959). *Eléménts de syntaxe structurale*, Klincksieck, Paris

van Valin, R.D. (1990). 'Semantic parameters of split intransitivity' in *Language 66*, 221-260

van Valin, R.D. e Foley, W.A. (1984). *Functional syntax and Universal grammar*, Cambridge, England: Cambridge University Press

van Valin, R.D. e La Polla, R.J. (1997). *Syntax. Structure, Meaning and Function*, Cambridge, Cambridge University Press

Vendler, Z. (1967). 'Verbs and Times' in *Linguistics in Philosophy*, Ithaca (N.Y.), Cornell University Press, 97-121

Vossen, P. (1995). *Grammatical and Conceptual individuation in the Lexicon*, PhD Thesis, Universiteit van Amsterdam

Widdows, D. (2004). *Geometry and Meaning*, Stanford California, CSLI Books

Zaenen, A. (1993). 'Unaccusativity in Dutch: Integrating Syntax and Lexical Semantics' in Pustejovsky, J. (ed.), *Semantics and the Lexicon*, Dordrecht, Kluwer, 129-161

# Appendice

*Tabella 4.1 Classificazione semantica dei 200 verbi italiani ripartiti in 40 classi*

| ID | Numero Classe | Classe | Sigla | Verbo |
|---|---|---|---|---|
| 1 | 1 | Aspect | AS# | cominciare |
| 1 | 1 | Aspect | AS# | continuare |
| 1 | 1 | Aspect | AS# | finire |
| 1 | 1 | Aspect | AS# | iniziare |
| 1 | 1 | Aspect | AS# | smettere |
| 1 | 1 | Aspect | AS# | terminare |
| 2 | 2 | Propositional Attitude | PROP# | considerare |
| 2 | 2 | Propositional Attitude | PROP# | credere |
| 2 | 2 | Propositional Attitude | PROP# | dubitare |
| 2 | 2 | Propositional Attitude | PROP# | pensare |
| 2 | 2 | Propositional Attitude | PROP# | ritenere |
| 2 | 2 | Propositional Attitude | PROP# | sapere |
| 3 | 2.1 | Desire (wish) | DEWI# | desiderare |
| 3 | 2.1 | Desire (wish) | DEWI# | sperare |
| 3 | 2.1 | Desire (wish) | DEWI# | volere |
| 4 | 2.2 | Desire (need) | DENE# | esigere |
| 4 | 2.2 | Desire (need) | DENE# | necessitare |
| 4 | 2.2 | Desire (need) | DENE# | richiedere |
| 5 | 3 | Transfer of Possession (obtaining) | POTA# | acquisire |
| 5 | 3 | Transfer of Possession (obtaining) | POTA# | acquistare |
| 5 | 3 | Transfer of Possession (obtaining) | POTA# | appropriare(si) |
| 5 | 3 | Transfer of Possession (obtaining) | POTA# | comprare |
| 5 | 3 | Transfer of Possession (obtaining) | POTA# | conseguire |
| 5 | 3 | Transfer of Possession (obtaining) | POTA# | impadronire(si) |
| 5 | 3 | Transfer of Possession (obtaining) | POTA# | ottenere |
| 6 | 3.1 | Transfer of Possession (giving-gift) | POGI# | distribuire |
| 6 | 3.1 | Transfer of Possession (giving-gift) | POGI# | fornire |
| 6 | 3.1 | Transfer of Possession (giving-gift) | POGI# | offrire |
| 6 | 3.1 | Transfer of Possession (giving-gift) | POGI# | regalare |
| 6 | 3.1 | Transfer of Possession (giving-gift) | POGI# | restituire |
| 6 | 3.1 | Transfer of Possession (giving-gift) | POGI# | vendere |
| 7 | 3.2 | Transfer of Possession (giving-supply) | POSU# | consegnare |
| 7 | 3.2 | Transfer of Possession (giving-supply) | POSU# | inviare |
| 7 | 3.2 | Transfer of Possession (giving-supply) | POSU# | mandare |
| 7 | 3.2 | Transfer of Possession (giving-supply) | POSU# | recapitare |
| 7 | 3.2 | Transfer of Possession (giving-supply) | POSU# | spedire |
| 8 | 4.1 | Motion (manner) | MOM# | camminare |
| 8 | 4.1 | Motion (manner) | MOM# | correre |
| 8 | 4.1 | Motion (manner) | MOM# | marciare |
| 8 | 4.1 | Motion (manner) | MOM# | passeggiare |
| 8 | 4.1 | Motion (manner) | MOM# | rotolare |
| 8 | 4.1 | Motion (manner) | MOM# | saltellare |
| 8 | 4.1 | Motion (manner) | MOM# | strisciare |
| 9 | 4.1.1 | Motion (manner-vehicle) | MOMV# | navigare |
| 9 | 4.1.1 | Motion (manner-vehicle) | MOMV# | pattinare |

