INDICE

Premessa

Il lavoro di tesi, condotto nell'ambito della linguistica computazionale ha come argomento l'intercoder agreement, cioè l'indice di accordo tra vari annotatori, nello specifico sarà oggetto di studio l'analisi del coefficiente di accordo su un'annotazione semantica.

E' stato indispensabile giungere gradualmente all'argomento fornendo al lettore delle nozioni preliminari importanti per comprendere nel miglior modo possibile l'argomento.

Introducendo la definizione generale di "percezione", successivamente quella di "semantica", si procede nel primo capitolo ad esporre il concetto di annotazione in linguistica computazionale proseguendo con una rassegna dei principali coefficienti di accordo, approfondendo il kappa di Cohen oggetto della tesi.

La panoramica è ulteriormente approfondita analizzando i problemi che possono presentarsi durante il calcolo dell'intercoder agreement.

Il secondo capitolo è incentrato sul lavoro di ricerca e studio svolto durante il tirocinio. Descritto inizialmente il progetto di lavoro e annotazione, si passa in segutio all'analisi del software utilizzato fornendo al lettore esempi sul tipo di annotazione svolta.

Nel terzo capitolo, i dati ricavati dall'annotazione svolta durante il periodo del tirocinio sarà calcolato l'intercoder agreement, utilizzando il coefficiente di accordo Kappa presentato nel capitolo precedente. Il capitolo sarà concluso da un'analisi sul lavoro svolto, analizzando i risultati ottenuti e motivandone le ragioni.

Introduzione

“Quando Le Porte della percezione sono spalancate, le cose appaiono come veramente sono: infinite ”

William Blake

-The Marriage of Heaven and Hell-

Cosi scriveva il poeta William Blake riferendosi alla percezione e al ruolo fondamentale che questa ha nell'interpretazione delle situazioni che ci circondano.
La percezione infatti, è un meccanismo psichico che permette all'uomo attraverso l'attivazione degli organi di senso di dare una spiegazione a ciò che lo circonda, di analizzare soggettivamente le informazioni provenienti dall'esterno. La percezione e conseguentemente l'interpretazione di una situazione sono il risultato delle conoscenze pregresse e del bagaglio culturale di ogni individuo, per questo motivo spesso capita che una stessa situazione venga interpretata e percepita diversamente da due o più individui.
In linguistica la percezione e l'interpretazione delle parole e dei testi trovano il corrispettivo nella semantica.
Un frame semantico è in linguistica computazionale la rappresentazione di una situazione o di un evento, è formato da unità lessicali che indicano l'evento in oggetto (esempio *incidente*) e da ruoli semantici detti anche Frame Elements (esempio *vittima*,*luogo,causa*).
I frames sono "evocati" da un verbo all'interno di una frase e permettono l'annotazione della frase (manualmente o con l'ausilio di programmi) con i ruoli semantici corrispondenti.
Presso l'International Computer Science Institute a Berkeley in California è nato FrameNet, un progetto atto alla creazione di una risorsa elettronica basata sui frame semantici.
Il database lessicale di FrameNet contiene circa 10000 unità lessicali (coppie di parole con il loro significato), 800 frame semantici e oltre 120000 frasi di esempio.
I testi contenuti in Framenet sono annotati manualmente e una condizione fondamentale è la verifica dell’affidabilità e replicabilità delle annotazioni, requisito

indispensabile per il calcolo dell'intercoder agreement.

I corpus annotati sono stati concessi da altri progetti già esistenti[3] :

- progetto Propbank (http://verbs.colorado.edu/~mpalmer/projects/ace.html). I testi comprendono una vasta gamma di argomenti. Lo scopo è valutare le relazioni esistenti tra le annotazioni effettuate da PropBank e le annotazione effettuate da FrameNet.
- progetto Masc (www.anc.org/MASC/Home.html). Testi integrali dalla American National Corpus, tra cui "LUcorpus", una raccolta di documenti che comprende trascrizioni di conversazioni telefoniche, e-mail, traduzioni dall'arabo.
- progetto Aquaint (www-nlpir.nist.gov/projects/aquaint). Testi da "Advanced Question & Answering", prodotti per la Nuclear Threat Initiative (http://www.nti.org).

Lo scopo del progetto è creare una serie di situazioni ed eventi atti ad aiutare la comprensione di un testo utilizzando tecniche di elaborazione del linguaggio naturale che permettano di associare ad ogni frase un frame e i corrispondenti ruoli semantici[4].

Di seguito un esempio di annotazione di frame semantico presente sul sito ufficiale del progetto[5]:

I ruoli semantici disponibile per il frame sono:

- Buyer: rappresenta il compratore che possiede il denaro per acquistare i beni.
- Goods: rappresenta i beni che vengono venduti
- Money: rappresenta il denaro utilizzato nella transazione
- Seller: rappresenta il venditore che possiede i beni da vendere

L'applicazione dei ruoli semantici avviene come segue:

The Buyer wants the Goods and offers Money to a Seller in exchange for them.

(traduz.L'acquirente vuole dei beni e offre denaro al venditore in cambio di questi).

[3] cfr. https://framenet.icsi.berkeley.edu/fndrupal/
[4] cfr. Wikipedia, L'Enciclopedia Libera.
[5] cfr. https://framenet.icsi.berkeley.edu/fndrupal/

Sono molti i ruoli semantici presenti in Framenet e utilizzabili, non è possibile elencarli tutti, di seguito i più frequenti[6]:

Agente o **attore**: l'entità che compie volontariamente un'azione;

Strumento: l'entità che serve involontariamente al compimento di un'azione;

Paziente o **oggetto**: obiettivo del compimento di un'azione;

Beneficiario: entità che riceve profitto o danno dall'esecuzione di un'azione;

Causa: entità che provoca il compimento di un processo;

Risultato: stato finale che segue il compimento di un processo;

Sorgente: stato iniziale all'esecuzione di un processo;

Scopo: stato verso cui mira un processo o un 'azione;

Luogo: entità spaziale in cui si manifesta un processo;

Tempo: circostanza temporale di un processo.

[6] cfr. Wikipedia, L'Enciclopedia Libera.

CAPITOLO I
L'INTERCODER AGREEMENT

1.1 L'annotazione in CL

Annotare significa aggiungere e arricchire un testo, nel caso specifico un corpus, con delle informazioni strutturali o semantiche utili per ricerche e analisi linguistiche.
L'annotazione semantica riguarda il significato delle espressioni linguistiche. Gli usi sono vari in linguistica anche se è possibile tracciare due tipi di annotazione semantica, una prima che riguarda le parole "lessicamente piene di un testo rispetto alle categorie semantico-concettuali predefinite"[7] e una seconda funzione che riguarda il riconoscimento e la codifica dei ruoli semantici, in cui, dato un certo costituente se ne esplicita il ruolo semantico all'interno della frase (paziente, agente, ecc).
L'annotazione di un testo si rivela particolarmente utile per indagini statistico-linguistiche; per questo motivo ci sono dei requisiti che devono essere soddisfatti da ogni schema di annotazione: *la compatibilità, la riproducibilità e l'usabilità*.
La prima caratteristica, la compatibilità di una annotazione è legata alla "compatibilità" con le teorie linguistiche già sviluppate nell'ambito di studio;
L'usabilità comprende i vari campi di applicazione di un testo (es. la ricerca);
Infine, la riproducibilità è la possibilità di " riprodurre" l'annotazione su più campi di interesse, meglio ancora se questi richiedono dei criteri di annotazione ben precisi che restringono l'arbitrarietà dell'annotazione consentendo cosi a diversi codificatori di annotare, anche se in momenti diversi, il testo in modo coerente.

1.2 Analisi dei coefficienti

Ponendo l'attenzione sull'annotazione di un testo o più in generale sullo studio di uno stesso oggetto da parte di due o più codificatori è possibile affermare che nell'ambito della ricerca in linguistica computazionale questo aspetto ha il nome di intercoder agreement.
L'intercoder agreement permette di valutare l'annotazione condotta da due persone o della stessa persona in momenti differenti osservando il loro lavoro sullo stesso testo

[7]cfr. Testo e computer, elementi di linguistica computazionale, Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli.

(intra-observed variation).
Questo permette di valutare il loro grado di accordo su uno stesso oggetto e capire se i giudizi forniti sono *affidabili, riproducibili* e *stabili*.
L'affidabilità è pertanto un requisito molto importante per dimostrare la validità dello schema di codifica e il grado di accordo tra gli annotatori.
L'affidabilità si correla a tre aspetti che devono essere verificati durante l'annotazione:

- *la stabilità (*o intra-coder agreement), permette di verificare la codifica di un oggetto a distanza di tempo; non è certo che la stessa persona a distanza di anni annoti un corpus allo stesso modo.
- la *riproducibilità,* il raggiungimento di uno stesso livello di codifica quando due o più codificatori lavorano anche se in modo indipendente allo stesso corpus.
- la *precisione,* nel caso ci fosse uno standard da rispettare, rappresenta la misura con cui i risultati ottenuti da processo di codifica si avvicinano allo standard.

