The *Dictionary of Italian Collocations*: Design and Integration in an Online Learning Environment

Stefania Spina
University for Foreigners Perugia
Piazza Fortebraccio 4, 06122 Perugia, Italy
stefania.spina@unistrapg.it

Abstract

In this paper, I introduce the DICI, an electronic dictionary of Italian collocations designed to support the acquisition of the collocational competence in learners of Italian as a second or foreign language.
I briefly describe the composition of the reference Italian corpus from which the collocations are extracted, and the methodology of extraction and filtering of candidate collocations. It is an experimental methodology, based on POS filtering, frequency and statistical measures, and tested on a 12-million-word sample from the reference corpus. Furthermore, I explain the main criteria for the composition of the dictionary, in addition to its integration with a Virtual Learning Environment (VLE), aimed at supporting learning activities on collocations. I briefly describe some of the main features of this integration with the VLE, such as the automatic recognition of collocations in written Italian texts, the possibility for students to obtain further linguistic information on selected collocations, and the automatic generation of tests for collocational competence assessment of language learners.
While the main goal of the DICI is pedagogical, it is also intended to contribute to research in the field of collocations.

1 Introduction

Multi-word units are widely recognized as playing an important role in various fields of language research; after the early interest in the domain of language teaching (Palmer, 1933), they occupy, ever more frequently, a central position in the field of lexicography (Benson, Benson & Ilson, 1986; Benson, 1990; Cowie, 1981, Granger & Meunier, 2008), natural language processing (Smadja, 1993; Calzolari et al., 2002; Sag et al., 2002), corpus linguistics (Sinclair, 1991) and language acquisition (Nesselhauf, 2005).
Due to the persistent lack of agreement on its definition, the Firthian term *collocation* has been applied to a wide variety of phenomena, ranging from idioms (*tirare le cuoia* "to die"), technical expressions (*sistema operativo* "operating system"), light verbs (*fare una domanda* "ask a question"), to restricted collocations (*particolare attenzione* "particular attention", *caffè bollente* "boiling coffee"), compounds (*parola chiave* "keyword"), and proper nouns (*Stati Uniti* "United States").
From the many different attempts to define these phenomena, in this article I intend to use the following definition, which considers "collocation" a broad term that does not rely on a particular theory, but can be applied (and narrowed) to any specific area or application: "A collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon" (Evert, 2005, p. 17).

1.1 Prototypical features

According to this definition, collocations can present different degrees of cohesion, ranging from frozen, semantically opaque and idiomatic expressions (*tagliare la corda* "to run away") to conventionally restricted combinations (*prendere una decisione* "to make a decision"). There are, however, at least three common properties that emerge as prototypical features in the linguistic analysis of collocations:

- their semantic non-compositionality;
- the non-substitutability of their components by semantically similar words (Evert, 2005, p. 16), due to their arbitrary nature: *camera oscura* but not **stanza oscura* "dark room";
- the vast range of their possible syntactic configurations: a manual tagging of an existing list of Italian collocations has found that they are composed by more than 150 different part-of-speech patterns (see section 2.2).

Thus, from this perspective, collocations are complex lexical items that show different syntactic and semantic profiles, which are placed along a continuum rather than falling into mutually-exclusive categories.

1.2 Motivation

After a period of oblivion (Kennedy, 2008), in recent years multi-word units are once again at the centre of interest of second and foreign language learning and SLA researchers (Nattinger & DeCarrico, 1992).

Collocational competence is widely recognized as playing a key role in two different aspects of language learning (Nesselhauf, 2005):

- in improving learners fluency, providing ready-to-use 'chunks' of language (Lewis, 2000; Sung, 2003);
- in facilitating their comprehension.

While native speakers have the innate ability to use and recognise such word combinations as the most appropriate way of expressing a given concept, collocational competence is often a "pain in the neck" for language learners, who produce linguistically-correct sentences in the target language, but are told that a native speaker would never produce the same sentence, simply because of the lack of collocational competence.

