# Exploiting corpus evidence for automatic sense induction

**REMA ROSSINI FAVRETTI FABIO TAMBURINI ANDREA ZANINELLO**

*Department of Linguistics and Oriental Studies*
*University of Bologna*
*via Zamboni 33, Bologna, Italy*

**Abstract (English)**

*In this paper, we discuss the application of a sense induction procedure to data from CORIS, a well-balanced reference corpus of Italian. The method considered discriminates between the different senses of a word by analysing the relationships between its collocates and suggesting collocate clusters, each of which corresponds to one sense of a word. The collocate clusters are represented as 3D-graphs in a semantic space. We show that for some examples the method can satisfactorily induce the senses of the chosen node; however, we also show that for some controversial instances human interpretation of the results is needed. We thus conclude that, although powerful, automated systems still require human knowledge both for the analysis and the interpretation of language phenomena, and that an integration of the two methodologies is desirable.*

Keywords: sense induction, corpora, CORIS, Italian, polysemy, 3D-graph

## 1 Introduction

The aim of this paper is to explore how statistical analysis of corpus evidence can contribute to sense disambiguation in non-annotated text. We focus on collocations as a source of surface evidence automatically extracted from corpora through positional and association-based procedures following probabilistic criteria.

Our basic assumption is in line with the Firthian tradition and the classical Harrisian distributional hypothesis, which assumes that 'similar' linguistic items (in particular, *semantically* similar items) will have similar distributions in naturally occurring texts.

More specifically, we hold that most characteristic collocates of a (potentially polysemic) word are a good indicator of its meaning(s) and therefore distributional 'closeness' between them can be seen as a hint of semantic similarity. Significant co-occurrence frequencies can be used to discriminate between the different senses of a word by grouping its collocates according to their distributional behaviour analysed through statistical association measures.

Our paper is organized as follows: firstly, we sketch the methodological background underlining our research and present a brief description of CORIS, the corpus of written Italian used in our study. Secondly, we describe the analysis tools and procedures exploited in our

research. Thirdly, we present some case studies focusing on polysemic words in Italian. Finally, we present and discuss our results and conclude with some remarks and perspective work.

## 2 Background and data

### *2.1 Background*

Our research is based on the traditional *contextual meaning theory* as exemplified in early works by Firth (1957) and further developed by his pupils (see, for example, Sinclair, 1991). In this framework, collocations are defined as *co-occurrences of words within a determined unit of information* (from next neighbours to whole texts) *where linguistic items appear together significantly,* in other words, with greater probability than one would expect if the relationship between them were completely random. In this framework we use the term "random" in its statistical sense, as we agree that in real language randomness is a nonsensical concept.

The meaning of a word is defined *contextually,* i.e. by the relations between the word itself and the other linguistic items that represent its context. This view dates back to the early Saussurian approach, where the function of an item is only defined *differentially* and the value ('*valeur*') of any sign only exists by virtue of the (differential) relationships that it holds with other items in the language.

This view has been variously applied in recent works and in particular the analysis of distributional similarity has been widely exploited in Word Space Models to measure the semantic or functional similarity between different words (e.g. Lenci, 2008). As Sahlgren (2008) points out, there exist two approaches to a distributional study of meaning: one consists in considering, as distributional evidence, the words that surround the target word, another is based on outlining distributional profiles based on the text regions where a word appears. These approaches, albeit often considered as equivalent, in fact rely on different types of distributional data exploiting, on the one hand, *paradigmatic relations,* and *syntagmatic relations,* on the other hand.

The methodology that we applied here is based on approaches of the first kind; however, differently form these approaches, we will exploit semantic similarities between the collocates of a node as evidence for discovering clusters of collocates, each of which should correspond to as many different senses of the node. As a computational method for our study,

we follow the work by Heyer (2001; 2002) for the construction of co-occurrence graphs which exploits the visual representation of a word's collocates to induce its different senses.

