Marco S. G. Senaldi[*†], Gianluca E. Lebani[*], Lucia Passaro[*] & Alessandro Lenci[*]

[*]CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa

[†]Laboratorio di Linguistica "G. Nencioni", Scuola Normale Superiore

SEMANT-IT

SPAZI SEMANTICI DISTRIBUZIONALI IN COLFIS

Con l'ausilio del toolkit DISSECT (Dinu *et al.* 2013), sono stati realizzati degli spazi distribuzionali per il corpus CoLFIS (Bertinetto *et al.* 2005). I dati ottenuti sono stati poi rappresentati sotto forma di grafi come il seguente, dove i nodi corrispondono alle parole del corpus e gli archi rappresentano le relazioni di similarità semantica tra le parole, calcolate in termini di coseno:

http://colinglab.humnet.unipi.it/Demo/COLFIS/colfis_10_2_top10K/

Segue una descrizione dettagliata del processo di estrazione dei dati distribuzionali dal corpus. In fondo alla pagina si trovano i link che rimandano ai grafi realizzati.

1) Estrazione delle statistiche di coricorrenza

Partendo dai file lemmatizzati dei testi autorizzati e non autorizzati, sono stati estratti tre tipi di statistiche di coricorrenza, che come finestra di contesto utilizzano rispettivamente:

1) le 2 parole contenuto[1] (nomi, verbi, aggettivi ed avverbi) immediatamente precedenti e immediatamente seguenti la parola target all'interno della stessa frase;
2) le 5 parole contenuto immediatamente precedenti e immediatamente seguenti la parola target all'interno della stessa frase;
3) le 20 parole contenuto immediatamente precedenti e immediatamente seguenti la parola target all'interno della stessa frase;

Nel caso 1) si otterranno, ad esempio, le seguenti statistiche, in ordine decrescente di frequenza:

essere-V	non-B	7000
essere-V	piu'-B	2722
essere-V	essere-V	2144
essere-V	suo-G	1371
essere-V	anche-B	1244

[1] Le parole contenuto sono così lemmatizzate: nomi comuni = S, nomi propri = E, verbi = V, aggettivi = G, avverbi = B. Per gli altri criteri di lemmatizzazione, si vedano le note di Cristina Burani al link http://linguistica.sns.it/CoLFIS/Lemmatiz/Criteri%20di%20lemmatizzazione.pdf

essere-V	potere-V	1220
essere-V	questo-G	1146
essere-V	molto-B	1037
essere-V	fare-V	983
etc.		

L'avverbio "non" ricorre quindi nel corpus 7000 volte alla distanza di 2 parole contenuto dal verbo target "essere", l'avverbio "più" 2722 volte alla distanza di 2 parole contenuto da "essere" e così via.

2) Selezione dei vettori target e delle parole contesto

Per costruire gli spazi distribuzionali, partendo dalle statistiche di coricorrenza sopra descritte sono state selezionate come vettori target le parole con una frequenza superiore ai 25 token e come dimensioni le parole con una frequenza superiore ai 10 token.

Da questa selezione sono state escluse anche alcune *stopwords*, comprendenti parole ad elevata frequenza e con un basso valore informativo e altre parole ritenute comunque non distintive ai fini di catturare la semantica delle parole target. Queste *stopwords* comprendono:

- i verbi *essere*, *avere*, *fare* e i verbi modali
- i nomi propri
- gli avverbi
- gli aggettivi dimostrativi e possessivi

Sono stati così ottenuti tre spazi distribuzionali, uno per ogni finestra di contesto utilizzata (± 2, ± 5, ± 20), ciascuno dei quali composto di **6465 vettori target** e **11273 vettori contesto**.

I tre spazi sono rappresentati sotto forma di matrici sparse. Nel caso dello spazio con finestra contestuale ± 2, si avrà una rappresentazione come la seguente:

dire-V	cosa-S	177
dire-V	sapere-V	174
dire-V	andare-V	145
dire-V	sentire-V	130
dire-V	stare-V	128
dire-V	tutto-G[2]	119
dire-V	vedere-V	118
dire-V	anno-S	102
dire-V	volta-S	90
etc.		

