Towards the Automatic Learning of Idiomatic Prepositional Phrases*

Sofía N. Galicia-Haro[1] and Alexander Gelbukh[2]

[1] Faculty of Sciences UNAM Universitary City, Mexico City, Mexico
sngh@fciencias.unam.mx
[2] Center for Computing Research, National Polytechnic Institute, Mexico
gelbukh@cic.ipn.mx, www.Gelbukh.com

Abstract. The objective of this work is to automatically determine, in an unsupervised manner, Spanish prepositional phrases of the type preposition - nominal phrase - preposition (P–NP–P) that behave in a sentence as a lexical unit and their semantic and syntactic properties cannot be deduced from the corresponding properties of each simple form, e.g., *por medio de* (by means of), *a fin de* (in order to), *con respecto a* (with respect to). We show that idiomatic P–NP–P combinations have some statistical properties distinct from those of usual idiomatic collocations. We also explore a way to differentiate P–NP–P combinations that could perform either as a regular prepositional phrase or as idiomatic prepositional phrase.

1 Introduction

Any computational system for natural language processing must cope with the ambiguity problem. One of the most frequent ambiguities in syntax analysis is the prepositional phrase attachment. A preposition can be linked to a noun, an adjective, or a preceding verb. There are certain word combinations of the type preposition - nominal group - preposition (P–NP–P) that can be syntactically or semantically idiosyncratic in nature or both, that we called IEXP. The automatic determination of such IEXP groups should reduce the prepositional phrase attachment problem by defining three or more simple forms (since the nominal group can contain more of a simple form) as one lexical unit.

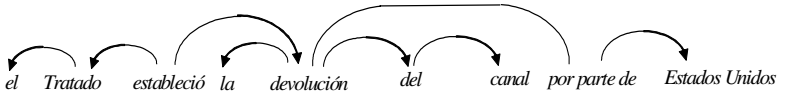
Spanish has a great number of prepositional phrases of the type P–NP–P more or less fixed. Among them: *a fin de* (in order to), *al lado de* (next to), *en la casa de* (in the house of), etc. The IEXP (*a fin de*, *al lado de*), can be analyzed assuming that these expressions behave as a syntactic unit and therefore could be included directly in a computational dictionary. Specifically, such combinations are frequently equivalent to prepositions, i.e., they can be considered as one multiword preposition: e.g., *in order to* is equivalent to *for* (or *to*) and has no relation with *order*; other examples: *in front of* (before), *by means of* (by), etc. Such dictionary can be useful in prepositional phrase attachment: given a compound preposition *in_order_to* is present in the

* Work partially supported by Mexican Government (CONACyT, SNI, CGPI-IPN, PIFI-IPN).

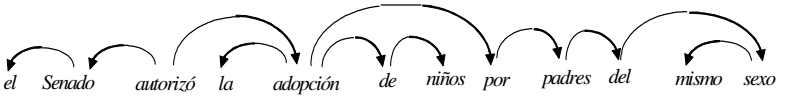
A. Gelbukh, A. de Albornoz, and H. Terashima (Eds.): MICAI 2005, LNAI 3789, pp. 780 – 789, 2005.

dictionary, the *to* in *John bought flowers in order to please Mary* would not be attached to *bought*.

For example, the phrase *El Tratado estableció la devolución del canal por parte de Estados Unidos* (The Treaty established the return of the canal by the United States) has the following structure according to Dependency Grammars [11]:



In opposition, regular P–NP–P are analyzed considering the initial combination P–NP like a unit, and the second preposition as a one introducing a complement, not always linked to the preceding NP. For example, the phrase *El Senado autorizó la adopción de niños por padres del mismo sexo* (The Senate authorized the adoption of children by same sex parents) that is very similar to the previous example in the POS of most words has the following structure:



We could observe that there are more links in a structure with regular prepositional phrases. Therefore it is necessary to distinguish which of the Spanish P–NP–P should be analyzed as a IEXP and which should be analyzed as a P–NP–P. But there is no a complete compilation of the IEXP groups. Nañez [12] compiled the most wide list but he himself considers that a prepositive relation study is something incomplete "susceptible of continuous increase". In addition, the main Spanish dictionaries [4], [14] do not contain the information necessary for a computational dictionary of this type.