*2.2 The Corpus*

The corpus used in our study is the 120-million word corpus CORIS - Corpus di Italiano Scritto - an electronically-based reference corpus of contemporary written Italian. The corpus is the result of a research carried out at the University of Bologna since 1998 and it was designed, developed and implemented according to linguistically motivated criteria aimed at assuring that the corpus is a representative and well-balanced sample of standard Italian (Rossini Favretti, Tamburini & De Santis, 2002).

The corpus contains written texts from the 1980s to the present and is updated through a monitor corpus every three years. It contains texts from a wide range of varieties of Italian, chosen by virtue of their representativeness. In particular, the selection and proportion of texts was based on external, internal, as well as comparability criteria and led to the following structure (see Table 1) and proportions between different subcorpora (see Table 2).

| Subcorpus | Sections | Text types (examples) |
|---|---|---|
| PRESS | newspapers, periodic, supplement | national, local<br>specialist, non-specialist<br>connotated, non-connotated etc. |
| FICTION | novels, short stories | Italian, foreign<br>for adults, for children<br>crime, adventure, science fiction, women's literature etc. |
| ACADEMIC PROSE | books, reviews | human sciences, natural sciences, experimental sciences, popular history, philosophy, arts, literary criticism, economy, biology, etc. |
| LEGAL AND ADMINISTRATIVE PROSE | books, reviews, documents | legal, bureaucratic, administrative etc. |
| MISCELLANEA | books, reviews, documents | books on religion, travel, cookery, hobbies, etc. |
| EPHEMERA | letters, leaflets, instructions | private, public<br>printed form, electronic form etc. |

Table 1. Structure of the CORIS corpus.

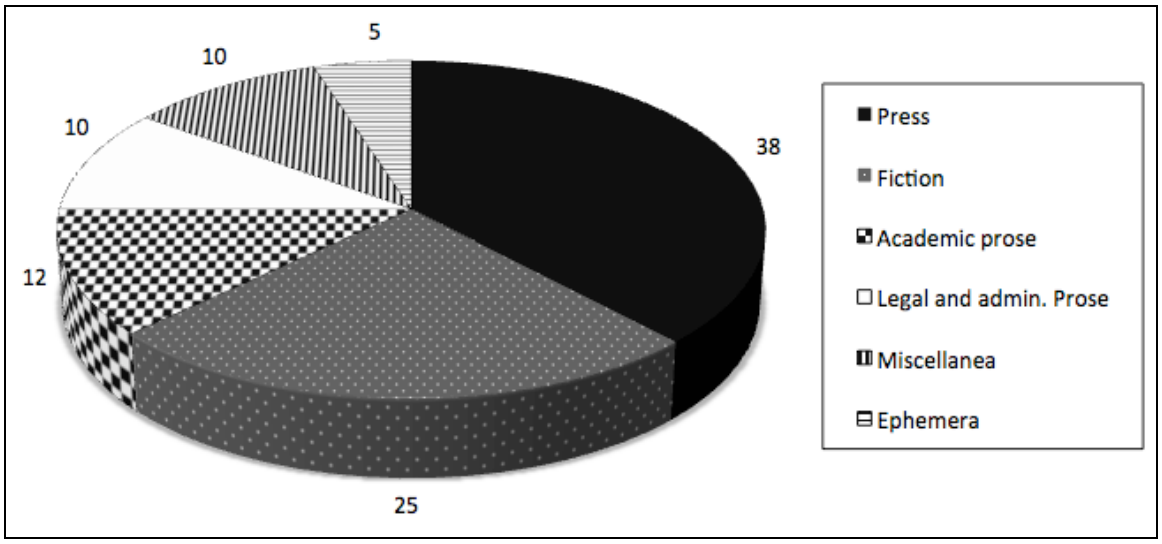


Table 2. Proportion of the subcorpora in CORIS (numbers represent percentages).