| 9 | 4.1.1 | Motion (manner-vehicle) | MOMV# | pedalare |
|---|---|---|---|---|
| 9 | 4.1.1 | Motion (manner-vehicle) | MOMV# | volare |
| 10 | 4.2 | Motion (directed) | MOD# | allontanare(si) |
| 10 | 4.2 | Motion (directed) | MOD# | arrivare |
| 10 | 4.2 | Motion (directed) | MOD# | entrare |
| 10 | 4.2 | Motion (directed) | MOD# | fuggire |
| 10 | 4.2 | Motion (directed) | MOD# | giungere |
| 10 | 4.2 | Motion (directed) | MOD# | scappare |
| 10 | 4.2 | Motion (directed) | MOD# | tornare |
| 10 | 4.2 | Motion (directed) | MOD# | uscire |
| 11 | 4.3 | Motion (cross) | MOC# | attraversare |
| 11 | 4.3 | Motion (cross) | MOC# | percorrere |
| 11 | 4.3 | Motion (cross) | MOC# | traversare |
| 12 | 5.1 | Emotion (objection) | EMO# | divertire(si) |
| 12 | 5.1 | Emotion (objection) | EMO# | impaurire(si) |
| 12 | 5.1 | Emotion (objection) | EMO# | infastidire(si) |
| 12 | 5.1 | Emotion (objection) | EMO# | irritare(si) |
| 12 | 5.1 | Emotion (objection) | EMO# | spaventare(si) |
| 13 | 5.2 | Emotion (origin) | EMOR# | amare |
| 13 | 5.2 | Emotion (origin) | EMOR# | entusiasmare |
| 13 | 5.2 | Emotion (origin) | EMOR# | esultare |
| 13 | 5.2 | Emotion (origin) | EMOR# | gioire |
| 13 | 5.2 | Emotion (origin) | EMOR# | odiare |
| 13 | 5.2 | Emotion (origin) | EMOR# | piacere |
| 13 | 5.2 | Emotion (origin) | EMOR# | soffrire |
| 14 | 6 | Facial Expression | FEX# | piangere |
| 14 | 6 | Facial Expression | FEX# | ridere |
| 14 | 6 | Facial Expression | FEX# | sbadigliare |
| 14 | 6 | Facial Expression | FEX# | sorridere |
| 14 | 6 | Facial Expression | FEX# | tossire |
| 15 | 7 | Perception | PER# | annusare |
| 15 | 7 | Perception | PER# | ascoltare |
| 15 | 7 | Perception | PER# | assaggiare |
| 15 | 7 | Perception | PER# | fiutare |
| 15 | 7 | Perception | PER# | guardare |
| 15 | 7 | Perception | PER# | scorgere |
| 15 | 7 | Perception | PER# | scrutare |
| 15 | 7 | Perception | PER# | sentire |
| 15 | 7 | Perception | PER# | vedere |
| 16 | 8 | Manner of Articulation | MAR# | bisbigliare |
| 16 | 8 | Manner of Articulation | MAR# | gridare |
| 16 | 8 | Manner of Articulation | MAR# | mormorare |
| 16 | 8 | Manner of Articulation | MAR# | sussurrare |
| 16 | 8 | Manner of Articulation | MAR# | urlare |
| 17 | 9 | Moaning | MOA# | deplorare |
| 17 | 9 | Moaning | MOA# | deprecare |
| 17 | 9 | Moaning | MOA# | lagnare(si) |
| 17 | 9 | Moaning | MOA# | lamentare(si) |
| 17 | 9 | Moaning | MOA# | rammaricare(si) |
| 18 | 10 | Communication | COM# | chiacchierare |
| 18 | 10 | Communication | COM# | conversare |
| 18 | 10 | Communication | COM# | dire |
| 18 | 10 | Communication | COM# | parlare |

| | | | | |
|---|---|---|---|---|
| 18 | 10 | Communication | COM# | raccontare |
| 18 | 10 | Communication | COM# | riferire |
| 19 | 10.1 | Communication (announcement) | COA# | affermare |
| 19 | 10.1 | Communication (announcement) | COA# | annunciare |
| 19 | 10.1 | Communication (announcement) | COA# | comunicare |
| 19 | 10.1 | Communication (announcement) | COA# | dichiarare |
| 19 | 10.1 | Communication (announcement) | COA# | informare |
| 20 | 10.2 | Communication (constitution) | COC# | decidere |
| 20 | 10.2 | Communication (constitution) | COC# | ordinare |
| 20 | 10.2 | Communication (constitution) | COC# | stabilire |
| 20 | 10.2 | Communication (constitution) | COC# | vietare |
| 21 | 10.3 | Communication (promise) | COP# | assicurare |
| 21 | 10.3 | Communication (promise) | COP# | garantire |
| 21 | 10.3 | Communication (promise) | COP# | impegnare(si) |
| 21 | 10.3 | Communication (promise) | COP# | promettere |
| 22 | 11 | Observation | OBS# | constatare |
| 22 | 11 | Observation | OBS# | evidenziare |
| 22 | 11 | Observation | OBS# | notare |
| 22 | 11 | Observation | OBS# | osservare |
| 22 | 11 | Observation | OBS# | registrare |
| 22 | 11 | Observation | OBS# | segnalare |
| 23 | 12 | Description | DES# | caratterizzare |
| 23 | 12 | Description | DES# | descrivere |
| 23 | 12 | Description | DES# | illustrare |
| 23 | 12 | Description | DES# | spiegare |
| 23 | 12 | Description | DES# | tratteggiare |
| 24 | 13 | Presentation | PRE# | manifestare |
| 24 | 13 | Presentation | PRE# | mostrare |
| 24 | 13 | Presentation | PRE# | presentare |
| 25 | 14 | Speculation | SPE# | fantasticare |
| 25 | 14 | Speculation | SPE# | immaginare |
| 25 | 14 | Speculation | SPE# | presumere |
| 25 | 14 | Speculation | SPE# | supporre |
| 26 | 15 | Insistence | INS# | insistere |
| 26 | 15 | Insistence | INS# | perseverare |
| 26 | 15 | Insistence | INS# | persistere |
| 27 | 16 | Learning | TEA# | apprendere |
| 27 | 16 | Learning | TEA# | imparare |
| 27 | 16 | Learning | TEA# | leggere |
| 27 | 16 | Learning | TEA# | studiare |
| 28 | 17.1 | Position (bring into position) | PSBR# | collocare |
| 28 | 17.1 | Position (bring into position) | PSBR# | levare |
| 28 | 17.1 | Position (bring into position) | PSBR# | mettere |
| 28 | 17.1 | Position (bring into position) | PSBR# | porre |
| 28 | 17.1 | Position (bring into position) | PSBR# | posizionare |
| 28 | 17.1 | Position (bring into position) | PSBR# | togliere |
| 29 | 17.2 | Position (be in position) | PSBE# | abitare |
| 29 | 17.2 | Position (be in position) | PSBE# | restare |
| 29 | 17.2 | Position (be in position) | PSBE# | rimanere |
| 29 | 17.2 | Position (be in position) | PSBE# | stare |
| 29 | 17.2 | Position (be in position) | PSBE# | vivere |
| 30 | 18 | Production | PROD# | costruire |
| 30 | 18 | Production | PROD# | creare |