These widely attested difficulties of non-native speakers are partly due to the way they acquire and mentally organise L2 vocabulary (Wray, 2002): SLA research demonstrates that, in contrast to native speakers, non-native speakers seem to start learning single words, and gradually build up more extended chunks (Schmitt & Underwood, 2004). Rather than store and use ready-made chunks of language, they "tend to overuse the creative combination of isolated words" (Jaén 2007, p. 131); following Sinclair (1991), learners rely on the open-choice principle more than on the idiom principle.

Given these considerations, it seems that collocations require specific pedagogical attention; to address this issue, a research project was initiated during the first half of 2009, the main goal of which was the creation of a *Dictionary of Italian Collocations* (DICI).

The DICI is a linguistic resource, based on natural language processing methodologies, which aims to support the processes of learning and testing the collocational competence of students of Italian as a second language within a Virtual Learning Environment (VLE); in reality, it is part of a group of tools projected to support second language acquisition. The goal of the on-going project is to enrich APRIL (*Ambiente personalizzato di rete per l'insegnamento linguistico* "Personalised web environment for language learning"), a VLE specifically devoted to language learning[1], with linguistic resources created from natural language processing methodologies (Spina, to be published). Natural language processing tools can be a valuable resource for language learning activities in online learning environments, which can benefit from large quantities of structured linguistic data, typically extracted from corpora and stored in databases, and from innovative computational methodologies (Tschichold, 2006).

Some of the main features of the dictionary, which will be discussed in the subsequent sections, are as follows:

- it is corpus-based (collocations are extracted from a reference corpus of Italian language);
- it is a learner-oriented tool, and, therefore, it is based on a list of the most common Italian collocations, classified on a frequency basis;
- it is also based on statistical methodologies (collocations extracted from the corpus are ordered by combining frequency with dispersion in the different textual genres represented in the corpus).

[1] More information about APRIL can be found at the web site of the projet: http://april.unistrapg.it.

This paper will describe the experiment undertaken with the goal of testing the methodology adopted for the creation of the dictionary; this methodology is divided into four distinct steps:

- extraction of candidate collocations from a reference corpus;
- filtering of the candidate collocations by their frequency and their dispersion within the textual genres included in the corpus;
- compilation of the dictionary;
- integration of the dictionary with the online learning environment, in order to support learning activities.

2 Methodology

2.1 Reference corpus

The collocations are extracted from the *Perugia Corpus* (PEC)[2], a reference corpus of Italian of the size of 18 million words. The corpus is composed of eight different sections, corresponding to eight different textual genres, as shown in table 1.

Textual genre	*Number of words*
fiction	3000000
non-fiction	2000000
web	5000000
academic prose	1000000
press	3000000
language of administration	1000000
television programs	1000000
spoken texts	2000000
TOTAL	18000000

Table 1: Composition of the PEC corpus

Each of the eight textual genres has different sub-genres, connected to the subject (for press, for example, the subjects are: culture, politics and current events) or to the different kinds of interaction (spoken texts are divided into speeches and dialogues).

As a result, despite its limited size, the PEC corpus allows the extraction of the most common collocations in a wide range of textual genres, which can be considered representative of contemporary Italian. In addition, the PEC corpus is tokenized, pos-tagged and lemmatized; this offers the possibility of filtering the collocations by part-of-speech (see section 2.2). The pos-tagging has been carried out using *TreeTagger* (Schmid, 1994), trained with ad-hoc resources[3].

In order to retrieve the collocations, the queries on the corpus have been executed using *IMS Corpus Workbench* (Christ, 1994).

2.2 POS filtering

It has been demonstrated (Pazos Bretaña & Pamies Bertrán, 2008) that pos-tagging and lemmatisation produce a higher degree of effectiveness in the automatic

[2] Details on the corpus composition can be found at the web site http://elearning.unistrapg.it/corpora/pec.html.