## 3 Procedure

The procedure exploited in this study and applied to CORIS was originally presented in Heyer *et al.* (2001) and was formerly used to modulate register variation (Rossini & Tamburini, 2009). The procedure creates collocation sets for the selected node through an automatic, iterated process of collocation analysis based on association measures and recursively applied to the collocates. The statistical measure used for collocation extraction is the log-likelihood ratio and the context window considered was the entire sentence.

The results are represented as co-occurrence graphs. This representation allows one to single out clusters of collocates connected at different strengths, and thus define different meaning areas providing a visualisation of polysemy through a representation of the collocates' distribution in a semantic space.

The visualization in the 3D-graph structure highlights areas where stronger relations between collocates are grouped together: homogeneity in the graph is shown when the collocates are interconnected (distributionally and therefore semantically), while separation in the representation is displayed when collocates are not connected.

In the next section we present some relevant examples of application of this procedure to some nodes, leading to different results, which will be eventually discussed.

## 4 Case studies and results

In this section we present the application of the procedure outlined in the previous section onto two case studies, and we present different results for the nodes considered.

### *4.1 Case study 1:* Risoluzione

We applied the above procedure to a highly polysemous word, '*risoluzione*', which can be variously translated into English by 'resolution', 'decision', 'cancellation' etc., according to its different senses. Below we present a snapshot of the concordances for the node, highlighting its possible uses (see Table 3):

| | |
|---|---|
| *STAMPAQuot*: | : " Dal punto di vista politico la **risoluzione del consiglio di sicurezza** dell ' |
| *STAMPAPeri*: | itiche e strategiche , la bozza di **risoluzione presentata al Consiglio** di sicure |
| *NARRATRoma*: | allo studio dei particolari e alla **risoluzione dei problemi logistici** , politici |
| *PRGAMMDocu*: | ti ) , salvo che il diritto alla risoluzione sia stato espressamente stipulato |
| *PRGAMMDocu*: | lità per l ' utente di ottenere la **risoluzione del contratto** entro un termine ra |
| *PRGAMMDocu*: | , con il sostegno dell ' UE , una **risoluzione d ' urgenza sul lavoro** forzati in |
| *PRGAMMDocu*: | ne rapporto , spettante in caso di **risoluzione del rapporto di lavoro** , è discip |
| PRGAMMDocu: | gli estremi della giusta causa di risoluzione , che davano diritto di ottenere |
| *PRGAMMVolu*: | o agli aeroporti dell ' area . 106 Risoluzione 753 ( 18 maggio 1992 ) Raccomanda |
| *MISCRivist*: | CCD , quindi è evidente che questa **risoluzione viene raggiunta** per interpolazion |
| *MISCRivist:* | asferimento di **immagini con alta risoluzione** e alta dinamica , quali radiograf |
| EPHEMIstru: | li le **dimensioni di pagina , la risoluzione** e così via , e non al suo contenu |

Table 3. A snapshot of the concordances for the node *risoluzione.*

The co-occurrence graph resulting from the application of the procedure to the node is displayed below (see Figure 1). The graph displays four distinct homogeneity areas that can be identified as clusters corresponding to as many senses of the word. In particular:

1. The first cluster (top-left) makes reference to *risoluzione* as the level of detail of an image (as in *alta risoluzione* = 'high resolution');
2. The second cluster (bottom-right) refers to the meaning of “solving” - e.g. of problems, cases, etc. (as in *risoluzione di problemi* = problem solving) ;
3. The third cluster (centre-right) makes reference to *risoluzione* as “decision” - e.g. of a formal body (as in *risoluzioni del Consiglio di Sicurezza* = Security Council Resolutions);
4. The fourth cluster (bottom-left) refers to *risoluzione* as “cancellation” – e.g. of a contract (as in *risoluzione di un contratto* = rescission of a contract).