| | | | | |
|---|---|---|---|---|
| 30 | 18 | Production | PROD# | erigere |
| 30 | 18 | Production | PROD# | produrre |
| 30 | 18 | Production | PROD# | realizzare |
| 31 | 19 | Renovation | RENO# | aggiustare |
| 31 | 19 | Renovation | RENO# | ammodernare |
| 31 | 19 | Renovation | RENO# | rinnovare |
| 31 | 19 | Renovation | RENO# | riparare |
| 32 | 20 | Support | SUP# | aiutare |
| 32 | 20 | Support | SUP# | incentivare |
| 32 | 20 | Support | SUP# | salvare |
| 32 | 20 | Support | SUP# | soccorrere |
| 33 | 21 | Quantum Change | QUCA# | abbassare |
| 33 | 21 | Quantum Change | QUCA# | accorciare |
| 33 | 21 | Quantum Change | QUCA# | allungare |
| 33 | 21 | Quantum Change | QUCA# | aumentare |
| 33 | 21 | Quantum Change | QUCA# | diminuire |
| 33 | 21 | Quantum Change | QUCA# | innalzare |
| 34 | 22 | Opening | OPE# | aprire |
| 34 | 22 | Opening | OPE# | chiudere |
| 34 | 22 | Opening | OPE# | schiudere |
| 34 | 22 | Opening | OPE# | socchiudere |
| 34 | 22 | Opening | OPE# | spalancare |
| 35 | 23 | Consumption | CONS# | bere |
| 35 | 23 | Consumption | CONS# | consumare |
| 35 | 23 | Consumption | CONS# | divorare |
| 35 | 23 | Consumption | CONS# | fumare |
| 35 | 23 | Consumption | CONS# | mangiare |
| 36 | 24 | Elimination | ELIM# | ammazzare |
| 36 | 24 | Elimination | ELIM# | annientare |
| 36 | 24 | Elimination | ELIM# | assassinare |
| 36 | 24 | Elimination | ELIM# | cancellare |
| 36 | 24 | Elimination | ELIM# | distruggere |
| 36 | 24 | Elimination | ELIM# | eliminare |
| 36 | 24 | Elimination | ELIM# | massacrare |
| 36 | 24 | Elimination | ELIM# | uccidere |
| 37 | 25 | Basis | BAS# | basare |
| 37 | 25 | Basis | BAS# | concernere |
| 37 | 25 | Basis | BAS# | imperniare |
| 37 | 25 | Basis | BAS# | vertere |
| 38 | 26 | Inference | INFE# | desumere |
| 38 | 26 | Inference | INFE# | dipendere |
| 38 | 26 | Inference | INFE# | evincere |
| 39 | 27 | Result | RESU# | derivare |
| 39 | 27 | Result | RESU# | emergere |
| 39 | 27 | Result | RESU# | risultare |
| 40 | 28 | Weather | WEAT# | fioccare |
| 40 | 28 | Weather | WEAT# | grandinare |
| 40 | 28 | Weather | WEAT# | nevicare |
| 40 | 28 | Weather | WEAT# | piovere |

*Tabella 4.2 Classificazione semantica dei 200 verbi italiani ripartiti in 24 classi*

| ID | Numero Classe | Classe | Sigla | Verbo |
|---|---|---|---|---|
| 1 | 1 | Aspect | AS# | cominciare |
| 1 | 1 | Aspect | AS# | continuare |
| 1 | 1 | Aspect | AS# | finire |
| 1 | 1 | Aspect | AS# | iniziare |
| 1 | 1 | Aspect | AS# | smettere |
| 1 | 1 | Aspect | AS# | terminare |
| 2 | 2 | Propositional Attitude | PROP# | considerare |
| 2 | 2 | Propositional Attitude | PROP# | credere |
| 2 | 2 | Propositional Attitude | PROP# | desiderare |
| 2 | 2 | Propositional Attitude | PROP# | dubitare |
| 2 | 2 | Propositional Attitude | PROP# | esigere |
| 2 | 2 | Propositional Attitude | PROP# | necessitare |
| 2 | 2 | Propositional Attitude | PROP# | pensare |
| 2 | 2 | Propositional Attitude | PROP# | richiedere |
| 2 | 2 | Propositional Attitude | PROP# | ritenere |
| 2 | 2 | Propositional Attitude | PROP# | sapere |
| 2 | 2 | Propositional Attitude | PROP# | sperare |
| 2 | 2 | Propositional Attitude | PROP# | volere |
| 3 | 3 | Transfer of Possession | POT# | acquisire |
| 3 | 3 | Transfer of Possession | POT# | acquistare |
| 3 | 3 | Transfer of Possession | POT# | appropriare(si) |
| 3 | 3 | Transfer of Possession | POT# | comprare |
| 3 | 3 | Transfer of Possession | POT# | consegnare |
| 3 | 3 | Transfer of Possession | POT# | conseguire |
| 3 | 3 | Transfer of Possession | POT# | distribuire |
| 3 | 3 | Transfer of Possession | POT# | fornire |
| 3 | 3 | Transfer of Possession | POT# | impadronire(si) |
| 3 | 3 | Transfer of Possession | POT# | inviare |
| 3 | 3 | Transfer of Possession | POT# | mandare |
| 3 | 3 | Transfer of Possession | POT# | offrire |
| 3 | 3 | Transfer of Possession | POT# | ottenere |
| 3 | 3 | Transfer of Possession | POT# | recapitare |
| 3 | 3 | Transfer of Possession | POT# | regalare |
| 3 | 3 | Transfer of Possession | POT# | restituire |
| 3 | 3 | Transfer of Possession | POT# | spedire |
| 3 | 3 | Transfer of Possession | POT# | vendere |
| 4 | 4 | Motion | MO# | allontanare(si) |
| 4 | 4 | Motion | MO# | arrivare |
| 4 | 4 | Motion | MO# | attraversare |
| 4 | 4 | Motion | MO# | camminare |
| 4 | 4 | Motion | MO# | correre |
| 4 | 4 | Motion | MO# | entrare |
| 4 | 4 | Motion | MO# | fuggire |
| 4 | 4 | Motion | MO# | giungere |
| 4 | 4 | Motion | MO# | marciare |
| 4 | 4 | Motion | MO# | navigare |
| 4 | 4 | Motion | MO# | passeggiare |
| 4 | 4 | Motion | MO# | pattinare |
| 4 | 4 | Motion | MO# | pedalare |

