# Syntax and Semantics vs. Statistics for Italian Multiword Expressions: Empirical Prototypes and Extraction Strategies

Luigi Squillante
Sapienza - Università di Roma
Email: luigi.squillante@uniroma1.it

## Abstract

In this work we present an empirical analysis performed on Italian nominal multiword expressions (MWEs) of the form [noun + adjective] that aims at studying quantitatively their syntactic and semantic features in order to improve their automatic identification and collection. Three indices are proposed, which are able to measure syntactic and semantic frozeness of the expressions on empirical basis in a corpus of about 1.8 million words, composed of Italian texts concerning the domain of physics. The combination of the three indices can be used to create a global measure, that we call Prototypicality Index (PI), which appears to be useful in the automatic extraction of terminological MWEs. The performance of PI at extracting true positives out of a candidate list is compared to those of the well-known statistical association measures Log-likelihood and Pointwise Mutual Information. Our results show how the performance of PI can be comparable to those of association measures, although it does not involve statistical calculations. Thus, PI can be seen as a new option for lexicographers and terminologists to integrate the already available statistical methods when identifying MWEs from texts.

**Keywords:** multiword expressions; terminology; prototype; extraction; empirical tests

## 1 Introduction

Nowadays multiword expressions (MWEs) represent one of the most studied phenomena in phraseological and lexicographic studies. They include a great variety of entities lying on a *continuum* between lexicon and syntax, whose typical features include morpho-syntactic fixedness, semantic restrictions, semantic unpredictability, constructions which differ from standard syntax, conventionality, institutionalization, etc. Their interpretation generally crosses the boundaries between words (Sag et al. 2002), and one of the best definitions to refer to such entities is proposed by Calzolari et al. (2002:1934), according to whom a MWE is "a sequence of words that acts as a single unit at some level of linguistic analysis".

Despite their apparent anomalous behavior, MWEs are a very important and frequent phenomenon in every language: in his famous *idiom principle* Sinclair (1991) states that idiomatic and morpho-syn-

tactically restricted combinations are as normal and natural in discourse as free combinations, while Jackendoff (1997) attests that the number of MWEs stored in the lexicon of any speaker is equal to that of simple words.

Throughout the twentieth century, linguists have developed a great amount of studies which examined the aspects of MWEs on a theoretical perspective, often leading to competing analyses, controversy on interpretations or overlapping terminology. In recent years, however, computational and corpus-based studies have become one of the dominant lines of research in this field, since quantitative features, such as the fact that MWE components tend to cooccur in text with higher frequencies, have proved to be very effective in the automatic treatment of MWEs, leading to the development and improvement of several association measures (AMs) in order to identify, study and automatically extract MWEs from texts (just to mention some works: Evert 2004; Evert 2008; Kilgarriff 2006; Ramisch et al. 2010; Seretan 2011).

When analyzing a corpus, by means of computational tools, linguists are usually able to create a list of candidate expressions of MWEs where each candidate has an association score assigned by AMs. In general, the primary goal is to identify the largest possible number of true positives (candidates that represent real MWEs) within a certain threshold of significance based on the score assigned to each candidate, e.g. to provide raw material for lexicography. In this process, AMs generally consider statistical quantities, such as the number of cooccurrences of the components, the number of occurrences of the single components, the size of the corpus, etc., often with no reference to any explicit linguistic behavior. Nevertheless, considering syntactic or semantic features of MWEs from a computational and corpus-linguistic point of view is useful to improve the performances of automatic extraction tools (as shown, for other languages, in Bannard 2007; Weller & Fritzinger 2010; Cap et al., 2013), as well as to develop a better understanding of the typical features of MWEs on empirical bases (cf. Squillante 2014), which are both aspects of preeminent interest for lexicographers dealing with multiword phenomena.