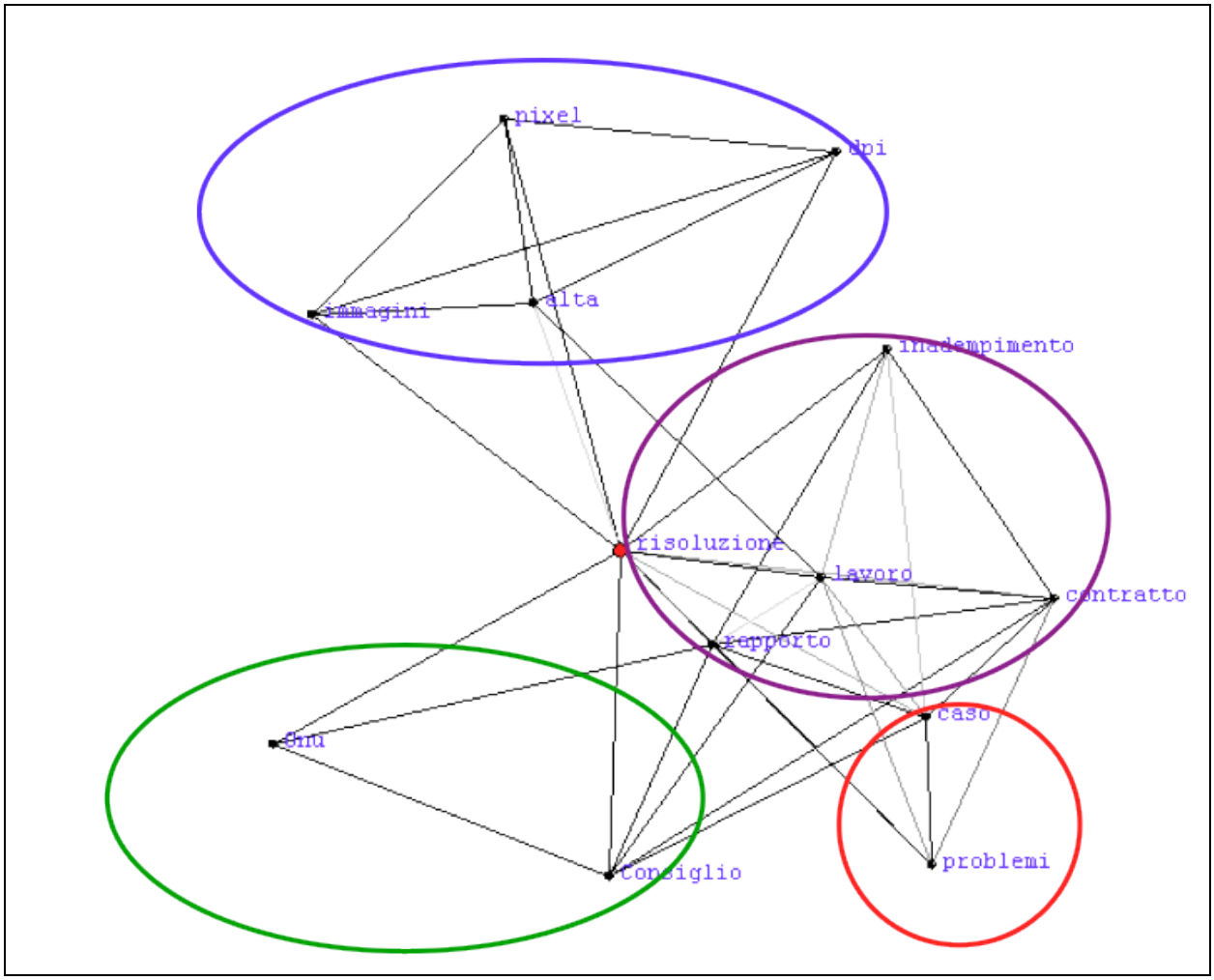


Figure 1. Co-occurrence graph for the node *risoluzione.*

As we can see in this example, the graph visualisation is able to separate between the different senses of the word and is fairly exhaustive, having identified all the possible uses that were highlighted by the analysis of the concordances as in Table 3.