The IEXP groups are used frequently in everyday language, therefore natural language applications need to be able to identify and treat them properly. Apart from syntactic analysis the range of applications where it is necessary to consider their specific non compositional semantic is wide: machine translation, question-answering, summarization, generation. But IEXP could present different uses and their meaning agree with context.

In this work, we mainly investigated corpus-based methods to obtain the prepositional phrases of the type IEXP (idiomatic expressions), and the form to differentiate their use as fixed forms from the literal ones, for example:

1. Idiomatic expression: *a fin de obtener un ascenso* (to obtain a promotion, literally: 'at end of'),
2. Free combination: *a fin de año obtendrá un ascenso* (at the end of the year she will be promoted),
3. Part of a larger idiom: *a fin de cuentas* (finally, literally: 'at end of accounts').

More precisely our aims are:

- To analyze the linguistic characteristics that differentiate IEXP groups from the regular prepositional phrases

- To find by statistical measures the IEXP groups that should be included in a computational dictionary for syntactic analysis
- To identify the different uses of the IEXP groups as idiomatic expression, free combination (literal), and part of a larger idiom.

In section 2 we present the linguistic analysis, then we present the characteristics of the corpus and some frequencies obtained from it. In section 4, we present the obtained results applying the statistical methods and our exploration to differentiate regular from idiomatic use for the same prepositional phrase.

2 Linguistic Analysis

It is important to emphasize some linguistics properties that place IEXP in a different group from the regular prepositional phrases. In Spanish grammar these groups are denominated adverbial locutions [15] or prepositional locutions [12] according to their function. As it is indicated in [15], locutions could be recognized by its rigid form that does not accept modifications and the noun that shows a special meaning, or by its global meaning, that is not the sum of the meanings of its components. For word combinations lexically determined that constitute particular syntactic structures, as it is indicated in [10], their properties are: restricted modification, non composition of individual senses and the fact that nouns are not substitutable.

1. Some features of prepositions and nouns in IEXP's

   Some IEXP groups are introduced by the preposition "so". This preposition is considered old fashioned by [15]. It has a restricted use or appears in fixed forms. We found 389 cases from the corpus: *so capa de* (1 case), *so pena de* (203 cases), *so pretexto de* (177), *so riesgo de* (5), *so peligro de* (1) and one case for P-NP.

   Other IEXP groups contain nouns that are rarely used outside the context of fixed expressions. Examples obtained from corpus are: *a fuer de* (in regard to, 44 cases), *a guisa de* (like, 54 cases) and *a la vera de* (to the side of, 187 cases).

2. Restricted modification

   Many of the nouns found in IEXP groups cannot be modified by an adjective. For example: *por temor a* (literally: by fear of) vs. *por gran temor a* (avoiding *vs.* by great fear of), *a tiempo de* vs. *a poco tiempo de* (at the opportune moment *vs.* a short time after), etc. In some cases, the modification forces to take a literal sense of the prepositional phrase, for example in the following sentences:

- … *por el gran temor a su estruendosa magia* (by the great fear to its uproarious magic) , *por el gran temor* has a literal meaning related to fear.
- … *denegó hoy la libertad bajo fianza por temor a una posible fuga.* (today denied the freedom on bail to avoid a possible flight), *por temor a* has a meaning related to avoid

3. Non substitutable nouns

   The noun inside the IEXP cannot be replaced by a synonym. For example in the phrase: *se tomará la decisión de si está a tiempo de comenzar la rehabilitación* (the decision will be taken on if it is the right time to begin the rehabilitation), where *a tiempo de* cannot be replaced by *a período de* (on period of), *a época de* (time of).

4. Variations in the nominal group

The nominal groups of certain IEXPs present variations due to inflection of the noun. For example, *a mano(s) de* :

- *... tras la traición que han sufrido a mano de sus antiguos protectores argelinos.* (... after the treason that, apparently, has undergone by cause and action of their old Algerian protectors)
- *... habitantes enardecidos por el asesinato de un profesor a manos de un policía...* (... inhabitants inflamed by the murder of a professor by cause and action of a policeman …)

5. Different uses, meaning agree with context

Some IEXP groups initiate fixed phrases or can be literal phrases according to the context, in addition to their use as idiomatic expressions. Examples:

"*al pie de*" It appears as idiomatic expression in:

- *La multitud que esperó paciente al pie de la ladera de la sede de la administración del canal, corrió hacia arriba ...* (The patient multitude that waited at the base of the slope of the seat of the channel administration, run upwards ....)

