

Recognising Complex Prepositions Prep+N+Prep as Negative Patterns in Automatic Term Extraction from Texts

CLAUDIA OLIVEIRA¹
MILENA GARRÃO²
LUIZ ALEXANDRE AMARAL³

¹ DEPARTAMENTO DE ENGENHARIA DE SISTEMAS, INSTITUTO MILITAR DE ENGENHARIA,
RIO DE JANEIRO, BRAZIL, CMARIA@DE9.IME.EB.BR

² DEPARTAMENTO DE LETRAS, PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO,
RIO DE JANEIRO, BRAZIL, MIGARRAO@UOL.COM.BR[†]

³ THE DEPARTMENT OF SPANISH AND PORTUGUESE, OHIO STATE UNIVERSITY,
COLUMBUS, OHIO, USA, AMARAL.1@OSU.EDU

Abstract. This work is a study of the delimitation of complex prepositions (CP) as lexical units, items of a computational lexicon that includes compounds and phrases. In addition, given the utmost importance of spotting noun phrases (NP) in document retrieval systems, parsing prepositional structures such as “*Prep₁ N Prep₂ X*” prevents the fragment “*N Prep₂ X*” from being detected as a noun phrase, i.e. the prepositional structure is a negative pattern to be used in the extraction of noun phrases from texts. We analyse the distinct attempts at the description of the class of CPs, in order to establish criteria that could be formalised and therefore made available in a computer system.

1. Introduction

Viewing the lexicon as a computational resource in a Natural Language Processing (NLP) system, more specifically, a text processing system, the motivation for this work is twofold. First, we address the delimitation of complex prepositions (CP) as lexical units, items of a computational lexicon. By and large, electronic dictionaries are word class-based. It is highly desirable that they also include compounds and phrases. In addition, given the utmost importance of spotting noun phrases (NP) in document retrieval systems, parsing prepositional structures such as “*Prep₁ N Prep₂ X*” prevents the fragment “*N Prep₂ X*” from being detected as a noun phrase, i.e. the prepositional structure is a negative pattern to be used in the extraction of noun phrases from texts.

Due to the steep increase in the amount of digital documents, in digital libraries as well as on the Internet, efficient access to documents became a key issue in the field of Information Retrieval (IR). The problems regarding search precision are related to indexing techniques (such as descriptors extraction and document structuring) and retrieval techniques. Text descriptors are represented by terms, which constitute the keys to document access and could be considered as linguistic equivalents of concepts. Spotting and extracting terms should be a highly valued task in text processing systems since the construction of the terminology pertaining to a domain of knowledge clearly improves search results as well as enables the categorization of documents.

According to Jacquemin em (Jacquemin, 2001), terms are predominantly NP in technical-scientific text bases, justifying the development and refinement of extraction

techniques for this type of expression. Spotting simple NPs (made up of a single word) consists of picking out the items, which are delimited by blanks or punctuation marks. The process may be followed by morphological treatment, frequently involving inflexional morphemes and derivational suffix stripping. The goal is to unify in the same term different words having an identical root. With respect to larger NPs, failure to recognise their syntactic structure hinders the system in several aspects. The most evident is the exclusion of complex NPs, which might have great impact on precision. For instance, the expression “stock exchange” will yield a better search result than separate items “stock” and “exchange”, in terms of a larger proportion of documents more closely related to the topic.

Another consequence of treating single words only, is the risk of treating as terms certain words or phrases which no longer hold their usual meaning, but are constituents of larger expressions. For instance, from the expression “in view of”, the term “view” should not be extracted as such. To address this problem it is necessary to establish negative patterns, i.e. to characterise the “non-NPs”.

In Portuguese and other romance languages, a type of expression that frequently contains nouns is the complex preposition (CP). Constructions such as “em virtude de”, “em face de” and “em favor de” (which could be translated as “in virtue of”, “in face of” and “in favor of”, respectively) are cases in which the noun that constitutes the CP lost its individual meaning. However, the automatic treatment of texts cannot rely on a simple list of CPs for two reasons. First, we believe that this set is an open class, particularly the CPs with a “*Prep₁ N Prep₂*” structure. Moreover, often a sequence of words that might be construed as a CP occurs, but parsing reveals the impossibility of such analysis. Such is the case of structures “*Prep₁ N₁ Prep₂ N₂*” where “*N₁ Prep₂ N₂*” is a NP.

The purpose of this work is to analyse the distinct attempts at the description of the class of CPs, in order to establish criteria that could be formalised and therefore made available in a computer system. With respect to the implementation of procedures to recognise CPs, we propose that some of them be classified as lexical items and thus included in the lexicon. Others, though, can only be treated syntactically or semantically. It is our aim that the criteria can be encoded by means of parsing rules, such that the computational procedure approximates the linguistic analysis of the phenomenon, particularly in non-obvious cases.

