UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Estrazione automatica di collocazioni da Wikipedia.it

Candidato: *Patrizia Ghilardi*

Relatore: *Prof. Alessandro Lenci*

Correlatore: *Prof.ssa Maria Simi*

Anno Accademico 2011-2012

Alla mia famiglia, tutta.

Indice generale

Introduzione

Lo scopo di questa tesi è l'estrazione automatica di collocazioni dal corpus costituito da tutti i testi della *Wikipedia* in lingua italiana. Come descritto in modo più approfondito nel capitolo 1, le collocazioni sono coppie di parole (ma possono essere anche più di due) che sono legate all'interno di una stessa frase da rapporti sintattici, si trovano molto frequentemente a ricorrere insieme nel linguaggio e hanno un significato complessivo che può essere anche diverso da quello delle singole parole che le compongono.

Per poter individuare le coppie che avranno più possibilità di essere poi riconosciute come vere collocazioni, si è utilizzata una misura statistica che indica il grado di associazione reciproco delle due parole. Si discuteranno le misure statistiche nella seconda parte del capitolo 1.

Nel primo capitolo, quindi, si ripercorrono a grandi linee le origini e l'evoluzione del pensiero degli studiosi di linguistica riguardo alle collocazioni e dei metodi scientifici con cui è possibile cercare di individuarle (ossia per mezzo della statistica).

Per poter applicare le misure statistiche, è necessario estrarre dal corpus di *Wikipedia* le coppie di parole di cui è necessario valutare il grado di associazione.

Per questo, si sono sviluppati degli script in Perl, che vengono trattati nel capitolo 2. Questi script sono serviti a estrarre dal file di origine del corpus solo alcune tipologie di informazioni. In particolare, si sono estratte solo certe categorie grammaticali di parole che svolgevano anche funzioni sintattiche ben precise all'interno delle frasi in cui risiedevano. Le parole in questione dovevano espressamente far parte della stessa frase ed essere riferite le une alle altre, ossia dovevano essere collegate da dipendenze sintattiche. Nell'ambito di questa tesi si è deciso di estrarre le seguenti categorie di coppie di parole:

- sostantivo e modificatore (aggettivo);
- verbo e soggetto;
- verbo e oggetto;
- verbo e sintagma preposizionale.

Una volta ottenuta l'estrazione delle coppie di parole in modo corretto, è stato necessario calcolare la frequenza con cui ciascuna coppia di parole ricorreva all'interno dei testi in esame, e, in seguito, per mezzo della misura di associazione statistica scelta, calcolarne il grado di associazione reciproca.

Nel capitolo 4 si vedrà come, per mezzo dei risultati dati dalla misura di associazione, si è proceduto nell'analisi dei dati e nell'individuazione di alcune collocazioni che si trovavano all'interno del corpus. Si è cercato di verificare l'efficacia della misura statistica utilizzata controllando un range di 100 coppie di parole per ciascuna categoria di coppie di parole. Una volta ordinato il range in senso decrescente secondo il punteggio della misura statistica usata, si è verificato: quante coppie effettivamente risultavano essere collocazioni, quante erano errori e quante non erano collocazioni nel senso stretto del termine. Inoltre, per agevolare la ricerca di collocazioni all'interno della totalità delle coppie estratte e sottoposte ad analisi statistica, vista la grande quantità dei dati, si è deciso di costruire un piccolo sito in PHP per "navigare" i risultati e procedere al riconoscimento di collocazioni. Il sito web si appoggia a una base di dati, ed è costituito da alcune pagine in cui, per mezzo di un form, è possibile selezionare i criteri della ricerca e digitare le parole di cui si vogliono cercare i collocati.

1. Le collocazioni

1.1 Dalle origini alle multiword expressions

Nel corso dei vari studi di linguistica e linguistica computazionale, molti studiosi hanno trattato le collocazioni e ne hanno dato una definizione delineando le caratteristiche che possono assumere per poterle riconoscere, analizzare e catalogare. Spesso queste definizioni sono rimaste generiche e pertanto è possibile incorrere in fraintendimenti. Tutti gli studiosi però convengono chiaramente su alcuni punti comuni: le collocazioni sono due o più parole che sono legate da rapporti sintattici e grammaticali, tendono a ricorrere insieme più spesso rispetto ad altre e, dal punto di vista semantico, veicolano informazioni aggiuntive rispetto ai singoli significati delle parole che le compongono.

Il primo a introdurre il termine collocazione è stato il linguista inglese J. R. Firth nel 1951. Da allora la definizione è stata ampliata e definita in maniera più specifica, ma è comunque importante vedere le origini di questo termine. Nel 1951, quindi, nell'ambito di uno studio sul significato delle parole, Firth introduce il nuovo concetto di "meaning by collocation" per spiegare che il significato di una parola può variare in base a quelle che la precedono o la seguono; dipende perciò dalle sue collocazioni.[1] In uno scritto successivo, chiarirà questo suo pensiero specificando che "one of the meaning of [a given word] is its habitual collocation" con le parole che si trovano collocate con essa più comunemente[2]. Da qui la sua celebre frase: "You shall know a word by the company it keeps!"[3]. Firth precisa che le collocazioni non vanno intese come contesto, sebbene esso sia molto rilevante ai fini dell'attribuzione di significato delle parole. Infine, egli tenta di dare una definizione della collocazione di una parola, associandola a una sorta di aspettativa reciproca, al fatto cioè che alcune parole si aspettano o in qualche modo prevedono la presenza di altre parole all'interno della stessa frase, manifestando quindi un'attrazione con certe parole piuttosto che con altre.

> "Collocations of a given word are statements of the habitual or customary places of that word (...). The collocation of a word or a 'piece' is not to be regarded as mere juxtaposition, it is an order of mutual expectancy. The words are mutually expectant and mutually prehended. (Firth 1968, p.12)"

1 Firth, J.R. (1957). *Papers in linguistics 1934-1951.* London, Oxford University Press. pp.194-196
2 Firth, J.R. (1968). *A synopsis of Linguistic theory, 1930-1955*. Oxford, Basil Blackwell. p.11
3 Ibidem

Manning e Schütze (1999), sostengono che le collocazioni consistono in due o più parole che corrispondono a un qualche modo convenzionale di esprimere un concetto.[4] Aggiungono però indirettamente, che non è necessario che queste si trovino in posizioni adiacenti all'interno della frase, opponendosi così a ciò che afferma Choueka (1988) secondo il quale le collocazioni sono composte da due o più parole che si trovano in posizione consecutiva all'interno della frase[5]. La definizione di Choueka è infatti limitante e per niente assoluta, poiché è vero che gran parte delle parole che compongono una collocazione spesso si trovano all'interno delle frasi in posizione adiacente, ma è altrettanto vero che possono essere considerate collocazioni anche quelle coppie o triplette di parole che all'interno della frase si trovano persino a una notevole distanza tra di loro. L'estrazione di collocazioni non adiacenti è possibile prendendo in considerazione corpora annotati dal punto di vista sintattico.

L'apporto di conoscenza più significativo che deriva dagli studi di Manning e Schütze è il tentativo di caratterizzazione dei tratti ricorrenti delle collocazioni. Essi identificano alcune caratteristiche che ritengono le collocazioni abbiano nella maggior parte dei casi; si tratta di non-composizionalità, non-sostituibilità e non-modificabilità[6].

La *limitata composizionalità* è intesa come totale o parziale impossibilità nel risalire al significato della collocazione apponendo uno accanto all'altro i significati delle parole che la compongono. Manning e Schütze adducono alcuni esempi per spiegare questa caratteristica: *white wine*, *white hair* e *white woman*. In tutte le tre coppie ricorre la parola "white", ma nei singoli casi l'accezione che viene data è differente per ciascun esempio. Potremmo fare un esempio simile con l'italiano: *energia pulita*, *coscienza pulita* e *faccia pulita* oppure anche *soldi puliti*. In tutti gli esempi, le derivazioni dell'aggettivo "pulito" non possono assumere l'accezione standard di "non sporco" o "lindo". Questo perché: l'energia pulita è energia non inquinante; la coscienza pulita è quella di colui che non ha fatto niente di male ed è tranquillo con se stesso; la faccia pulita è sì una faccia pulita nel senso di non sporca, ma può anche essere intesa come un volto fresco di una persona senza troppi fronzoli; infine, i soldi puliti sono quelli che non provengono dal riciclaggio della malavita.

È chiaro quindi, come nel significato globale delle collocazioni, ci sia spesso un

4 Manning, C. e H. Schütze. (1999). *Foundations of statistical natural language processing.* Cambridge, MIT Press. p.151
5 Ivi. p.183
6 Ivi. p.184

significato aggiuntivo che non è presente nel significato delle singole parti.

La *non sostituibilità* implica che le parole che compongono la collocazione non possano essere sostituite con sinonimi, anche se nel contesto veicolano lo stesso significato; questo è specialmente vero ad esempio per: *acqua dolce*, *dolce attesa* o *dolce metà*, dove l'aggettivo "dolce" non può essere sostituito con melenso o caramelloso, così come in questi esempi anche i sostantivi non possono essere sostituiti con sinonimi.

La *non modificabilità* invece comporta la non alterabilità dal punto di vista grammaticale dei componenti della collocazione e l'impossibilità di includere, all'interno dell'espressione, ulteriori elementi lessicali. È il caso delle espressioni idiomatiche. Per esempio l'espressione *tutto fa brodo* se si modificasse in *tutti fanno brodo** o in *tutto fa buon brodo**, non avremmo più il significato di "tutto serve" e all'orecchio del parlante nativo suonerebbe come sbagliato.

Evert (2009) associa le collocazioni a una tendenza che le parole hanno di ricorrere vicine ad altre. Egli dice infatti che: "it is based on a compelling, widely-shared intuition that certain words have a tendency to occurr near each other in natural language"[7]; il motivo per cui tendono ad occorrere vicino ad altre è che esse "sono caratterizzate da un forte legame di associazione reciproca" (Lenci et al. 2005).

1.2 Multiword expressions

Evert constata come nel corso dei lunghi studi sulle collocazioni, il significato legato a questo concetto abbia assunto un carattere sempre più controverso e ambiguo. Egli ritiene necessario fare una distinzione tra due gruppi di coppie di parole. Il primo gruppo è costituito da coppie di parole che per mezzo di rilevazioni empiriche risultano ricorrere spesso insieme e hanno, dal punto di vista statistico, un'alta probabilità di essere associate nella stessa frase. Del secondo gruppo fanno parte le multiword expressions (MWE), ossia combinazioni di più parole che possono avere valenza idiosincratica, sono lessicalizzate e il loro significato è riconosciuto in modo naturale e automatico da un parlante nativo della lingua in esame[8]. Il termine MWE è di più recente origine rispetto a quello di collocazione e viene utilizzato frequentemente nell'ambito della linguistica computazionale, Nel 2002 le MWE sono state ulteriormente

7 Evert, S. (2009). *Corpora and collocations.* In Corpus Linguistics. An International Handbook (Vol.2). Berlino, Mouton de Gruyter. p.1212

8 Ivi. pp.1213-1214

classificate tra espressioni lessicalizzate ed espressioni istituzionalizzate. A sua volta, la prima tipologia di espressioni, viene ulteriormente suddivisa in:

- *espressioni fisse*, che non possono subire né variazioni morfosintattiche, né aggiunte di modificatori aggettivali o avverbiali (es:*botteghe oscure*, *scatola nera*, *farmaco generico, buon senso*);
- *espressioni semi-fisse*, che restano immutabili per quanto riguarda l'ordine delle parole che le costituiscono, ma possono subire un certo grado di variazione lessicale che comprende l'inflessione e la forma riflessiva (es: *salvare per un pelo*, *conto salato*, *dare una/la mano*, *corpo armato*);
- *espressioni sintatticamente flessibili*, che sono modificabili, oltre che con la flessione, anche con variazioni sintattiche (es: *prendere tra due fuochi*, *essere all'altezza*, *mettere alla prova*, *cogliere di sorpresa*).[9]

Secondo la suddivisione appena citata, le espressioni lessicalizzate sono caratterizzate da almeno una parte semantica o sintattica idiosincratica oppure contengono parole che non occorrono singolarmente ma sono sempre dipendenti o associate ad altre. Le espressioni istituzionalizzate, invece, sono semanticamente e sintatticamente composizionali, hanno acquisito un certo grado di convenzionalità e sono caratterizzate da un'alta frequenza (es: *dente avvelenato*, *benzina verde*, *tetto basso*, *letto matrimoniale*).

In base a questa categorizzazione, Evert ritiene che le combinazioni di parole che si trovano a metà strada tra quelle completamente fisse e quelle completamente flessibili, siano le coppie di parole che corrispondono al termine fraseologico di collocazione: sono semi-composizionali, una delle parole è determinata lessicalmente da quella o quelle che la accompagnano e il suo significato è in qualche modo modificato[10], portando quindi un'accezione leggermente o totalmente diversa dal significato letterale delle parole.

Le tipologie di collocazioni che si possono individuare nei corpora, dipendono strettamente dal registro linguistico e dal dominio del corpus stesso.