*4.2 Case study 2:* Calcio

As a second example, we analysed the collocates of the node "*calcio*" which, according to the analysis of the concordances, are organised around three main axes, corresponding to as many senses of the word (cf. Table 4).

```
STAMPAQuot:      straordinaria come per i mondiali di calcio ' 90 " . I comunisti aggiungono di
STAMPAQuot: questa annata . [BEGINDOC] 09/06/1997 CALCIO FLASH PLAYOFF E PLAYOUT DI C . Play
STAMPAQuot: gli stadi , cedendoli alle società di calcio ; ma nello stesso tempo i club chie
STAMPAQuot: nuove mosse . Ma se si preferisce il calcio , con Fifa 2000 ci si può cimentare
STAMPAPeri: Qualche nome : Ferrarelle ( 368 mg di calcio per litro ) , Sangemini ( 327,8 ) ,
STAMPASupp: lizzano l ' automotivatore squadre di calcio come il Milan , campioni come Alber
NARRATTrRo:   vantarsi di aver atterrato con un calcio sulle zampe un giovane rinoceronte
NARRATTrRa:   superiori perché sostengono che il calcio consiste in ventidue imbecilli che
PRACCVolum:  transitorio innalzamento del tasso di calcio , la tubulina si libera anche dalla
MISCVolumi:      ( 400 mg % ) ; fosforo ( 800 mg % ) ; calcio ( 700 mg % ) . Non è presente lo
io
```

Table 4. A snapshot of the concordances for the node *calcio.*

The node presents three main uses: *football* (as in *mondiali di calcio* = world champions), *kick* (as in *calcio sulle zampe* = kick on the paws), *calcium* (as in *tasso di calico* = level of calcium).

However, when we applied the co-occurrence graph visualisation procedure to the node (see Figure 2), different results were returned. In particular, the meaning clusters induced by the procedure are organised around three main clusters:

1. the first cluster (top-left) corresponds to the '*football'* meaning, as in *squadra di calcio* = 'football team'.
2. the second cluster (centre-right and across) makes reference to *calcio* as the chemical element 'calcium', as in *carbonato di calcio* = 'calcium carbonate'
3. the third cluster (bottom-right) corresponds to a very specific use of the node as it refers to a 'cranberry collocation' making reference to the title of a popular TV show about football *Quelli che il calcio...* (lit. 'those who football...').