"*al pie de*" It initiates a larger idiom: *al pie de la letra* (exactly) in:

- *Nuestro sistema de procuración de justicia se ha transformado y en vez de observar al pie de la letra las leyes …* (Our system of justice care has been transformed and instead of observing the laws exactly …)

"*al pie de*" It initiates a free combination in:

- *El anillo estaba junto al pie de María* (The ring was next to the foot of Maria)

3 Characteristics of the Corpus

For our analysis, we selected four Mexican newspapers that are daily published in the WEB with a considerable part of their complete publication. The texts correspond to diverse sections: economy, politics, culture, sport, etc. from 1998 to 2002. The text collection has approximately 60 million words [7].

Initially, to analyze the word groups of the type P–NP–P we used a POS tool to assign morphological annotation. We developed a program to extract the P–NP–P patterns where prepositions correspond to a very wide list of simple prepositions that was obtained from [12] which includes prepositions with liberality.

We extracted all word strings corresponding to P–NP–P using the following grammar:

```
PP      → P NP P

NP      → N | D N | V-Inf | D V-Inf
```

where P stands for preposition, N for noun, D for determinant, and V-inf for infinitive verb (in Spanish, infinitives can be modified by a determinant: *el fumar está prohibido*, literally: 'the to-smoke is prohibited').

In the complete text collection, we found 2,590,753 strings of the type P–NP–P, with 372,074 different types. The different groups with frequency greater than two were 103,009. In these strings, the above described cases for syntactic analysis were

recognized: compound words of type IEXP, and the combination P–NP followed by a preposition that introduces a second prepositional phrase.

Since many IEXP are traditionally considered as complex prepositions and functional words have high frequency we obtained the frequencies of P–NP–P groups recognized with the grammar as a first approximation to determine them. Table 1 shows the frequencies of the 20 more frequent groups, where we could observe that 17 groups are IEXP type and only three phrases: *en la ciudad de* (literally: in the city of), *del gobierno de* (literally: of the government of), *del estado de* (literally: of the state of), do not correspond to IEXP type. The groups *en la ciudad de*, *del estado de* have a high score since there are two Mexican names that include the nouns 'city' (ciudad) and 'state' (estado) to differentiate them from the country: *ciudad de Mexico* (Mexico City) and *estado de México* (Mexico State). The group *del gobierno de* has a high score since 'government' has immediately an attribute related to persons and countries linked with preposition 'of'. Most of the 17 IEXPs are equivalent to a single preposition.

The sequences of words that compose the nominal groups are distributed of the following way in the text collection: NP constituted by a noun: 44,.646; NP constituted by several words: 58,363. It is observed that more than 50% contain determinants.

Table 1. Frequencies of P–NP–P patterns

Frequency	P GN P
11612	a pesar de
11305	a partir de
10586	de acuerdo con
10418	en contra de
7982	por parte de
7240	en la ciudad de
6819	a fin de
6512	en materia de
5904	en el caso de
5758	por lo menos

Frequency	P GN P
5403	en favor de
4998	a partir del
4899	en el sentido de
4882	con el fin de
4779	a lo largo de
4510	en caso de
4469	del gobierno de
4186	en cuanto a
4124	del estado de
4114	por medio de

4 Statistical Procedure

From the linguistic analysis we noted that IEXP could be classified as collocations. The extracting criteria is based on such assumption applying general methods for collocation extraction in addition to frequency. Diverse statistical measures have been used to identify lexical associations between words from corpus [2], [5], [6], [16]. The measures that we used to determine the lexical association of words are:

- Frequency (Freq),
- Point-wise mutual information (PMI),
- Log-likelihood (LL),
- Pearson measure (χ^2, or Chi-2).

We obtained these four measures for the complete collection of texts using the statistical program NSP, developed by [1]. Because of the total size of the collection we

first split it in sixty parts and we applied the NSP ngram module to each part. Then we combined the results and we extracted automatically all P–NP–P groups determined by the grammar. Finally we applied each of the above four statistical measures by the corresponding NSP module and we obtained the ranked measures for them.