Nella ricerca di collocazioni è possibile che tra i risultati delle analisi si vengano a

9 Per tutta la categorizzazione. Sag Ivan A., Timoty Baldwin, Francis Bond, Ann Coperstake e Dan Flickinger. (2002) *Multiword Expressions: A Pain in the Neck for NLP* In Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2002), volume 2276 di (Lecture Notes in Computer Science), Berlino, Springer-Verlag. pp.3-8

10 Evert, S. op. cit. pp. 1213-1214

trovare in questa categoria alcune costruzioni sintattiche e fraseologiche particolari. Alcune di queste possono essere le costruzioni a verbo supporto, ossia coppie di verbi e sostantivi che presentano un nesso tra di loro, dove il sostantivo è un nome deverbale o un aggettivo sostantivato[11] (es. *fare uso*, *fare ritorno*, *avere la meglio*, *avere importanza*); nomi propri composti; espressioni idiomatiche, quindi costruzioni fisse di una lingua. Le ultime due tipologie di collocazioni sono nella maggior parte dei casi 'congelate' in un'unica forma, con poche possibilità di modifica senza provocare uno stravolgimento del significato o risultare anomale. Da questo punto di vista si può dire, aggiungendo un nuovo punto alla categorizzazione di Manning e Schütze, che le collocazioni hanno anche un certo grado di convenzionalità, "sono espressione di usi convenzionali o stereotipati"[12]. Inoltre, come accennato anche in precedenza, se la ricerca viene effettuata su testi di dominio tecnico, è molto prevedibile ottenere collocazioni che possono essere classificate come termini tecnici e sintagmi terminologici[13]. Il sintagma terminologico è formato da un minimo di due unità legate dal punto di vista sintattico che formano un costituente minimo all'interno della frase ed esprimono un concetto appartenente a domini tecnici e scientifici (Collet 1997, p.195).[14] Di seguito vengono proposte alcune collocazioni estratte automaticamente dal corpus di Wikipedia.it che sono esempi di termini tecnici e sintagmi terminologici: *destinare all'abitazione*, *eleggere per mandato*, *andare in scena*, *condannare all'ergastolo*, *processare in contumacia*, *firmare un contratto*, *prestare servizio*, *ricoprire la carica*, *vestire la maglia*, *sistema operativo*, *casa editrice*, *casa discografica*, *stazione meteorologica*, *quartier generale*, *sistema solare*, *anidride carbonica*.

1.3 Le misure di associazione

Le collocazioni, da quanto sopra emerso, sono coppie di parole che hanno una forte tendenza a comparire in posizioni ravvicinate all'interno delle frasi. Per trovare queste coppie di parole nei corpora ci sono vari metodi, il più intuitivo è il semplice conteggio delle volte che due parole si trovano a co-occorrere.

11 Wikipedia, voce *verbo supporto*
12 Lenci, A., S. Montemagni e V. Pirrelli. (2005). *Testo e computer: elementi di linguistica computazionale*. Roma, Carocci. p.197
13 Manning, C. e H. Schütze. op. cit. p.152
14 Questa definizione Collet la dà per i sintagmi terminologici francesi, ma in parte è applicabile anche all'italiano.

Evert distingue tre tipologie di cooccorrenza: *cooccorrenza superficiale, cooccorrenza testuale* e *cooccorrenza sintattica*. Si parla di cooccorrenza superficiale se due parole si trovano vicine all'interno di un testo, ma la "vicinanza" che esse possono avere tra di loro è stabilita da una finestra di contesto per la quale è necessario decidere la grandezza (span). La finestra di contesto può essere intesa come una maschera che permette di mostrare solo una porzione della frase per volta, lasciando tutto il resto in secondo piano. Se lo span della finestra di contesto è fissato a 4, la maschera della finestra di contesto scorre il testo parola per parola mostrando sempre quattro parole adiacenti e verranno messe in risalto le porzioni di testo che contengono una parola di interesse. Si propone un esempio con alcune righe di testo tratte dal libro Eragon, cercando le cooccorrenze della parola "freccia":

> "Si fermò e scoccò una **freccia** contro la zoppicante femmina in fuga. La mancò di un soffio; la **freccia** si perse sibilando nel buio. Il ragazzo imprecò e si volse per prendere un'altra **freccia**."

Nel testo è stata posta in grassetto la parola focus della ricerca e sono state sottolineate le tre parole precedenti e successive che verranno visualizzate nella maschera della finestra di contesto. I dati che verranno estratti perciò saranno molti, ma saranno molti di più utilizzando la cooccorrenza testuale che invece viene ricercata senza l'uso di finestre di contesto e i limiti per indicare l'inizio e la fine di una possibile cooccorrenza sono i confini di frase, ma a volte anche di interi testi e documenti. Risulteranno così collocazioni possibili anche coppie di parole che hanno deboli legami di attrazione e di dipendenza reciproca. Con la ricerca di cooccorrenze sintattiche, invece, il numero delle coppie di parole candidate ad essere collocazioni sarà in numero inferiore rispetto alle altre due tipologie di cooccorrenza. Quelle sintattiche, infatti, vengono ricercate tra parole che all'interno della stessa frase evidenziano un rapporto di tipo sintattico, ad esempio il nome con l'aggettivo corrispondente, oppure il verbo con il nome che è il soggetto o l'oggetto relativo[15].

Secondo la definizione data di collocazione, una qualunque coppia di parole che ricorra almeno due volte nello stesso corpus sarebbe da considerare come una potenziale collocazione[16]. È per questo che insieme al calcolo della frequenza o della misura di associazione in questione vengono utilizzate delle soglie di frequenza che scremano i

15 Per le tre tipologie di cooccorrenze: Evert S. op. cit. pp.1220-1223
16 Ivi. p.1215

risultati eliminando quei dati con una frequenza tale da non poter incidere sull'analisi per la ricerca di collocazioni.

Le soglie di frequenza vengono stabilite in base alla grandezza del corpus di riferimento, si può comunque far presente come una soglia uguale a >=2 sia spesso non sufficiente, poiché un alto numero di coppie di parole corrisponderanno al valore di frequenza richiesto e verranno incluse anche coppie che ricorrendo soltanto due volte nel corpus, non forniscono dati rilevanti ai fini della ricerca collocazionale.

1.3.1 La frequenza

Il modo più semplice per calcolare la forza di attrazione delle parole all'interno di un corpus è verificare quante volte all'interno di esso le varie parole si vengono a trovare in posizioni vicine tra di loro. Una volta ottenuti i risultati dei calcoli si dovranno mettere in ordine decrescente le frequenze ottenute, ma al vertice della lista di frequenze non si potranno osservare esclusivamente coppie di parole che costituiscono collocazioni, e anzi, sarà molto probabile che la stragrande maggioranza di esse non lo siano affatto, poiché formate per la maggior parte da parole funzionali (articoli, preposizioni, congiunzioni ecc.). Questo effetto è spiegato dalla legge di Zipf, la quale in modo teorico dice che all'aumentare del rango di una parola (ossia la posizione che occupa nella lista decrescente delle frequenze) diminuisce anche la frequenza in modo progressivo, si deduce quindi che “poche parole sono ripetute molto spesso, molte parole hanno invece frequenza 1 o poco maggiore”[17]. Perciò, dall'ordinamento decrescente delle frequenze di tutti i bigrammi che fanno parte del corpus in analisi, quelli che si trovano ai vertici non sono affatto delle collocazioni poiché, come affermano Lenci et al (2005), "la frequenza assoluta di un bigramma non quantifica il grado di associazione lessicale di due parole. (...) É plausibile ipotizzare che due parole siano tanto più fortemente associate quanto più spesso si presentano insieme *rispetto alle volte in cui ricorrono l'una indipendentemente dall'altra.*"[18]

La misura della frequenza con cui ricorrono insieme due parole all'interno del testo, da sola, non è sufficiente a indicare la forza di attrazione che esiste tra le due parole e a fornire una prova che esse costituiscano una collocazione.

É importante quindi conoscere le frequenze assolute delle singole parole all'interno del corpus, per poter confrontare il numero di volte che due parole ricorrono insieme con il

17 Lenci, A., S. Montemagni e V. Pirrelli. op. cit. p. 142
18 Ivi. pp. 199-200

numero di volte che le stesse due parole ricorrono singolarmente. É necessario valutare se il fatto che due parole ricorrano con una certa frequenza possa essere attribuita al caso oppure se sia una manifestazione del loro grado di attrazione. Se il numero di volte che una delle parole ricorre da sola è molto più grande del numero con cui ricorre insieme all'altra parola, la "disparità suggerisce che non c'è un legame di associazione diretto"[19] tra le due parole.

1.3.2 Hypothesis Testing

Come si è visto, attraverso il solo calcolo della frequenza, non è possibile ottenere risultati validi per il nostro scopo, poiché, come sottolineano Manning e Schütze, ciò che è importante riuscire a sapere è se due parole occorrono più spesso del caso[20]. Per risolvere il quesito è necessario ricorrere alla statistica formulando un'ipotesi da assumere come vera (la cosiddetta ipotesi nulla), stabilendone in seguito l'accettabilità o la rigettabilità con le misure statistiche o i test di significatività. Se si formula l'ipotesi nulla nella quale il verificarsi dell'evento x sia dovuto solo al caso si procederà calcolando la probabilità con la quale è previsto che l'evento x accada. Nel caso la probabilità calcolata sia inferiore ad una certa soglia l'ipotesi nulla verrà accettata, nel caso contrario verrà rifiutata.

Evert, per verificare l'ipotesi nulla, confronta la frequenza osservata (O) del bigramma con quella che egli chiama frequenza attesa (E)[21], che calcola in base alla seguente formula:

$$E = f_1 \cdot (f_2 / N) = \frac{f_1 f_2}{N} \qquad \textit{Formula 1}$$

dove f_1 e f_2 sono le frequenze delle due parole di cui si analizza la cooccorrenza e N è la grandezza del corpus.

Il risultato ottenuto dal calcolo della frequenza attesa viene utilizzato da Evert come ordine di grandezza per verificare l'indipendenza dell'ipotesi nulla. Nel caso in cui la frequenza osservata sia significativamente maggiore della frequenza attesa possiamo rigettare l'ipotesi nulla e valutare la coppia come potenziale collocazione poiché la ricorrenza delle due parole non sarebbe casuale ma dotata di una certa forza di attrazione reciproca.

19 Lenci, A., S. Montemagni e V. Pirrelli. op. cit. p.200
20 Ivi. p.162
21 Evert, S. op. cit. pp.1224-1225

1.3.3 Test di significatività

I test di significatività sono test statistici volti a quantificare la forza con cui i dati falsificano o verificano l'ipotesi nulla[22].

Fanno parte dei test di significatività alcune misure di associazione che sono state usate nell'ambito della ricerca di collocazioni: lo z-test e il t-test che fanno parte delle misure di associazione più semplici.

Questi test di significatività hanno come caratteristica comune il fatto che i valori risultanti vengono confrontati con dei valori critici elencati in tabelle specifiche. Tuttavia, è necessario precisare che, nell'ambito della linguistica e della lessicografia computazionale, i valori risultanti dai test statistici di significatività non vengono confrontati nelle tabelle corrispondenti, ma impiegati per stilare una classifica da usare come lista di possibili collocazioni da sottoporre a valutazione linguistica. Questa decisione è stata presa in seguito a vari esperimenti con corpora in cui anche il 70% dei valori dei test statistici, effettuati sulle coppie di parole dei corpora analizzati, risultavano superare i valori critici delle tabelle di riferimento; non forniscono quindi risultati abbastanza accurati per procedere nello studio e nell'analisi dei dati.[23]

Nella statistica inferenziale, per poter effettuare il confronto dei dati ottenuti dai test di significatività nella tabella e valutare di conseguenza l'ipotesi nulla, è necessario sapere qual è il grado di libertà dei dati di cui si è in possesso e decidere un livello di significatività a cui fare riferimento.

Il grado di libertà non è altro che la quantità dei dati a cui è consentito di variare. Questo è calcolato come n - 1, dove n è il totale dei dati. Se n è un numero molto grande, allora il grado di libertà si approssima a infinito.

Il livello di significatività è un dato che viene scelto a priori da chi si accinge ad effettuare il test statistico, è un dato espresso in termini probabilistici e corrisponde alla quantità di incertezza con cui si sta valutando l'ipotesi nulla[24], ossia la probabilità che ci si possa trovare nella condizione di rifiutare l'ipotesi nulla quando invece sia da accettare e viceversa. Da questo punto di vista, quindi, costituisce anche una quantificazione del grado di errore. I valori più comuni che vengono presi in

22 Per la parte sui test di significatività mi sono informata su siti web didattici universitari: vedi sezione siti web consultati in Bibliografia

23 Evert, S. op. cit. p.1227

24 Colagrande V. *"Alcuni elementi di verifica di ipotesi statistiche"* tratto dal sito web: http://www.biostatistica.unich.it/mat_didattica/Odont/Verif_IPOTESI_odonto.pdf

considerazione sono: 0.05 (5%), 0.01 (1%), 0.005 (0.5%), 0.001 (0.1%). Più piccolo è il livello di significatività che si assume e minore sarà la probabilità di incorrere in errori di valutazione dell'ipotesi nulla.

Tornando quindi ai test di significatività, se il risultato che si ottiene dal calcolo del test è minore del valore critico corrispondente al livello di significatività scelto, l'ipotesi nulla viene rifiutata. Al verificarsi dell'evento dell'ipotesi nulla, perciò, si attribuisce una valenza statisticamente significativa, dovuta cioè difficilmente al caso.

1.3.4 z-test e t-test

Lo z-test e il t-test sono due metodi di calcolo del test statistico di significatività che si basano sul valore medio della popolazione degli eventi.

Lo z-test è dato dalla seguente formula:

$$z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad \textit{Formula 2}$$

dove X è la media campionaria, ossia la media calcolata su un campione della popolazione, μ è la media aritmetica della popolazione, σ^2 è la varianza (di quanto possono distaccarsi i valori estratti dal valore calcolato dalla media, quanto cioè possono essere vari gli eventi rispetto alla media) la quale, una volta estratta dalla radice, diventa la deviazione standard o scarto quadratico medio, e n è il totale della popolazione.

Nella statistica inferenziale, il valore di z viene confrontato con il valore critico di riferimento e, nel caso risulti maggiore di quest'ultimo, si può rifiutare l'ipotesi nulla di indipendenza.