[3] The POS tagset used for the PEC corpus is the same used for *La Repubblica* corpus (Baroni et al., 2004).

extraction of collocations. A methodology based on a POS filter has thus been adopted for the extraction of candidate collocations in the PEC corpus.
Accordingly, the actual phase of retrieval was preceded by the analysis and the manual tagging of an existing list of collocations, extracted from the *LIP*, a 500.000 words Italian spoken corpus (De Mauro et al., 1993) and from the Italian *Wordnet* (Roventini et al., 2000). From the 150 different POS sequences found in the list, the 10 most frequent have been selected. This selection has only considered collocations with a referential meaning, excluding combinations that only have grammatical or textual functions.
The 10 most frequent POS sequences with referential meaning, which cover 75% of all the referential collocations of the list, are listed in table 2.

ADJ ADV N	nudo come un verme "as naked as a worm"
ADJ CONG ADJ	bianco e nero "black and white"
ADJ N	terzo mondo "third world"
N ADJ	cassa comune "common fund"
N CONG N	andata e ritorno "back and forth"
N N	caso limite "borderline case"
N PRE N	abito da sera "evening dress"
V ADJ	stare zitto "keep quiet"
V ART N	fare la doccia "take a shower"
V N	avere paura "be afraid"

Table 2: The 10 POS sequences of collocations extracted from the PEC corpus

2.3 Collocations extraction

For the extraction of collocations, an experiment was conducted on a 12-million-word sample of the PEC corpus (corresponding to the fiction, press, academic prose and web sections) and on six of the ten selected POS sequences: ADJ-CONG-ADJ, N-CONG-N, N-N, N-PRE-N, V-ART-N and V-N.
The candidate collocations were retrieved in the four corpus sections using the *Corpus Query Processor* (CQP), which allows the execution of queries in large text corpora with linguistic annotations.
The extraction of the collocations with the POS sequence V-N, for example, was executed using the following query:
cwb-scan-corpus -C PEC lemma+0 ?pos+0=/VER/ lemma+1 ?pos+1=/NOUN/
Once the lists of candidate collocations were obtained, they were filtered by removing all the candidates with frequency = 1.
The result of this operation of filtering is listed in Table 3.

	N° of collocations
ADJ CONG ADJ	1492
N CONG N	1376
N N	2365
N PRE N	16637
V ART N	14701
V N	5072

Table 3: Number of collocations per POS sequence

In the following part of the experiment, the resulting 41643 candidate collocations received further processing in order to:

- remove the combinations that are not collocations;
- obtain a list suitable to account for actual frequency across the different textual genres.

2.4 Collocations filtering

Once automatically extracted, the candidate collocations were filtered upon a selection based on a statistical measure of dispersion, in order to account for their distribution in the different corpus sections.
This balance between pure frequency and degree of distribution across the different genres is crucial in order to avoid disproportions due to high frequency values in a single section of the corpus. The collocation *aggrottare la fronte* "to frown", for example, has a fairly high frequency in the corpus, but all its 71 occurrences belong to the fiction section, and consequently its dispersion value is 0.
Accordingly, the value of pure frequency is re-calculated on the basis of the distribution within the textual genres, obtaining a reduced frequency as a result. As a measure of dispersion the Juilland's D value has been chosen (Juilland & Chang-Rodriguez, 1964; Bortolini et al., 1971; Oakes, 1998):

$$D = 1 - \frac{\sigma}{\mu\sqrt{n-1}}, \quad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n}}.$$

Furthermore, given the need to account for both dispersion and frequency, the value of usage (U) has been calculated, following the Juilland & Chang-Rodriguez (1964) procedure:

$$U = \sum_{i=1}^{n} x_i \left(1 - \frac{\sigma}{\mu\sqrt{n-1}}\right).$$

The threshold for the inclusion in the dictionary has been set to a usage value of ≥ 2; all the word combinations with a U value lower than 2 were therefore excluded.
The result of this filtering process is listed in Table 4.