## 2 Motivations

Our work presents an empirical study conducted on the Italian language which, unlike other major languages like English or German, still lacks well-founded computational studies in lexicography dealing with complex expressions like MWEs. Although "GRADIT - Grande Dizionario Italiano dell'Uso" (De Mauro, 1999-2007), known as the most comprehensive lexicographic resource for Italian, has a highly corpus-oriented perspective and explicitly focuses on the quantitative presence of MWEs in Italian, no explicit computational methods were involved in identifying the expressions. Similarly, the most recently published Italian collocation dictionaries (Urzì 2009; Lo Cascio 2012; Tiberii 2012) still rely mostly on intuition and only partly replicate data collection strategies, without considering a defined and explicit methodology based on corpora. Thus, there is a need to investigate computational techniques for lexicographic analyses of Italian MWEs, especially because Italian morpho-syntax differs from those of the above-mentioned Germanic languages.

The nature of our study is twofold: on the one hand we focus on empirical evidences in order to study the prototypical concept of MWE; on the other hand we compare the traditional statistical measures with syntactic or semantic tests for the identification and the extraction of MWEs from texts.
Finally, our work is focused on terminology. In fact, especially in technical domains, MWEs appear in high number even in small corpora, since specialized languages are a powerful source of multiword terminology and we see it as a matter of importance that they are identified and collected so that they can be included in the respective dictionaries and multiword terminology collections.

# 3 Methodology

## 3.1 Corpus and Prototype of MWE

As a first approach, in our study we opted to focus on the field of physics. The choice of physics is interesting since its lexicon, unlike other scientific domains such as that of medicine, is still primarily composed of highly polysemous every-day words which are put together in MWEs to form technical expressions, pursuing the established tradition started with Galileo Galilei in the seventeenth century, as recalled by Migliorini (1994:398).
In order to have an empirical base to perform our analysis, we built a corpus of about 1.8 million words collecting Italian texts concerning physics, including educational books (6,2% of the total), Wikipedia pages (34,5%), academic textbooks (20,7%), theses and dissertations (38,6%).
Our corpus was POS-tagged with TreeTagger (Schmid, 1994) and enhanced by means of a semi-automatic and manual post-tagging process in order to improve the tagging quality, e.g. to correct macroscopic systematic errors and include unrecognized technical lemmas in the dictionary. The final accuracy of the tagged corpus is evaluated at around 96% by manually checking 300 random sentences of the corpus.
We chose to analyze only nominal MWEs of the form [noun + adjective] in a first approach, representing the unmarked Italian nominal phrase. In physics, in fact, the use of nominal phrases is dominant and nominalization is often attested to be a standard feature of special languages. This is also supported by the fact that the majority of MWEs labeled by GRADIT as part of the special language of physics are nominal (2668), while only 9 belong to any other grammatical category.
Although MWEs can exhibit a great variability of behaviors, as it has been mentioned in the introduction, we chose to focus on features which could be investigated and tested on corpora, and we started with the initial hypothesis that the prototype of a MWE is an expression:

- that does not allow for interruptions or insertions of other words between its components;
- whose word order is not modifiable;
- whose components cannot be substituted by their synonyms.

The expression *relatività generale* 'general relativity' is a clear example of a terminological MWE which satisfies these three conditions, since it cannot be interrupted (cf. **relatività più generale* 'more general relativity'), it does not allow a modification in the order of its components, although this is possible for Italian nominal phrases (cf. **generale relatività*) and it cannot be modified by substituting one of its components with a synonym (cf. **relatività universale* 'universal relativity' or **relatività totale* 'total relativity').

However, although these features involving fixedness are typically associated to nominal MWEs in Italian, they do not always appear together in all expressions. For example, interruptibility is allowed for *punto debole* 'weak point', which admits *punto più debole* 'weaker point'; *infrarosso lontano* 'far infrared' is attested together with *lontano infrarosso*; while *gas ideale* 'ideal gas' can be substituted by *gas perfetto* 'perfect gas'. Because of this, the concept of prototype is thought of just as a model which could help to order the expressions on a continuous scale from a maximum grade of fixedness on several levels (adhesion to the prototype) to more flexible expressions.