In order to prove the methods of idiomatic expression identification, we eliminated the P–NP–P groups that appeared less than three times, because when applying statistical methods to sparse data poor results are obtained. Since in the text collection the prepositional groups are not annotated as composed words, we compared the results against the most wide list of prepositional locutions available (LPL), that of [12].

In Figure 1 we present the learning curves automatically obtained, the horizontal axis correspond to the number of the top P–NP–P ranked by each method and the vertical axis correspond to the number of IEXP detected. In the first step we considered the top one hundred prepositional phrases ranked by the statistical measure and automatically were searched the groups considered in LPL, then we considered the top ten thousand to apply the same detection, then we considered the top twenty thousand prepositional phrases and so on augmenting ten thousand P–NP–Ps each time.

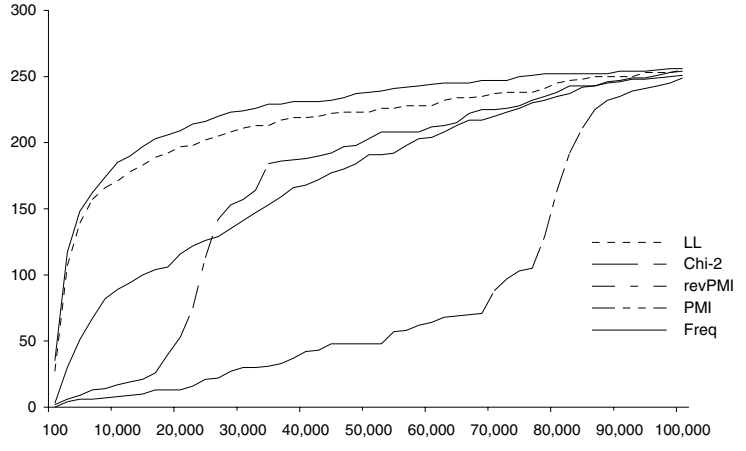


Fig. 1. Statistical measures for the P-NP-P with a minimum occurrence of 3

We considered 252 cases of LPL that correspond to specific IEXP groups, since the author considers very general cases as “a ... en” (literally: to … in), “a ... por” (literally: to … by), etc. An extract of LPL is the following group:

197. con ganas de	204. con motivo de
198. con idea de	205. con objeto de
199. con independencia de	206. con occasion de
200. con intención de	207. con omission de
201. con la mira en	208. con... para...
202. con miras a	209. con peligro de
203. con motivo a	210. con... por...

Since there is no generally accepted definition of the latter three measures for three elements, we applied pair-wise measures to [P–NP] and [P] of the whole combination P–NP–P: e.g., the PMI assigned to *a fin de* was the PMI of the two strings (1) *a fin* and (2) *de*. Evaluation of other possible ways of calculating the dependency between the three elements was left for our future work.

4.1 Discussion

The best measure proved to be simple frequency, and log-likelihood shows nearly the same performance. On the first 100 items the precision obtained with frequency measure was 50% (which is 20% recall on LPL).

However, manual inspection of the results revealed among the top elements new IEXP, such as *a las afueras de* (literally: 'to the outskirts of'), *bajo los auspicios de* (literally: 'under the auspices of'), *en el marco de* (literally: 'in the frame of'), *con el objetivo de* (literally: 'with the objective of'), *a efecto de* (literally: 'to effect of'), *por concepto de* (literally: 'by concept of'), etc. and variants of existing ones in LPL such as: *por la vía de* (literally: 'by the way of'), *en las manos de* (literally: 'in the hands of'). Thus, real precision of our method is better than that measured by comparison with LPL, and the method allows detection of new combinations.

PMI gives very poor results, which is in accordance with the general opinion that PMI is not a good measure of dependency (though it is a good measure of independency) [10]. What is more, PMI seems to have inverse effect: it tends to group the idiomatic examples nearer the end of the list. With this, ordering the list in *reverse* order by PMI (revPMI in Figure 1) gives better results. However, such order is much worse than Freq and LL orders: it groups most of the combinations in question around the positions from 25,000 to 35,000 in the list of 100,000. Pearson measure also shows much worse performance than LL. For a detailed comparison of the log-likelihood and chi-squared statistics, see [13].