	N° of collocations with U ≥ 2
ADJ CONG ADJ	49
N CONG N	36
N N	92
N PRE N	545
V ART N	876
V N	449

Table 4: Number of collocations with $U \geq 2$

The resulting 2047 candidate collocations, ordered by usage, are the most frequent and evenly dispersed combinations found in the sample corpus.
A final filtering operation consists in the manual exclusion of non-collocations (for example fully compositional and predictable combinations, such as *è un ragazzo* "he is a boy", or part of larger word

combinations). Table 5 shows the result of this manual filtering.
The results show that the number of non-collocations varies upon the different POS sequences. In some cases it is a question of tagging errors, especially in the N-N sequence, where the tagger happens to mix up nouns and adjectives. In other, more frequent cases (especially with the verb *essere* "to be") the multi-word unit is not a conventional combination of words.

	N° of manually filtered non-collocations	*%*
ADJ CONG ADJ	0	0,0
N CONG N	3	8,3
N N	29	31,5
N PRE N	116	21,3
V ART N	263	30,0
V N	83	18,5
TOTALE	494	24,1

Table 5: Number and percentage of manually excluded non-collocations

2.5 Compilation of the dictionary

The final result of the experimental procedure described above, aimed at the extraction and selection of collocations from a reference corpus, is a list of 1553 word combinations, which correspond to the dictionary entries.
Thus this final list is the starting point of the dictionary, which is then enriched with other structured information connected with lexical, syntactic, semantic, contextual and statistical aspects of collocations.
The dictionary is built up as a lexical database where each field provides a specific type of information and can have relationships with other fields; the structure and compilation of the database was guided by three key points:

- the aim of the dictionary is its integration with an online learning environment. Accordingly, the database should permit this integration and data exchange with other web applications, and thus adopt an open and flexible format[4];
- for the same reason, the data included in the dictionary should be organized and described using a detailed formalization, in order to be processed by other software. Due to the great variability of multi-word units, this research of formalization can represent a real challenge (Tschichold, 2008);
- finally, due to the pedagogical goal of the dictionary, all the data which will be provided to the final users, who are non-native speakers of Italian, should be accurately calibrated.

The information provided by the database is categorized using three different criteria:

1. data visible to the learners. This is typically the semantic and syntactic information (definition, examples, part-of-speech, syntactic context of occurrence of collocations). The compilation of this section of the database is performed with the support of existing lexicographic resources (definitions are extracted from electronic dictionaries); with regard to examples of use, authentic instances are provided by the reference corpus;
2. data to be processed by other applications. The first goal in the integration with an online learning environment is the automatic recognition of collocations; to achieve this goal, data must be grammatically and syntactically analyzed. The internal syntactic configuration of collocations must be formally described, as well as the syntactic context of occurrence of combinations; it must be specified, for example, whether they are completely invariable or whether they can be interrupted by elements not belonging to the word combination. This information is crucial for the retrieval of collocations across texts. Table 6 shows three examples of collocations with different syntactic configuration; in the first case, *fare la doccia* "to take a shower", both the verb and the noun can vary their morphosyntactic inflection, and an optional element (an adverb) can be inserted after the verb. In the second case, *abito da sera* "evening dress", only the first noun can take the plural form, and no other element can be placed within the collocation. The third example, *alti e bassi* "highs and lows", is the easiest case, where the whole collocation is invariable;

collocation	*syntactic configuration*
fare la doccia	[V$fare][ADV]? la\|una\|NUM [N$doccia]
abito da sera	[N$abito] da_sera
alti e bassi	alti_e_bassi

Table 6: Three examples of syntactic configurations

3. statistical information acquired from previous filtering processes. For each collocation, the values of frequency (total and relative to each corpus section) are stored in the database, as well as the values of dispersion and usage. This information is very important for the attribution of each collocation to a specific level of linguistic competence of learners (see section 3).

To summarize the structure of the database, table 7 shows an example of a dictionary entry.