É possibile trovarsi nella condizione di non disporre del valore della varianza della popolazione, in quel caso è necessario ricorrere al t-test. Il t-test è sostanzialmente uguale allo z-test con la differenza che la varianza σ^2 viene sostituita con la stima della varianza s^2 basata anziché sull'intera popolazione, su un campione di essa, secondo la seguente formula:

$$t = \frac{\overline{X} - \mu}{\sqrt{s^2/n}} \quad dove \quad s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \qquad \textit{Formula 3}$$

Queste formule possono risultare di difficile comprensione e applicazione, per questo si sceglie di utilizzare le formule che riporta Evert[25]. Come si nota nella formula seguente,

25 Evert, S. op. cit. p.1227

egli propone l'approssimazione[26] dei termini che compongono le due formule z e t test, con le già citate frequenza osservata (*O*) e frequenza attesa (*E*).

$$z-score = \frac{O-E}{\sqrt{E}} \qquad t-score = \frac{O-E}{\sqrt{O}} \qquad \textit{Formula 4}$$

Secondo Evert[27], lo z test applicato sui bigrammi fornisce dei valori poco attendibili con le frequenze attese (*E*) inferiori a 1; questi bigrammi, infatti, otterranno dei valori finali "gonfiati" rispetto alla realtà, risultando nell'ordinazione della lista decrescente di punteggi, in posizioni più alte rispetto al dovuto. Per questo motivo ritiene sia preferibile utilizzare il t-test, che peraltro, tra i due è il più usato proprio per questi scopi.

1.3.5 Misure effect-size

Le misure di associazione statistica di cui si è discusso fino a qui sono tutte misure di significatività; di seguito si vedrà una tra le più famose misure di associazione nel campo della linguistica computazionale che fa parte delle misure effect-size (misure di dimensione dell'effetto). Le misure di dimensione dell'effetto danno meno peso alla grandezza del campione[28] e cercano di quantificare la forza di attrazione che si rileva tra le coppie di parole[29].

La Mutual Information (MI) è la misura di associazione a cui si è appena fatto riferimento, deriva dalla teoria dell'informazione (ed è più conosciuta come come Pointwise Mutual Information) e misura la quantità di informazione che due eventi aleatori forniscono reciprocamente, cioè quanto il primo evento dice sul verificarsi del secondo. Nel caso dei bigrammi è stimata come la probabilità che c'è, dopo aver trovato la parola w_1, che la seconda sia w_2 e viceversa[30]. La formula con cui si calcola l'Informazione Mutua di due parole w_1 e w_2 è la seguente:

$$MI = \log_2 \frac{p(w_1, w_2)}{p(w_1) \cdot p(w_2)} \qquad \textit{Formula 5}$$

dove $p(w_1, w_2)$ è la probabilità condizionata di trovare le due parole insieme; $p(w_1)$ è la probabilità di trovare la prima parola e $p(w_2)$ è la probabilità di trovare la seconda

26 Evert trae l'approssimazione da: Baroni, M e S. Evert. (2009). *"Statistical methods for corpus exploitation"*. In A. Lüdeling e M. Kytö (eds.), *"Corpus linguistics: An international handbook"*. (Vol.2). Berlino, Mouton de Gruyter: pp. 777-802

27 Evert, S. op. cit. p.1227

28 Bachmann C., Luccio R. e Salvadori E. (2005). *La verifica della significatività dell'ipotesi nulla in psicologia*. Firenze, Firenze University Press. p.53

29 Evert, S. op. cit. p.1234

30 Lenci, A., S. Montemagni e V. Pirrelli. op. cit. p.200

parola.

Alternativamente, si può definire la MI come il logaritmo in base 2 del rapporto tra le frequenze osservate e le frequenze attese[31].

$$\log_2 \frac{O}{E} \qquad \textit{Formula 6}$$

Se le probabilità della formula in (Formula 5) sono sostituite con le frequenze relative, è possibile derivare la formula in (Formula 6). Come risulta evidente dall'espressione riportata di seguito, portando al denominatore la totalità delle parole del campione analizzato (*N*), si ottiene al denominatore la formula della frequenza attesa e al nominatore quella della frequenza osservata, esattamente la stessa formula di Evert:

$$MI = \log_2 \frac{f(w_1, w_2) \cdot N}{f(w_1) \cdot f(w_2)} = \log_2 \frac{f(w_1, w_2)}{\frac{f(w_1) \cdot f(w_2)}{N}} = \log_2 \frac{O}{E} \qquad \textit{Formula 7}$$

Non essendo una misura di significatività, il valore calcolato sulla base di questa formula non viene confrontato con le tabelle dei valori critici come nelle misure precedentemente affrontate. Con i risultati che si ottengono viene stilata una lista decrescente composta dalle coppie di parole analizzate con i relativi valori di Mutual Information.

La Mutual Information rispetta le due convenzioni che ogni misura di associazione dovrebbe rispettare. La prima è che alti punteggi indicano una grande associazione e bassi punteggi indicano bassa associazione. In presenza di un punteggio di associazione estremamente basso sarebbe possibile parlare anche di repulsione tra due parole. La seconda convenzione è che i punteggi possono avere valori positivi o negativi a seconda che la frequenza osservata superi o meno la frequenza attesa: nel caso venga superata il punteggio dovrebbe risultare positivo, nel caso sia invece inferiore il punteggio dovrebbe essere negativo.

La MI si basa sull'assunto che nel caso in cui la probabilità congiunta di due parole fosse uguale alla probabilità relativa delle singole parole saremmo di fronte all'indipendenza degli eventi, qualora il valore uscente fosse molto alto si avrebbe dipendenza degli eventi e se fosse negativo avremmo repulsione.

Secondo quanto appena espresso, quando si ha uguaglianza tra frequenza osservata e frequenza attesa, il punteggio della Mutual Information sarà uguale a zero, questo

31 Evert, S. op. cit. p.1226

risultato indica una natura non collocazionale del bigramma in quanto le due parole ricorrono tanto quanto ricorrono per volere del caso e si dicono indipendenti[32]. Un punteggio positivo, ossia quando O è molto maggiore di E, indica lo status di collocazione, quando invece E è maggiore di O si ottiene un punteggio negativo di MI, che indica che ci troviamo di fronte a un bigramma anti-collocazione, due parole che si repellono.

Il problema che si pone con la Mutual Information è la sensibilità nei confronti delle basse frequenze e degli hapax. Gli hapax sono parole che compaiono una volta sola all'interno del corpus in analisi. Questo loro manifestarsi una volta sola, produce dei punteggi molto alti nella Mutual Information[33]; dall'altro lato, due parole che occorrono una sola volta nel corpus, anche se si trovano a ricorrere insieme, non possono essere di fatto considerate collocazioni poiché potrebbero essere associate solo per caso.[34]

Per provare a sopperire a questo si possono filtrare le frequenze con una soglia minima di occorrenza, ma questa opzione migliora il risultato non risolvendone il problema[35].

Un modo per tentare di ridurre le basse frequenze è utilizzare la local-MI che consiste semplicemente nel moltiplicare tutta la formula della MI per la frequenza osservata O[36].

$$local\ MI\ =\ O \cdot \log_2 \frac{O}{E} \qquad Formula\ 8$$

Dopo aver discusso lungamente di tutti questi metodi statistici, si può già indicare quale sarà quella che verrà utilizzata in questa ricerca di collocazioni. Una volta estratte dal corpus di Wikipedia.it tutte le coppie di parole per le quali si intende analizzare la collocatività, su di esse verrà applicata la local-MI per ottenere l'indice che quantifica il loro grado di associazione e procedere in seguito all'analisi delle collocazioni candidate. Nel prossimo capitolo verrà esposta la metodologia per l'estrazione delle coppie di parole dal corpus che si intende analizzare.

32 Ibidem.
33 Vedere in appendice le collocazioni candidate dal corpus di wikipedia.it che risultano in cima alla lista dopo aver applicato la MI
34 Lenci, A., S. Montemagni e V. Pirrelli. op. cit. p.203
35 Manning, C. e H. Schütze. op. cit. p.182
36 Vedere in appendice le collocazioni candidate dal corpus di wikipedia.it che risultano in cima alla lista dopo aver applicato la Local-MI

2 Metodologia per l'estrazione di coppie di parole

Prima di poter applicare le misure associative e quindi iniziare a individuare le collocazioni, è necessario estrarre dai testi di Wikipedia.it le coppie di parole su cui verteranno le analisi.

Non verranno estratte tutte le coppie di parole indiscriminatamente, ma solo quelle che rispondono ad alcune funzioni sintattiche e grammaticali all'interno della struttura delle singole frasi. Di questo si parlerà più approfonditamente in seguito, intanto si procederà a illustrare il modo in cui si presenta il file di input del corpus da cui dovranno essere estratte le coppie di parole.

2.1 Analisi del file di in input

Il file di input è costituito dall'insieme dei testi che fanno parte della Wikipedia italiana. Questo corpus è stato prodotto dalla pipeline linguistica del progetto Semawiki[1a]. Il corpus è stato lemmatizzato e annotato a livello morfologico e sintattico con la catena di strumenti Tanl[1b] (Text Analytics and Natural Language), infine parsato con il parser DESR[1c].

L'aspetto iniziale del file di input, quindi, era il seguente:

```
<s score="LogLikelihood: -2.29105  0.409152  0.10116">
1   Ma  ma  C   CC  _   5   con _   _
2   altre   altro   D   DI  num=p|gen=f 3   mod _   _
3   caratteristiche caratteristica  S   S   num=p|gen=f 5   subj _   _
4   hanno   avere   V   VA  num=p|per=3|mod=i|ten=p 5   aux _   _
5   fatto   fare    V   V   num=s|mod=p|gen=m  0   ROOT    _   _
6   in  in  E   E   _   5   comp    _   _
7   modo    modo    S   S   num=s|gen=m 6   prep _   _
8   che che C   CS  _   5   arg _   _
9   si  si  P   PC  num=s|gen=n 10  clit _   _
10  inserisse   inserire    V   V   num=s|per=3|mod=c|ten=i  8   sub _   _
11  ugualmente  ugualmente  B   B   _   10  mod_temp    _   _
12  nel in  E   EA  num=s|gen=m 10  comp    _   _
13  contesto    contesto    S   S   num=s|gen=m 12  prep _   _
14  della   di  E   EA  num=s|gen=f 13  comp    _   _
15  musica  musica  S   S   num=s|gen=f 14  prep _   _
16  indiana indiano A   A   num=s|gen=f 15  mod _   _
17  (   (   F   FB  _   19  punc_   _
18  anche   anche   B   B   _   19  mod _   _
19  di  di  E   E   _   5   comp    _   _
20  quella  quello  P   PD  num=s|gen=f 19  prep _   _
21  "   "   F   FB  _   22  punc_   _
22  classica    classico    A   A   num=s|gen=f 20  mod _   _
23  "   "   F   FB  _   22  punc_   _
24  )   )   F   FE  _   5   punc_   _
25  .   .   F   FS  _   5   punc_   _
</s>
<s score="LogLikelihood: -2.78355  0.312174  0.0618188">
1   La  il  R   RD  num=s|gen=f 2   det _   _
2   principale  principale A   A   num=s|gen=n 3   subj _   _
```

Immagine 1: Il file di input

1a Progetto sviluppato dal Dipartimento di Informatica e il Dipartimento di Linguistica dell'Università di Pisa e dall'ILC-CNR di Pisa http://medialab.di.unipi.it/wiki/SemaWiki

1b G. Attardi, S. Dei Rossi, M. Simi. (2010). *The Tanl Pipeline*. In "Proc. of LREC Workshop on WSPP". Malta

1c G. Attardi.(2006) *Experiments with a Multilanguage Non-Projective Dependency Parser.* In "Proc. of the Tenth Conference on Natural Language Learning". New York

Come si può notare, ogni frase è racchiusa da un tag <s></s> ed è separata dalla successiva con un ritorno a capo.
Ciascun componente della frase, sia le parole che la punteggiatura, è analizzato su singole righe che, idealmente, si possono scomporre in una tabella di 8 campi separati ciascuno da una tabulazione. Nella tabella sottostante viene riportata una riga del file di input inserita nella tabella ideale di cui si è appena parlato, rendendola quindi più chiara dal punto di vista visivo.

ID	FORMA	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
3	caratteristiche	caratteristica	S	S	num=p|gen=f	5	subj

Tabella 1

Ogni campo, con i relativi "codici" e abbreviazioni, veicola diverse informazioni della struttura sintattica e di quella morfologica delle frasi, (struttura morfo-sintattica).
Il primo campo, quello dell'id, è un un indice univoco che parte da 1 con il primo token della frase e si incrementa di uno per ogni token successivo fino alla fine della frase.
Il secondo e il terzo campo contengono informazioni di tipo morfologico: nel secondo si ha la forma flessa della parola o il segno di interpunzione, nello stato originario della frase, mentre nel terzo la forma grafica viene ricondotta al lemma, ossia "la forma rappresentativa di tutte le forme flesse"[2] la forma con cui possiamo rintracciare una parola su un dizionario (di norma, ad esempio, i verbi si riconducono all'infinito presente, i sostantivi al maschile singolare ecc..).
I campi 4, 5 e 6 forniscono le informazioni che riguardano la categoria grammaticale del token, nel dettaglio:
- il CPOSTAG (*Coarse-grained part-of-speech tag*), quarto campo, esplicita se si tratta di un nome con l'annotazione "S", di un aggettivo con "A", di un verbo con "V", di un segno di interpunzione con "F", di un articolo con "R", ecc. ;
- il POSTAG (Fine-grained part-of-speech tag), quinto campo, specifica, dove possibile, il campo precedente. Se si è di fronte a un nome proprio, infatti, l'annotazione sarà "SP", se invece a un aggettivo possessivo "AP", un articolo determinativo "RD", indeterminativo "RI", ecc. ;
- il sesto campo contiene tratti morfologici che completano il campo della Part of Speech (della categoria grammaticale). Ad esempio per nomi e aggettivi è possibile

2 Wikipedia alla voce"lemma" http://it.wikipedia.org/wiki/Lemma_(linguistica)

trovare il numero “num=” e il genere”gen=”, per i verbi il numero “num=”, il modo verbale “mod=”, il tempo verbale “ten=” e la persona ”per=”.

Con il settimo campo, che viene chiamato HEAD, si ottengono informazioni di tipo sintattico a dipendenze. Dal momento che ogni componente della frase è identificato da un id, si possono rendere esplicite le dipendenze dei singoli componenti all'interno della frase con una corrispondenza di codici numerici. In altre parole all'interno dell'analisi di ciascun token, il campo head conterrà il numero dell'id del componente da cui dipende, o a cui è legato da un qualche rapporto.