The reason for considering the hypothesis of such a prototype comes from studies like those of Masini (2009) and Squillante (2014), which show how the nucleus of the prototype seems to include those expressions that are generally referred to as *polirematiche* in the Italian lexicographic tradition and exhibit syntagmatic *and* paradigmatic frozeness, needing the cooccurrence of their components in order to acquire their specific meaning (e.g. *luna di miele* 'honeymoon'; *essere al verde* 'to have no money', lit. 'to be at green'). Terminological expressions are generally part of this group.

When fixedness becomes less strict and modification is allowed, the *continuum* of MWEs moves towards those expressions that we can call *lexical collocations,* which show only preference for the cooccurrence of their components (e.g. *capelli castani* 'chestnut brown hair' or *compilare un modulo* 'to fill a form'), being «not fixed but recognizable phraseological units» (Tiberii, 2012).

## 3.2 Three Indices for the Measure of Empirical Frozeness

Following Squillante (2014), we implemented a computational tool that performs empirical tests concerning the above-mentioned features of modifiability for each candidate expression. Each of the features is quantified by an index whose value is computed on the basis of the comparison between the occurrences of the modified expression and those of the regular basic unmarked form in the corpus, i.e. the lemmatized form, regardless of inflection (which our analysis proved to be not a relevant feature in discriminating MWEs from standard expressions). All the queries are made on surface forms or POS categories, depending on the test, and do not involve syntactic structures as they would arise from parsing.

Given an expression, the index of interruptibility ($I_i$) counts the number of the occurrences of the sequence in its basic form [noun + adjective], say $n_i$, and the occurrences of the same sequence with one word occurring between the two components ($n_{bf}$), calculating the following ratio:

$$I_i = \frac{n_i}{n_{bf} + n_i}$$

In this way, a high number of interrupted expressions with respect to those which are not interrupted let the index acquire a high value. The sum in the denominator let the index be limited between 0 and 1.

In an analogous way, the index concerning the reverse order ($I_o$) compares the number of occurrences of the inverted sequence [adjective + noun] ($n_o$) with those of the basic form $n_{bf}$, according to the formula:

$$I_o = \frac{n_o}{n_{bf} + n_o}$$

Finally, the index concerning the feature of substitutability compares the number of occurrences of the basic form with the occurrences of all the sequences in which one of the two components is replaced by one of its synonyms (if present). If the number of occurrences of the *i*-th synonym of the first and the second component are called respectively $n_{s1,i}$ and $n_{s2,i}$, the total number of substituted sequences for the expression is:

$$n_s = \sum_i n_{s1,i} + \sum_i n_{s2,i}$$

and the index $I_s$ is given by the formula:

$$I_s = \frac{n_s}{n_{bf} + n_s}$$

The calculation of $I_s$ is subjected to the availability of an external synonym list. In our study, as a first approach, we chose the GNU-OpenOffice Thesaurus for the Italian language[1] for practical reasons, since it was immediately available, easily manageable and proved to be good enough for our purpose. However, one can integrate the tool with other more specific resources in the future, in order to improve the quality of the results.

The values of the three indices can be merged into a single function that we call Prototypicality Index (PI), representing the adherence of the expression to the hypothesized prototype. We consider the following formula:

$$PI = \frac{n_{bf}}{n_{bf}^{max}} \cdot \frac{1}{1 + I_i + I_o + I_s}$$

whose value increases when the values of the three indices decrease (thus, a high PI value means high fixedness), and in which the three features are weighted in the same way by the operation of

1 http://linguistico.sourceforge.net/pages/thesaurus_italiano.html.

sum. In this way an expression with a very high value for just one of the indices can have a resulting PI value similar to that of an expression with average values distributed on all the three indices. Therefore, this structure is useful to take into account the flexibility of the nature of MWEs. Finally, the PI considers a correction factor, given by the normalized ratio between the frequency of the expression and that of the most frequent candidate expression $n_{bf}^{max}$. This correction factor, which is bounded between 0 and 1, is needed to take into account the fact that low occurrences for the basic form in the corpus reduce the reliability of the empirical tests, since the presence or the absence of modifications cannot be tested on a large set of expressions.