2.6 Integration in a learning environment

A great amount of research has been carried out in recent years in the field of electronic dictionaries of collocations (Santos Pereira & Mendes, 2002; Alonso Ramos, 2003; Bolshakov & Miranda-Jiménez, 2004, all references relating only to languages other than English). The applications of such lexicographic tools are numerous, in the domains of language learning, natural language processing (tagging, parsing, word sense disambiguation,

[4] The consequent choice has been the use of *MySql*, one of the most popular open source relational databases.

for example), information retrieval, automatic translation, and many others.

Collocation	numero di telefono
POS	N
Definition	gruppo di numeri da comporre per telefonare a qualcuno
Example	se mi dai il tuo numero di telefono ti richiamo fra 5 minuti
Syntactic context	V$comporre\|V$digitare|V$fare ART
Internal structure	N PRE N
Syntactic configuration	[N$numero] di_telefono
Total frequency	106
Frequency fiction	69
Frequency press	14
Freq. academic	5
Frequency web	18
Dispersion	0,4
Usage	48,2

Table 7: The example entry *numero di telefono* "telephone number"

In this context, the specificity of the *Dictionary of Italian Collocations* lies in its integration with a Virtual Learning Environment. By "virtual learning environment" (VLE), I intend here a web application specifically devoted to language learning, delivered in an e-learning modality.
From the many available e-learning platforms, the one selected to be used within the *APRIL* project is *Moodle*[5]. The final goal of the DICI is to be integrated in *Moodle* through the use of an additional module called LELE (*Linguistically-Enhanced Learning Environment*).
A further experiment is thus aimed at testing the integration of LELE with *Moodle*, as means to provide language learners with additional NLP resources, in order to improve their collocational competence.
Through this integration, in a specific area of our VLE devoted to the study of vocabulary, students of Italian as a second or foreign language can perform receptive and productive learning activities concerning the recognition and the active use of collocations, with the support of all the information stored in the lexical database described above.
Some of the features of the dictionary in its integration with the VLE are:

- to automatically recognize and highlight multi-word units in written Italian texts;
- to show additional linguistic information about the selected collocations;
- to generate collocation tests for collocational competence assessment of second or foreign language learners.

The lexical database is the core of LELE; through the use of the additional module it can be connected to a pos-tagger; once again, *Treetagger* has been chosen for this task.
The VLE contains a number of web pages with written Italian texts, specifically selected to support linguistic activities; each student can filter these texts to automatically highlight specific lexical items, like academic vocabulary (Spina, to be published), or, in this case, collocations.
When a text is filtered, it is processed by the LELE system; the process of collocations retrieval encompasses three different steps:

1. firstly the text is tagged and lemmatized, producing a three-column list for each word, composed of token - POS – lemma rows;
2. the first phase is invisible to the end user, as is the second, in which the lemma-POS sequences are compared to all the lemma-POS sequences of the database;
3. if an exact match is found, the corresponding word sequence is highlighted in the final filtered version of the text, which is then visible to the student.

At the end of this process, students can perform linguistic activities with the original texts, where all the collocations (or, amongst them, only some specific grammatical categories) are explicitly presented (Figure 1); they can also obtain further linguistic information about their meaning, usage and syntactic configuration.

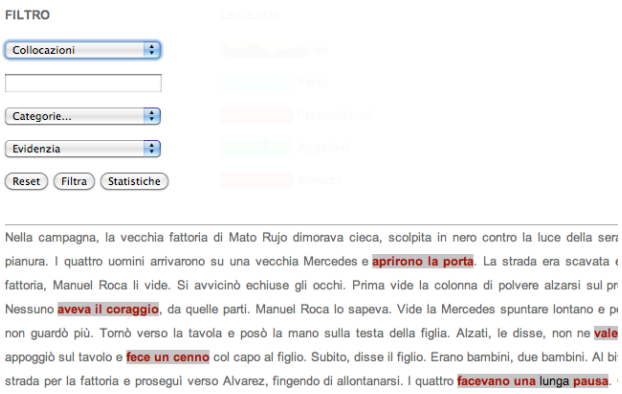

Figure 1 – A sample of a text processed by the LELE system.