Con il campo numero 8, denominato DEPREL, viene indicata la tipologia di relazione di dipendenza con cui si trova nei confronti del token corrispondente. Le tipologie di relazioni che vengono prese in considerazione dall'analizzatore sono varie: soggetto (subj), oggetto (obj), verbo (aux) radice (ROOT), modificatore (mod), complemento predicativo del soggetto o dell'oggetto a seconda della sua dipendenza (pred), preposizioni che in genere introducono vari complementi (comp), complementi che sono introdotti da preposizioni (prep), ecc..

2.2 Selezione dei tipi di dipendenze sintattiche

Avendo presente la struttura del file di input, si può iniziare a dire che in questo modo non è necessario limitarsi a cercare i bigrammi adiacenti o usare le finestre di contesto, che darebbero come risultato qualsiasi coppia di parole, anche senza che abbiano rilevanza per il nostro obiettivo. Con un file di input annotato dal punto di vista morfo-sintattico, è possibile estrarre solo quelle categorie grammaticali che vengono ritenute più rilevanti (evitando quindi nella maggior parte dei casi, congiunzioni, aggettivi possessivi, articoli...) e contemporaneamente che sono legate sintatticamente. Con questa tipologia di corpus si potranno quindi estrarre quelle che, come si è illustrato nel precedente capitolo, Evert chiama “cooccorrenze sintattiche”.

Per quanto riguarda l'obiettivo di ricerca di questa tesi si cercheranno:

aggettivi che hanno funzione sintattica di modificatore con i sostantivi corrispondenti che svolgono qualunque funzione sintattica (es: *strumento musicale*);

verbi con i sostantivi soggetto (es: *piano prevedere*) e complemento oggetto corrispondenti (es: *innalzare pressione*);

coppie formate da un verbo e da un sintagma preposizionale. Questo ultimo a sua volta

è composto da: un token (in genere si tratta di preposizioni) con DEPREL uguale a "comp" e un sostantivo con DEPREL uguale a "prep" (es: *mettere in funzione*).

L'estrazione di queste coppie è possibile proprio perché ciascun token, grazie all'annotazione sintattica, è collegato alla testa da cui dipende, per mezzo del numero presente nel campo HEAD.

Si propone di seguito una frase estratta dal testo di input per evidenziare, con un esempio, i punti focali della ricerca. Si sono inseriti i dati della frase in una tabella, come illustrato anche in precedenza, eliminando però i campi 9 e 10 perché non portano informazioni:

ID	FORMA	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
1	Ma	ma	C	CC	_	5	con
2	altre	altro	D	DI	num=p|gen=f	3	mod
3	caratteristiche	caratteristica	S	S	mun=p|gen=f	5	subj
4	hanno	avere	V	VA	num=p|per=3|mod=i|ten=p	5	aux
5	fatto	fare	V	V	num=s|mod=p|gen=m	0	ROOT
6	in	in	E	E	_	5	comp
7	modo	modo	S	S	num=s|gen=m	6	Prep
8	che	che	C	CS	_	5	Arg
9	si	si	P	PC	num=s|gen=n	10	Clit
10	inserisse	inserire	V	V	num=s|per=3|mod=c|ten=i	8	Sub
11	ugualmente	ugualmente	B	B	_	10	mod_temp
12	nel	in	E	EA	num=s|gen=m	10	comp
13	contesto	contesto	S	S	num=s|gen=m	12	prep
14	della	di	E	EA	num=s|gen=f	13	comp
15	musica	musica	S	S	num=s|gen=f	14	prep
16	indiana	indiano	A	A	num=s|gen=f	15	mod
17	.	.	F	FS	_	5	punc

Tabella 2

Il file di output che risulterà in seguito all'esecuzione dello script per estrarre i sostantivi con i loro aggettivi modificatori, dovrà quindi contenere le due parole in rosso: "musica indiano". Come evidenziato dal colore, infatti, tra queste due parole c'è la corrispondenza tra l'ID del sostantivo con l'HEAD dell'aggettivo (ovvero 15) che ovviamente sta a significare la dipendenza dell'aggettivo nei confronti del sostantivo.

Per quanto riguarda invece il file di output risultante dallo script per l'estrazione di verbo e sintagma preposizionale, in questo caso dovrà contenere la coppia "inserire in_contesto" che sono state evidenziate in verde. Anche tra queste coppie c'è la

corrispondenza tra ID del sostantivo e HEAD della preposizione del sintagma preposizionale (ovvero 12). Per unire i componenti del sintagma preposizionale sono stati utilizzati ID della preposizione e HEAD del verbo, con la stessa logica precedente (ovvero 10). All'interno di questa frase, c'è anche un'altra coppia di questo tipo: "fare in_modo". Infine per mezzo dello script soggetto-verbo si otterrebbe l'estrazione della coppia evidenziata in verde "caratteristica fare", anche questa per mezzo della corrispondenza ID-HEAD (ovvero 5). Nelle immagini che seguiranno ciascuna sezione, sarà possibile ritrovare i sopracitati esempi (Immagini 2, 3, 6).
Al fine di poter analizzare con le misure di associazione le coppie di parole di interesse, è necessario ottenere un file con le suddette coppie di parole separate da una tabulazione e seguite da un ritorno a capo.

2.3 Procedimento di estrazione delle dipendenze sintattiche

Per estrarre le coppie di parole è necessario sviluppare un piccolo programma che prendendo il testo annotato di Wikipedia in input, stampi solo e soltanto i componenti della frase che rispondono a certe caratteristiche. Il linguaggio di programmazione che ho usato per fare lo script è il Perl, che è molto efficace nella manipolazione di dati testuali poiché supporta efficacemente il pattern matching per mezzo di regular expressions (RE). È un linguaggio di programmazione interpretato. La sintassi del Perl è molto flessibile e per questo è un linguaggio di programmazione semplice da imparare e da usare.

2.3.1 Coppie Sostantivo - Modificatore

Per ottenere le coppie Sostantivo - Modificatore, questo è l'algoritmo che ho elaborato:

- scorrere tutto il file di input riga per riga;
- memorizzare ciascuna frase in un hash che ha una corrispondenza key-value uguale a: ID del token-intera riga del token;
- controllare per ogni frase se c'è un token che è aggettivo (A) e modificatore (mod), nel caso abbia riscontro positivo, memorizzare l'ID, il numero dell'HEAD e la forma del lemma ciascuno in una variabile e "salvare" tutte le variabili in un nuovo hash usando come chiave l'ID del token;
- controllare se ci sono token che sono sostantivi (S) e nel caso di riscontro

positivo memorizzare l'ID e la forma del lemma in una variabile e salvare entrambe le variabili in un altro hash usando, anche qui, l'ID del token come chiave;

- infine controllare per ogni chiave di entrambi gli ultimi due hash creati, se un ID corrisponde con l'HEAD dell'altro e, nel caso corrisponda, far stampare entrambe le variabili contenenti i lemma dei token corrispondenti separati ciascuno da una tabulazione e seguiti da un ritorno a capo.

L'intero script che estrae coppie sostantivo-modificatore, implementato sulle basi dell'algoritmo, è consultabile in appendice.

Nella prima parte dello script viene scorso il file di input riga per riga, viene tagliata ogni riga dove ci sono spazi o tabulazioni ed ogni elemento viene inserito in un array. A questo punto ogni riga del file di input sarà un array e ciascuna parte della riga un elemento distinto dell'array. Viene posta la condizione che l'inizio delle righe di interesse debbano iniziare con un numero, in modo da trascurare quelle che fanno parte della formattazione del file, e quindi i tag <s></s>.

Il programma, mentre scorre le righe, "memorizza" ogni frase in un nuovo hash. Infatti, nel caso in cui il programma riscontri che l'ID della riga successiva corrisponda al valore 1, allora resetta l'hash; ha dunque terminato una frase ed ha iniziato a scorrerne una nuova.

Nella parte di codice a cui viene attribuito il label MOD[3], si chiede di individuare solo e soltanto quei token che hanno funzione modificatore e sono aggettivi. Nelle frasi contenenti i titoli delle entrate di wikipedia, il valore dell'HEAD dell'aggettivo, nonostante abbia funzione di modificatore e quindi lasci presumere che si riferisca a una qualche altra parola, era uguale a 0 e quindi non aveva una testa da cui dipendeva. Questo procurava delle coppie formate dall'aggettivo e un sostantivo che ovviamente era vuoto, perciò si richiede anche che l'HEAD sia maggiore di 0.

Per mezzo di alcune RE vengono ripuliti i risultati non prendendo in considerazione parte dei dati non rilevanti. Una volta individuati i token che rispondono alle richieste, vengono inizializzate alcune variabili per "memorizzare" solo gli elementi fondamentali e si inseriscono le stesse in un nuovo hash, che quindi contiene tutti gli aggettivi con funzione modificatore.

3 Appendice cap. 4.2.1 righe 15-30

Nella parte di codice a cui viene attribuito il label NOME[4], analizza lo stesso hash "%frase" che è stato analizzato nelle righe di codice del label MOD. Questa parte di script deve individuare solo e soltanto quei token che sono dei sostantivi. Come per i modificatori, vengono ripuliti i risultati con una RE, vengono inizializzate alcune variabili per "memorizzare" gli elementi fondamentali e vengono inserite in un nuovo hash per poter incrociare successivamente i dati con quelli del modificatore.
Nella parte finale dello script, identificato con il label ID[5], verranno confrontati i dati presenti nei due hash appena creati. Basterà verificare l'eventuale presenza di una corrispondenza tra l'ID di ciascun sostantivo con l'HEAD di ciascun modificatore e, nel caso questa corrispondenza esista, si chiede che vengano stampati nel file di output il lemma del sostantivo e il lemma del modificatore separati da una tabulazione e seguiti da un ritorno a capo. In questo modo si otterrà un file con le coppie di parole che si aveva intenzione di ottenere.
Quando l'interprete finisce di elaborare lo script Perl, si otterrà il file di output che verrà analizzato con le misure di associazione per trovare le eventuali collocazioni presenti.
Nell'immagine 2 viene riportato un campione delle coppie ottenute.

```
nota costante
musica   classico
musica   indiano
caratteristica  importante
contesto   indiano
strumento   occidentale
```

Immagine 2: Output N - Agg

2.3.2 Coppie Verbo - Soggetto / Verbo - Oggetto

Per estrarre queste due tipologie di coppie di parole si utilizzerà lo stesso metodo usato per quelle dei sostantivi-modificatori. Sfruttando nuovamente la corrispondenza HEAD-ID e sulla base del precedente algoritmo si cercheranno: tutti i sostantivi con funzione soggetto e tutti i verbi per la prima tipologia di coppie, tutti i sostantivi con funzione oggetto e tutti i verbi per la seconda.

4 Ivi. righe 31-37
5 Appendice cap 4.2.1 righe 38-43

Come è possibile notare con la visione dell'intero script posto in appendice, le differenze introdotte si trovano nelle parti di codice che, nello script sostantivo-modificatore, erano etichettate con i label MOD e NOME. Al loro posto, per le coppie verbo-soggetto, vengono inseriti i label SUBJ e VERB, e per le coppie verbo-oggetto i label OBJ e VERB.

Nel blocco di codice SUBJ[6] si chiede di controllare il DEPREL del token in analisi e per poter procedere nel codice seguente viene imposto che il suo contenuto sia uguale a "subj", deve perciò avere funzione soggetto. Se questa condizione è soddisfatta si richiede che il soggetto sia espressamente un sostantivo per mezzo di un controllo sul POSTAG e si filtrano i risultati per eliminare eventuali token il cui lemma è composto da numeri o caratteri che non veicolano informazioni rilevati ai fini della ricerca.

Il blocco di codice VERB[7] è uguale per entrambi gli script di queste due tipologie di coppie di parole, viene controllato il CPOSTAG del token e viene richiesto che esso sia uguale a "V", deve perciò essere un verbo. Anche qui vengono filtrati i dati con un'espressione regolare per eliminare i dati non rilevanti.

Il codice con il label OBJ[8] è sostanzialmente uguale al corrispondente codice per la ricerca dei soggetti, ovviamente la condizione sul campo del DEPREL sarà diversa,; viene imposto, infatti, che esso sia uguale a "obj" e che quindi svolga una funzione di complemento oggetto.

Nelle immagini seguenti si propongono due campioni dei risultati ottenuti con gli script relativi: coppie verbo-soggetto (Immagine 3) e coppie verbo-oggetto (Immagine 4).

fare caratteristica
fondare musica
essere tecnica
affidare accompagnamento
essere fatto
richiedere pezzo

Immagine 3: Output V - Subj

sostituire corda
spingere suonatore
creare circolo
preferire suonatore
avere ruolo
trovare ruolo

Immagine 4: Output V - Ogg

2.3.3 Coppie Verbo – Sintagma preposizionale

6 Appendice cap. 4.2.2 righe 15-2
7 Ivi. righe 28-29
8 Appendice cap. 4.2.3 riga 15

Nel caso della ricerca di “coppie” Verbo – Sintagma preposizionale, si può parlare più in generale di sintagmi verbali SV che contengono al loro interno un sintagma preposizionale SP. In questo caso viene minimizzata l'estrazione eliminando l'eventuale articolo o il determinante; perciò le combinazioni di parole che verranno cercate avranno la forma V_Prep_N. Sarà possibile procedere all'estrazione delle sole triplette in cui:

- V potrà essere un verbo qualsiasi,
- Prep avrà la funzione comp (che di solito veicola un complemento)
- N potrà avere la sola funzione di prep, che nella tipologia di analisi del file di input indica esattamente il sintagma preposizionale.

Una volta estratte le triplette, si uniranno i token Prep e N con un underscore, così da ottenere le coppie V-SP .

Sempre in base all'algoritmo pensato per le coppie sostantivo-modificatore, si procederà cercando per primi tutti i token con funzione “prep”, poi tutti i token con funzione “comp” e infine tutti i token che sono verbi. In seguito, sfruttando di nuovo la corrispondenza HEAD-ID, si procederà a concatenare i due token che costituiscono il sintagma preposizionale e infine ad associarlo al verbo corrispondente, qualora ne avesse uno. Per fare questo, si è prodotto il codice di cui è possibile prendere visione in appendice[9].