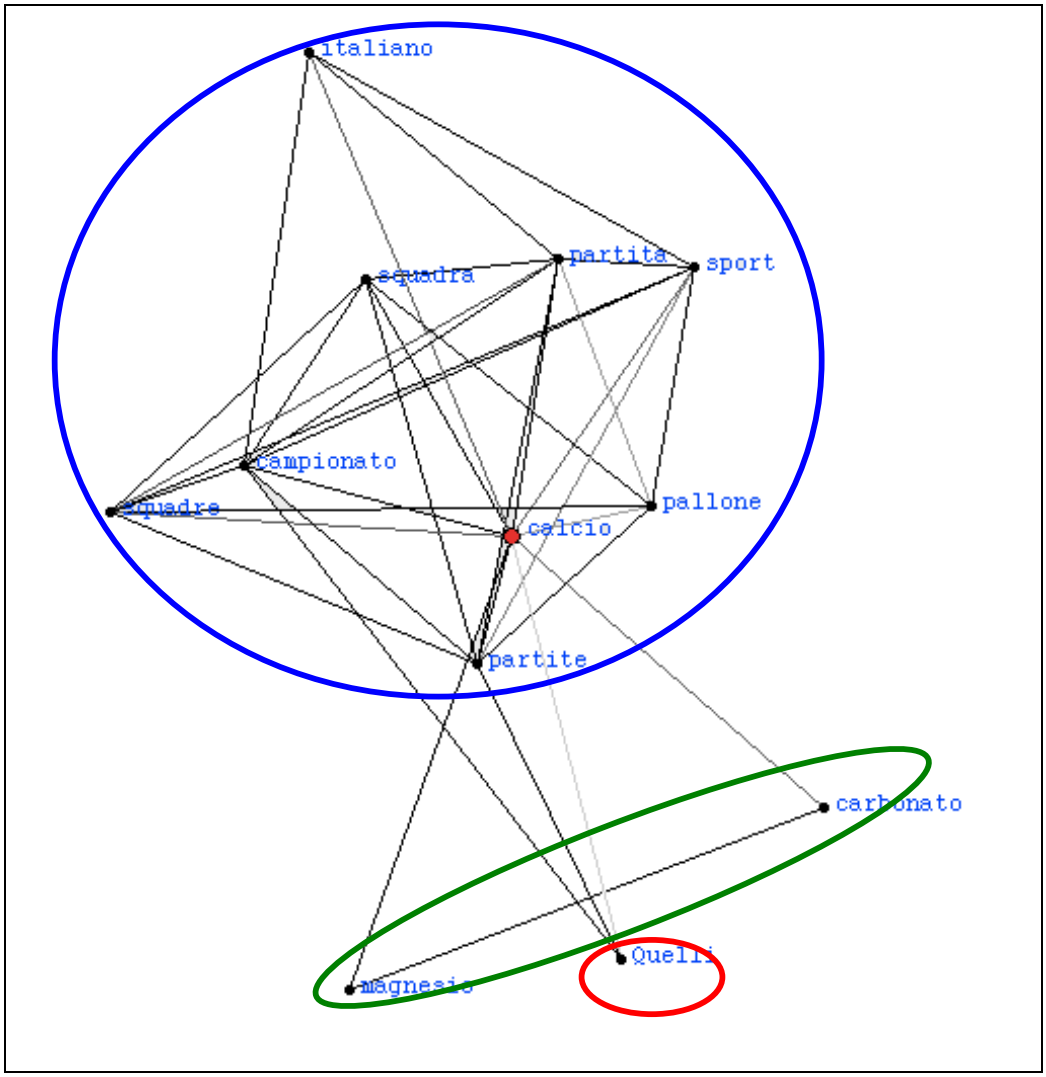


Figure 2. Co-occurrence graph for the node *calcio*.

## 6 Discussion, conclusions and future work

As we outlined above, the 3D-graph in the first case study returned all four senses of the word previously observed in the analysis of concordances and can thus be said to provide an accurate representation of the polysemy of the node. The second case study, on the other hand, showed a mismatch between the analysis of concordances and the 3D-graph. In the first case, three different senses had been found, including the 'basic' sense ('kick') of *calcio*. However, the 'kick' sense is surprisingly absent in the graph, a result that depends on a too low representation of this use within the computational methodology applied. In return, another 'meaning' nucleus is highlighted in the graph and corresponds to a very specific use of the word *calcio* (the title of a TV show about football) that is represented as a separate cluster form the football one

(although weakly connected to it) because it appears in a 'cranberry collocation', i.e. a very idiosyncratic structure which does not conform to the normal rules of syntactic composition in Italian.

It could be said that it is controversial whether this use actually represents another sense of the word *calcio,* (since, taken alone, it has got the same reference to football as the other cluster), however, if we consider the whole context where the term appears, we believe that it is correct to separate it from the 'football' cluster, since the word is used in a very idiosyncratic way and represents one sense on its own (which we might label the 'TV' sense). Therefore, we showed that the methodology can in many cases correctly identify the main senses of a node through the observation of the resulting graphs.