Conclusions

In this article, I have argued that collocational competence needs specific pedagogical attention in second and foreign language learning. I have then introduced the DICI, an electronic dictionary of Italian collocations, designed to support collocational competence acquisition of learners of Italian. While the main goal of the DICI project is pedagogical, it is also intended to contribute to research in the field of multi-word units.
I have then briefly described the composition of the reference Italian corpus from which the collocations were extracted, and the methodology of extraction and filtering of candidate collocations. This experimental methodology identified 1553 collocations in a 12-million-word sample of the reference corpus, for 6 of the 10 POS sequences selected to be included in the dictionary.
Furthermore, I have explained the main criteria for the composition of the lexical database of collocations, as well as its integration with a Virtual Learning Environment.

[5] *Moodle* is a web application for producing e-learning courses. It is a global development project designed to support a social constructionist framework of education. For more information about *Moodle*, see the web site http://moodle.org/.

In the next step of the project we plan to apply the same methodology to the entire corpus, for all of the 10 selected POS sequences, and then to build up the final version of the dictionary which will then be integrated within the VLE.

There are however some issues concerning the methodology that future research should address:

- with regard to statistical measure of usage, the opportunity to choose a value lower that 2 as a threshold for inclusion in the dictionary should be verified; after the test, in fact, less than 5% of initial collocations fell within this range;
- further research is also needed to assign collocations to the different levels of linguistic competence of learners. It is not obvious that the usage value is the only valid criterium to attribute a lexical item to learners of a specific level (beginner, intermediate or advanced). A correlation between collocations and levels of competence would have both pedagogical and technical benefits, because it would allow the division of the database into different competence-based sections and consequently an improvement in the speed of the LELE system;
- with regard to the integration within the VLE, we plan to dedicate further research to add other tools aimed at supporting learning activities on collocations, particularly concerning productive tasks (a resource to guide and assist the students in writing activities, for example, is in the early stages of development).

Acknowledgements

The research reported in this paper was supported by a grant by the Fondazione Cassa di Risparmio di Perugia and the University for Foreigners Perugia.

I wish to express my thanks to Francesco Scolastra and Luigi Pinca for their precious collaboration.

References

Alonso Ramos, M. (2003). Hacia un Diccionario de colocaciones del español y su codificación. In M. A. Martí et al. (Eds.), Lexicografía computacional y semántica. Barcelona: Universitat de Barcelona, pp. 11-34.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. & Mazzoleni, M. (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper Italian. In Proceedings of LREC 2004.

Benson, M. (1990). Collocations and general-purpose dictionaries. International Journal of Lexicography, 3(1), pp. 23-35.

Benson, M., Benson, E. & Ilson, R. (1986). The BBI Dictionary of English Word Combinations. Amsterdam: John Benjamins.

Bolshakov, I.A & Miranda-Jiménez, S. (2004). A Small System Storing Spanish Collocations. In A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing. Berlin: Springer, pp. 248-252.

Bortolini, U., Tagliavini, C. & Zampolli, A. (1971). Lessico di frequenza della lingua italiana contemporanea. Milano: Garzanti.

Calzolari, N. et al. (2002). Towards Best Practice for Multiword Expressions in Computational Lexicons. In Proceedings of the Third International Conference on Language Resources and Evaluation (Las Palmas, Canary Islands; Spain, 29 May – 31 May 2002), pp. 1934-1940.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In Proceedings of COMPLEX' 94, Budapest.

Cowie, A. (1981). The treatment of collocations and idioms in learners' dictionaries. Applied Linguistics, 2, pp. 223-235.

De Mauro, T., Mancini, F., Vedovelli, M. & Voghera, M. (1993). Lessico di frequenza dell'italiano parlato. Milano: Etas.

Evert, S. (2005). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, IMS, University of Stuttgart.