[2] Si tenga presente che nella lemmatizzazione adottata in CoLFIS vengono annotati come aggettivi ('-G') anche i quantificatori.

In questa rappresentazione, le parole della prima colonna sono i vettori target, mentre le parole della seconda colonna costituiscono i contesti. I valori della terza colonna rappresentano ancora le frequenze di coricorrenza grezze.

3) Pesatura delle coricorrenze e riduzione della dimensionalità

Le coricorrenze grezze sono state poi pesate mediante la PPMI (Positive Pointwise Mutual Information; Niwa & Nitta 1994), una misura di associazione statistica che rileva se due parole coricorrono con una frequenza superiore al caso. Nel caso specifico della PPMI, i valori di associazione negativi vengono portati a 0:

$$\text{PPMI(t; c)} \equiv log \frac{p(t,c)}{p(t)p(c)}$$

Nella formula, *t* rappresenta una parola target e *c* una parola contesto.

I dati sopra riportati dello spazio con finestra ± 2 sono stati ad esempio convertiti nella rappresentazione sottostante, con la frequenza grezza sostituita dalla PPMI:

dire-V	cosa-S	1.687587
dire-V	sapere-V	1.352569
dire-V	andare-V	0.991734
dire-V	sentire-V	1.537674
dire-V	stare-V	0.718565
dire-V	tutto-G	0.217043
dire-V	vedere-V	0.784856
dire-V	volta-S	0.659078
etc.		

Le dimensioni della matrice sono state poi ridotte a 100 mediante SVD (Singular Value Decomposition; Deerwester *et al.* 1990).

4) Calcolo dei coseni

La similarità semantica tra i vettori target è stata calcolata con il coseno:

$$sim_{cos}(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2} \sqrt{\sum_{i=1}^{N} y_i^2}}$$

Per ciascuno dei tre spazi, sono state estratte le dieci parole più simili ad ogni parola target in termini di coseno.

A titolo di esempio, vengono qui di seguito elencate le dieci parole che risultano più vicine a "proiettile" nello spazio con finestra contestuale ± 2, ordinate per valori decrescenti di coseno:

proiettile-S	calibro-S	0.85510579507
proiettile-S	pallottola-S	0.842365463724
proiettile-S	pistola-S	0.835189138589
proiettile-S	sparare-V	0.828515192148
proiettile-S	fucile-S	0.774163305693
proiettile-S	torace-S	0.76851562214
proiettile-S	colpo-S	0.756969933026
proiettile-S	centrare-V	0.704648028778
proiettile-S	impugnare-V	0.703718878166
proiettile-S	granata-S	0.702945574337

5) Realizzazione dei grafi

Per rendere agevole la consultazione dei dati, gli spazi distribuzionali sono stati rappresentati sotto forma di grafi mediante *Gephi* (Bastian *et al.* 2009), un software *open-source* per la creazione e l'esplorazione di reti. Nei grafi ottenuti, i nodi rappresentano le parole, mentre gli archi che connettono i nodi rappresentano le relazioni di similarità semantica misurate attraverso i coseni.

Si ritorni all'esempio della parola "proiettile" nello spazio con finestra contestuale ± 2. Nell'immagine sottostante (Figura 1), si può vedere la parola in esame, rappresentata dal rispettivo nodo, connessa ad altre parole attraverso degli archi provvisti di un peso pari alla loro distanza di coseno. Sul lato sinistro dell'interfaccia vengono elencate le dieci parole più vicine a "proiettile" in termini di coseno (*10 nearest neighbours*) e tutte le parole che presentano "proiettile" tra le dieci parole ad esse più vicine (*In the 10 nearest neighbours of*).