## 4 Analysis and Results

As a first analysis, we considered the whole set of nominal MWEs labeled as part of the lexicon of physics in GRADIT. The considered set consists of a total amount of 1.551 MWEs, 595 of which are attested to occur in our corpus.

The resulting values of the three indices (considered separately) indicate that 73% of the attested expressions are never interrupted, 93% never appear in reverse order and 64% do not attest any substitution of their components. The empirical evidence, hence, suggests that the syntactic fixedness, more than paradigmatic frozeness, seems to be relevant in outlining the prototype of nominal MWEs in physic Italian terminology. It must be underlined that the absence of modifications in the corpus does not mean that the expression does not allow them in general, nevertheless the empirical evidence can be considered a good approximation in our computational perspective.[2]

Since the list of physics-related MWEs extracted from GRADIT is supposed to include only terminological expressions with a completely definite phraseological status, we can consider them as a gold standard for further analyses.

In fact, the PI can be used as a new measure for the automatic extraction of MWEs from texts.

On the basis of the PI values, it is possible to assign each expression of a list of candidates a score and order the expressions according to it.

In order to have empirical evidence of the performance of the PI, we considered an input list from our corpus, composed of all the bigrams of the form [noun + adjective] which were extracted automatically, forming a set of about 22.700 expressions.

If we order the list according to PI we obtain results which appear analogous to those generally produced by statistical AMs, since PI is able to filter out most non-MWE candidates, which get very low scores and are pushed to the end of the list. At the same time, expressions appearing with very high

2 It must be said that some noise in this kind of approach is unavoidable, since it can happen that few expression can exhibit modifications, but the modified expressions are not MWEs anymore, as in the case of *forza debole* 'weak force' meaning one of the four fundamental interactions, which is attested together with *debole forza*, meaning just that the intensity of a generic force is weak.

scores at the top of the list have high probability of representing true MWEs. Table 1 and Table 2 show, respectively, the top and the end of the list sorted according to PI.

| Rank | MWE candidate | English translation | PI value |
|---|---|---|---|
| 1 | Campo magnetico | Magnetic field | 0.9565 |
| 2 | Campo elettrico | Electric field | 0.6133 |
| 3 | Momento angolare | Angular momentum | 0.5717 |
| 4 | Meccanica quantistica | Quantum mechanics | 0.5205 |
| 5 | Calorimetro elettromagnetico | Electromagnetic calorimeter | 0.4748 |
| 6 | Modello standard | Standard model | 0.4259 |
| 7 | Valore medio | Mean value | 0.4206 |
| 8 | Massa invariante | Rest mass | 0.3683 |
| 9 | Energia cinetica | Kinetic energy | 0.3630 |
| 10 | Campo gravitazionale | Gravitational field | 0.3423 |
| 11 | Campo elettromagnetico | Electromagnetic field | 0.3314 |
| 12 | Relatività generale | General relativity | 0.3155 |
| 13 | Buco nero | Black hole | 0.2997 |
| 14 | Meccanica classica | Classic mechanics | 0.2591 |
| 15 | Carica elettrica | Electric charge | 0.2395 |

**Table 1: Top-15 of the candidate list made of all the [noun + adjective] bigrams attested in our corpus sorted according to the Prototyicality Index values.**

| Rank | MWE candidate | English translation | PI value |
|---|---|---|---|
| 22686 | Entità indipendente | Independent entity | $7.0651 \cdot 10^{-6}$ |
| 22687 | Caso tale | Case such | $6.8689 \cdot 10^{-6}$ |
| 22688 | Fotone due | Photon two | $4.8725 \cdot 10^{-6}$ |
| 22689 | Condizione fondamentale | Fundamental condition | $4.4757 \cdot 10^{-6}$ |
| 22690 | Sistema vivente | Living system | $4.3961 \cdot 10^{-6}$ |
| 22691 | Parte maggiore | Bigger part | $4.3766 \cdot 10^{-6}$ |
| 22692 | Ambito magnetico | Magnetic range | $3.9057 \cdot 10^{-6}$ |
| 22693 | Condizione finale | Final condition | $2.6518 \cdot 10^{-6}$ |
| 22694 | Dimensione media | Average dimension | $2.5493 \cdot 10^{-6}$ |
| 22695 | Forma standard | Standard shape | $2.2079 \cdot 10^{-6}$ |