E' possibile delineare tre punti principali per una buona valutazione dell'accordo:

- gli annotatori devono lavorare in modo indipendente, evitando consultazioni e consigli durante lo svolgimento del lavoro di annotazione.
- le categorie per la classificazione devono essere esaustive e quanto più possibile chiare ed esclusive;
- non ci devono essere restrizioni per l'attribuzione delle categorie.

Infatti un elemento può essere classificato -e conseguentemente assegnato ad una categoria- da due o più annotatori in maniera arbitraria. Questo implica il verificarsi di un certo numero di giudizi coincidenti per effetto della casualità.
E' importante considerare infatti l'ipotesi di *casualità* nell'assegnazione di un oggetto a una categoria.
Non è infatti da escludere la possibilità che due annotatori pur lavorando senza accordo assegnino casualmente un oggetto ad una stessa categoria, producendo dunque giudizi coincidenti casuali.
Per la stima dell'accordo si utilizzano dei coefficienti che, dato il numero di

annotatori e categorie da poter assegnare ne permettono la valutazione.

I principali coefficienti utilizzati nell'ambito della ricerca statistica in linguistica sono: S (Bennett, Alpert, and Goldstein 1954), π (Scott 1955) , *k* (Cohen 1960) e α (Krippendorff 1980)[8].

Ogni coefficiente presuppone che l'annotazione si basi su una serie di categorie, annotatori, oggetti analizzati e probabilità di accordi attesi e accordi osservati.

I tre coefficienti hanno delle similarità e delle diversità, in particolare condividono il calcolo dell'accordo osservato come misura dell'accordo tra gli annotatori.

Tutti e tre i coefficienti presuppongono l'indipendenza di lavoro degli annotatori e assumono che la probabilità che ogni annotatore assegni indipendentemente un elemento ad una determinata categoria è rappresentato dalla somma del prodotto di ogni assegnazione di un oggetto ad una categoria.

La differenza tra S, π e k è nell'assunzione del calcolo della probabilità che un annotatore assegni un oggetto a una determinata categoria.

Per capire meglio la differenza di calcolo con i tre coefficiente si riporta di seguito l'esempio fatto da Zwick[9] che illustra chiaramente gli effetti delle differenze di distribuzione tra i vari coefficienti. Assumendo che ad annotare ci siano due annotatori ci sono tre possibili situazioni che si possono verificare:

- gli annotatori assegnano valori uguali a tutte le categorie, in questo caso, tutti e tre i coefficienti di accordo hanno lo stesso valore.
- i due annotatori pur non assegnano lo stesso valore alle categorie, il coefficiente k e π hanno ancora lo stesso valore.
- i codificatori non concordano sulla percentuale di oggetti appartenenti ad una determinata categoria.

I coefficienti π e k, sono generalmente utilizzati per studi di affidabilità con più di due annotatori e in questo caso l'accordo osservato non può più essere definito come la percentuale di elementi su cui si è d'accordo, dal momento che inevitabilmente ci saranno elementi su cui alcuni annotatori concorderanno e altri no.

[8] cfr. Inter-Coder Agreement for Computational Linguistics, Ron Artstein, Massimo Poesio
[9] cfr. Inter-Coder Agreement for Computational Linguistics, Ron Artstein, Massimo Poesio

Per questo motivo talvolta le analisi vengono effettuate per “coppia di annotatori”;
Fleiss definisce l'accordo come la proporzione tra il numero di coppie di annotatori e il numero di giudizi espressi per quella determinata categoria.
Per rappresentare i giudizi espressi dagli annotatori utilizza una tabella che elenca per ogni elemento con il numero di giudizi ricevuti per ciascuna categoria; questo schema sarà definito da Di Eugenio e Glass(2004) “tavolo di accordo”.
Di seguito si analizza il lavoro svolto da sei psichiatri che classificano i disturbi di dieci pazienzi in cinque possibili categorie[10]:

- depressione
- disturbi della personalità
- schizofrenia
- neurosi
- altro

	Depressione j=1	**Disturbi della personalità j=2**	**Schizofrenia j=3**	**Neurosi** j=4	**Altro j=5**	***Pi***
1	0	0	0	0	14	1,000
2	0	2	6	4	2	0,253
3	0	0	3	5	6	0,308
4	0	3	9	2	0	0,440
5	2	2	8	1	1	0,330
6	7	7	0	0	0	0,462
7	3	2	6	3	0	0,242
8	2	5	3	2	2	0,176
9	6	5	2	1	0	0,286
10	0	2	2	3	7	0,286
Totale	20	28	39	21	32	
Pj	0,142	0,200	0,279	0,150	0,229	

Indicando con *N* il numero totale di soggetti, *n* il numero di voti per ogni soggetto, e *k* il numero di categorie in cui sono fatte le assegnazioni, nella situazione proposta i dati sono cosi rappresentati:

[10] cfr. Fleiss, J. Measuring nominal scale agreement among many raters. Psychological Bulletin.

N=10; k=5;

Si calcola *P*j, la proporzione di tutte le assegnazioni date alla *j-esima* categoria, dividendo il totale delle assegnazioni che sono state date alle categorie per il totale complessivo delle assegnazione fatte, nel caso della prima colonna "Depressione", si ha:

Totale = 20+28+39+21+32= 140

$$\boldsymbol{Pj} = \frac{20}{140} = 0.142$$

Si calcola *Pi,* , elevando a potenza ogni valore dato a ciascuna categoria diviso il numero dei medici -1. Nel caso della prima riga si ha:

$$\mathbf{Pi} = \frac{(0^2+0^2+0^2+0^2+14^2)-14}{14(14-1)} = \frac{(0^2+0^2+0^2+0^2+14^2)-14}{182} = \frac{196-14}{182} = \frac{182}{182} = 1$$

Si calcola successivamente la media tra Pi e Pj per calcolare il coefficiente kappa.
E' necessario effettuare la somma dei valori di Pi e quelli di Pj.

Pi $=0.143^2+ 0.200^2+ 0.279^2+ 0.150^2+ 0.229^2$

$=0.020+0.020+0.027+0.093+0.052=$ **0.213**

Pj = 1.00+0,253+0,308+0,440+0,330+0,462+0,242+0,176+0,286+0,286=**3.780**

$$\mathbf{P} = \frac{1}{(10)}(3780) = \mathbf{0.378}$$

$$\mathbf{k} = \frac{0.378-0.213}{1-0.213} = \mathbf{0.20}$$

Il valore del kappa, seguendo la tabella proposta da Landis e Koch (1977) indica un lieve accordo tra i medici.

Se l'accordo osservato è misurato sulla base di un accordo a coppie, ha senso misurare l'accordo atteso in termini di confronti a coppie, cioè la probabilità che una qualsiasi coppia di annotatori lavorando indipendentemente avrebbe lo stesso giudizio per un determinato elemento pur non avendo delle categorie prestabilite ma

basandosi sulla casualità.

Nell'ambito del calcolo dell'intercoder agreement il coefficiente più utilizzato è sicuramente il kappa.

Il coefficiente ***k*** è stato proposto da Cohen nel 1960 con l'articolo "A coefficient of agreement for nominal scales" pubblicato su Educational and Psychological Measurement (Vol. XX, No. 1, pp. 37-46).

L'uso di questo coefficiente risponde all'esigenza di valutare l'accordo in varie situazioni ed eventi, un caso di esempio nell'ambito medico è rappresentato dall'analisi della decisione di due chirurghi che decidono sulla necessità di operare un paziente e forniscono risposte concordanti o in psicologia se due o più psicologi condividono la stessa diagnosi.

Il coefficiente permette di valutare non solo il lavoro di due o più individui ma di tener traccia della riproducibilità del lavoro di un solo individuo in due momenti differenti (ad esempio se un medico darà la stessa diagnosi dopo aver visionato un'analisi clinica).

Di seguito si ripropone il test eseguito da Cohen[11] che analizza nell'articolo una situazione caratteristica della ricerca psicologica, due medici che hanno analizzato separatamente e in modo indipendente il comportamento delle stesse 200(duecento) persone, classificandole in tre differenti tipologie:

A = disordini della personalità

B = neurosi

C =psicosi

con i seguenti risultati:

		MEDICO1			
	CATEGORIE	A	B	C	**Totale**
MEDICO 2	A	50	26	24	100
	B	24	4	32	60
	C	6	30	4	40
	Totale	80	60	60	200

[11] cfr. J.Cohen, A coefficient of agreement for nominal scales Educational and Psychological Measurement, Vol. XX, No. 1, pp. 37-46.