In seguito alla parte di codice comune a tutti gli script qui presentati, in cui si scorre una frase alla volta inserendo ciascuna frase nell'apposito hash, si procede alla ricerca dei token di interesse.

Nella parte di codice PREP[10], viene richiesto che il valore contenuto nell'indice dell'array corrispondente al DEPREL sia uguale a “prep”. Nel caso in cui ci siano dei riscontri positivi, si procede a salvare i dati del token relativo nel nuovo hash %prep i valori corrispondenti all'id, l'head e il lemma.

La stessa cosa viene fatta nella parte di codice con il label COMP[11], in cui, a differenza del precedente, viene chiesto che il valore del DEPREL sia uguale a “comp” e i dati vengono inseriti nel nuovo hash %comp.

Nel codice con il label VERB[12] viene richiesto che il valore contenuto nel CPOSTAG sia

9 Appendice cap. 4.2.4
10. Ivi. righe 16-25
11. Ivi. righe 26-33
12. Ivi. righe 34-40

uguale a “V” e quindi che il token corrispondente sia un verbo. I dati relativi vengono inseriti nel nuovo hash %verbs, in cui vengono inseriti solamente i valori dell'id e del lemma.

Nella parte di codice etichettata come SINT_PREP[13] si procede a confrontare tutti gli ID dei token contenuti nell'hash %comp, con tutte le HEAD dell'hash %prep. Nel caso di corrispondenza, viene inserito il valore dell'HEAD del token che si trova nell'hash %comp e il LEMMA del rispettivo hash %prep in un nuovo hash chiamato %sint. A questo punto nell'hash %sint si trovano tutti i sintagmi preposizionali che sono stati trovati nella frase presa in analisi.

Infine, nel codice con il label STAMPA[14], viene fatto un controllo per verificare l'eventuale corrispondenza di HEAD-ID tra l'hash %sint e l'hash %verbs. In questo modo viene verificata l'esistenza di dipendenza sintattica tra i token e se questa esistesse, verrebbe stampato il LEMMA del verbo, seguito dal LEMMA del sintagma preposizionale. Nell'immagine 5, si può osservare un campione dell'output risultante.

```
fare  in_modo
inserire  in_contesto
passare  da_tonalità
fondare su_tonica
applicare  al_musica
affidare  al_sarangi
```

Immagine 5: Output V - SP

13 Appendice cap. 4.2.4 righe 41-47
14 Ivi righe 48-51

3 Analisi dei dati

Le liste di coppie di parole ottenute con gli script illustrati nel capitolo precedente sono state analizzate statisticamente per mezzo di altri script che calcolavano le frequenze delle singole parole, la frequenza con cui le stesse ricorrono insieme e la Local Mutual Information (LMI).

Nella tabella 1 si riportano le quantità delle coppie di parole estratte dal corpus divise per categorie e il numero delle stesse che superano con la LMI la soglia di 10.

Categoria	Totale coppie	Coppie con LMI > 10
Sostantivo-Modificatore	1.795.205	568.398
Verbo-Soggetto	2.019.693	274.866
Verbo-Oggetto	853.970	168.697
Verbo-Sint. Preposizionale	2.051.392	466.946

Tabella 1

Per ricerca delle collocazioni, in questo caso, si intende un'analisi delle coppie di parole trovate e l'individuazione delle collocazioni utilizzando la conoscenza del linguaggio italiano come parlante nativo, unito alla verifica di una LMI alta (in questo caso >> 10). Spesso si è fatto uso anche di una ricerca dei significati delle parole su dizionari italiani e anche la ricerca delle due parole sul web. Nei casi in cui si era in dubbio sulla classificazione delle coppie o quando si riteneva di essere di fronte ad un errore si è ricorsi a una verifica del contesto da cui erano state estratte le coppie. Questo è stato possibile per mezzo di un motore di ricerca chiamato *Deepsearch v.0*[1]. Deepsearch incorpora tutti i livelli di analisi linguistica prodotti dalla pipeline TANL e che consente, attraverso l'uso di un indice arricchito e un linguaggio di interrogazione ad hoc, di fare ricerche per lemma e per relazione di dipendenza.

3.1 Analisi delle prime100 coppie con maggiore LMI

In primo luogo si è deciso di analizzare tutte le coppie di parole che si trovano nelle prime 100 posizioni per ciascuna categoria di estrazione. Lo scopo è verificare l'efficacia della misura statistica utilizzata; infatti nelle prime 100 posizioni dovrebbero trovarsi coppie di parole con un'alta probabilità di costituire collocazioni. Come vedremo, questa è variabile, poiché tutta la procedura di estrazione è stata realizzata automaticamente.

Per tutte le tipologie di coppie di parole è possibile prendere visione delle 100 coppie con punteggio di LMI maggiore in appendice di questa tesi (cap.4.3). Nella sezione 3.2 di questo capitolo si analizzeranno invece alcune coppie, estratte dalla totalità dei dati

1 Raggiungibile all'indirizzo http://semawiki.di.unipi.it/search.html

per ciascuna categoria che contengono alcune parole scelte arbitrariamente.

3.1.1 Sostantivo-Modificatore

Dopo aver analizzato attentamente le prime 100 coppie di parole[2] con il punteggio di LMI più alto tra la categoria sostantivo-modificatore, si possono fornire alcuni dati: 41 coppie risultano essere collocazioni; 6 errori di parsing o di annotazione; 13 coppie di parole che sicuramente devono la loro alta posizione nella lista alla grande quantità di testi di quel genere; mentre le restanti 40 coppie non sono collocazioni.

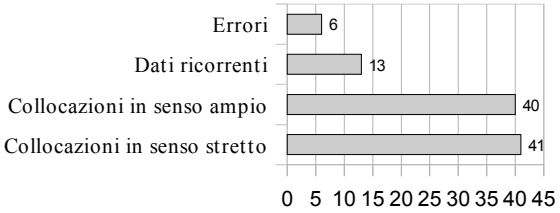

Grafico 1

Tra le 41 collocazioni individuate si trovano soprattutto termini tecnici, seguiti per numerosità da coppie di parole che fanno parte del linguaggio comune, un nome proprio e un'espressione idiomatica.

Tra i termini tecnici individuati è importante far notare la presenza di alcune collocazioni in particolare: *fascia principale* (regione del sistema solare compresa tra le orbite di Marte e Giove che contiene la maggiore concentrazione di asteroidi), *sistema operativo*, *colonna sonora*, *nome proprio*, *campo magnetico*, *carro armato*, *elezioni politiche*, *quartier generale*, *sistema solare, etichetta discografica.*

Alcune delle collocazioni che fanno parte del linguaggio comune che si trovano in questo insieme di coppie di parole sono: *gruppo musicale*, *opinione pubblica*, *cartoni animati*, *scuola elementare*, *fratello maggiore*, *acqua dolce*.

L'unico nome proprio che è stato trovato è *Giochi olimpici* e l'espressione idiomatica è *buona parte*.

Gli errori in questo gruppo di potenziali collocazioni sono elencati di seguito e per alcuni di essi si tenterà di ipotizzarne il motivo.

Al 10° posto si trova *spagnolo comune* che potrebbe essere spagnolo in quanto lingua o in quanto abitante della Spagna, ma si ritiene che sia troppo frequente come combinazione con questi significati. Al contrario, se la parola comune fosse sostantivo e spagnolo aggettivo, una presenza così forte all'interno dei testi di Wikipedia sarebbe più

2 Si può prendere visione delle coppie in Appendice cap. 4.3.1

giustificabile. Al suo interno, infatti, esistono entrate per i comuni italiani come anche per quelli stranieri. Non per niente la coppia di parole che ha la LMI più alta in assoluto in questa tipologia di coppie è proprio *comune francese*, seguita al 31° posto da *comune tedesco*. In seguito al controllo del contesto di utilizzo dei due termini con Deepsearch si può confermare l'ipotesi fatta: spagnolo è stato erroneamente taggato come sostantivo e comune come aggettivo, è perciò un errore di postagging.

Alla posizione 21 si trova *battezzato corrispondente* a cui non si riesce a dare alcun senso all'interno del linguaggio italiano, e non risulta nemmeno che ad una persona battezzata possa corrispondere qualcosa o qualcuno (tranne il padrino e la madrina) ma non si spiegherebbe comunque una presenza così considerevole all'interno dei testi. Dopo un controllo del contesto con Deepsearch si è individuato il problema nel mancato riconoscimento di "corrispondente" come participio presente anziché aggettivo.

Subito dopo si trova la coppia *facente parte*. Nonostante spesso il participio presente di alcune forme verbali abbia funzione aggettivale, in questo caso non si ritiene lo sia.

La coppia *arto marziale,* invece, presenta un errore di diversa tipologia: è un errore di lemmatizzazione. Nel testo doveva essere sicuramente sotto forma di "arti marziali", ma la parola arti è stata riconosciuta come forma plurale di arto e non di arte.

Nella coppia *sito web*, la prima parola è stata riconosciuta come aggettivo, che però viene usato solo in ambito burocratico. In realtà entrambe le parole sarebbero dei sostantivi, ma talvolta *web* viene usato anche come aggettivo per indicare qualcosa relativo al web. Ad ogni modo queste due parole, sebbene non riconosciute in modo esatto in questo contesto, costituiscono una collocazione.

Alla posizione 100 si trova un errore palese, siamo di fronte a *di euro*, dove ovviamente la parola "di" non è un sostantivo ma una preposizione semplice, ed "euro" non è un aggettivo bensì un nome.

Tra le coppie di parole che devono la loro alta LMI alla quantità di testi con stessa struttura linguistica sono: *comune francese*, *territorio comunale*, *centro storico*, *cenno biografico*, *comune tedesco*, *stato federato*, *città natale*, *gruppo etnico*.

Questo genere di coppie si ritrovano anche nelle altre categorie analizzate e fanno tutte parte di un ambito istituzionale o storico-biografico.

Da quanto emerge, si ritiene che la LMI sia sufficientemente attendibile per la ricerca di collocazioni tra sostantivi e aggettivi. Di fatto si ha un 41% circa di collocazioni contro un 40% di coppie che non lo sono.

Con le altre categorie vedremo che non sarà altrettanto buona la percentuale.

3.1.2 Verbo-Soggetto/Oggetto

Tra le prime 100 posizioni[3] dei verbi-soggetti si trovano 27 collocazioni e 16 errori.

Tra le collocazioni individuate, le seguenti possono rientrare nella categoria dei termini tecnici: *album uscire*, *scoppiare guerra*, *svolgere gara/edizione/campionato*. Alcune coppie di parole sono riconducibili all'ambito storico-geografico: *sorgere/trovare cattedrale/chiesa/stazione*, *comprendere arcidiocesi*.

Altre collocazioni individuate non attribuibili a categorie particolari sono: *leggenda narrare*, *termine indicare/riferire*, *leggenda/tradizione volere*, *origine/notizia risalire*, *edizione vedere*.

Alla posizione 65 si trova la coppia *valere pena* che costituisce una collocazione che fa parte delle espressioni idiomatiche. Questa coppia è riconducibile sicuramente infatti a modi di dire del tipo: "vale la pena ricordare"oppure "non ne vale la pena".

Il problema principale nell'individuazione delle coppie di questa categoria è che il parser non riesce a distinguere correttamente tutte le forme passive dei verbi, di conseguenza non sempre riesce ad attribuire il tag di subj_pass ai soggetti delle frasi passive. Nelle frasi passive, rispetto alle frasi attive, c'è una sorta di inversione tra soggetto e oggetto: il soggetto della frase passiva resta, dal punto di vista semantico, l'oggetto della stessa frase in forma attiva e il complemento d'agente il soggetto. Di conseguenza, la maggior parte degli errori che si sono individuati sono dovuti al mancato riconoscimento del soggetto passivo.

Facendo un'analisi dei 16 errori si possono distinguere 5 errori di postagging e lemmatizzazione, mentre i restanti sono tutti dovuti al mancato riconoscimento della forma passiva. Alcuni esempi delle coppie contenenti forme passive non riconosciute sono: *erigere diocesi/prefettura*, *eleggere presidente/deputato*, *ordinare sacerdote*.

I 5 errori sono: *decidere di*, *dotare di*, *iniziare de l*e *comporre da*, dove in tutti i casi per soggetto è stata riconosciuta una preposizione; in *Olimpiade Nuoto* invece la prima parola è stata riconosciuta come un verbo anziché un nome.

Analizzando le coppie verbo-oggetto[4], si sono individuate 29 collocazioni e 3 errori.

Tra le collocazioni ci sono prevalentemente termini tecnici dell'ambito lavorativo e sportivo. Alcuni esempi sono: *segnare gol/rete*, *prestare servizio*, *ricoprire ruolo/carica/incarico*, *vestire maglia*, *svolgere ruolo/attività/funzione*, *rassegnare dimissione*. Altre collocazioni trovate sono locuzioni, espressioni idiomatiche e

3 Si può prendere visione delle coppie in Appendice cap. 4.3.2
4 Ivi cap. 4.3.3

costruzioni a verbo supporto: *fare fronte*, *dovere nome*, *dare luogo*, *stringere amicizia*, *prendere posto* (prendere il posto per qualcuno, prendere il posto di qualcuno, prendere posto nel senso di sedersi), *avere luogo*, *dare origine*, *tenere conto* (nel senso proprio di contare, ma anche di considerare), *rendere conto* (rendere conto nel senso di giustificare o spiegare e nella forma riflessiva come accorgersi di qualcosa), *prendere parte* (prendere le parti di qualcuno), *fare parte* (fare la parte a qualcuno), *dare vita, vedere luce*.

Due collocazioni che meritano di essere menzionate, ma che non fanno parte delle categorie sopra indicate sono: *deporre uovo* e *dichiarare guerra*.