**Table 2: End of the candidate list made of all the [noun + adjective] bigrams attested in our corpus sorted according to the Prototyicality Index values.**

In order to evaluate the performance of the PI, we chose to compare its results on our candidate list with two well-known statistical association measures, Log-likelihood (Dunning 1993), hereafter LL, and Pointwise Mutual Information (Church & Hanks 1990), hereafter PMI, which are widely used in corpus-linguistics to identify MWEs. Both AMs can be seen as representatives of two general groups of measures which quantify two different aspects of word combinations: LL measures how unlikely it is that the two words are independent while PMI investigates "how much the observed cooccurrence frequency exceeds expected frequency" as stated in Evert (2008: 1128). In this way, their use can provide two different perspectives on the statistical extraction of MWEs.

By means of the computational tool "mwetoolkit" (Ramisch et al. 2010), each bigram of our candidate list is assigned a LL and a PMI value, so that all the expressions can be ordered according to their statistical scores. The performance of PI and the two measures is evaluated on the basis of the rate of the retrieval of true positives in the lists: we compare how many true MWEs are detected while going through the lists, according to the ordering established by the scores of statistical measures and PI.

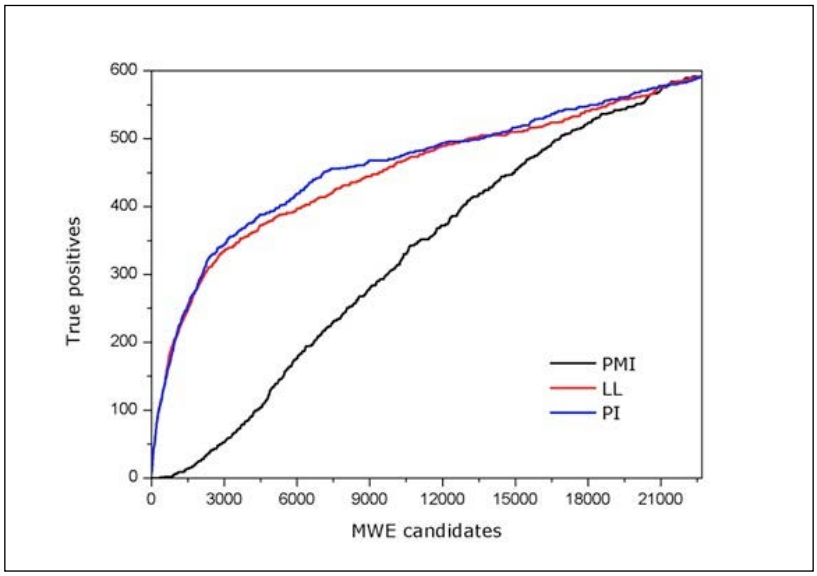


**Figure 1: Comparison between the extraction rates of true positives of Pointwise Mutual Information (black), Log-likelihood (red) and Prototipicality Index (blue).**

Figure 1 shows the curves representing the extraction rates for the three measures. As one can see, LL and PI had quite similar performances at identifying true positives, thus indicating that syntactic and semantic tests on empirical data can provide good results when used in extraction tasks. The poorer result of PMI can be justified by the fact that no frequency threshold was applied at the beginning and this AM is known for overestimating low-frequency expressions which are often false positives (Evert 2008).

We noted that for the first 1.800 candidates (corresponding to a 40% of true MWEs retrieved) LL obtained slightly better results with respect to PI, but for the remaining 20.900 candidates, the PI was almost always the better choice. This seems to indicate that on large scales the PI can be more useful

to lexicographers, who are generally interested in retrieving the largest possible number of MWEs and not only those in the first positions of the lists generated by statistics.