Le frequenze ricavate sono assolute e per ricavare l'indice k proposto da Cohen, è importante trasformare in frequenze relative (con totale uguale a 1,0).

		MEDICO1			
	CATEGORIE	A	B	C	**Totale**
MEDICO 2	A	**0,25** (*0,20*)	**0,13**(*0,15*)	**0,12**(*0,15*)	0,50
	B	**0,12**(*0,12*)	**0,02**(*0,09*)	**0,16**(*0,09*)	0,30
	C	**0,03**(*0,08*)	**0,15**(*0,06*)	**0,02**(*0,06*)	0,20
	Totale	0,40	0,30	0,30	1,00

In grassetto sono riportate le proporzioni osservate; ad esempio, nella casella 1,1(A,A) si ha 0,25 = 50/200 (presi dalla tabella precedente con le frequenze assolute), in corsivo quelle attese, calcolate moltiplicando il totale dell'attribuzione alle categorie delle frequenze assolute di ogni medico, es.0,50*0,40=0,20, nella condizione che l'attribuzione dell'annotatore alla categoria sia stata casuale.

Nella formula proposta Cohen standardizza la proporzione osservata e proporzione totale attesa, dividendola per la massima differenza possibile non casuale.

Nelle ultime due tabelle dei dati, l'informazione utile è fornita dalle frequenze collocate lungo la diagonale principale (nella tabella 3 x 3, le caselle 1,1; 2,2; 3,3).

Nel caso dell'esempio, con le proporzioni la somma della diagonale principale

0,25 + 0,02 + 0,02 = 0,29 è la proporzione totale osservata **Po** = 0,29

0,20 + 0,09 + 0,06 = 0,35 è la proporzione totale attesa **Pe** = 0,35.

quindi l'indice proposto da Cohen è:

$$k = \frac{Po-Pe}{1-Ae} = \frac{0,29-0,35}{1-0,35} = \frac{-0,06}{0,65} = \mathbf{-\ 0,0923}$$

Con le frequenze assolute è possibile una stima più semplice e rapida.

Dopo aver calcolato

- le frequenze osservate **Fo** = 50 + 4 + 4 = 58 (nella prima tabella)
- e quelle attese **Fe** = 40 + 18 + 12 = 70 (nella tabella successiva), calcolate moltiplicando il totale dell'attribuzione delle categorie delle frequenze assolute e dividendo per il totale; per la prima cella 1,1(A,A) il valore sarà dato da 100*80=8000/200=40, per la seconda 2,2(B,B) da 60*60=3600/200=18, per la terza 3,3(C,C) da 40*60=2400/200=12.

		MEDICO1			
	CATEGORIE	A	B	C	**Totale**
MEDICO 2	A	**40**	30	30	100
	B	24	**18**	18	60
	C	16	12	**12**	40
	Totale	80	60	60	200

utilizzando solo i valori collocati sulla diagonale principale il calcolo dell'indice k diventa:

$$k = \frac{Fo-Fe}{N-Fe} = \frac{58-70}{200-70} = \frac{-12}{130} = -0{,}0923$$

In entrambi i casi il valore ottenuto è **k= - 0,0923**. Ciò significa che i due medici si trovano d'accordo su una proporzione di casi che è minore di quella che si sarebbe ottenuta con una attribuzione casuale dei pazienti alle varie categorie.
In conclusione, i due medici forniscono valutazioni tendenzialmente contrapposte.
Cohen definisce i coefficienti π, k, α "coefficienti di accordo per scale nominali", ciò significa che sono coefficienti con valori compresi tra -1 e +1.
E' facilmente deducibile che l'indice k è rilevante solo quando ha un valore positivo, in questo caso con **k= - 0,0923** l'accordo tra i medici è molto basso.

1.3 Problemi relativi l'intercoder agreement

Precedentemente è già stata espressa l'importanza dell'annotazione linguistica.
Un corpus annotato è quindi simile al risultato di un esperimento scientifico e l'annotazione può essere considerata valida solo se è *riproducibile* - cioè, se gli stessi risultati annotati possono essere replicati in un esercizio indipendente di codifica. Krippendorff[12] sostiene pertanto che qualsiasi ricerca utilizzi un accordo osservato come misura della riproducibilità deve soddisfare i seguenti requisiti[13]:

- ➔ deve assumere un esaustivo schema di annotazione, definendo regole e criteri

[12] cfr. Content Analysis: An introduction to Its Methodology,second edition,Sage,Thousand Oaks CA
[13] cfr. Inter-Coder Agreement for Computational Linguistics, Ron Artstein, Massimo Poesio

da adottare durante l'annotazione.

- ➔ deve fondarsi su criteri chiaramente specifici in merito alla scelta degli annotatori.
- ➔ deve garantire l'indipendenza del lavoro tra gli annotatori.

Nell'ambito dell'annotazione linguistica queste linee guida non sempre sono rispettate, nello specifico:

1. Il primo requisito è violato dall'uso di inserire nuove regole di annotazione nello schema di codifica già esistente e definito prima dell'inizio del lavoro di annotazione, ciò rappresenta un limite per la riproducibilità e l'affidabilità dello schema.
2. Il secondo requisito è spesso violato, utilizzando annotatori esperti che conoscono lo scopo dello studio, il che rende praticamente impossibile per altri riprodurre i risultati sulla base dello stesso schema di annotazione.
3. Pratiche che violano il terzo requisito (indipendenza) includono annotatori che si consultano durante il lavoro di annotazione discutendo le scelte anche quando sorgono problemi non previsti nelle istruzioni di annotazione.

La violazione di queste regole non permette a volte di raggiungere dei risultati soddisfacenti.

Tuttavia, ci sono delle tabelle di valutazione per stimare l'accordo e la riproducibilità tra vari annotatori.

La tabella seguente è stata proposta da Landis e Koch nel 1977.

Kappa	Agreement
< 0.00	Nessun accordo
0.00–0.20	Lieve accordo
0.21–0.40	Accordo equo
0.41–0.60	Accordo moderato
0.61–0.80	Accordo sostanziale
0.81–1.00	Accordo perfetto

Dopo aver stabilito i valori per ogni accordo è necessario interpretare i valori ottenuto dallo studio. Nel caso del test effettuato da Cohen con k= - 0,0923 non risulta nessun accordo tra i medici.

1.4 Polisemia e annotazione

La polisemia è stata al centro di diversi studi ed esperimenti che hanno evidenziato l'influenza che questa caratteristica può avere nella valutazione dell'intercoder agreement.

Gli studi riportati di seguito sono stati condotti in tre edizioni dall'organizzazione internazionale per la valutazione di Sistemi di Disambiguazione Semantica, SENSEVAL. [14]

In un primo momento sei annotatori sono stati invitati ad analizzare circa 600 parole francesi (duecento nomi, duecento verbi, duecenti aggettivi) utilizzando il repertorio dei sensi fornito dal Petit Larousse.

I risultati riportati da questo primo studio e analizzati da Veronis (1998) non sono stati soddisfacenti, infatti è stato osservato un accordo percentuale di 0,68 per i nomi, 0,74 per i verbi, e 0,78 per gli aggettivi, corrispondenti ai valori k di 0,36, 0,37 e 0,67.

In un secondo studio sulla polisemia, condotto sempre da Veronis[15], le parole percepite dai soggetti come maggiormente polisemiche sono state nuovamente analizzate dai sei annotatori che hanno avuto la possibilità di assegnare ai termini più significati.

Nonostante questa possibilità anche in questo caso utilizzando il coefficiente k per il calcolo dell'accordo i valori riscontrati risultavano bassi.

I valori riscontrati sono stati di 0,63 per i verbi, 0,71 per gli aggettivi e 0,73 per i nomi, tradotto in valori di k 0,41 (verbi), 0,41 (aggettivi) e 0,46 (nomi).

Anche nelle due successive edizioni di SENSEVAL, gli studiosi Veronis in un primo momento e Mihalcea Chklovski e Kilgarriff in un secondo pur utilizzando come tag per i verbi i sensi forniti dalla risorsa WordNet hanno evidenziato nel primo caso un accordo (calcolato utilizzando il coefficiente k) del 70% e nel secondo del 67,3%

[14] cfr. Inter-Coder Agreement for Computational Linguistics, Ron Artstein, Massimo Poesio
[15] cfr. Véronis, J. 1998. A study of polysemy judgments and inter-annotator agreement. In Proc. of SENSEVAL-1-

con un valore medio del k intorno allo 0,58.
Il problema dei bassi valori di accordo non ha convinto gli studiosi e inizialmente sono state proposte due soluzioni:

- la prima proposta è stata avanzata da Kilgarriff che consigliava di assumere lessicografi professionisti, in quanto ciò poteva consentire un rialzo dell'accordo percentuale di quasi il 95,5%.
- la seconda proposta è l'introduzione di schemi di classificazione con categorie più ampie che raggruppano i sensi del dizionario, per faciltare gli annotatori e "limitare" il campo di scelta di un "senso" per un termine.