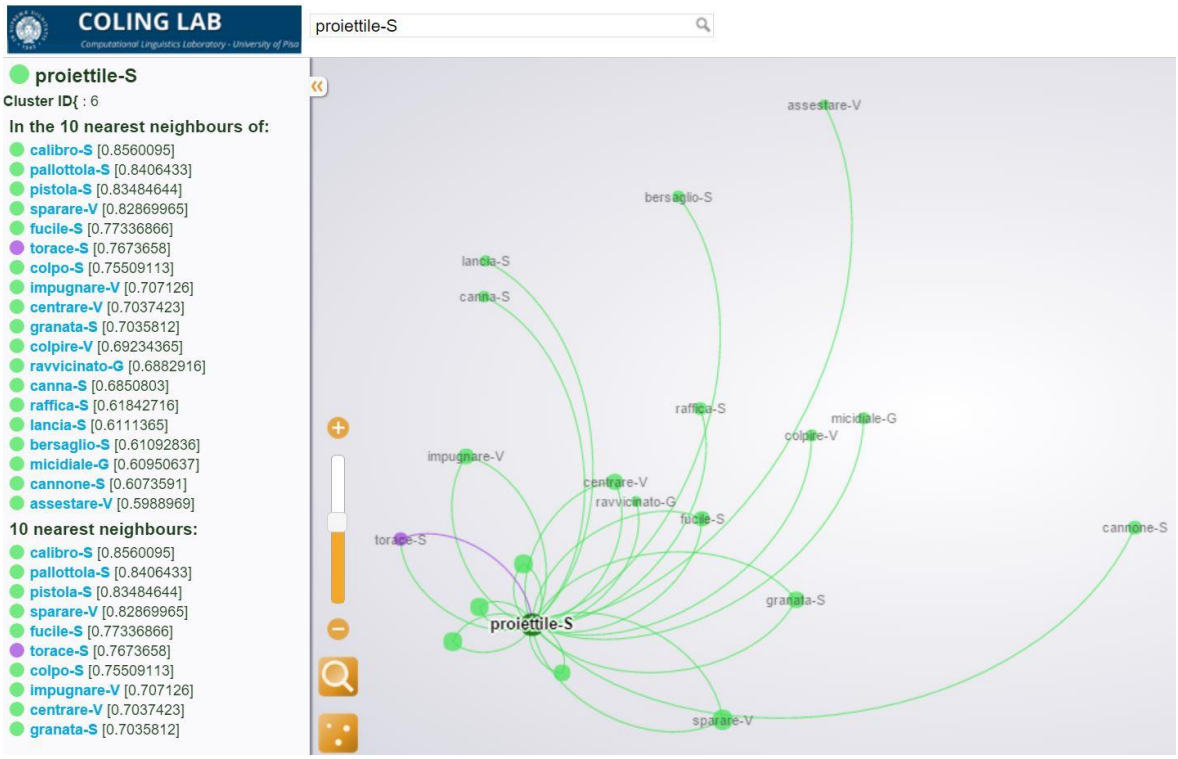

Figura 1: rappresentazione della parola “proiettile” nel grafo ottenuto dallo spazio con finestra contestuale ± 2

L’organizzazione spaziale del grafo viene determinata da *Force Atlas* (Jacomy *et al.* 2014), un algoritmo di visualizzazione che simula un sistema fisico in cui i nodi si respingono come particelle cariche e gli archi attraggono i nodi fino al raggiungimento di uno stato di equilibrio. In questo caso, la forza di attrazione tra i nodi è costituita dal coseno, mentre la forza di repulsione è rappresentata da una costante determinata empiricamente per impedire la sovrapposizione dei nodi.

Per permettere una navigazione efficace del grafo, è stato successivamente applicato l’algoritmo *Label Adjust* (https://gephi.org/tutorials/gephi-tutorial-visualization.pdf) che aggiusta la disposizione delle etichette dei nodi per evitarne la sovrapposizione (Figura 2).