| | | | | |
|---|---|---|---|---|
| 4 | 4 | Motion | MO# | percorrere |
| 4 | 4 | Motion | MO# | rotolare |
| 4 | 4 | Motion | MO# | saltellare |
| 4 | 4 | Motion | MO# | scappare |
| 4 | 4 | Motion | MO# | strisciare |
| 4 | 4 | Motion | MO# | tornare |
| 4 | 4 | Motion | MO# | traversare |
| 4 | 4 | Motion | MO# | uscire |
| 4 | 4 | Motion | MO# | volare |
| 5 | 5 | Emotion | EM# | amare |
| 5 | 5 | Emotion | EM# | divertire(si) |
| 5 | 5 | Emotion | EM# | entusiasmare |
| 5 | 5 | Emotion | EM# | esultare |
| 5 | 5 | Emotion | EM# | gioire |
| 5 | 5 | Emotion | EM# | impaurire(si) |
| 5 | 5 | Emotion | EM# | infastidire(si) |
| 5 | 5 | Emotion | EM# | irritare(si) |
| 5 | 5 | Emotion | EM# | odiare |
| 5 | 5 | Emotion | EM# | piacere |
| 5 | 5 | Emotion | EM# | soffrire |
| 5 | 5 | Emotion | EM# | spaventare(si) |
| 6 | 6 | Facial Expression | FEX# | piangere |
| 6 | 6 | Facial Expression | FEX# | ridere |
| 6 | 6 | Facial Expression | FEX# | sbadigliare |
| 6 | 6 | Facial Expression | FEX# | sorridere |
| 6 | 6 | Facial Expression | FEX# | tossire |
| 7 | 7 | Perception | PER# | annusare |
| 7 | 7 | Perception | PER# | ascoltare |
| 7 | 7 | Perception | PER# | assaggiare |
| 7 | 7 | Perception | PER# | fiutare |
| 7 | 7 | Perception | PER# | guardare |
| 7 | 7 | Perception | PER# | scorgere |
| 7 | 7 | Perception | PER# | scrutare |
| 7 | 7 | Perception | PER# | sentire |
| 7 | 7 | Perception | PER# | vedere |
| 8 | 8 | Communication | COM# | affermare |
| 8 | 8 | Communication | COM# | annunciare |
| 8 | 8 | Communication | COM# | assicurare |
| 8 | 8 | Communication | COM# | bisbigliare |
| 8 | 8 | Communication | COM# | chiacchierare |
| 8 | 8 | Communication | COM# | comunicare |
| 8 | 8 | Communication | COM# | constatare |
| 8 | 8 | Communication | COM# | conversare |
| 8 | 8 | Communication | COM# | decidere |
| 8 | 8 | Communication | COM# | deplorare |
| 8 | 8 | Communication | COM# | deprecare |
| 8 | 8 | Communication | COM# | dichiarare |
| 8 | 8 | Communication | COM# | dire |
| 8 | 8 | Communication | COM# | evidenziare |
| 8 | 8 | Communication | COM# | garantire |
| 8 | 8 | Communication | COM# | gridare |
| 8 | 8 | Communication | COM# | impegnare(si) |
| 8 | 8 | Communication | COM# | informare |