La seconda proposta si è rivelata efficace in quanto un esperimento condotto da Palmer, Dange Fellbaum, proponendo uno schema di raggruppamento sensi per il verbo "call" (riferiti alla risorsa WordNet 1.7) ha portato ad un incremento dell'accordo dell'82% in contrapposizione al 71% delle precedenti annotazioni.

Si riporta di seguito lo schema di raggruppamento utilizzato per lo studio[16].

SENSE	DESCRIPTION	EXAMPLE	HYPERNYM
WN1	name, call	"They named their son David"	**LABEL**
WN3	call, give a quality	"She called her children lazy and ungrateful"	**LABEL**
WN19	call, consider	"I would not call her beautiful"	**SEE**
WN22	address, call	"Call me mister"	**ADDRESS**

Tuttavia il calcolo del coefficiente di accordo in caso di polisemia ha sollevato molte discussioni. Diverse proposte e metodologie sono state avanzate, tra le quali si ricorda lo studio di Melamed e Resnik (2000) che hanno sviluppato un metodo per il calcolo con K considerando tagsets gerarchici.

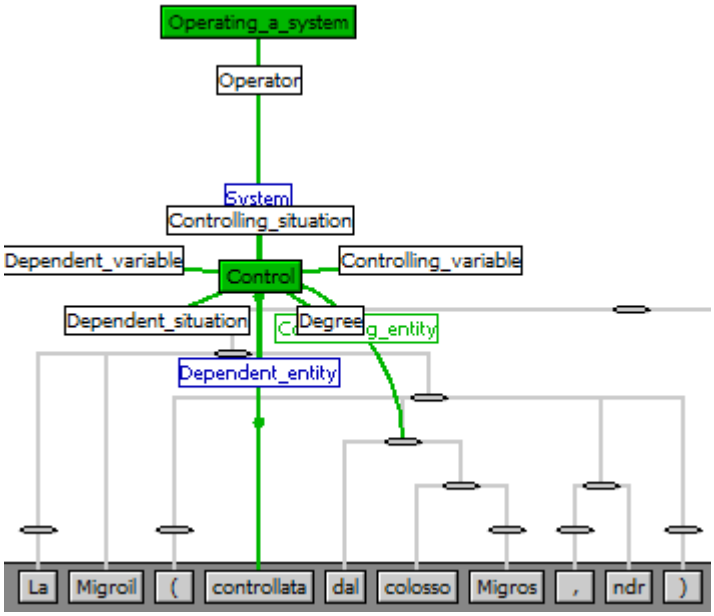
[16] cfr. Inter-Coder Agreement for Computational Linguistics, Ron Artstein, Massimo Poesio

CAPITOLO II
ITALIAN FRAMENET

2.1 Presentazione lavoro

Il lavoro di tirocinio si è concentrato sull'annotazione semantica di brevi sub-corpus in ambito giornalistico di circa cento frasi ciascuno tratti dal principale corpus ISST.
Il corpus ISST (Italian Syntactic-Semantic Treebank) sviluppato tra il 1999 e il 2001 rappresenta una risorsa molto importante nel campo del natural language processing.
Il corpus, già annotato a livello morfosintattico e sintattico a dipendenze, contiene una selezione di articoli giornalistici provenienti da diverse testate italiane (La Repubblica, il Corriere della Sera, il Sole 24Ore) e periodici che sono stati selezionati per coprire una varietà elevata di argomenti quali la politica, l'economia, la cultura, la salute, lo sport e il tempo libero. Gli articoli selezionati sono stati pubblicati nelle rispettive testate tra il 1985 e il 1995.
Il lavoro svolto all'interno del tirocinio prevedeva l'annotazione semantica di circa cinquecento frasi, estratte dal corpus ISST e annotate in stile FrameNet utilizzando il software di annotazione SALTO, un software sviluppato appositamente per l'annotazione manuale di ruoli e classi semantiche.
Il software contiene già un elenco di frames disponibili per l'assegnazione; in caso di assenza nell'elenco su SALTO, è possibile aggiungerlo manualmente utilizzando il database presente sul sito ufficiale di FrameNet.
Annotare in stile FrameNet significa associare al frame proposto i ruoli semantici corrispondenti. Il sito ufficiale di FrameNet è stata una risorsa essenziale e preziosa per la ricerca e la scelta dei ruoli semantici da associare a ciascuna frase.
Scegliere un frame o un ruolo semantico non è un compito semplice, in particolare quando non è ben chiaro il contesto in cui il ruolo deve essere inserito e si può cadere nell'errore di percepire il senso della frase diversamente da ciò che realmente esprime.
In questo senso FrameNet è stato illuminante poichè associa al frame e al ruolo semantico una breve descrizione dello stesso chiarendo all'annotatore eventuali dubbi e aiutandolo a scegliere in modo accurato un frame o ruolo semantico anziché un altro.

Nell'esempio sottostante l'annotatore ha associato al verbo due frame:



Prima di effettuare una scelta è opportuno consultare le descrizioni di ciascun frame sul sito di Framenet, in questo caso:

Control

Lexical Unit Ir

Definition:

A Controlling_entity, Controlling_situation, or Controlling_variable control a Dependent_entity, Dependent_situation, or Dependent_variable. The latter, dependent, element or some aspect of it is not just influenced, but determined by the controlling element.

Can you CONTROL the robot?

Although you may not be able to CONTROL every situation, you can control your reaction.

Students do not *have* CONTROL over the temperature of the heating water.

Everyone knows that the price of oil CONTROLS the price of asphalt.

Operating_a_system

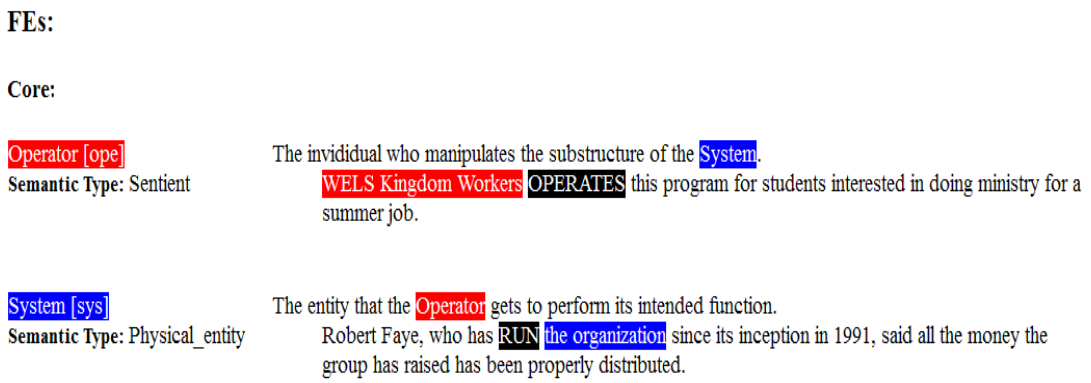
Definition:

An Operator manipulates the substructure of a System such that the System performs the function it was created for. This frame di frame does not profile the purpose of an agent but rather their manipulation of an entity (System/Instrument). In the case of using, Instrument at all, as in: Mrs. Adams used the room as a place to hang the family laundry to dry.

This true story of domestic bliss was related to me by a friend of mine who for a time OPERATED a bar in San Francisco.

A California man has plead guilty to federal charges that he RAN a pyramid scam which defrauded nearly 7000 investors.

un esempio di ruoli semantici per il frame “operating_a_system”



Dopo aver letto le descrizioni dei frame e aver visionato i ruoli semantici disponibili risulta più semplice effettuare una scelta.

Prima dell'inizio del lavoro sono stati stabiliti dei criteri da adottare per l'annotazione, in particolare per l'assegnazione dei frames -sono state assegnate classi semantiche solo ai verbi- e dei ruoli semantici (scelte particolari per l'annotazione di pronomi riflessivi e clitici).

Gli annotatori che hanno preso parte al progetto sono tre, l'individualità e l'indipendenza nello svolgimento dell'annotazione è stato un elemento fondamentale e importante per la valutazione dei risultati raggiunti poiché ogni annotatore ha singolarmente avuto modo di riflettere sulle frasi da annotare, scegliendo i ruoli semantici da associare al frame in base alla propria percezione e seguendo un proprio ragionamento logico.