Il primo errore che troviamo all'interno di questo insieme è un errore di postagging e riguarda la coppia *particolar modo*, dove la parola "particolar" è stata riconosciuta come verbo, mentre si tratta ovviamente di un aggettivo. Il seguente errore è *potere essere*, dove la parola "essere" non può essere un sostantivo ma un verbo. L'ultimo errore può avere varia natura: si tratta delle parole *andare incontro*. È sicuramente un errore perché la parola "incontro" può essere sia avverbio che sostantivo, ma in entrambi i casi non sono complemento oggetto del verbo corrispondente. Nel caso dell'avverbio, in quanto il complemento oggetto doveva essere formato da un sostantivo; nel caso del sostantivo, perché il modo con cui "andare" e "incontro" si usano insieme è per esplicitare un complemento di moto a luogo (andare all'incontro) e quindi non si tratterebbe di un complemento oggetto.

3.1.3 Verbo-Sintagma preposizionale

Tra le prime 100 coppie[5] verbo-sintagma preposizionale si possono individuare 34 collocazioni e 2 errori.

Tra le coppie che risultano essere collocazioni, un piccolo gruppo è riconducibile al mondo dello spettacolo: *dirigere da regista*, *andare/mandare in onda*, *mettere in scena*.

Dell'ambito legislativo-politico-amministrativo invece si trovano: *aderire al partito/gruppo*, *reggere da vescovo/arcivescovo*, *cadere in mano* (sotto il potere o dominio), *entrare in vigore*, *salire al trono*, *entrare in servizio*, *radere al suolo.* Come si può notare dai due esempi precedenti, le triplette che sono composte dalla preposizione "da" in genere sono poi composte da un complemento di agente.

Altre collocazioni che non possono essere raggruppate in ambiti di utilizzo comuni: *mettere in mostra* (nella forma riflessiva), *scendere in campo* (dirsi pronti ad affrontare una sfida), *mettere in discussione* (anche riflessivo), *mettere in atto*, *entrare in contatto*,

5 Si può prendere visione delle coppie in Appendice cap. 4.3.4

dare alla luce, *mettere in luce* (dove la parola "luce" è molto spesso modificata dagli aggettivi buono o cattivo), *fregiare di titolo* (nella forma riflessiva).

I due errori che troviamo in questa tipologia di potenziali collocazioni sono "portare con sé" e "durare fino al". In entrambe le triplette il problema risiede nell'individuazione del sintagma preposizionale. La prima ha una preposizione seguita da un pronome personale e la seconda addirittura è costituita da un avverbio e una preposizione. Entrambe quindi sono errori.

Come si può notare, in queste tre ultime categorie, l'analisi delle potenziali collocazioni con esito positivo non supera la quantità di 30 su 100. La misura statistica della LMI risulta quindi essere maggiormente problematica in questi casi.

3.1.4 Le "non-collocazioni"

Da quanto emerso fino a qui, perciò, in tutte e quattro le categorie risultano esserci più della metà delle coppie che non sono collocazioni e non sono errori. Come è possibile? E che cosa sono? Innanzitutto è un risultato che era stato preventivato, perché la LMI non indica l'essere o meno collocazione, ma quantifica il grado di associazione che hanno due parole. Queste coppie sono perciò costituite da parole che ricorrono molto spesso nel linguaggio comune o in linguaggi di domini specifici, senza però avere tutte le caratteristiche tipiche delle collocazioni (ad esempio la non-composizionalità, la non-sostituibilità e la non-modificabilità). Possiamo fare un'ulteriore osservazione riguardo alla composizione dei testi che fanno parte di questo corpus preso in analisi. Wikipedia contiene al suo interno delle entrate che hanno una stessa struttura di base e che si differenziano spesso solamente per pochi dati. Di conseguenza si hanno intere frasi che differiscono da un'entrata all'altra soltanto per una o due parole. Questa alta frequenza può aver quindi inciso sulla presenza di queste coppie tra le prime 100 posizioni con LMI più alta. La natura enciclopedica del corpus perciò potrebbe essere causa del non superamento del 50% di collocazioni con l'utilizzo della LMI.

Di seguito si riportano alcuni esempi delle coppie di parole che hanno un largo utilizzo nel linguaggio comune ma non si possono considerare collocazioni nel senso stretto del termine.

- (Sostantivo-Modificatore) *gran parte*, *alto livello*, *temperatura media*, *programma televisivo*, *parco nazionale*, *vita privata* ecc. ;
- (Verbo-Soggetto) *esempio/risultato/obiettivo/scopo essere*, *nome significare*, *debutto avvenire*, *parola derivare* ecc. ;

- (Verbo-Oggetto) *prendere nome*, *avere figlio*, *porre fine*, *chiedere aiuto*, *suonare chitarra*, *scrivere libro* ecc. ;
- (Verbo-Sintagma preposizionale) *partire da fine*, *cedere in prestito*, *derivare da fatto*, passare *di tempo*, *dedicare al studio*, *dividere in parte* ecc. .

3.2 Analisi su estrazione arbitraria di parole

Per procedere alla seconda metodologia di analisi dei dati, vista la grande mole, si è reso necessario sviluppare un piccolo sito web per agevolare la ricerca delle collocazioni.
Il sito è stato implementato utilizzando il PHP come linguaggio di programmazione, anche perché il sito necessitava di appoggiarsi a una base di dati.
È stato utilizzato il pacchetto di EasyPHP per far girare in locale il sito, poiché è un ambiente di sviluppo che comprende un server web Apache, un database MySql e un interprete PHP.
Per mezzo dell'interfaccia grafica di phpMyAdmin, con la quale è possibile gestire ed amministrare il database, sono state create le 5 tabelle di cui necessita il sito internet. La tabella a cui si fa capo per effettuare il login è costituita da 8 campi: nickname, password, nome, cognome, città, nazione, e-mail, motivo di utilizzo. Il campo nickname è la Primary Key, poiché è necessario essere registrati per accedere al form di ricerca dati, questa è una tabella fondamentale nella quale non può esserci più di una persona con lo stesso pseudonimo. Sono stati creati i campi con le informazioni dell'utente per poter, eventualmente, monitorare la provenienza dell'utenza e i motivi per i quali chiedono l'accesso allo strumento.
Le altre 4 tabelle hanno tutte una struttura simile. Sono costituite da due campi testuali in cui sono racchiusi i componenti delle coppie, un campo per la frequenza della prima parola, un campo per quella della seconda parola, un campo per la frequenza delle due parole quando ricorrono insieme e uno per la LMI.
I dati ottenuti dall'analisi statistica sono stati inseriti nelle tabelle del database per mezzo della funzione “importa” presente in phpMyAdmin. Con questa funzione, infatti, è possibile inserire nel database grandi quantità di dati da file di varie tipologie, nel caso in esame erano file testuali con estensione txt. I dati non sono stati normalizzati per quanto riguarda le lettere maiuscole al momento dell’inserimento nel database, per permettere di riconoscere i nomi propri di cosa e di persona.
Il file di origine dei dati non conteneva un indice univoco per ciascuna coppia di parole,

perciò non è stata inserita nessuna chiave primaria, che sarebbe comunque possibile inserire a posteriori. In questo caso non sono stati aggiunti i campi con gli indici, perché, essendo semplici tabelle che non hanno ulteriori dati ad essi collegati in tabelle separate, non è stato ritenuto necessario.

3.2.1 La ricerca per mezzo del sito

Il sito, a cui è stato dato il nome "Collocationary", ha una semplice interfaccia utente e si apre con una pagina iniziale nella quale si chiede venga effettuato il login per poter accedere al menu di navigazione (Immagine 1).

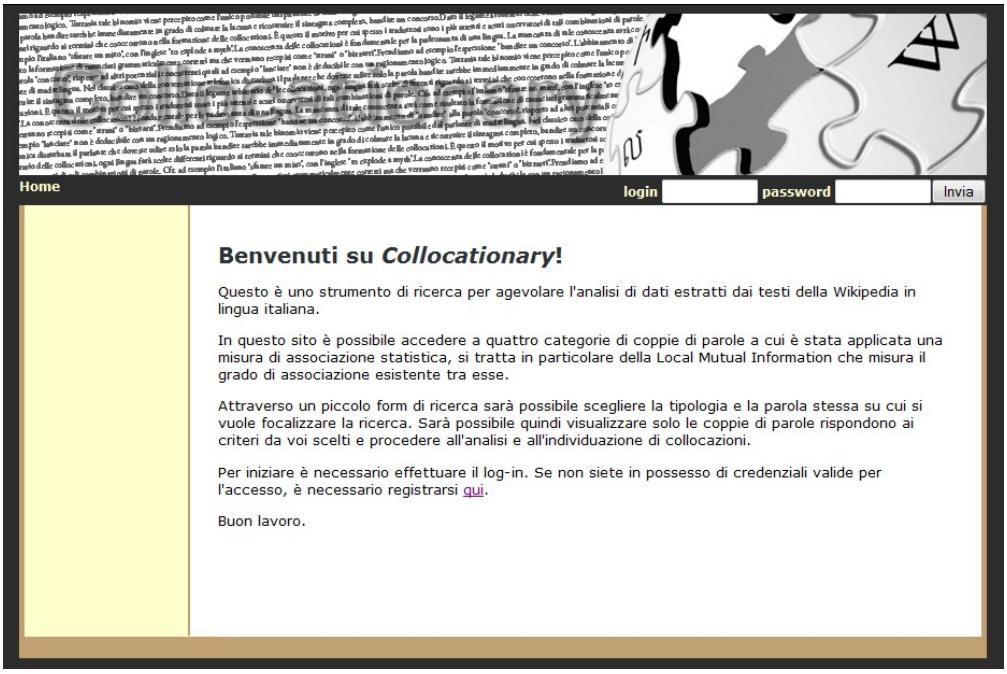

Immagine 1: Home Page Collocationary

Una volta effettuato il login viene visualizzato il menu da cui è possibile scegliere la tipologia di ricerca che si desidera effettuare. Le possibilità sono ovviamente: sostantivo-modificatore, soggetto-verbo, verbo-oggetto e verbo-sintagma preposizionale. Facendo click sulla tipologia di ricerca desiderata si apre la pagina corrispondente, nella quale è possibile scegliere i criteri di ricerca. Se l'utente ha la necessità di fare una ricerca tra le coppie sostantivo-modificatore, potrà scegliere se desidera ricercare un sostantivo o un aggettivo (viene dato in default il campo del sostantivo) e poi dovrà digitare la parola di cui vuole verificare l'esistenza all'interno del database. Affinché la ricerca abbia buon fine, sarà necessario inserire la parola nella forma maschile singolare se si tratta di un sostantivo o di un aggettivo, mentre se si tratta di un verbo dovrà essere inserita nel modo infinito; in pratica come se si cercasse l'entrata di un dizionario.

I risultati della ricerca verranno visualizzati all'interno di una tabella in ordine decrescente secondo il valore della LMI. Scorrendo le liste, che adesso sono molto più ristrette, sarà possibile trovare collocazioni con il solo utilizzo delle facoltà linguistiche del ricercatore, oppure aiutandosi con dizionari di lingua o, di nuovo, una ricerca sul web.

Per ogni tipologia di ricerca si è proceduto stilando una lista delle parole che potevano avere più di un significato o che erano molto ricorrenti nel linguaggio italiano; in seguito con le suddette parole è stata iniziata la ricerca nel form del sito web.

Sono state trovate moltissime collocazioni, ma ne vedremo solo una piccola parte cercando di inserirle in gruppi come è stato fatto nell'esposizione della precedente metodologia. Le collocazioni che al momento dell'analisi si trovavano tra le prime 10 posizioni delle tabelle di uscita della ricerca del form, verranno segnalate con il seguente simbolo: §.

3.2.2 Analisi per categoria

Nella categoria sostantivo-modificatore è risultato più semplice trovare collocazioni, probabilmente anche perché è quella con una quantità di dati maggiore.

Sono state scelte arbitrariamente 4 parole, sulle quali si è svolta la ricerca di collocazioni tra sostantivi-modificatori; si tratta di 2 aggettivi e 2 sostantivi: grande, nero, casa, colpo.

Con l'aggettivo “grande” si sono rilevati molti nomi propri, termini tecnici ed espressioni che sono entrate nel linguaggio comune. Tra i nomi propri si evidenziano: *Rio grande* §, *grande Slam* §, *grande Guerra*, *grande Mela*, *grande Gatsby*, *grande Lebowski*, *grande Fratello*. Si possono inserire nella categoria dei termini tecnici le seguenti collocazioni: *grande carro* che potrebbe benissimo far parte dei nomi propri, visto che indica una costellazione in particolare, *grande distribuzione*, *grande schermo* §, *grande pubblico* §. Tra le espressioni idiomatiche sono state individuate invece collocazioni del tipo: *grande salto*, *grande spessore*, *grande passo*, *grande svolta*.

Tra le collocazioni composte dall'aggettivo “nero” non ci sono molti nomi propri, i termini tecnici sono in numero molto maggiore seguiti dai termini di uso comune. Gli unici nomi propri sono *Africa nera*, *Peste nera*, *Pinot nero*. Un'osservazione da fare è che la coppia *peste nera* si trova tra le prime 10 coppie estratte ma senza la lettera maiuscola, mentre la stessa coppia come nome proprio è rilevabile alla posizione 81. Le collocazioni *giovedì nero* e *martedì nero* potrebbero essere inserite anch'esse tra i nomi

propri, poiché indicano esattamente un giorno e un anno preciso. Tra i termini tecnici è possibile inserire queste collocazioni: *buco nero* §, *cintura nera* §, *cronaca nera* §, *scatola nera* §, *fumata nera*, *lavoro nero*, *mercato nero*. Alcune delle espressioni che sono entrate nel linguaggio comune e che sono state riconosciute come collocazioni sono: *pecora nera*, *bestia nera*, *lista nera*.

Cercando il sostantivo "casa", le collocazioni che sono emerse sono per lo più termini tecnici: *casa editrice* §, *casa discografica* §, *casa automobilistica* §, *casa produttrice* §, *casa regnante* §, *casa farmaceutica*, *casa popolare*, *casa circondariale*, *casa chiusa*. Il primo nome proprio che si incontra è *Casa bianca* e lo troviamo alla posizione 103.