| | | | | |
|---|---|---|---|---|
| 8 | 8 | Communication | COM# | lagnare(si) |
| 8 | 8 | Communication | COM# | lamentare(si) |
| 8 | 8 | Communication | COM# | mormorare |
| 8 | 8 | Communication | COM# | notare |
| 8 | 8 | Communication | COM# | ordinare |
| 8 | 8 | Communication | COM# | osservare |
| 8 | 8 | Communication | COM# | parlare |
| 8 | 8 | Communication | COM# | promettere |
| 8 | 8 | Communication | COM# | raccontare |
| 8 | 8 | Communication | COM# | rammaricare(si) |
| 8 | 8 | Communication | COM# | registrare |
| 8 | 8 | Communication | COM# | riferire |
| 8 | 8 | Communication | COM# | segnalare |
| 8 | 8 | Communication | COM# | stabilire |
| 8 | 8 | Communication | COM# | sussurrare |
| 8 | 8 | Communication | COM# | urlare |
| 8 | 8 | Communication | COM# | vietare |
| 9 | 9 | Description | DES# | caratterizzare |
| 9 | 9 | Description | DES# | descrivere |
| 9 | 9 | Description | DES# | fantasticare |
| 9 | 9 | Description | DES# | illustrare |
| 9 | 9 | Description | DES# | immaginare |
| 9 | 9 | Description | DES# | manifestare |
| 9 | 9 | Description | DES# | mostrare |
| 9 | 9 | Description | DES# | presentare |
| 9 | 9 | Description | DES# | presumere |
| 9 | 9 | Description | DES# | spiegare |
| 9 | 9 | Description | DES# | supporre |
| 9 | 9 | Description | DES# | tratteggiare |
| 10 | 10 | Insistence | INS# | insistere |
| 10 | 10 | Insistence | INS# | perseverare |
| 10 | 10 | Insistence | INS# | persistere |
| 11 | 11 | Learning | TEA# | apprendere |
| 11 | 11 | Learning | TEA# | imparare |
| 11 | 11 | Learning | TEA# | leggere |
| 11 | 11 | Learning | TEA# | studiare |
| 12 | 12 | Position (bring into position) | PSBR# | collocare |
| 12 | 12 | Position (bring into position) | PSBR# | levare |
| 12 | 12 | Position (bring into position) | PSBR# | mettere |
| 12 | 12 | Position (bring into position) | PSBR# | porre |
| 12 | 12 | Position (bring into position) | PSBR# | posizionare |
| 12 | 12 | Position (bring into position) | PSBR# | togliere |
| 13 | 13 | Position (be in position) | PSBE# | abitare |
| 13 | 13 | Position (be in position) | PSBE# | restare |
| 13 | 13 | Position (be in position) | PSBE# | rimanere |
| 13 | 13 | Position (be in position) | PSBE# | stare |
| 13 | 13 | Position (be in position) | PSBE# | vivere |
| 14 | 14 | Production | PROD# | costruire |
| 14 | 14 | Production | PROD# | creare |
| 14 | 14 | Production | PROD# | erigere |
| 14 | 14 | Production | PROD# | produrre |
| 14 | 14 | Production | PROD# | realizzare |
| 15 | 15 | Renovation | RENO# | aggiustare |

| | | | | |
|---|---|---|---|---|
| 15 | 15 | Renovation | RENO# | ammodernare |
| 15 | 15 | Renovation | RENO# | rinnovare |
| 15 | 15 | Renovation | RENO# | riparare |
| 16 | 16 | Support | SUP# | aiutare |
| 16 | 16 | Support | SUP# | incentivare |
| 16 | 16 | Support | SUP# | salvare |
| 16 | 16 | Support | SUP# | soccorrere |
| 17 | 17 | Quantum Change | QUCA# | abbassare |
| 17 | 17 | Quantum Change | QUCA# | accorciare |
| 17 | 17 | Quantum Change | QUCA# | allungare |
| 17 | 17 | Quantum Change | QUCA# | aumentare |
| 17 | 17 | Quantum Change | QUCA# | diminuire |
| 17 | 17 | Quantum Change | QUCA# | innalzare |
| 18 | 18 | Opening | OPE# | aprire |
| 18 | 18 | Opening | OPE# | chiudere |
| 18 | 18 | Opening | OPE# | schiudere |
| 18 | 18 | Opening | OPE# | socchiudere |
| 18 | 18 | Opening | OPE# | spalancare |
| 19 | 19 | Consumption | CONS# | bere |
| 19 | 19 | Consumption | CONS# | consumare |
| 19 | 19 | Consumption | CONS# | divorare |
| 19 | 19 | Consumption | CONS# | fumare |
| 19 | 19 | Consumption | CONS# | mangiare |
| 20 | 20 | Elimination | ELIM# | ammazzare |
| 20 | 20 | Elimination | ELIM# | annientare |
| 20 | 20 | Elimination | ELIM# | assassinare |
| 20 | 20 | Elimination | ELIM# | cancellare |
| 20 | 20 | Elimination | ELIM# | distruggere |
| 20 | 20 | Elimination | ELIM# | eliminare |
| 20 | 20 | Elimination | ELIM# | massacrare |
| 20 | 20 | Elimination | ELIM# | uccidere |
| 21 | 21 | Basis | BAS# | basare |
| 21 | 21 | Basis | BAS# | concernere |
| 21 | 21 | Basis | BAS# | imperniare |
| 21 | 21 | Basis | BAS# | vertere |
| 22 | 22 | Inference | INFE# | desumere |
| 22 | 22 | Inference | INFE# | dipendere |
| 22 | 22 | Inference | INFE# | evincere |
| 23 | 23 | Result | RESU# | derivare |
| 23 | 23 | Result | RESU# | emergere |
| 23 | 23 | Result | RESU# | risultare |
| 24 | 24 | Weather | WEAT# | fioccare |
| 24 | 24 | Weather | WEAT# | grandinare |
| 24 | 24 | Weather | WEAT# | nevicare |
| 24 | 24 | Weather | WEAT# | piovere |