2.2 Software di annotazione SALTO

SALTO è un software che permette l'annotazione e la gestione di corpora.

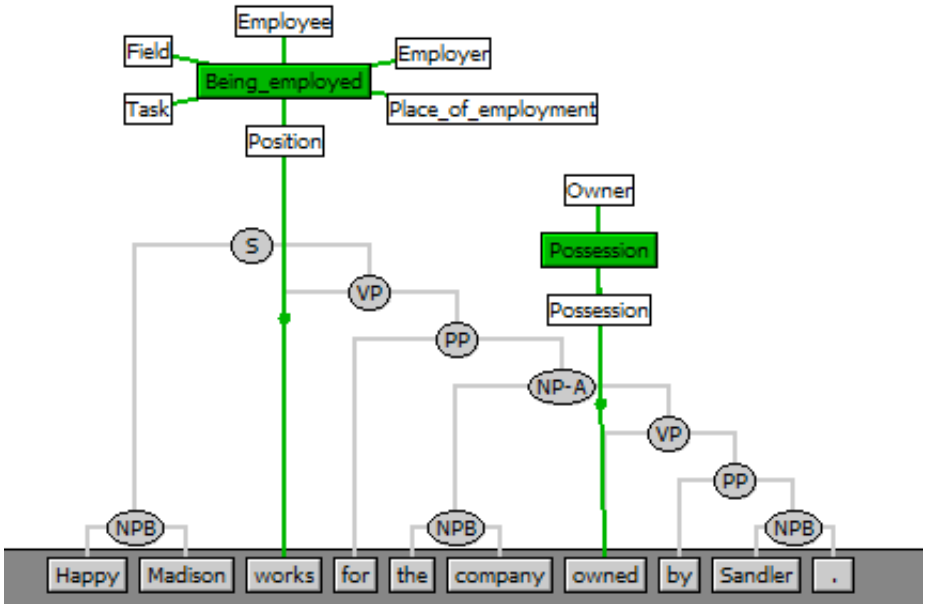
E' stato sviluppato e utilizzato all'interno del progetto progetto SALSA (2002) -Saarbrucken Lexical Semantics Annotation and Analysis- i cui scopi principali sono:

- ➔ creare un'esaustiva risorsa di annotazione semantica utilizzando FrameNet e i frame semantici che fornisce.
- ➔ Favorire lo sviluppo di modelli per l'analisi semantica e la loro applicazione nell'ambito del Natural Language Processing (NLP).

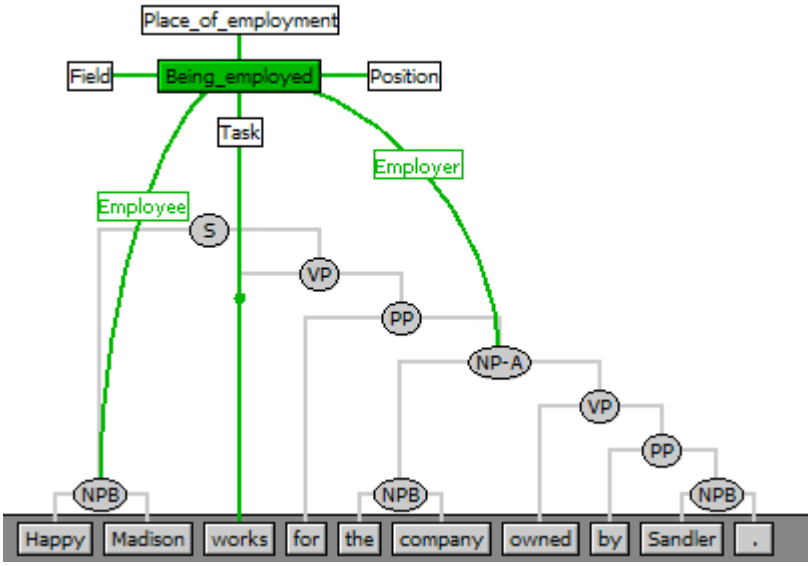
Il software SALTO permette, dato un corpus di dati in input di aggiungere ad un primo esistente livello di annotazione sintattico (tramite annotazione manuale o automatica) un secondo livello di annotazione semantica.
Le caratteristiche principali del programma permettono[17]:

- ➔ creazione di subcorpora per l'annotazione;
- ➔ distribuzione dei corpora di diversi annotatori;
- ➔ definizione di oggetti e classi per l'annotazione;
- ➔ annotazione con editor visuale;
- ➔ gestione corpus e analisi dei dati contenuti.

L'annotazione avviene tramite un ambiente grafico in cui data una serie di frasi, -ognuna identificata da id univoci- è possibile aggiungerne il frame e attraverso la tecnica drag-and-drop assegnare i corrispondenti ruoli semantici.
Di seguito due esempi di annotazione, il primo solo con frame e il secondo con l'aggiunta dei ruoli semantici[18].



[17] cfr.Documentation of SALTO-Salsa project Saarbrucken
[18] cfr.Example_corpus - SALTO



Nell'esempio riportato è utilizzato il frame “Being_employed” per il verbo works che costituisce il nodo principale e dai ruoli semantici “Employee” e “Employer” riferiti a “Happy Madison” e “the company owned by Sandler”.

SALTO non si limita solamente alle annotazioni individuali ma, nei casi in cui il database contiene più annotazione permette di confrontarle tra loro e visualizzare le differenze di annotazioni tra i vari annotatori (inter-annotator disagreement).

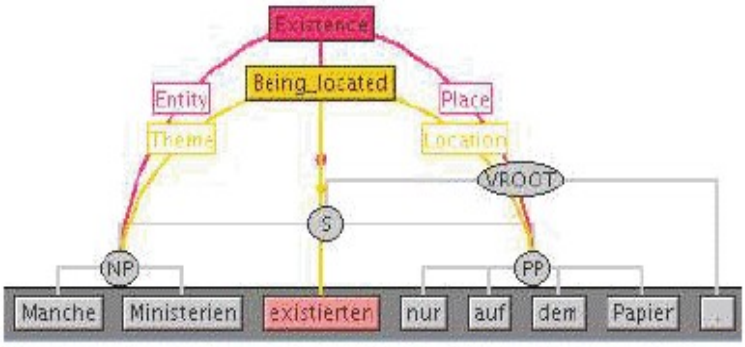
Questo processo si divide in due fasi.

- ➔ in una prima fase si crea un solo corpus integrando le annotazioni dei due annotatori evidenziandone le differenze con colori diversi.

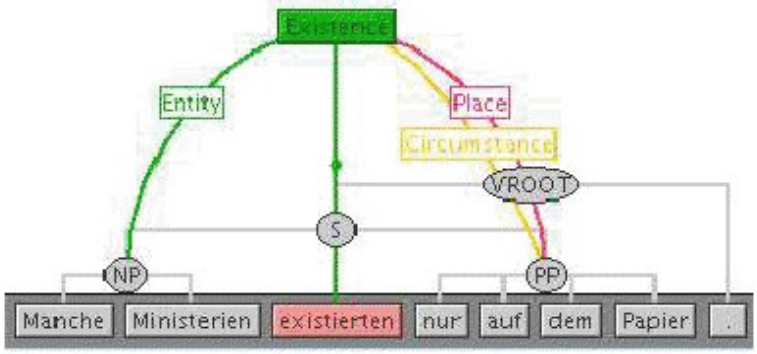
  Le annotazioni identiche in entrambi i corpora sono rappresentate allo stesso modo nel corpus risultante.
- ➔ nella seconda fase “di correzione” è possibile effettuare una selezione sui frames proposti e valutare se mantenerle entrambe o sceglierne solo uno.Nel caso in cui nessuno dei frames proposti risulti accettabile, entrambe le versioni possono essere cancellate, è a discrezione degli annotatori scegliere la soluzione più idonea.

Nelle immagini proposte è possibile vedere come nel primo caso i due annotatori hanno interpretato e annotato la stessa frase utilizzando frame e ruoli semantici

diversi[19]:



Nel secondo caso seppur con lo stesso frames hanno utilizzato ruoli semantici diversi, “Place” e ”Circumstance”:



Talvolta, nel caso in cui fosse necessario aggiungere un frame per mancanza nella lista fornita dal software o perchè nessuno dei frame presenti risulta idoneo è possibile inserirlo manualmente attingendo alla risorsa FrameNet che, anche in questo caso si rivela preziosa poichè è possibile analizzare attraverso le dettagliate descrizioni molti frame e scegliere quello più idoneo alla frase da annotare in base al contesto e all'interpretazione che se ne da.