A possible explanation for the fact that simple frequency performs in our case better than statistical dependency measures is as follows. A usual collocation is often a new term formed out of existing words, so it is more specific, and of more restricted use, than each of the two words separately. However, in our case the idiomatic (IEXP) combinations are equivalent to functional words—prepositions—which are of much more frequent use than the corresponding NP used in its literal meaning. Also, it is suggested in [9] that frequent word chains of some specific POSs (such as P–NP–N) tend to be terminological; our study can be considered as a particular case of this method.

This P–NP–P construction seems to be common across various languages, mainly among Romance languages. The work of [16] is dedicated to Dutch where less quantity of P–NP–P (34% compared with Spanish) and uses were analyzed; so the same method could be used for other languages.

The obtained results could be used to create the input databases employed by tokenizer processors such as that of [8] where authors view IEXP identification as a tokenization ambiguity problem.

4.2 Differentiating Idiomatic Expressions

We consider heuristic or filters as a form to differentiate the idiomatic expressions from the free combinations and the fixed ones. We considered to identify the grammar categories of the IEXP neighbors, based on sequences of 5 and 6 contiguous words, obtained of the total text collection with the purpose of measuring the results for the variants use.

For the group “a fin de” we obtained in an automatic form the following groups:

1. a fin de V_inf	70.3%
2. a fin de Conj (que)	21.92%
3. a fin de S	
(*cuenta, cuentas,*	2.8%
año,	1.6%
abril, mayo, julio, agosto, noviembre	0.066%
mes,	0.69%
milenio,	0.1%
semana	0.044%
sexenio,	0.055%
siglo)	0.38%
4. a fin de Adv (*no*)	1.66%

Where: **Conj** means conjunction and **Adv** means adverb. Variants 1 and 4 can be grouped since the fourth is the denied version of the first variant. In the line of "año" (year) the corresponding Spanish noun phrases to the following noun phrases are considered: year, the year, this year, the present year. In the line of "siglo” (century) the noun phrases correspond to: century, the century, this century. In "sexenio" and "milenio" the noun phrases considered are: sexenio, the sexenio, millennium, the millennium. In the line of "mes” (month) the nouns are: month, this month.

Additionally three groups exist: 1) punctuation mark (0.32%) although immediately an infinitive verb appears, 2) number (0.022%) referred to a year: 1997, 3) adverb (0.011%) followed by an infinitive verb and 4) anomalous constructions (0.033%).

As it is reported in [3] idiomatic expressions combine with a more restricted number of neighboring words than free combinations. They computed three indices on the basis of three-fold hypothesis: a) idiomatic expressions should have few neighbors, b) idiomatic expressions should demonstrate low semantic proximity between the words composing them, and c) idiomatic expressions should demonstrate low semantic proximity between the expression and the preceding and subsequent segments

Considering the first hypothesis we observed the opposite effect: *a fin de* as IEXP combines with almost any infinitive verb, whereas *a fin de* as free combination combines with few neighbors: nominal groups with semantic mark of time (year, month). The explanation could be the one suggested in the previous section. Also *a fin de* combines with few neighbors to form the fixed phrases: *a fin de cuenta* y *a fin de cuentas* (literally: at end of account-s).

Considering the third hypothesis: the semantic proximity between the expression and its neighbors will be high if the expression has a literal meaning, and low if it is figurative. For the example, we found in group 3 the literal variants in the context of nouns with semantic mark of time since they are semantically near to *fin* (end). We considered a simple semantic proximity to the IEXP noun, using EurowordNet[1] and looking for similar terms (synonymous) in their glosses. The forms *a fin de cuenta* and *a fin de cuenta*s were analyzed in the same form satisfying the third hypothesis.

For *fin* (end) and *año* (year) the following nouns were found: "time" (*tiempo*), "period" (*periodo*), then they form an expression with literal meaning.

Data for *fin*

09169786n 11 conclusión_8 [99%] terminio_1 [99%] terminación_4 [99%] final_10 [99%] fin_4 [99%] finalización_6 [99%]	Tiempo en el que se acaba una cosa: "el final del año académico"; "la finalización del periodo de garantía"

Data for *año*

09125664n 4 año_1 [99%]	Tiempo que tarda un planeta en completar su vuelta alrededor del sol
09127492n 10 año_2 [99%]	Periodo de tiempo que comprende 365 días

But we did not find similar terms in the glosses for *cuenta* (account) and *fin* (end), then they form an expression with figurative meaning.