*Tabella 4.3 Classificazione semantica dei 200 verbi italiani ripartiti in 10 classi*

| ID | Numero Classe | Classe | Sigla | Verbo |
|---|---|---|---|---|
| 1 | 1 | Aspect | AS# | cominciare |
| 1 | 1 | Aspect | AS# | continuare |
| 1 | 1 | Aspect | AS# | finire |
| 1 | 1 | Aspect | AS# | iniziare |
| 1 | 1 | Aspect | AS# | smettere |
| 1 | 1 | Aspect | AS# | terminare |
| 2 | 2 | Cognition | COGN# | affermare |
| 2 | 2 | Cognition | COGN# | amare |
| 2 | 2 | Cognition | COGN# | annunciare |
| 2 | 2 | Cognition | COGN# | annusare |
| 2 | 2 | Cognition | COGN# | apprendere |
| 2 | 2 | Cognition | COGN# | ascoltare |
| 2 | 2 | Cognition | COGN# | assaggiare |
| 2 | 2 | Cognition | COGN# | assicurare |
| 2 | 2 | Cognition | COGN# | bisbigliare |
| 2 | 2 | Cognition | COGN# | caratterizzare |
| 2 | 2 | Cognition | COGN# | chiacchierare |
| 2 | 2 | Cognition | COGN# | comunicare |
| 2 | 2 | Cognition | COGN# | considerare |
| 2 | 2 | Cognition | COGN# | constatare |
| 2 | 2 | Cognition | COGN# | conversare |
| 2 | 2 | Cognition | COGN# | credere |
| 2 | 2 | Cognition | COGN# | decidere |
| 2 | 2 | Cognition | COGN# | deplorare |
| 2 | 2 | Cognition | COGN# | deprecare |
| 2 | 2 | Cognition | COGN# | descrivere |
| 2 | 2 | Cognition | COGN# | desiderare |
| 2 | 2 | Cognition | COGN# | dichiarare |
| 2 | 2 | Cognition | COGN# | dire |
| 2 | 2 | Cognition | COGN# | divertire(si) |
| 2 | 2 | Cognition | COGN# | dubitare |
| 2 | 2 | Cognition | COGN# | entusiasmare |
| 2 | 2 | Cognition | COGN# | esigere |
| 2 | 2 | Cognition | COGN# | esultare |
| 2 | 2 | Cognition | COGN# | evidenziare |
| 2 | 2 | Cognition | COGN# | fantasticare |
| 2 | 2 | Cognition | COGN# | fiutare |
| 2 | 2 | Cognition | COGN# | garantire |
| 2 | 2 | Cognition | COGN# | gioire |
| 2 | 2 | Cognition | COGN# | gridare |
| 2 | 2 | Cognition | COGN# | guardare |
| 2 | 2 | Cognition | COGN# | illustrare |
| 2 | 2 | Cognition | COGN# | immaginare |
| 2 | 2 | Cognition | COGN# | imparare |
| 2 | 2 | Cognition | COGN# | impaurire(si) |
| 2 | 2 | Cognition | COGN# | impegnare(si) |
| 2 | 2 | Cognition | COGN# | infastidire(si) |
| 2 | 2 | Cognition | COGN# | informare |
| 2 | 2 | Cognition | COGN# | insistere |

| | | | | |
|---|---|---|---|---|
| 2 | 2 | Cognition | COGN# | irritare(si) |
| 2 | 2 | Cognition | COGN# | lagnare(si) |
| 2 | 2 | Cognition | COGN# | lamentare(si) |
| 2 | 2 | Cognition | COGN# | leggere |
| 2 | 2 | Cognition | COGN# | manifestare |
| 2 | 2 | Cognition | COGN# | mormorare |
| 2 | 2 | Cognition | COGN# | mostrare |
| 2 | 2 | Cognition | COGN# | necessitare |
| 2 | 2 | Cognition | COGN# | notare |
| 2 | 2 | Cognition | COGN# | odiare |
| 2 | 2 | Cognition | COGN# | ordinare |
| 2 | 2 | Cognition | COGN# | osservare |
| 2 | 2 | Cognition | COGN# | parlare |
| 2 | 2 | Cognition | COGN# | pensare |
| 2 | 2 | Cognition | COGN# | perseverare |
| 2 | 2 | Cognition | COGN# | persistere |
| 2 | 2 | Cognition | COGN# | piacere |
| 2 | 2 | Cognition | COGN# | presentare |
| 2 | 2 | Cognition | COGN# | presumere |
| 2 | 2 | Cognition | COGN# | promettere |
| 2 | 2 | Cognition | COGN# | raccontare |
| 2 | 2 | Cognition | COGN# | rammaricare(si) |
| 2 | 2 | Cognition | COGN# | registrare |
| 2 | 2 | Cognition | COGN# | richiedere |
| 2 | 2 | Cognition | COGN# | riferire |
| 2 | 2 | Cognition | COGN# | ritenere |
| 2 | 2 | Cognition | COGN# | sapere |
| 2 | 2 | Cognition | COGN# | scorgere |
| 2 | 2 | Cognition | COGN# | scrutare |
| 2 | 2 | Cognition | COGN# | segnalare |
| 2 | 2 | Cognition | COGN# | sentire |
| 2 | 2 | Cognition | COGN# | soffrire |
| 2 | 2 | Cognition | COGN# | spaventare(si) |
| 2 | 2 | Cognition | COGN# | sperare |
| 2 | 2 | Cognition | COGN# | spiegare |
| 2 | 2 | Cognition | COGN# | stabilire |
| 2 | 2 | Cognition | COGN# | studiare |
| 2 | 2 | Cognition | COGN# | supporre |
| 2 | 2 | Cognition | COGN# | sussurrare |
| 2 | 2 | Cognition | COGN# | tratteggiare |
| 2 | 2 | Cognition | COGN# | urlare |
| 2 | 2 | Cognition | COGN# | vedere |
| 2 | 2 | Cognition | COGN# | vietare |
| 2 | 2 | Cognition | COGN# | volere |
| 3 | 3 | Transfer of Possession | POT# | acquisire |
| 3 | 3 | Transfer of Possession | POT# | acquistare |
| 3 | 3 | Transfer of Possession | POT# | appropriare(si) |
| 3 | 3 | Transfer of Possession | POT# | comprare |
| 3 | 3 | Transfer of Possession | POT# | consegnare |
| 3 | 3 | Transfer of Possession | POT# | conseguire |
| 3 | 3 | Transfer of Possession | POT# | distribuire |
| 3 | 3 | Transfer of Possession | POT# | fornire |
| 3 | 3 | Transfer of Possession | POT# | impadronire(si) |