Durante l'annotazione può capitare di non trovare il frame adatto all'annotazione o al contrario notarne più di uno idoneo; in questo caso per evitare una scelta sbagliata SALTO permette l'annotazione con più frame cosi da consentire all'annotatore in un secondo tempo di poter valutare e scegliere.

Questo aspetto è rilevante non solo per la qualità l'annotazione ma anche per indagini statistiche poiché per avere delle buone valutazioni è importante che alla base ci sia

[19] cfr.Documentation of SALTO-Salsa project Saarbrucken

una buona annotazione.

Le funzionalità di SALTO offrono la possibilità di tracciare delle statistiche sul lavoro svolto da due o più annotatori "confrontando" le varie annotazioni, è possibile sapere quanti e quali frame sono stati utilizzati dagli annotatori, quante frasi sono state annotate allo stesso modo e quante in modo differenti e scegliere l'ordine di visualizzazione.

La raccolta di questi dati è finalizzata anche ad altri scopi statistici quali ad esempio calcolare l'indice di accordo tra i vari annotatori (intercoder agreement).

CAPITOLO III
ESPERIMENTO SU ITALIAN FRAMENET

3.1 Calcolo del coefficiente di accordo Kappa

Il lavoro di annotazione è stato realizzato da tre annotatori che hanno annotato utilizzando frame semantici un minimo di cinquecento frasi ciascuno. Ai fini del calcolo del coefficiente di accordo è stato selezionato un campione di frasi su cui fare l'annotazione congiunta dei tre annotatori.

Il coefficiente utilizzato per calcolare l'intercoder agreement è il kappa di Cohen.
I dati emersi dallo studio delle annotazioni sono i seguenti:

- ➔ 420 frasi sono state annotate dai tre i codificatori.
- ➔ 1606 i frames utilizzati per le annotazioni verbali suddivisi per tipologia:
  - 16 frame che indicano stati mentali
  - 82 frame che indicano il compimento di azioni
  - 8 frame che indicano situazioni e spostamenti

Nella tabella sono rappresentati cinque frame di esempio appartenenti alle diverse categorie del campione di dati esaminato. Per semplicità sono cosi suddivisi:
Tipo 1: frame che indicano stati mentali
Tipo 2: frame che indicano il compimento di azioni
Tipo 3: frame che indicano situazioni e spostamenti

Tipo 1	**Tipo 2**	**Tipo3**
Cogitation	Chatting	Encounter
Desiring	Discussion	Arriving
Expectation	Statement	Departing
Feeling	Communication	Hostile_encounter
Memory	Seeking	Robbery

Ad ogni frame è stato assegnato ai fini della valutazione dei dati una categoria e punteggio, corrispondente al numero di annotatori che lo hanno assegnato.

Di seguito riportiamo un esempio del procedimento svolto per l'analisi esamimando due situazioni differenti:

- ➔ i tre annotatori hanno assegnato lo stesso frame alla stessa categoria;
- ➔ i tre annotatori hanno assegnate frame differenti;

Nel primo caso i tre annotatori hanno assegnato ad un verbo lo stesso frame:
Il primo annotatore ha assegnato tre istanze di frame:
lex_unit_1->"Abandonment"
lex_unit_2->"Activity"
lex_unit_3->"Age"

Il secondo annotatore due istanze:
lex_unit_1->"Abandonment"
lex_unit_2->"Activity_ongoing"

Il terzo annotatore, quattro istanze:
lex_unit_1->"Abandonment"
lex_unit_2->"Activity_ongoing"
lex_unit_3->"Age"
lex_unit_4->"Arriving"

E' evidente che i tre annotatori hanno delle lexical unit "in comune" e altre annotate in maniera diversa. Ogni lexical unit rappresenta una categoria il cui valore sia numerico sia letterale (le categorie rappresentano sempre un nuovo frame) cambia ogni qualvolta si analizza una nuova frase; il valore di ogni lexical unit assegnata da un annotatore è 1, il massimo valore ottenibile per ogni categoria è 3, in questo caso significa che i tre annotatori hanno assegnato lo stesso frame allo stesso verbo.
Per il lavoro oggetto di analisi le categorie sono cinque poiché il numero massimo di lexical unit utilizzati per ogni frase è stato questo.

In questo esempio i valori assegnati per ciascuna categoria saranno i seguenti:
Categoria 0: Abandonment

Categoria 1: Activity

Categoria 2: Activity_ongoing

Categoria 3: Age

Categoria 4: Arriving

La matrice risultante è formata dalle categorie e dalle lexical unit assegnate. Considerando che ogni lexical unit assegnata da un annotatore vale 1 si ha:

lexical unit 1: Abandonment, è stata assegnata da tutti e tre gli annotatori il suo valore sarà 3, e sarà posizionato nella corrispondente cella.

lexical unit 2: **Activity_ongoing**, è stato assegnato da due annotatori su tre, il valore di questo frame è 2; il terzo annotatore ha assegnato alla stessa lexical unit "**Activity**" quindi si assegna il valore 1.

lexical unit 3: Age, solo due annotatori hanno assegnato questo frame quindi il valore è 2.

lexical unit 4: **Arriving**, solo un annotatore ha assegnato la lexical unit 4 per questo il valore è 1 e si posiziona in corrispondenza dalla lexical unit 4.

	Categoria 0	Categoria 1	Categoria 2	Categoria 3	Categoria 4	Totale
LU 1	**3**					
LU 2		**2**	**1**			
LU 3				**2**		
LU 4					**1**	
Totale						

Si supponga che nella frase successiva la situazione cambi e i tre annotatori abbiano annotato per la stessa lexical unit frames differenti:

Il primo annotatore ha assegnato una istanza di frame:

lex_unit_1->"Memory"

Il secondo annotatore una istanza:
lex_unit_1->"Remembering_experience"

Il terzo annotatore, una istanza:
lex_unit_1->"Remebering_information"

Rispetto alla situazione precedente il valori di assegnazione alla categoria 0 cambia:
Categoria 0: Memory

Cambia il valore della lexical unit poiché:
lexical unit 1: Memory per il primo annotatore
Remembering_experience per il secondo
Remembering_information per il terzo

In questo caso non avendo assegnato lo stesso frame il valore da inserire nella matrice non sarà più 3 come per il caso precedente ma 1-1-1 in quanto pur annotando la stessa lexical unit sono state fatte assegnazioni differenti. I valori in matrice sono cosi assegnati:

	Categoria 0	Categoria 1	Categoria 2	Categoria 3	Categoria 4	Totale
LU 1	**1**	**1**	**1**			
LU 2						
LU 3						
LU 4						
Totale						

Il procedimento di analisi e assegnazione illustrato per i due casi di esempio è lo stesso utilizzato per tutte le frasi. I valori di ogni cella mutano costantemente in base alle assegnazioni di frame poiché può capitare che in una frase gli annotatori assegnino lo stesso frame quindi il valore di una lexical unit sarà 3 mentre nella frase successiva annotino frame diversi.

Ai fini del calcolo dell'intercoder agreement i valori delle celle tra le varie fasi si sommano. Per questo motivo il risultato per le due frasi di esempio analizzate è:

	Categoria 0	Categoria 1	Categoria 2	Categoria 3	Categoria 4	Totale
LU 1	**4**	**1**	**1**			
LU 2		**2**	**1**			
LU 3				**2**		
LU 4					**1**	
Totale						

Nella prima cella categoria 0-lexical unit 1 il valore 4 è dato dalla somma della lexical unit assegnata nel primo esempio (3) e della seconda (1). La stessa metodologia si applica per i valori contenuti nelle altre celle.
I valori inseriti per ogni categoria sono dati dal totale di frame che una coppia o un singolo annotatore ha assegnato per quella determinata categoria.
Il metodo utilizzato per stimare questi valori è stato il confronto di ogni singola frase annotata da tutti e tre gli annotatori e dai frame assegnati.

Utilizzando questa metodologia e queste linee guida nel lavoro oggetto di analisi i risultati sono stati i seguenti:

	Categoria 0	Categoria 1	Categoria 2	Categoria 3	Categoria 4	Totale
LU 1	**787**	332	45	0	0	1164
LU 2	0	**230**	80	3	0	313
LU 3	0	0	**80**	21	0	101
LU 4	0	0	0	**25**	3	28
Totale	787	562	205	49	3	1606

E' necessario trasformare i valori ottenuti in frequenze osservate e attese, queste ultime riportate in corsivo. Le prime si ottengono dividendo il valore di ogni cella per il totale, ad esempio per la categoria 0 sarà 787/1606= 0,49; le frequenze attese si ottengono moltiplicando il totale dell'attribuzione alle categorie delle frequenze

assolute nella condizione che l'attribuzione dell'annotatore alla categoria sia stata casuale, per la categoria 0 sarà 0,72*0,49=0,35.