As an additional analysis, we considered also a frequency threshold on the input candidate lists, in order to minimize the problems related to low-frequency expressions, which especially affect PMI. Thus, we filtered our list, keeping only expressions with a frequency $f \geq 30$ (for a total of 301 expressions) and performed the same procedure as above.

Since a frequency of at least 30 occurrences can provide a good empirical basis for the tests, we decided to consider in this case also a "pure" variant of the PI, which is not corrected by the frequency information and is given by the following formula:

$$PI_p = \frac{1}{1 + I_i + I_o + I_s}$$

Figure 2 shows the extraction rates for the four measures. Once again LL and PI are the best choices and their performances are almost equal. This time PMI appears to be more useful, as one could expect, although its extraction rate is less effective than LL or PI. Finally $PI_p$ shows an extraction rate which is clearly better than that of PMI for the first 80 candidates, while for the remaining candidates its performance can be comparable to PMI. At the end of the process the number of true positives retrieved was 101 for LL and PMI, 99 for PI and 98 for $PI_p$.

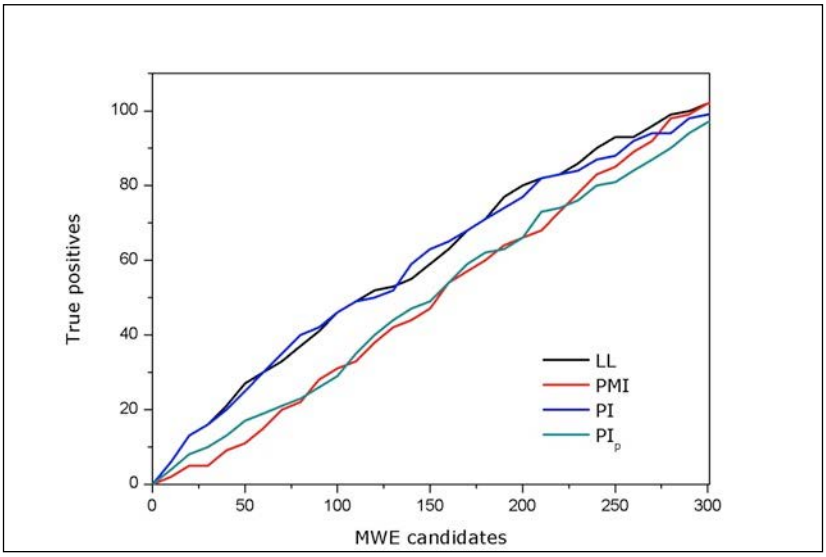


**Figure 2: Comparison between the extraction rates of true positives for Log-likelihood (black), Pointwise Mutual Information (red), Prototypicality Index (Blue) and the pure version of PI (green) on a candidate list with a frequency threshold of 30 occurrences.**

# 5 Conclusions and future work

In this work we have shown how syntactic and semantic features can play an important role in studying MWEs from a computational perspective. In the case of Italian nominal MWEs of the form [noun + adjective] belonging to the special language of physics, empirical tests performed on a corpus of 1.8 million words suggested that syntactic and semantic frozeness are effective features when outlining the prototype of this kind of expressions, although semantic substitutions are more tolerated than syntactic modifications.

The three indices that quantify empirical frozeness considered in this work proved effectiveness in extraction tasks of MWEs when merged in a function that we called Prototypicality Index, which produced results that can be considered comparable to those of statistical association measures.

Such results show how our methodology can be seen as a new option for lexicographers and terminologists, to integrate the already available statistical methods when identifying MWEs from texts, thus providing one more perspective in the extraction task which can be useful to have a more complete and general overview of the phenomena as well as to create complete terminological dictionaries or resources.