| | | | | |
|---|---|---|---|---|
| 3 | 3 | Transfer of Possession | POT# | inviare |
| 3 | 3 | Transfer of Possession | POT# | mandare |
| 3 | 3 | Transfer of Possession | POT# | offrire |
| 3 | 3 | Transfer of Possession | POT# | ottenere |
| 3 | 3 | Transfer of Possession | POT# | recapitare |
| 3 | 3 | Transfer of Possession | POT# | regalare |
| 3 | 3 | Transfer of Possession | POT# | restituire |
| 3 | 3 | Transfer of Possession | POT# | spedire |
| 3 | 3 | Transfer of Possession | POT# | vendere |
| 4 | 4 | Motion | MO# | allontanare(si) |
| 4 | 4 | Motion | MO# | arrivare |
| 4 | 4 | Motion | MO# | attraversare |
| 4 | 4 | Motion | MO# | camminare |
| 4 | 4 | Motion | MO# | correre |
| 4 | 4 | Motion | MO# | entrare |
| 4 | 4 | Motion | MO# | fuggire |
| 4 | 4 | Motion | MO# | giungere |
| 4 | 4 | Motion | MO# | marciare |
| 4 | 4 | Motion | MO# | navigare |
| 4 | 4 | Motion | MO# | passeggiare |
| 4 | 4 | Motion | MO# | pattinare |
| 4 | 4 | Motion | MO# | pedalare |
| 4 | 4 | Motion | MO# | percorrere |
| 4 | 4 | Motion | MO# | rotolare |
| 4 | 4 | Motion | MO# | saltellare |
| 4 | 4 | Motion | MO# | scappare |
| 4 | 4 | Motion | MO# | strisciare |
| 4 | 4 | Motion | MO# | tornare |
| 4 | 4 | Motion | MO# | traversare |
| 4 | 4 | Motion | MO# | uscire |
| 4 | 4 | Motion | MO# | volare |
| 5 | 5 | Position | POS# | abitare |
| 5 | 5 | Position | POS# | collocare |
| 5 | 5 | Position | POS# | levare |
| 5 | 5 | Position | POS# | mettere |
| 5 | 5 | Position | POS# | porre |
| 5 | 5 | Position | POS# | posizionare |
| 5 | 5 | Position | POS# | restare |
| 5 | 5 | Position | POS# | rimanere |
| 5 | 5 | Position | POS# | stare |
| 5 | 5 | Position | POS# | togliere |
| 5 | 5 | Position | POS# | vivere |
| 6 | 6 | Change | CHAN# | abbassare |
| 6 | 6 | Change | CHAN# | accorciare |
| 6 | 6 | Change | CHAN# | aggiustare |
| 6 | 6 | Change | CHAN# | allungare |
| 6 | 6 | Change | CHAN# | ammazzare |
| 6 | 6 | Change | CHAN# | ammodernare |
| 6 | 6 | Change | CHAN# | annientare |
| 6 | 6 | Change | CHAN# | aprire |
| 6 | 6 | Change | CHAN# | assassinare |
| 6 | 6 | Change | CHAN# | aumentare |
| 6 | 6 | Change | CHAN# | bere |

| | | | | |
|---|---|---|---|---|
| 6 | 6 | Change | CHAN# | cancellare |
| 6 | 6 | Change | CHAN# | chiudere |
| 6 | 6 | Change | CHAN# | consumare |
| 6 | 6 | Change | CHAN# | costruire |
| 6 | 6 | Change | CHAN# | creare |
| 6 | 6 | Change | CHAN# | diminuire |
| 6 | 6 | Change | CHAN# | distruggere |
| 6 | 6 | Change | CHAN# | divorare |
| 6 | 6 | Change | CHAN# | eliminare |
| 6 | 6 | Change | CHAN# | erigere |
| 6 | 6 | Change | CHAN# | fumare |
| 6 | 6 | Change | CHAN# | innalzare |
| 6 | 6 | Change | CHAN# | mangiare |
| 6 | 6 | Change | CHAN# | massacrare |
| 6 | 6 | Change | CHAN# | produrre |
| 6 | 6 | Change | CHAN# | realizzare |
| 6 | 6 | Change | CHAN# | rinnovare |
| 6 | 6 | Change | CHAN# | riparare |
| 6 | 6 | Change | CHAN# | schiudere |
| 6 | 6 | Change | CHAN# | socchiudere |
| 6 | 6 | Change | CHAN# | spalancare |
| 6 | 6 | Change | CHAN# | uccidere |
| 7 | 7 | Facial Expression | FEX# | piangere |
| 7 | 7 | Facial Expression | FEX# | ridere |
| 7 | 7 | Facial Expression | FEX# | sbadigliare |
| 7 | 7 | Facial Expression | FEX# | sorridere |
| 7 | 7 | Facial Expression | FEX# | tossire |
| 8 | 8 | Argument | ARG# | basare |
| 8 | 8 | Argument | ARG# | concernere |
| 8 | 8 | Argument | ARG# | derivare |
| 8 | 8 | Argument | ARG# | desumere |
| 8 | 8 | Argument | ARG# | dipendere |
| 8 | 8 | Argument | ARG# | emergere |
| 8 | 8 | Argument | ARG# | evincere |
| 8 | 8 | Argument | ARG# | imperniare |
| 8 | 8 | Argument | ARG# | risultare |
| 8 | 8 | Argument | ARG# | vertere |
| 9 | 9 | Support | SUP# | aiutare |
| 9 | 9 | Support | SUP# | incentivare |
| 9 | 9 | Support | SUP# | salvare |
| 9 | 9 | Support | SUP# | soccorrere |
| 10 | 10 | Weather | WEAT# | fioccare |
| 10 | 10 | Weather | WEAT# | grandinare |
| 10 | 10 | Weather | WEAT# | nevicare |
| 10 | 10 | Weather | WEAT# | piovere |