	Categoria 0	Categoria 1	Categoria 2	Categoria 3	Categoria 4	Totale
LU 1	**0,49***(0,35)*	0,20	0,02	0,000	0,000	0,72
LU 2	0,000	**0,14** *(0,06)*	0,04	0,001	0,00	0,19
LU 3	0,000	0,00	**0,04** *(0)*	0,01	0,00	0,06
LU 4	0,000	0,00	0,00	**0,01** *(0)*	0,001	0,01
Totale	0,49	0,34	0,12	0,03	0,00	1,00

Sommando le frequenze ottenute si ha:

Frequenze osservate (Fo): 0,49+0,14+0,04+0,01=**0,68**

Frequenze attese (Fe): 0,35+0,06+0+0= **0,41**

L'equazione per calcolare l'indice è:

$$k=\frac{Fo-Fe}{1-Fe}$$

applicando i valori ottenuti si ha:

$$k=\frac{0,68-0,41}{1-0,41}=\frac{0,27}{0,59}=\mathbf{0,45}$$

Il valore del coefficiente kappa è 0,45.

Confrontando il valore con la tabella proposta da Landis e Koch nel 1977 il valore risultante da questo esperimento rispecchia un accordo moderato fra i tre annotatori.

Kappa	Agreement
< 0.00	Nessun accordo
0.00–0.20	Lieve accordo
0.21–0.40	Accordo equo
0.41–0.60	Accordo moderato
0.61–0.80	Accordo sostanziale
0.81–1.00	Accordo perfetto

I valori di k sono stati calcolati solo sulle frasi annotate da tutti e tre gli annotatori. Sono state ecluse dalla valutazione quelle annotate da soli due annotatori su tre, per due motivi principali, sia per non alterare il valore del coefficiente considerando che includere le frasi annotate da sole due persone e frasi annotate da tre non era coerente con lo scopo dell'esperimento e sia perchè un metodo simile non avrebbe evidenziato il valore reale delle annotazioni.

3.2 Analisi dei risultati ottenuti

Oltre il calcolo dell'intercoder agreement è possibile analizzare e quantificare diversi aspetti relativi alle annotazioni effettuate e ai risultati ottenuti.

E' importante considerare per il valore risultante dall'analisi del coefficiente la quantità non indifferente delle frasi annotate. Il campione di frasi utilizzato per il calcolo si limita alle sole frasi annotare da tutti e tre i codificatori ma quelle annotate singolarmente variano da un minimo di 500 frasi a un massimo di 800.

Di seguito riportiamo alcuni dati statistici dei dati relativi al campione di frasi annotato da tutti e tre i codificatori:

- ➔ 420 frasi sono state annotate in totale dai tre i codificatori. Tra queste:
  208 frasi sono state annotate allo stesso modo da tutti e tre i codificatori;
  212 frasi sono state annotate utilizzando frame diversi.
- ➔ 1606 i frames totali utilizzati per le annotazioni verbali. Di cui:
  1164 classificati come lexical unit 1
  313 classificati lexical unit 2
  101 lexical unit 3
  28 lexical unit 4

Volendo rappresentare i dati appena analizzati in percentuale si ha:

49,53% delle frasi sono state annotate allo stesso modo da tutti i codificatori**;**

50,47% delle frasi sono state annotate utilizzando frame diversi;

I risultati raggiunti nel primo caso (frasi annotate o meno allo stesso modo) permettono tre riflessioni sull'annotazione effettuata.

La prima riflessione è **sull'interpretazione** che i codificatori hanno dato alle frasi, i risultati evidenziano una lieve differenza tra le frasi annotate allo stesso modo e quelle annotate in modo differente, ciò significa che per quasi la metà della frasi analizzate pur lavorando indipentemente i codificatori le hanno interpretate allo stesso modo.

La seconda riflessione riguarda il valore del kappa ottenuto (0,45). Dalla tabella proposta da Landis e Koch nel 1977 risulta un accordo moderato fra i tre annotatori.

E' importante analizzare il risultato in misura ai dati sopra proposti; probabilmente con una lieve differenza evidenziata dalle analisi sopra prososte tra le frasi annotate allo stesso modo e quelle annotate in modo differente ci si aspetterebbe un accordo maggiore tra i codificatori.

Quest'affermazione può essere veritiera valutando la quantità dei dati del campione in oggetto. Probabilmente calcolando l'intercoder agreement con la stessa situazione sopra proposta su un campione di dati minore il valore sarebbe stato maggiore.

La grandezza del campione di dati influenza tutta l'analisi del calcolo è dimostrato anche nello studio effettuato da Veronis (1998). In un primo momento sei annotatori sono stati invitati ad analizzare circa 600 parole francesi (duecento nomi, duecento verbi, duecenti aggettivi) utilizzando il repertorio dei sensi fornito dal Petit Larousse. I risultati riportati da questo primo studio e analizzati da Veronis (1998) non sono stati soddisfacenti, infatti è stato osservato un accordo percentuale di 0,68 per i nomi, 0,74 per i verbi, e 0,78 per gli aggettivi, corrispondenti ai valori k di 0,36, 0,37 e 0,67.

La terza riflessione riguarda la complessità intrinseca della nozione stessa di frame semantico.Poiché il frame semantico è la rappresentazione di una situazione, di uno stato o di un evento, la scelta di un frame anziché un altro è a discrezione dell'annotatore e basata sull'interpretazione personale che se ne da. Questo aspetto determina maggior difficoltà nella scelta che viene fatta in ciascuna frase.

E' chiaro come l'annotazine semantica, diversamente dagli altri tipi di annotazione che risultanto essere più “oggettivi” (per esempio in un'annotazione sintattica la componente personale è del tutto assente poiché non bisogna interpretare una situazione ma analizzare sintatticamente un testo) sia influenzata dalla soggettiva interpretazione.

Un'ulteriore analisi sul campione delle annotazioni effettuate riguarda i frames più e meno utilizzati. Per semplicità si riportano di seguito i primi dodici risultati di entrambe le categorie.
Sono considerati più utilizzati i frame ricorrenti in più di quattro frasi:

Appearance è stato il frame più utizzato, ricorre in 40 frasi.
Statement è stato utilizzato in 32 frasi
Cogitation 29 frasi
Desiring 29 frasi
Categorization 27 frasi
Opinion 16 frasi
Trust 16 frasi
Memory 15 frasi
Referring name 10 frasi
Remembering information 10 frasi
Arriving 9 frasi
Becoming-aware 9 frasi
Remembering experience 9 frasi
Being name 9 frasi.
[...]

I frame meno utilizzati (da una a quattro frasi) risultano essere:

Evoking utilizzato in 1 sola frase
Making face 1 frase
Placing 1 frase
Questioning 1 frase
Event 1 frase
Visiting 1 frase
Robbery 1 frase
Discussion 2 frasi
Evidence 2 frasi

Waiting 2 frasi
Assistence 2 frasi
Hear 2 frasi
[….]

E' chiaro che la scelta del frame da utilizzare è stata influenzata dalla percezione e dall'intepretazione della frase. Dall'analisi è evidente che tra i frame più utilizzati nelle prime sei posizioni ci sono frames che indicano condizioni mentali, quali pensare(cogitation)e desiderare(desiring) mentre i frame meno utilizzati sono stati per lo più quelli che indicano azioni, ad esempio rubare(robbery) e aspettare(waiting).
E' evidente che nel campione di dati analizzato sono stati riscontrate delle difficoltà nell'assegnare i frame su alcuni verbi, gli annotatori hanno interpretato in modo differente il verbo nel contesto della frase assegnando conseguentemente frame diversi. Un esempio è la frase seguente:

E' improbabile che gli abitanti di Knin , sui quali prima della resa pioveva un proiettile d' artiglieria ogni dieci secondi, abbiano ***ricordato*** *l' anniversario dell'atomica di Hiroshima.*

In questo caso il verbo annotato è stato “ricordare”, tutti e tre gli annotatori hanno cosi assegnato i frame:
Annotatore 1: Memory
Annotatore 2: Remembering_experience
Annotatore 3: Remembering_information
Dall'analisi emerge che tutti e tre hanno interpretato il verbo “ricordare” come un elemento che evoca un ricordo, tutti e tre i frame evocano questa esperienza ma in modo diverso. “Memory” e “Remembering_experience” sono frame che evocano esperienze sulla sfera personale mentre “Remebering_information” evoca un ricordo di informazioni quindi ha un significato più generale e distaccato.
Su questa base è possibile delineare -utilizzando i dati del campione di annotazioni in oggetto- una serie di casi analoghi in cui ad uno stesso verbo sono stati assegnati tre

frame diversi. Di seguito una tabella riassuntiva dei casi, con le voci “Assegnazione 1”, “Assegnazione 2” e “Assegnazione 3” si indicano le assegnazioni fatte dai tre annotatori. Per facilità nella lettura ogni verbo è stato riportato al modo infinito.