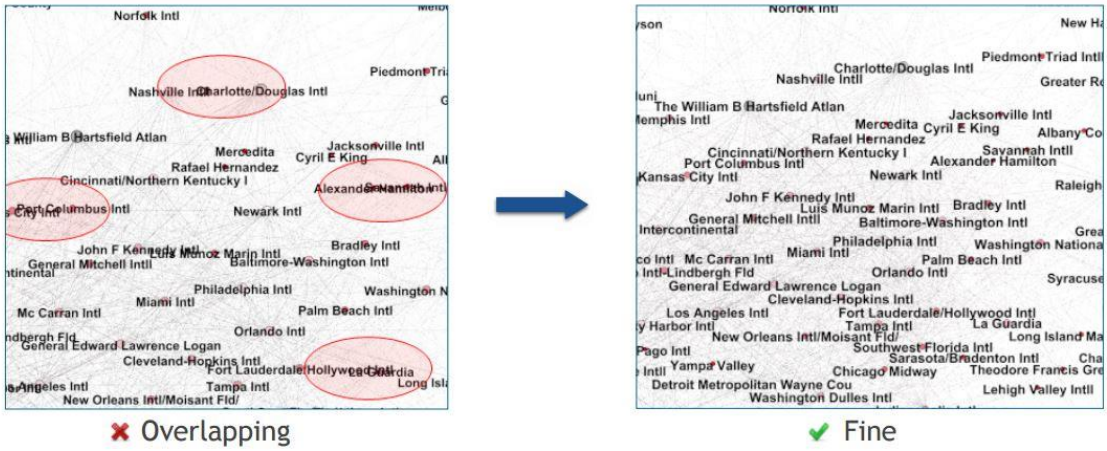

Figura 2: esempio di applicazione dell’algoritmo *Label Adjust*

Attraverso il *metodo di Louvain* (Blondel *et al.* 2008), i nodi del grafo vengono infine raggruppati in *cluster* (o *comunità*), corrispondenti in linea di massima a differenti aree semantiche. Nel grafo creato dallo spazio con finestra ± 2, è possibile ad esempio osservare un *cluster* che riunisce parole inerenti alla sfera politica (Figura 3), un *cluster* relativo alle emozioni (Figura 4), un *cluster* che racchiude parole indicanti relazioni di parentela (Figura 5) e via dicendo. Ognuno dei *cluster* è contraddistinto da un numero e da uno specifico colore dei nodi. Nella Figura 1, sul lato sinistro dell’interfaccia viene riportato sotto *Cluster ID* il numero del cluster a cui appartiene, ad esempio, la parola “proiettile”.

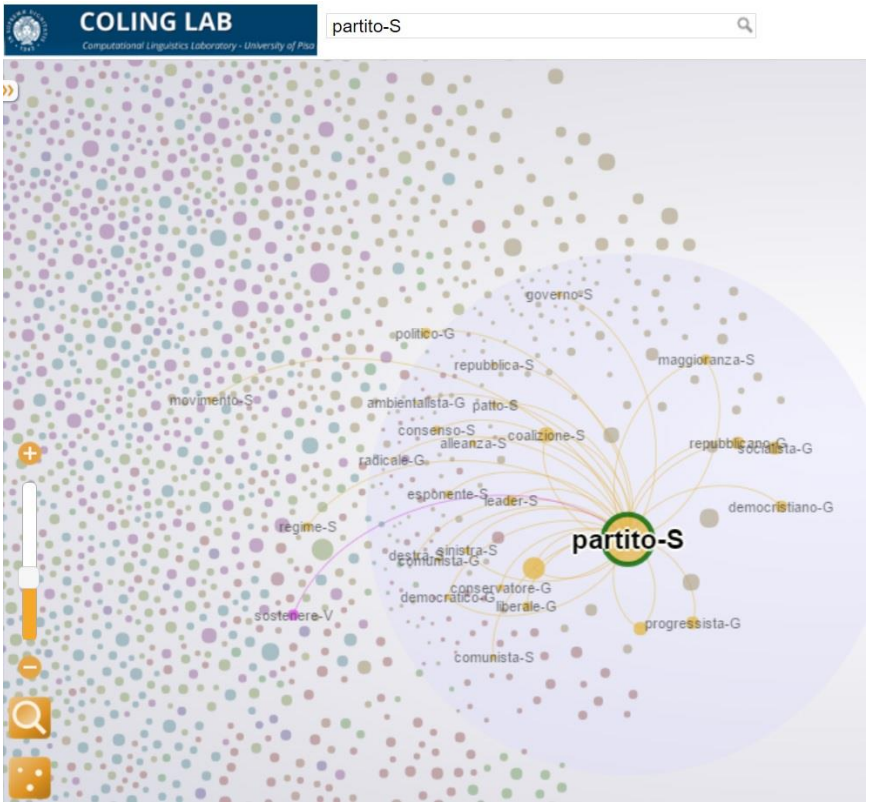

Figura 3: *cluster* di parole appartenenti alla sfera politica nel grafo dello spazio con finestra ± 2