Moreover, as mentioned above, the PI works better on larger scales and appears to be useful to lexicographers who are interested in retrieving more efficiently MWEs when considering a high coverage, thus dealing with expressions spanning throughout the candidate list and not focusing only on its top. This feature of PI can be explained by the fact that syntax and semantics, unlike statistical features, show more strength and reliability when dealing with less frequent
expressions.

Nevertheless, the fact that a simplified version of the PI, which does not involve frequency information, produced worse results (but still similar to AMs) on a limited candidate list composed by expressions with more than 30 occurrences, shows that frequency inevitably plays a role in helping the retrieval of true positives.

However, the empirical results presented in this work must be tested on larger and more general corpora, as well as on corpora of other specialized domains, in order to evaluate the usefulness of the PI for general and specialized lexicography.

Future works must include the development of tools which can deal with other pattern of nominal MWEs as well as other grammatical categories, such as verbal or adverbial MWEs, where the above-mentioned features of modifiability are to be used in different ways when defining the prototype.

Lastly, the tools developed are to be made available, e.g. as a part of corpus research workbenches, for lexicographers and terminologists.

# 6 References

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions: Analysis, Acquisition and Treatment (ACL 2003)*. Sapporo, Japan.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C. & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002).* Las Palmas, Canary Islands.

Cap, F., Weller, M. & Heid, U. (2013). Using a Rich Feature Set for the Identification of German MWEs. In *Proceedings of Machine Translation Summit XIV.* Nice, France.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, 16(1), pp. 22-29.

De Mauro, T. (1999-2007). GRADIT, Grande Dizionario Italiano dell'Uso. Torino: UTET.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. In *Computational Linguistics*, 19(1), pp. 61-74.

Evert, S. (2004). The Statistics of Word Cooccurrencies: Word Pairs and Collocations. PhD Thesis. University of Stuttgart.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling, M. Kytö (eds.) Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter, pp. 1212-1248.

Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge: MIT Press.

Kilgarriff, A. (2006). Collocationality (and how to measure it). In *Proceedings of the 12th EURALEX International Congress.* Torino: Dell'Orso, pp. 997-1004.

Lo Cascio, V. (2012). *Dizionario combinatorio compatto italiano*. Amsterdam: John Benjamins Publishing Company.

Masini, F. (2009). Combinazioni di parole e parole sintagmatiche. In M. Catricalà, P. Pietrandrea, E. Lombardi Vallauri, P. Di Giovine, D. Cerbasi, L. Mereu, L. Gaeta, G. Fiorentino, P. D'Achille, M. Grossmann, E. Jezek, F. Masini, A. Pompei, E. Bonvino, F. Orletti, M. Frascarelli (eds.) Spazi linguistici. Studi in onore di Raffaele Simone. Roma: Bulzoni, pp. 191-209.

Migliorini, B. (1994). *Storia della lingua italiana.* Milano: Bompiani [I. ed 1960].

Ramisch, C., Villavicencio, A. & Boitet, C. (2010). Mwetoolkit: a Framework for Multiword Expression Identification. In *Proceedings of the 7th International Conference on Language Resources and Evluation (LREC 2010).* Valletta, Malta.

Sag, I., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd CICLing (CICLing-2002), vol. 2276/2010 of LNCS.* Mexico City, Mexico.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing.* Manchester, UK.

Seretan, V. (2011). *Syntax-based Collocation Extraction*. Berlin: Springer.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Squillante, L. (2014). Towards an Empirical Subcategorization of Multiword Expressions. To appear in *Proceedings of the EACL 10th Workshop on Multiword Expressions.* Gothenburg, Sweden.

Tiberii, P. (2012). Dizionario delle Collocazioni. Le combinazioni delle parole in italiano. Bologna: Zanichelli.

Urzì, F. (2009). *Dizionario delle Combinazioni Lessicali.* Luxembourg: Convivium.

Weller, M. & Fritzinger, F. (2010). A hybrid approach for the identification of multiword expressions. In *Proceedings of the SLCT 2010 Workshop on Compounds and Multiword Expressions.* Linköping, Sweden.