Granger, S. & Meunier, F. (2008). Phraseology. An interdisciplinary perspective. Amsterdam: John Benjamins.

Jaén, M.M. (2007). A Corpus-driven Design of a Test for Assessing the ESL Collocational Competence of University Students. International Journal of English Studies, 7(2), pp. 127-147.

Juilland, A & Chang-Rodriguez, E. (1964). Frequency Dictionary of Spanish Words. The Hague: Mouton & Co.

Kennedy, G. (2008). Phraseology and language pedagogy. In F. Meunier and S. Granger (Eds). Phraseology in Foreign Language Learning and Teaching. Amsterdam: John Benjamins, pp. 21-41.

Lewis, M. (2000). Teaching collocation. Further developments in the lexical approach. Hove: Language Teaching Publications.

Meunier, F. & Granger S. (2008). Phraseology in foreign language learning and teaching. Amsterdam: John Benjamins.

Nattinger, J.R. & DeCarrico, J.S. (1992). Lexical phrases and language teaching. Oxford: Oxford University Press.

Nesselhauf, N. (2005). Collocations in a learner corpus. Amsterdam: John Benjamins.

Nation, P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.

Oakes, M. (1998). Statistics for corpus linguistics. Edinburgh: Edinburgh University Press.

Palmer, H.E. (1933). Second Interim Report on English Collocations. Tokyo: Kaitakusha.

Pazos Bretaña, M. & Pamies Bertrán, A. (2008). Combined statistical and grammatical criteria. In S. Granger & F. Meunier (Eds), Phraseology. An interdisciplinary perspective. Amsterdam: John Benjamins, pp. 391-406.

Roventini, A., Alonge, A., Calzolari, N., Magnini, B. & Bertagna, F. (2000). ItalWordNet: a Large Semantic Database for Italian. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000, Athens, Greece, 31 May – 2 June 2000). Paris: The European Language Resources Association (ELRA), pp. 783-790.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002), Mexico City, pp. 1–15.

Santos Pereira, L. A. & Mendes, A. (2002). An electronic dictionary of collocations for European Portuguese: methodology, results and applications. In A. Braasch, C. Povlsen (Eds.), Proceedings of the Tenth EURALEX International Congress, Copenhagen: Center for Sprogteknologi, II, pp. 841-849.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing (http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf).

Schmitt, N. & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N. Schmitt (Ed.), Formulaic Sequences. Amsterdam: John Benjamins, pp. 173–189.

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

Smadja, F. (1993). Retrieving collocations form text: Xtract. Computational Linguistics, 19(1), pp. 143-177.

Spina, S. (2010). The Dici Project: towards a Dictionary of Italian Collocations integrated with an online language learning platform. In Proceeding of eLexicography in the 21st century: new challenges, new applications (Louvain-La-Neuve, 22-24 octobre 2009), Louvain-La-Neuve, Cahiers du Cental.

Spina, S. (to be published). Building a suite of online resources to support academic vocabulary learning. Paper presented at the EUROCALL 2009 conference, Universidad Politécnica de Valencia, September 2009.

Sung, J. (2003). English lexical collocations and their relation to spoken fluency of adult non-native speakers. Unpublished doctoral dissertation, Indiana University of Pennsylvania, Pennsylvania.

Tschichold, C. (2006). Intelligent CALL: The magnitude of the task. In P. Mertens, C. Fairon, A. Dister & P. Watrin (Eds). Verbum ex machina. Actes de la 13e conférence sur le Traitement automatique des langues naturelles. Louvain-la-Neuve: Presses universitaires de Louvain (Cahiers du Cental 2), pp. 806-814.

Tschichold, C. (2008). A computational lexicography approach to phraseologisms. In S. Granger & F. Meunier (Eds), Phraseology. An interdisciplinary perspective. Amsterdam: John Benjamins, pp. 361-376.

Wray, A. (2002). Formulaic language and the lexicon. Cambridge: Cambridge University Press.