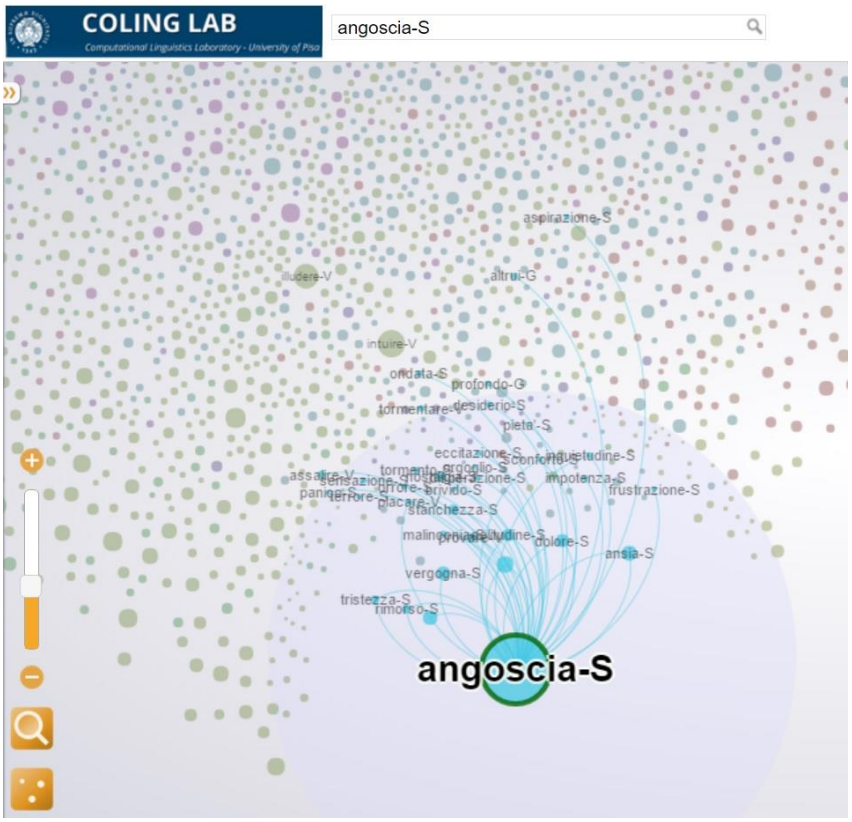

Figura 4: *cluster* di parole appartenenti alla sfera delle emozioni nel grafo dello spazio con finestra ± 2