Con il sostantivo "colpo" si sono individuate le seguenti espressioni idiomatiche: *duro colpo* §, *colpo secco*, *brutto colpo*, *colpo basso*, *colpo sicuro*.

Nella ricerca di collocazioni tra soggetto-verbo si sono trovate espressioni del linguaggio comune i cui soggetti sono nella maggioranza dei casi nomi astratti, di conseguenza il verbo prendeva spesso il significato figurato. Gli esempi che vengono proposti sono con i sostantivi "sogno" e "fortuna": *sogno infrangere* §; *sogno sfumare* §; *sogno coronare* §; *sogno tramontare; sogno coltivare*; *fortuna sorridere* §; *fortuna girare*; *fortuna baciare.*

In seguito alla ricerca nella categoria verbo-oggetto, si possono mettere in evidenza alcune collocazioni interessanti ottenute estraendo le coppie contenenti i verbi "gettare" e "mettere". Le seguenti collocazioni sono tutte espressioni idiomatiche: *gettare spugna* §; *gettare maschera*; *gettare fango*; *gettare benzina* (sul fuoco); *mettere mano* §; *mettere zizzania*; *mettere pulce* (nell'orecchio); *mettere pietra* (metterci una pietra sopra); *mettere naso*.

Per la categoria verbo-sintagma preposizionale si propongono alcuni risultati ottenuti con la ricerca dei verbi "prendere" ed "entrare". Le collocazioni di maggiore interesse sono: *prendere su serio* §; *prendere sotto ala* (dove "sotto" è una preposizione impropria); *prendere al volo*; *prendere per buono*; *entrare in carica*; *entrare in grazie*; *entrare in rosa*; *entrare in ballo*.

Due "errori" che vale la pena di menzionare sono: *prendere di posizione* e *prendere di corrente*. In entrambi i casi la forma che, di norma, dovrebbe assumere la prima parola all'interno del testo è "presa", di conseguenza avrebbero dovuto essere sostantivi e non verbi. Entrambe le triplette sono comunque collocazioni, nonostante non siano state riconosciute nell'esatta categoria.

Da quanto emerso dalle analisi svolte, si può quindi dire che la LMI è comunque una

buona misura statistica per scremare la grande quantità di dati che portano in genere le analisi linguistiche sui corpora di grandi dimensioni. Questa misura statistica permette quindi una velocizzazione del lavoro manuale di coloro che si apprestano a svolgere analisi di questo genere.

4. Appendice

4.1 Misure statistiche

Di seguito sono riportate le prime 30 coppie di parole che fanno parte della categoria sostantivo soggetto-modificatore che risultano trovarsi in cima alle liste decrescenti per PointwiseMI e LocalMI. Le coppie saranno accompagnate dalle rispettive frequenze singole e dalla frequenza delle due parole quando si trovano a ricorrere insieme, seguite appunto dalla misura statistica applicata. Si noterà immediatamente la differenza dei dati che emergono. Nella lista PMI troviamo infatti, per la stragrande maggioranza, parole straniere che non sono quasi mai riconosciute nella categoria grammaticale adeguata. Nella lista LMI troviamo invece coppie di parole che sono collocazioni in senso stretto e coppie di parole che nel linguaggio italiano sono molto utilizzate insieme e manifestano quindi una forte attrazione reciproca.

N°	Sostantivo soggetto	Modificatore	F1F2	F1	F2	PMI
1	zunzhe	jiaoxinglu	1	1	1	20.4668
2	zonta	furon	1	1	1	20.4668
3	ziemia	lubuska	1	1	1	20.4668
4	zhaoshi	qingxi	1	1	1	20.4668
5	zerro	musillo	1	1	1	20.4668
6	zendàle	sendale	1	1	1	20.4668
7	zekutsu	dachi	1	1	1	20.4668
8	Ze	leive	1	1	1	20.4668
9	zdobywcą	kosmosu	1	1	1	20.4668
10	zajdi	jasno	1	1	1	20.4668
11	Yt	Urimi	1	1	1	20.4668
12	Yo	hecho	1	1	1	20.4668
13	yini	lukousauro	1	1	1	20.4668
14	yidishe	avtonome	1	1	1	20.4668
15	yeshiva	ketana	1	1	1	20.4668
16	yanka	nyen	1	1	1	20.4668
17	xilene	cianolo	1	1	1	20.4668
18	xeroderma	pigmentoso	1	1	1	20.4668
19	xenocristalli	mantellici	1	1	1	20.4668
20	Xe	republica	1	1	1	20.4668
21	would	disagree	1	1	1	20.4668
22	wop	doo	1	1	1	20.4668
23	woogie	boogie	1	1	1	20.4668
24	wine	critical	1	1	1	20.4668
25	wind	winter	1	1	1	20.4668
26	wickets	cancelletti	1	1	1	20.4668
27	where	phase	1	1	1	20.4668

28	were	alive	1	1	1	20.4668
29	welter	superwelter	1	1	1	20.4668
30	we	dansu	1	1	1	20.4668

N°	Sostantivo soggetto	Modificatore	F1F2	F1	F2	LMI
1	parte	maggior	10005	22830	11310	58141.5216
2	anno	successivo	4864	11135	14014	26738.8669
3	cenno	storico	3246	5633	11696	19988.2171
4	bambino	uguale	2067	2461	2432	18535.0655
5	cenno	biografico	2179	5633	2799	16660.2734
6	geografia	fisico	1781	1980	4591	14514.0136
7	anno	seguente	2189	11135	5213	12635.1525
8	parte	gran	2065	22830	3363	10912.5282
9	colonna	sonoro	1243	2017	2179	10787.9128
10	sede	vescovile	1179	4115	1422	9655.6953
11	stazione	meteorologico	1198	3567	2149	9372.2349
12	Vita	privato	1225	3712	2777	9099.3504
13	caratteristica	tecnico	1266	6386	4361	7648.7784
14	territorio	comunale	1038	3241	3449	7340.9151
15	guerra	mondiale	1081	4442	3957	7002.4297
16	chiesa	parrocchiale	881	4990	1306	6707.9362
17	collegamento	esterno	860	1634	3476	6688.6979
18	Dato	climatologici	721	2894	724	6461.5500
19	Stazione	meteorologico	674	771	2149	6203.0481
20	guerra	civile	886	4442	2704	5971.6908
21	Chiesa	cattolico	710	1848	2033	5749.0824
22	temperatura	medio	854	2713	4390	5721.1008
23	campionato	mondiale	881	4316	3957	5483.4279
24	coordinata	geografico	568	762	1769	5256.3567
25	velocità	massimo	741	1735	4567	5248.0186
26	Casa	editore	606	3355	717	5158.2812
27	vicariato	apostolico	508	532	1586	4962.6483
28	prefettura	apostolico	506	531	1586	4941.6041
29	suddivisione	amministrativo	591	940	3043	4861.5579
30	stazione	ferroviario	714	3567	2603	4855.2662

4.2 Script

Di seguito sono riportati per intero gli script elaborati. Ciascuna riga di codice inizia con un numero che ovviamente non fa parte dello script, ma facilita l'individuazione delle porzioni spiegate nel capitolo relativo.

4.2.1 Estrazione coppie sostantivo-modificatore

```
1  while(<>) {
2    chomp;
```

```
3    @line = split;
4    if ($line[0]=~/^[0-9][0-9]?/) {
5     $val=$line[0];
6     $a=1;
7      FRASE:for ($val==$a; $a++;) {
8       $r=\@line;
9       if ($val==1) {
10          %frase = ();
11          %nome = ();
12          %mod = ();
13      }
14      %frase = ($r[0] => $r);
15    MOD:if ($frase{$r[0]}[7]=~/^mod$/){
16        if ($frase{$r[0]}[4]=~/^A$/){
17          if ($frase{$r[0]}[6]>0){
18            if ($frase{$r[0]}[2]=~/^[A-Za-z]+[^@#\/%\=\$\^\*][-]?/){
19                if ($frase{$r[0]}[2]=~/^tale$|^Tale$|^tali$|^Tali$|
^stesso$|^primo$|^secondo$/){
20                }
21                else{
22                    $lemmamod=\$frase{$r[0]}[2];
23                    $cod=$frase{$r[0]}[0];
24                    $head=$frase{$r[0]}[6];
25                    %mod = ($cod=> [$head,$$lemmamod]);
26              }
27            }
28          }
29        }
30      }
31    NOME:if ($frase{$r[0]}[3]=~/^S$/){
32          if ($frase{$r[0]}[2]=~/^[A-Za-z]+[^@#\/%\=\$\^\*][-]?/){
33              $id=$frase{$r[0]}[0];
34              $lemmanome=\$frase{$r[0]}[2];
35              %nome= ($id=> [$id,$$lemmanome]);
36          }
37        }
38    ID: {
39        if($nome{$id}[0]==$mod{$cod}[0]){
40          print "$nome{$id}[1]\t$mod{$cod}[1]\n";
41            delete $mod{$cod};
42        }
```

```
43      }
44      %frase = ();
45      last FRASE;
46      }
47    }
48  }
```

4.2.2 Estrazione coppie verbo-soggetto

```
1  while(<>) {
2    chomp;
3    @line = split;
4    if ($line[0]=~/^[0-9][0-9]?/) {
5     $val=$line[0];
6     $a=1;
7      FRASE:for ($val==$a; $a++;) {
8       $r=\@line;
9       if ($val==1) {
10          %frase = ();
11          %verb = ();
12          %subj = ();
13      }
14      %frase = ($r[0] => $r);
15    SUBJ:if ($frase{$r[0]}[7]=~/^subj$|^subj_pass$/){
16        if ($frase{$r[0]}[4]=~/^S$/){
17          if ($frase{$r[0]}[2]=~/^[A-Za-z]+[^@#\/%\=\$\^\*][-]?/){
18              $lemmasubj=\$frase{$r[0]}[2];
19              $cod=$frase{$r[0]}[0];
20              $head=$frase{$r[0]}[6];
21              $head=int $head;
22              $cod=int $cod;
23              %subj = ($cod=> [$head,$$lemmasubj]);
24          }
25        }
26    }
27    VERB:
28      if ($frase{$r[0]}[3]=~/^V$/){
29          if ($frase{$r[0]}[2]=~/^[A-Za-z]+[^@#\/%\=\$\^\*][-]?/){
30            $id=$frase{$r[0]}[0];
31            $lemmaverb=\$frase{$r[0]}[2];
32            %verb = ($id=> [$id,$$lemmaverb]);
```

```
33          }
34       }
35     ID: {
36         if($verb{$id}[0]==$subj{$cod}[0]){
37           print "$verb{$id}[1]\t$subj{$cod}[1]\n";
38           delete $subj{$cod};
39         }
40       }
41       %frase = ();
42       last FRASE;
43       }
44     }
45 }
```

4.2.3 Estrazione coppie verbo-oggetto

```
1  while(<>) {
2    chomp;
3    @line = split;
4    if ($line[0]=~/^[0-9][0-9]?/) {
5     $val=$line[0];
6     $a=1;
7      FRASE:for ($val==$a; $a++;) {
8       $r=\@line;
9       if ($val==1) {
10          %frase = ();
11          %verb = ();
12          %obj = ();
13       }
14       %frase = ($r[0] => $r);
15     OBJ:if ($frase{$r[0]}[7]=~/^obj$/){
16          if ($frase{$r[0]}[4]=~/^S$/){
17           if ($frase{$r[0]}[2]=~/^[A-Za-z]+[^@#\/%\=\$\^\*][-]?/){
18               $lemmaobj=\$frase{$r[0]}[2];
19               $cod=$frase{$r[0]}[0];
20               $head=$frase{$r[0]}[6];
21               $head=int $head;
22               $cod=int $cod;
23               %obj = ($cod=> [$head,$$lemmaobj]);
24           }
25         }
26     }
```

```
27    VERB:if ($frase{$r[0]}[3]=~/^V$/){
28        if ($frase{$r[0]}[2]=~/^[A-Za-z]+[^@#\/%\=\$\^\*][-]?/){
29          $id=$frase{$r[0]}[0];
30          $lemmaverb=\$frase{$r[0]}[2];
31          %verb = ($id=> [$id,$$lemmaverb]);
32        }
33    }
34    ID: {
35        if($verb{$id}[0]==$obj{$cod}[0]){
36          print "$verb{$id}[1]\t$obj{$cod}[1]\n";
37          delete $obj{$cod};
38        }
39      }
40      %frase = ();
41      last FRASE;
42       }
43     }
44 }
```

4.2.4 Estrazione coppie verbo-sintagma preposizionale

```
1  while(<>) {
2      chomp;
3      @line = split;
4    if ($line[0]=~/^[0-9][0-9]?/) {
5        $val=$line[0];
6        $a=1;
7      FRASE:for ($val==$a; $a++;) {
8          $r=\@line;
9        if ($val==1) {
10         %frase = ();
11         %prep = ();
12         %comp = ();
13         %verbs = ();
14       }
15        %frase = ($r[0] => $r);
16      PREP:if ($frase{$r[0]}[7]=~/^prep$/){
17        if ($frase{$r[0]}[2]=~/^[A-Za-z]+[^@#\/%\=\$\^\*][-]?/){
18          $lemmaprep=\$frase{$r[0]}[2];
19          $cod=$frase{$r[0]}[0];
20          $head=$frase{$r[0]}[6];
21          $head=int $head;
```

```
22            $cod=int $cod;
23            %prep = ($cod=> [$head,$$lemmaprep,$cod]);
24          }
25        }
26        COMP:if($frase{$r[0]}[7]=~/^comp$/){
27          if ($frase{$r[0]}[2]=~/^[A-Za-z]+[^@#\/%\=\$\^\*][-]?/){
28            $id=$frase{$r[0]}[0];
29            $rif=$frase{$r[0]}[6];
30            $lemmacomp=\$frase{$r[0]}[2];
31            %comp = ($id=> [$id,$$lemmacomp,$rif]);
32          }
33        }
34        VERB:if($frase{$r[0]}[3]=~/^V$/){
35          if ($frase{$r[0]}[2]=~/^[A-Za-z]+[^@#\/%\=\$\^\*][-]?/){
36            $verb=$frase{$r[0]}[0];
37            $lemmaverb=\$frase{$r[0]}[2];
38            %verbs = ($verb=> [$verb,$$lemmaverb]);
39          }
40        }
41        SINT_PREP:if($comp{$id}[0]==$prep{$cod}[0]){
42           $head_comp=$comp{$id}[2];
43           $sint_prep="$comp{$id}[1]_$prep{$cod}[1]";
44           %sint=($head_comp=>[$head_comp,$sint_prep]);
45           delete $comp{$id};
46           delete $prep{$cod};
47        }
48        STAMPA:if($sint{$head_comp}[0]==$verbs{$verb}[0]){
49           print "$verbs{$verb}[1]\t$sint{$head_comp}[1]\n";
50           delete $verbs{$verb};
51        }
52        %frase = ();
53        last FRASE;
54      }
55    }
56 }
```

4.3 Coppie dei primi 100 risultati con LMI maggiore

Nelle tabelle delle sezioni seguenti, viene mostrato il numero di ranking seguito dai singoli componenti della coppia nell'ordine indicato nel titolo della sezione corrispondente. Inoltre vengono sottolineate le coppie di parole che sono state

considerate collocazioni e quelle considerate errori sono segnalate con il simbolo *.