In addition, we conclude our work with some final remarks and observations on the results obtained.

i) Firstly, as exemplified by the second case study (in particular by the mismatch between the senses found in the concordances and those induced in the graph), we would like to point out that, since our methodology is based on the observation of naturally occurring data, results of the clustering procedure depend on the distribution of the data because computational methods are sensible to the frequency of the phenomena under investigation. The absence of a phenomenon does not necessarily ensure that the latter does not exist in the language, but we believe that negative evidence is a very important clue about the significance of language facts, based on and validated by statistical significant measures applied on naturally occurring texts (and thus not induced through deductive reasoning).

ii) Secondly, the case study on the node *calcio* showed that the system creates different clusters, and therefore indicates two separate meanings for the 'TV' and the 'football' sense. From an extra-linguistic point of view, of course, the senses in both examples are semantically connected, as they refer to the same *referential* meaning of the word. However, we believe that it is desirable to keep them separate, as explained before, because we are mainly interested in the different *uses* of one node and thus it is very useful to be able to distinguish between a very common, unmarked instance of the word (as in the football sense) as opposed to its idiosyncratic use.

iii) Thirdly, it is worth pointing out that the sense induction procedure does not separate between different kinds of meaning relations and identifies uniquely general macro-relations of *semantic similarity/dissimilarity.* In particular, we showed that, on the one hand, the methodology treats in a unified manner semantic compatibility relations, such

as polysemy (as in the case of *risoluzione*), homonymy (as in the case of *calcio*), synonymy etc., as it does, on the other hand, for semantic incompatibility relations such as antonymy, complementarity etc. As a matter of example, the kind of semantic relation between the sense clusters of *calcio* would probably be interpreted by linguists as an instance of homonymy rather that polysemy (especially for the 'football' and the 'chemistry' senses), as the word happens to have the same form in the two senses but its meanings are not connected in any way, as opposed to the meanings of *resolution.*

As a suggestion for future work, we believe that the system could be positively implemented by the integration of linguistic as well as extra-linguistic information. This could be done exploiting human knowledge as well as making use of language resources such as electronic dictionaries, ontologies etc. This would allow one to *label the senses* as well as to overcome some of the problems discussed above (such as a separation between true polysemy and usage idiosyncrasies). Moreover, we believe that this procedure may be expanded to the historical dimension in order to study the evolution of a word's senses across time. We plan to do so by applying the procedure to the diachronic corpus DiaCORIS (cf. Onelli *et al.*, 2006), a representative and balanced collection of Italian written language ranging from the National Unification of Italy - 1861 - to the end of the Second World War - 1945.

## 7 References


Firth, J. R. (1957). Papers in Linguistics 1934-1951. OUP. Oxford.

Heyer, G., Lauter, M., Quasthoff, U., Wittig, T., Wolff, C. (2001). Learning relations using collocations, Proceedings of the IJCAI Workshop on Ontology Learning, Seattle, USA.

Heyer, G., Quasthoff, U., & Wolff, C. (2002). Automatic analysis of large text corpora - A contribution to structuring WEB communities. Lecture Notes in Computer Science, 2346, 15-26.

Lenci, A. (ed.) (2008). From context to meaning: distributional models of the lexicon in linguistics and cognitive science, Italian Journal of Linguistics, 20/1.

Onelli, C., Proietti, D., Seidenari, C., Tamburini, F. (2006). The DiaCORIS Project: a diachronic corpus of written Italian. Proceedings of the 5th International Conference on Language Resources and Evaluation - LREC 2006. Genoa: 1212-1215.

Rossini Favretti, R., Tamburini, F., (2009). Exploring register variation through corpus evidence, in Abstracts of DGfS 2009 Workshop on Corpus, Colligation, Register Variation, Osnabruck, p. 155.

Rossini Favretti, R., Tamburini, F., De Sanctis, C., (2001). A corpus of written Italian: a defined and dynamic model. In Proceedings of Corpus Linguistics 2001 Conference, Lancaster,UK.

Sahlgren, M. (2008). The Distributional Hypothesis. From context to meaning: Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), Rivista di Linguistica, volume 20, numero 1, 2008.

Sinclair. J., (1991). Corpus, Concordance, Collocation. Oxford University Press.