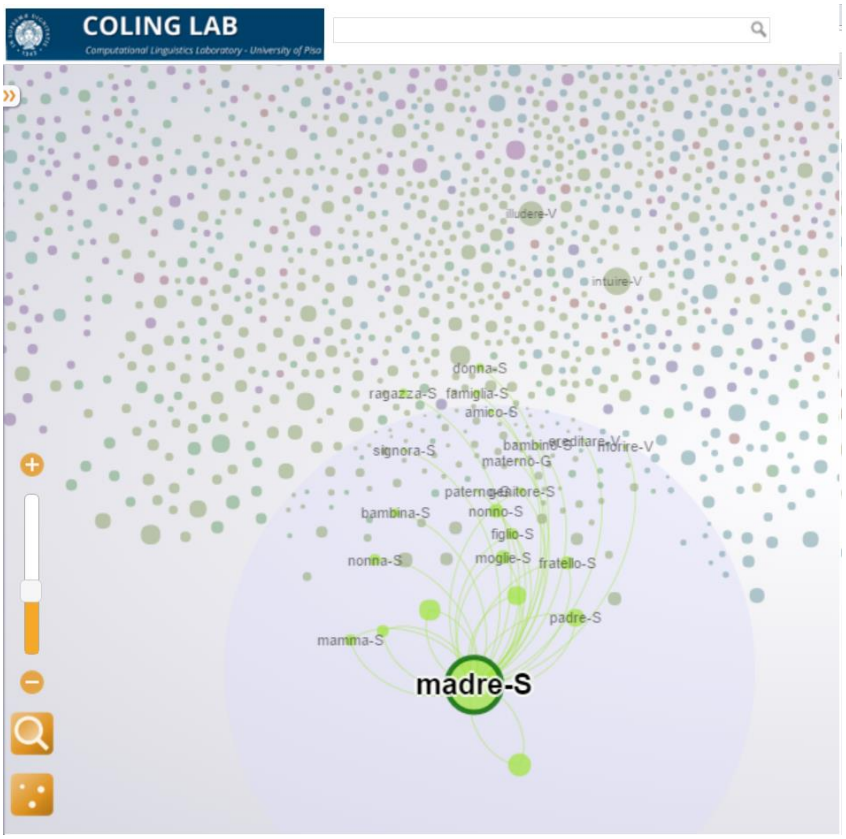

Figura 5: *cluster* di parole indicanti relazioni di parentela nel grafo dello spazio con finestra ± 2

6) Link

Per ciascuno dei tre spazi sono stati realizzati due grafi. Il primo riporta tutti i dati nella loro interezza, fornendo quindi tutte le dieci parole più vicine in termini di coseno ad ogni parola target. Nella seconda versione si è deciso di privilegiare l'efficacia visiva rispetto alla completezza dei dati, riportando solo i diecimila coseni più elevati estratti dai dati di partenza. Seppur sprovvista di parte dei dati, questa seconda versione offre un grafo meno affollato e più agilmente navigabile, in cui i *cluster* emergono con maggiore definitezza.

Grafo dello spazio con finestra ± 2 con tutti i coseni
http://colinglab.humnet.unipi.it/Demo/COLFIS/colfis_10_2_full/

Grafo dello spazio con finestra ± 2 con i 10.000 coseni più alti
http://colinglab.humnet.unipi.it/Demo/COLFIS/colfis_10_2_top10K/

Grafo dello spazio con finestra ± 5 con tutti i coseni
http://colinglab.humnet.unipi.it/Demo/COLFIS/colfis_10_5_full/

Grafo dello spazio con finestra ± 5 con i 10.000 coseni più alti
http://colinglab.humnet.unipi.it/Demo/COLFIS/colfis_10_5_top10K/

Grafo dello spazio con finestra ± 20 con tutti i coseni

http://colinglab.humnet.unipi.it/Demo/COLFIS/colfis_10_20_full/

Grafo dello spazio con finestra ± 20 con i 10.000 coseni più alti
http://colinglab.humnet.unipi.it/Demo/COLFIS/colfis_10_20_top10K/

Bibliografia

Bastian, Mathieu; Heymann, Sebastian & Jacomy, Mathieu 2009. Gephi: an open source software for exploring and manipulating networks. *ICWSM 8*. 361-362.

Bertinetto, Pier Marco; Burani, Cristina; Laudanna, Alessandro; Marconi, Lucia; Ratti, Daniela; Rolando, Claudia & Thornton, Anna M. 2005. *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*. http://linguistica.sns.it/CoLFIS/Home.htm

Blondel, Vincent D.; Guillaume, Jean-Loup; Lambiotte, Renaud & Lefebvre, Etienne 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 10. P10008.

Deerwester, Scott C.; Dumais, Susan T.; Landauer, Thomas K.; Furnas, George W. & Harshman, Richard A. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science (JASIS)* 41(6). 391–407.

Dinu, Georgiana; Pham, Nghia T. & Baroni, Marco 2013. DISSECT: DIStributional SEmantics Composition Toolkit. *Proceedings of the System Demonstrations of ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*. East Stroudsburg, PA: ACL. http://clic.cimec.unitn.it/composes/toolkit/index.html

Jacomy, Mathieu; Venturini, Tommaso; Heymann, Sebastian & Bastian, Mathieu 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9(6). e98679.

Niwa, Yoshiki & Nitta, Yoshihiko 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. *Proceedings of the 15th International Conference On Computational Linguistics*. 304-309.