Data for *cuenta*

00380975n 7 numeración_1 [99%] enumeración_1 [99%] cómputo_1 [99%] cuenta_1 [99%] conteo_1 [99%]	Acción de contar
02130199n 1 cuenta_2 [99%]	Pequeña bola atravesada por un agujero
04270113n 0 cuenta_3 [99%] nota_4 [99%]	Factura o recibo en un restaurante
08749769n 0 cuenta_4 [99%]	Dinero que se debe
08318627n 3 cuenta_5 [99%] cómputo_4 [99%] recuento_2 [99%]	El número total contado: "recuento globular"

Future work will comprise acquiring the context of all IEXP groups and we will carry out the above process. In addition, we will try to distinguish the diverse IEXP from the P–NP–P with high frequency using the same method.

5 Conclusions

Idiomatic word combinations of IEXP type, usually functioning as compound prepositions, have statistical properties distinct from those of usual idiomatic collocations. In particular, they combine with a greater number of words than usual idioms. For

[1] http://nipadio.lsi.upc.edu/cgi-bin/wei4/public/wei.consult.perl

their unsupervised determination from a corpus, a simple frequency measure performs better than other statistical dependence measures. In particular, among most frequent P-NP-P word chains, about 50% are idiomatic. Inspection of the most frequent chains of this type permits to detect idiomatic combinations not present in existing dictionaries. We presented an approach to differentiate them from fixed phrases and free combinations by means of semantic proximity.

References

1. Banerjee, S. & Pedersen, T.: The Design, Implementation, and Use of the Ngram Statistic Package. Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, México (2003)
2. Church, K.W. and P. Hanks.: Word association norms, mutual information, and lexicography. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (1989) 76-83
3. Degand, Liesbeth & Bestgen, Yves.: Towards automatic retrieval of idioms in French newspaper corpora. Literary and Linguistic Computing (2003) 18 (3), 249-259
4. Diccionario de María Moliner.: Diccionario de Uso del Español. Primera edición versión electrónica (CD-ROM) Editorial Gredos, S. A. (1996)
5. Dunning, Ted.: Accurate methods for the statistics of surprise and coincidence" Computational Linguistics. (1993) 19(1):61-74
6. Evert, Stefan & Brigitte Krenn.: Methods for the Qualitative Evaluation of Lexical Association In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse, France. (2001) pp. 188-195
7. Galicia-Haro, S. N.: Using Electronic Texts for an Annotated Corpus Building. In: 4th Mexican International Conference on Computer Science, ENC-2003, Mexico, pp. 26–33.
8. Graña Gil, J., Barcala Rodríguez, F.M., Vilares Ferro, J.: Formal Methods of Tokenization for Part-of-Speech Tagging. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2002) Lecture Notes in Computer Science vol. 2276 Springer-Verlag (2002) pp. 240-249
9. Justeson, J. S., S. M. Katz. Technical Terminology: Some Linguistic properties and an algorithm for identification in text. Natural Language Engineering (1995) 1:9–27
10. Manning, C. D. & Schutze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts. (1999)
11. Mel'cuk, Igor.: Dependency Syntax: Theory and Practice, New York: State University of New York Press. (1988)
12. Nañez Fernández, Emilio.: Diccionario de construcciones sintácticas del español. Preposiciones. Madrid, España, Editorial de la Universidad Autónoma de Madrid. (1995)
13. Rayson, P., Berridge, D., Francis, B. Extending the Cochran rule for the comparison of word frequencies between corpora. In Vol. II of Purnelle G. et al. (eds.) Le poids des mots: Proc. of 7th International Conf. on Statistical analysis of textual data (JADT 2004), Presses universitaires de Louvain (2004) pp. 926–936.
14. Real Academia Española.: Diccionario de la Real Academia Española, 21 edición (CD-ROM), Espasa, Calpe (1995)
15. Seco, Manuel.: Gramática esencial del español, introducción al estudio de la lengua, Segunda edición revisada y aumentada, Madrid, Espasa Calpe. (1989)
16. Villada, Begoña & Bouma, Gosse. A corpus-based approach to the acquisition of collocational prepositional phrases. Proceedings of EURALEX 2002, Copenhagen. Denmark (2002) pp. 153–158