Verbo	**Assegnazione 1-E**	**Assegnazione 2-F**	**Assegnazione 3-M**
VEDERE	Seeking	Perception_active	Becoming_aware
CONOSCERE	Make_acquitance	Becoming_aware	Perception_active
PARLARE	Chatting	Discussion	Speak_on_topic
CONOSCERE	Name_conferral	Being_named	Familiarity
RICORDARE	Evoking	Similarity	Remembering information
PARLARE	Communication	Speak in topic	Statement
RICORDARE	Memory	Evoking	Remembering experience
CHIAMARE	Name_conferral	Referring_by_name	Being_named
VINCERE	Finish_competition	Getting	Finish_game
RICORDARE	Memory	Remebering information	Remembering experience
PENSARE	Categorization	Opinion	Cogitation
PARLARE	Topic	Speak_on_topic	Statement
CHIAMARE	Name_conferral	Referring_by_name	Being_named
PRESENTARE	Cause_to_perceive	Have_associated	Presence
NASCERE	Coming_to_be	Achieving_first	Origin
SPARARE	Hit_target	Use_firearm	Shoot_projectiles
CREARE	Intentionally_create	Building	Creating
PARLARE	Chatting	Discussion	Communication

Alcuni verbi sono stati riportati due volte poiché ripresentandosi in contesti diversi hanno assunto significati diversi e conseguentemente anche frame diversi.

Dall'analisi è evidente la difficoltà di assegnare dei frame comuni ad alcuni verbi, si notino verbi come “parlare” e ”ricordare” che pur ripresentandosi più volte hanno evidenziato come gli annotatori non sono riusciti ad assegnare dei frame comuni pur evidenziando la sfera di appartenenza del verbo.

A esempio per il verbo “ricordare” tutti e tre gli annotatori concordano nella natura evocata dal verbo, il ricordo, eppure pur concordando su ciò sono stati assegnati tre

frame diversi.

RICORDARE	Memory	Evoking	Remembering experience

E' lecito chiedersi il perchè pur risultando un senso comune per il verbo l'attribuzione risulti diversificata.

Un primo motivo può essere senza dubbio l'interpretazione data ad una prima lettura della frase.

Un secondo motivo potrebbe essere un'iniziale indecisione nella scelta chiarita successivamente da un confronto e da uno studio dei frame singolarmente. Ad esempio si supponga che l'annotatore 1 abbia ritenuto ad una prima lettura della frasi idonei due frame “Memory” e “Evoking”, consultando il sito ufficiale di FrameNet abbia letto le rispettive descrizioni e abbia scelto un frame piuttosto che un altro.

In entrambi i casi l'interpretazione e la percezione sono di fondamentale importanza ma mentre nel primo caso l'annotatore si basa completamente su questa nel secondo caso ha associato alla percezione l'analisi del frame operando una scelta più oggettiva rispetto al primo caso.

Un terzo motivo è che molti dei frame discordanti sono estremamente simili tra di loro, ovvero cercano di catturare sfumature semantiche molto sotttili, tra le quali è dunque facile discordare. Infatti, non capita quasi mai che gli annotatori differiscano tra frames molto diversi tra di loro, es. Motion, vs. Speaking.

CONCLUSIONI

L'intercoder agreement è la misura dell'indice di accordo tra due o più giudici (annotatori nel contesto linguistico) calcolata utilizzando il coefficiente di accordo kappa proposto da Cohen. L'applicazione del coefficiente come dimostrato non si limita ad analisi nel campo linguistico ma anche a quello medico o ambientale o qualsiasi altra situazione in cui si ritenga opportuno elaborare delle statistiche per valutare il grado di accordo.

Il coefficiente kappa nonostante non sia l'unico mezzo per effettuare analisi statistiche tra i principali coefficienti utilizzati nell'ambito della ricerca statistica (S,π e α) risulta essere il più utilizzato. Con l'ausilio di questo coefficiente è possibile analizzare la riproducibilità dei dati nel corso del tempo.

Il calcolo dell'intercoder agreement effettuato utilizzando le annotazioni del progetto Italian FrameNet ha evidenziato i punti salienti dell'equazione di questo coefficiente. Attraverso le categorie e le assegnazioni degli annotatori è stato possibile tener traccia e "schematizzare" un lavoro quantitativamente grande e complesso.

Gli studi possibili da effettuare su un progetto linguistico sono svariate e il calcolo dell'intercoder agreement ne rappresenta solo una piccola parte; il risultato raggiunto dimostra come tra i tre annotatori ci sia un moderato accordo, frutto di un metodo di lavoro e di un rispetto delle regole di annotazione (indipendenza e autonomia nello svolgimento del lavoro) attuate durante tutto il periodo di annotazione.

L'analisi delle annotazioni ha reso possibile uno studio statistico sulle stesse.

Dai risultati emersi è evidente come la grande quantità di dati da un lato abbia influenzato il risultato del coefficiente dall'altro ha permesso di delineare una "classifica" dei frame più utilizzati e scoprire che tra questi per quanto riguarda l'annotazione verbale si tende molto ad utilizzare quelli che indicano uno stato mentale.

Dall'ultima analisi riguardante i verbi che hanno rispecchiato minor accordo tra gli annotatori è evidente come l'assegnazione di frame su verbi come "ricordare" e "parlare" è particolarmente complessa e pone gli annotatori in una condizione di difficoltà nella scelta del frame.

Gli esperimenti di interannotator agreement sono fondamentali per stimare

l’affidabilità e la replicabilità di annotazioni linguistiche, soprattutto per livelli di codifica così complessi come quello semantico.

Questo è determinante per creare risorse annotate affidabili da usare per l’addestramento di strumenti per il trattamento automatico del linguaggio.

BIBLIOGRAFIA

Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli *Testo e computer, elementi di linguistica computazionale, Carocci editore,2005.*

Ron Artstein, Massimo Poesio *Inter-Coder Agreement for Computational Linguistics.*

J.Cohen, *A coefficient of agreement for nominal scales Educational and Psychological Measurement, Vol. XX, No. 1, pp. 37-46.*

Krippendorff *,Content Analysis: An Introduction to Its Methodology, Sage, Beverly Hills, CA. 1980.*

Krippendorff *Content Analysis: An introduction to Its Methodology, Second edition, Sage, Thousand Oaks CA.*

Véronis*, J. 1998. A study of polysemy judgments and inter-annotator agreement. In Proc. of SENSEVAL-1.*

Aljoscha Burchardt, Katrin Erk, Anette Frank,Andrea Kowalski, Sebastian Pado,Manfred Pinkal *Using FrameNet for the semantic analysis of German: annotation, representation and automation pp. 209-240*

Aljoscha Burchardt, Katrin Erk, Anette Frank,Andrea Kowalski, Sebastian Pado *Salto Versatile Multi-Level Annotation Tool.*

Rebecca Passonneau, Nizar Habash, Owen Rambow *Inter-annotator Agreement on Multilingual Semantic Annotation Task.*

Hwee Tou Ng, Chung Yong Lim, Shou King Foo *A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation, DSO National Laboratories 20 Science Park Drive, Singapore 118230.*

SALSA Project Saarbrücken *Documentation of SALTO aka SALSA Tool, September 4,2007*

Melamed, I. Dan and Philip Resnik. *Tagger evaluation given hierarchical tagsets. Computers and the Humanities, 2000, pp.79–84.*

Di Eugenio, Barbara and Michael Glass.. *The kappa statistic: A second look. Computational Linguistics, 2004,* pp**.** *95–101.*

Fleiss, Joseph L. *Measuring nominal scale agreement among many raters. Psychological Bulletin, 1971 , pp. 378–382.*

Alessandro Lenci, Martina Johnson, Gabriella Lapesa, *Building an Italian FrameNet through Semi-automatic Corpus Analysis.*

Alishahi e Stevenson, *A Computational Model for Early Argument Structure Acquisition, 2007.*

SITI WEB

Wikipedia, *L'Enciclopedia Libera Online.*

FrameNet, *https://framenet.icsi.berkeley.edu/fndrupal/.*

Grazie

Alla mia famiglia, per avermi insegnato il valore della cultura e dello studio motivandomi sempre alla conoscenza e all'approfondimento.

Ai miei amici, compagni di viaggio e di emozioni in questo percorso importante di crescita personale e culturale, grazie per aver reso questi anni indimenticabili.

Ad Antonio, per aver sempre creduto in me e per avermi incoraggiata ad andare sempre avanti con forza e tenacia.