4.3.1 Sostantivo-Modificatore

1	comune	francese
2	guerra	mondiale
3	parte	maggior
4	semiasse	pari
5	anno	successivo
6	semiasse	maggiore
7	fascia	principale
8	parte	gran
9	comunità	autonomo
10	spagnolo	comune*
11	colonna	sonoro
12	anno	ultimo
13	serie	televisivo
14	anno	seguente
15	Chiesa	cattolico
16	casa	editore
17	guerra	civile
18	sistema	operativo
19	successo	grande
20	territorio	comunale
21	battezzato	corrispondente*
22	casa	discografico
23	centro	storico
24	cenno	storico
25	essere	umano
26	cenno	biografico
27	forza	armato
28	parte	facente*
29	stazione	ferroviario
30	suffraganea	cattolico
31	comune	tedesco
32	regione	storico
33	parte	buono
34	periodo	breve
35	chiesa	parrocchiale
36	stato	federato
37	tempo	breve
38	numero	gran
39	stazione	meteorologico
40	famiglia	nobile
41	gruppo	musicale
42	arto	marziale*
43	città	natale

51	secolo	scorso
52	campionato	mondiale
53	anno	solo
54	etichetta	discografico
55	strada	statale
56	vicariato	apostolico
57	serie	massimo
58	altare	maggiore
59	linea	ferroviario
60	divisione	amministrativo
61	catena	montuoso
62	posto	terzo
63	età	inferiore
64	nome	proprio
65	stagione	successivo
66	vita	privato
67	campo	magnetico
68	ruolo	importante
69	velocità	massimo
70	opinione	pubblico
71	cantone	francese
72	periodo	lungo
73	carro	armato
74	impero	romano
75	epoca	romano
76	compagnia	aereo
77	coordinata	geografico
78	elezione	politico
79	quartier	generale
80	cartone	animato
81	album	terzo
82	parco	nazionale
83	politica	estero
84	programma	televisivo
85	sistema	solare
86	scuola	elementare
87	fratello	minore
88	suddivisione	amministrativo
89	gruppo	etnico
90	anno	tardo
91	acqua	dolce
92	temperatura	medio
93	serie	animato

44	cinta	murario
45	diametro	medio
46	mitologia	greco
47	vittoria	finale
48	geografia	fisico
49	Giochi	olimpico
50	sede	vescovile

94	web	Sito*
95	circondario	rurale
96	secolo	successivo
97	livello	alto
98	anno	precedente
99	fratello	maggiore
100	di	Euro*

4.3.2 Verbo-Soggetto

1	derivare	nome
2	contare	superficie
3	assommare	bambino
4	contare	diocesi
5	assommare	popolazione
6	comprendere	diocesi
7	rientrare	città
8	essere	capoluogo
9	vedere	edizione
10	erigere	diocesi*
11	situare	città*
12	essere	Cantone
13	essere	esempio
14	essere	scopo
15	risalire	origine
16	eleggere	presidente*
17	significare	nome
18	volere	tradizione
19	derivare	termine
20	basare	economia
21	estendere	territorio
22	trovare	cattedrale
23	giocare	partita
24	nascere	figlio
25	iniziare	carriera
26	indicare	termine
27	essere	caratteristica
28	trasferire	famiglia
29	decidere	di*
30	certificare	Bureau
31	volere	leggenda
32	esistere	versione
33	uscire	album
34	avvenire	debutto
35	festeggiare	onomastico*
36	ordinare	sacerdote*

51	dotare	di*
52	deporre	femmina
53	reggere	territorio*
54	trovare	stazione
55	avvenire	esordio
56	trovare	chiesa
57	derivare	parola
58	svolgere	attività
59	contenere	album
60	celebrare	festa*
61	prevedere	progetto
62	iniziare	costruzione
63	fallire	tentativo
64	contare	arcidiocesi
65	valere	pena
66	comprendere	distretto
67	narrare	storia
68	essere	capocannoniere
69	giocare	squadra
70	fare	arrondissement
71	iniziare	del*
72	raccontare	storia
73	identificare	cratere*
74	svolgere	gara
75	morire	padre
76	esistere	tipo
77	comporre	dipartimento
78	tenere	elezione
79	scorrere	fiume
80	comporre	da*
81	svolgere	edizione
82	partecipare	squadra
83	comprendere	arcidiocesi
84	militare	squadra
85	risalire	notizia
86	svolgere	campionato

37	essere	contea
38	sposare	figlio
39	usare	termine*
40	essere	sede
41	essere	risultato
42	eleggere	deputato*
43	iniziare	lavorio
44	essere	obiettivo
45	narrare	leggenda
46	attestare	trentennale
47	scoppiare	guerra
48	ammontare	popolazione
49	essere	Dioecesis
50	riferire	termine

87	nominare	presidente
88	andare	cosa
89	Olimpiade	Nuoto*
90	sciogliere	gruppo
91	risalire	fondazione
92	derivare	toponimo
93	essere	differenza
94	giocare	stagione
95	tenere	edizione
96	arrivare	successo
97	erigere	prefettura*
98	sorgere	chiesa
99	mancare	giorno
100	ritenere	studioso

4.3.3 Verbo-Oggetto

1	fare	parte
2	presentare	orbita
3	prendere	nome
4	contare	battezzato
5	assommare	femmina
6	prendere	parte
7	contare	abitanti
8	giocare	partita
9	dare	vita
10	iniziare	carriera
11	avere	figlio
12	vincere	premio
13	rendere	conto
14	riscuotere	successo
15	vincere	campionato
16	segnare	gol
17	tenere	conto
18	particolar	modo*
19	firmare	contratto
20	interpretare	ruolo
21	ottenere	successo
22	porre	fine
23	cambiare	nome
24	risolvere	problema
25	vincere	medaglia
26	prestare	servizio
27	fare	riferimento
28	vincere	titolo
29	pubblicare	album

51	svolgere	ruolo
52	potere	essere*
53	svolgere	attività
54	ricevere	premio
55	svolgere	funzione
56	attirare	attenzione
57	prendere	posto
58	dare	inizio
59	fare	uso
60	avere	inizio
61	suonare	chitarra
62	disputare	partita
63	assumere	nome
64	dare	nome
65	ricoprire	incarico
66	avere	sede
67	conseguire	laurea
68	scrivere	libro
69	avere	bisogno
70	vincere	scudetto
71	comprendere	città
72	aprire	porta
73	portare	nome
74	stringere	amicizia
75	vendere	copia
76	dare	luogo
77	girare	film
78	totalizzare	presenza
79	uscire	album

30	avere	successo
31	ricavare	territorio
32	frequentare	scuola
33	segnare	rete
34	collezionare	presenza
35	ricoprire	ruolo
36	dare	origine
37	fare	ritorno
38	ricoprire	carica
39	vincere	coppa
40	esistere	francesi
41	cedere	porzione
42	vedere	vittoria
43	ottenere	risultato
44	raggiungere	posizione
45	intraprendere	carriera
46	avere	luogo
47	vestire	Maglia
48	chiedere	Aiuto
49	raccontare	Storia
50	disputare	Campionato

80	giocare	ruolo
81	registrare	album
82	avere	origine
83	pubblicare	libro
84	avere	scopo
85	dovere	nome
86	parlare	lingua
87	andare	incontro*
88	fare	fronte
89	deporre	uovo
90	fare	apparizione
91	assumere	denominazione
92	prendere	possesso
93	rassegnare	dimissione
94	salvare	vita
95	avere	compito
96	dichiarare	guerra
97	vedere	luce
98	ottenere	vittoria
99	avere	ruolo
100	distare	chilometro

4.3.4 Verbo-Sintagma preposizionale

1	situare	in_dipartimento
2	caratterizzare	da_semiasse
3	situare	in_comunità
4	ubicare	in_distretto
5	risalire	al_secolo
6	situare	in_land
7	dirigere	da_regista
8	andare	in_onda
9	suddividere	in_parrocchia
10	reggere	da_vescovo
11	appartenere	al_famiglia
12	estrarre	da_album
13	formare	da_unione
14	rientrare	in_classe
15	partire	da_anno
16	consentire	in_provincia
17	mettere	in_luce
18	prendere	in_considerazione
19	elevare	al_rango
20	conoscere	con_nome
21	entrare	in_servizio
22	portare	con_sé*

51	nascere	da_famiglia
52	dare	al_luce
53	interpretare	da_attore
54	trarre	da_romanzo
55	caratterizzare	da_presenza
56	partire	da_metà
57	scendere	in_campo
58	tradurre	in_lingua
59	pubblicare	da_Comics
60	laureare	in_giurisprudenza
61	durare	fino_al*
62	cadere	in_mano
63	reggere	da_arcivescovo
64	battere	in_finale
65	iscrivere	al_facoltà
66	situare	in_contea
67	dedicare	al_studio
68	elevare	in_concistoro
69	tornare	in_patria
70	entrare	in_vigore
71	dotare	di_motore
72	fregiare	di_titolo

23	situare	in_distretto
24	dovere	al_fatto
25	radere	al_suolo
26	entrare	in_contatto
27	partire	da_secolo
28	derivare	da_latino
29	situare	in_provincia
30	mettere	in_evidenza
31	salire	al_trono
32	ricavare	da_diocesi
33	passare	di_tempo
34	aderire	al_gruppo
35	mettere	in_scena
36	derivare	da_greco
37	iniziare	di_secolo
38	partecipare	al_campionato
39	apparire	per_volta
40	derivare	da_parola
41	realizzare	da_pittore
42	dividere	in_parte
43	tradurre	in_italiano
44	sconfiggere	in_battaglia
45	recitare	in_film
46	risalire	al_periodo
47	mettere	in_atto
48	pubblicare	da_Records
49	mettere	in_discussione
50	derivare	da_nome

73	partecipare	al_guerra
74	militare	in_campionato
75	riuscire	in_impresa
76	risalire	al_anno
77	arruolare	in_esercito
78	aprire	al_pubblico
79	progettare	da_architetto
80	trovare	di_fronte
81	iniziare	di_anno
82	aderire	al_partito
83	posporre	al_nome
84	mettere	in_mostra
85	derivare	da_fatto
86	riuscire	in_intento
87	apparire	in_film
88	ricavare	da_arcidiocesi
89	cedere	in_prestito
90	situare	in_regione
91	nascere	in_famiglia
92	estendere	su_superficie
93	sposare	in_nozze
94	trasferire	con_famiglia
95	costringere	al_ritiro
96	mandare	in_onda
97	morire	al_età
98	risalire	al_metà
99	inserire	in_album
100	partire	da_fine

5. Bibliografia

Bachmann Christina, Luccio Riccardo e Salvadori Emilia 2005. La verifica della significatività dell'ipotesi nulla in psicologia. Firenze, Firenze University Press.

Baroni, Marco, e Stefan Evert. 2009. *Statistical methods for corpus exploitation*. In (a cura di) A. Lüdeling e M. Kytö. "Corpus Linguistics. An International Handbook". (Vol.2). Berlino, Mouton de Gruyter. pp. 777-803.

Collet, Tanja. 1997. *La réduction des unités terminologiques complexes de type syntagmatique.* In "META". (XLII,1). pp.193-206.

Evert, Stefan. 2007. Corpora and collocations. In (a cura di) A. Lüdeling e M. Kytö. "Corpus Linguistics. An International Handbook". (Vol.2) Berlino, Mouton de Gruyter. pp. 1212-1248.

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

Firth, John Rupert. 1968. *A synopsis of Linguistic theory, 1930-1955*. Oxford, Basil Blackwell.

Lenci, Alessandro, Simonetta Montemagni e Vito Pirrelli. 2005. *Testo e computer: Elementi di linguistica computazionale*. Roma, Carocci.

Manning, Christopher. e Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MIT Press.

Sag, Ivan A., Timoty Baldwin, Francis Bond, Ann Coperstake e Dan Flickinger. 2002 *Multiword Expressions: A Pain in the Neck for NLP* In (a cura di) Alexander Gelbukh. "Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2002)". Volume 2276 di (Lecture Notes in Computer Science). Berlino, Springer-Verlag. pp.1-15.

Soliani, Lamberto 2005. *Statistica Univariata e Bivariata, Parametrica e Non Parametrica per le Discipline Ambientali e Biologiche*. Parma, UNI.NOVA.

Siti web consultati

Colagrande, Vittorio "Alcuni elementi di verifica di ipotesi statistiche". http://www.biostatistica.unich.it/mat_didattica/Odont/Verif_IPOTESI_odonto.pdf

Quaderno di epidemiologia veterinaria.
http://www.quadernodiepidemiologia.it/epi/assoc/t_stu.htm

Materiale corso Intelligenza artificiale.

http://www.psico.unitn.it/didattica/corsi/1023/

Laboratorio virtuale di probabilità e statistica.

http://www.ds.unifi.it/VL/VL_IT/index.html