*Tabella 4.4 Elenco dei 105 frames di sottocategorizzazione selezionati*

ogg_d#
0#
pred#
comp_a#
comp_in#
ogg_d#comp_a#
cla_che#
ogg_d#comp_in#
si#0#
inf_a#
comp_di#
inf_di#
comp_con#
comp_da#
comp_su#
si#ogg_d#
comp_per#
ogg_d#comp_con#
si#comp_a#
si#comp_in#
ogg_d#comp_per#
ogg_d#comp_su#
ogg_d#comp_da#
si#comp_di#
ogg_d#inf_a#
ogg_d#pred#
cla_se#
pred_in#
comp_a#comp_in#
ogg_d#comp_di#
si#inf_a#
si#pred#
ogg_d#cla_che#
comp_come#
si#comp_con#
comp_a#inf_di#
si#comp_su#
si#cla_che#
si#inf_di#
ogg_d#inf_di#
ogg_d#comp_come#
comp_tra#
inf_da#
si#ogg_d#comp_a#
si#comp_da#
comp_a#comp_con#
cla_come#
comp_a#cla_che#
comp_a#comp_per#
comp_a#comp_da#
si#comp_per#
comp_a#comp_di#
comp_con#comp_in#
si#ogg_d#comp_in#
comp_in#comp_per#
comp_contro#
comp_di#comp_in#
comp_sotto#
ogg_d#inf_da#
si#comp_a#comp_in#
comp_in#inf_a#
comp_a#inf_a#
si#cla_se#
comp_a#cla_se#
comp_da#comp_in#
comp_verso#
comp_a#comp_su#
ogg_d#cla_come#
comp_in#comp_su#
ogg_d#comp_sotto#
si#ogg_d#comp_con#
ogg_d#comp_tra#
ogg_d#comp_contro#
si#comp_come#
comp_attraverso#
comp_in#inf_di#
si#ogg_d#comp_di#
si#ogg_d#comp_su#
si#comp_a#comp_di#
ogg_d#comp_verso#
si#comp_a#comp_con#
comp_di#comp_per#
comp_con#comp_di#
si#ogg_d#comp_da#
comp_con#comp_per#
si#comp_di#comp_in#
si#ogg_d#comp_per#
si#comp_con#comp_in#
comp_a#comp_come#
si#comp_tra#
comp_come#comp_in#
comp_da#comp_per#
si#comp_contro#
si#comp_verso#
comp_con#comp_da#
ogg_d#comp_attraverso#
si#comp_a#comp_per#
comp_dentro#
si#comp_a#comp_da#
comp_con#comp_su#
si#inf_da#
comp_per#comp_su#
si#comp_in#comp_per#
comp_da#comp_di#
comp_di#comp_su#

*Tabella 4.5 Elenco dei 50 frames di sottocategorizzazione selezionati*

```
ogg_d#
0#
pred#
comp_a#
comp_in#
ogg_d#comp_a#
cla_che#
ogg_d#comp_in#
si#0#
inf_a#
comp_di#
inf_di#
comp_con#
comp_da#
comp_su#
si#ogg_d#
comp_per#
ogg_d#comp_con#
si#comp_a#
si#comp_in#
ogg_d#comp_per#
ogg_d#comp_su#
ogg_d#comp_da#
si#comp_di#
ogg_d#inf_a#
ogg_d#pred#
cla_se#
pred_in#
comp_a#comp_in#
ogg_d#comp_di#
si#inf_a#
si#pred#
ogg_d#cla_che#
comp_come#
si#comp_con#
comp_a#inf_di#
si#comp_su#
si#cla_che#
si#inf_di#
ogg_d#inf_di#
ogg_d#comp_come#
comp_tra#
inf_da#
si#ogg_d#comp_a#
si#comp_da#
comp_a#comp_con#
cla_come#
comp_a#cla_che#
comp_a#comp_per#
comp_a#comp_da#
```

*Tabella 4.6 Elenco dei 25 frames di sottocategorizzazione selezionati*

```
ogg_d#
0#
pred#
comp_a#
comp_in#
ogg_d#comp_a#
cla_che#
ogg_d#comp_in#
si#0#
inf_a#
comp_di#
inf_di#
comp_con#
comp_da#
comp_su#
si#ogg_d#
comp_per#
ogg_d#comp_con#
si#comp_a#
si#comp_in#
ogg_d#comp_per#
ogg_d#comp_su#
ogg_d#comp_da#
si#comp_di#
ogg_d#inf_a#
```

*Tabella 4.7 Elenco dei 34 frames di sottocategorizzazione non preposizionali selezionati*

```
ogg_d#
0#
pred#
comp#
ogg_d#comp#
cla_che#
si#0#
inf_a#
inf_di#
si#ogg_d#
si#comp#
ogg_d#inf_a#
ogg_d#pred#
cla_se#
pred_in#
comp#comp#
si#inf_a#
si#pred#
ogg_d#cla_che#
comp#inf_di#
si#cla_che#
si#inf_di#
ogg_d#inf_di#
inf_da#
si#ogg_d#comp#
cla_come#
comp#cla_che#
ogg_d#inf_da#
si#comp#comp#
comp#inf_a#
si#cla_se#
comp#cla_se#
ogg_d#cla_come#
si#inf_da#
```