The remainder of this paper is organised as follows. In section 2 we review the grammatical notion of compound expressions, emphasising the linguistic concept of CP. The linguistic approaches to the characterisation of this class of expressions serve as a counterpoint to the computational view, which we call the operational definition of CP and present in section 3, by means of an objective set of recognition criteria. In section 4 we describe the data that have been used to test the operational criteria and discuss some of the most peculiar findings in the corpus. Final comments and conclusions are presented in section 5.

2. Brief overview of the concept of CP

2.1 Traditional Grammar

Prescriptive grammarians of the Portuguese language have not treated the concept of CP systematically. They resort to using lists to describe the phenomenon that is not restricted to such a simple formal representation.

The universally accepted list of simple prepositions, also called the list of essential prepositions, is easy to characterise, because it is a finite set – *a, até, após, contra, para, por, de, desde, ante, perante, trás, sob, sobre, com, em, entre, sem*. Listing CPs, however, seems to generate at least two immediate problems that are clearly related. One, of a practical nature, reveals the discrepancies between the listings of different grammarians. The other, of a more theoretical nature, confirms the position that CPs constitute an open class.

It is important to highlight that most grammarians do not refer to the phenomenon as *complex prepositions* per se but as *prepositional locutions*. A locution could be defined as a combination of two or more words with a specific semantic function. For the sake of terminology coherence we will opt for the term *complex prepositions* (CP). We will return to this terminology issue in the next section.

The following table presents some grammarians considerations regarding Portuguese CPs (Dias, 1995:194):

Author	Definition	Some Examples
(Said Ali, 1969:101-2)	It is a preposition variation; it could be a single word or a combination of words. In the latter case it is called a prepositional locution (CP). It is normally made up of an adverb or an adverbial locution combined with the word <i>de</i> (more rarely with <i>a</i> or <i>com</i>).	<i>em conformidade com;</i> <i>de acordo com;</i> <i>em harmonia com</i>
(Bechara, 1980:155) (Bechara, 1999:301)	This grammarian adopts a similar definition to Said Ali: he claims the CP is made up of an adverb or adverbial locution followed by the preposition <i>de, a</i> or <i>com</i> .	<i>atrás de;</i> <i>em frente a;</i> <i>por causa de;</i> <i>de acordo com</i>
(Cunha and Cintra, 1976:510)	It is a combination of two or more words, in which the latter is a simple preposition (usually <i>de</i>).	<i>a respeito de;</i> <i>em lugar de</i>
(Gama Kury et alii, 1981:201)	The CP is usually made up of an adverb combined with a preposition. It semantically stands as a preposition.	<i>de acordo com;</i> <i>por causa de</i>

These definitions agree upon two main aspects. Under the formal aspect, they all present a preposition as the last element of what is considered a CP. Under a functional aspect, they claim the CP is applied as a preposition.

However, the adverb or adverbial locution presence, considered a CP defining factor according to Bechara, is a drawback to the class as a whole, since it would exclude expressions such as **por causa de**, **em virtude de** (made up of nouns) and **a partir de** (made up of a verb). Therefore, it seems to be clear the tension between the CPs' concept and their examples, since some of the listed instances clash with the phenomenon's theoretical explanation.

Thus, one may say that the definitions do not exhaustively comprehend the phenomenon, not only because they fail to provide a consensual and definitive criterion to it but mainly because they fail to present an operational criterion to it.

2.2- Other views of the phenomenon: the functional criterion hegemony

Brazilian grammarian Mattoso Câmara Jr (Camara Jr, 1984) considers a locution as “a conjunction of two words which maintain their phonetic and morphemic individuality, but make up a signifying unit for a specific function”. The locutional character of an expression relies on the fact that it is a signifying block with a given role. He emphasizes that the nouns forming the prepositional locution or, as we mentioned in the previous section, the CPs (such as **em virtude de**) have already undergone grammaticalization; in other words, they have been through a “process that consists on turning simple, lexical semantically full words into grammatical words”. What should be considered, therefore, is the signifying block and not the meaning of each one of the items belonging to the CP.

Although this fact seems trivial at first, it is extremely relevant for a CP classification criterion. Actually, the presence of verbs and nouns in the CPs does not compromise their adverbial character. The signifying block meaning does not refer to the noun or the verb, but to the circumstance, as in *em frente de* (“in front of”, an idea of space) and *devido a* (“due to”, an idea of cause).

Bomfim (Bomfim, 1999), in a study focusing the delimitation of lexical unit, highlights the little attention given to CPs and proposes an analysis’ strategy which should take into consideration:

- a) the signifying content, the morphological properties and the syntactic behaviour of the elements that make up the phenomenon;
- b) the level of internal cohesion of the entire complex and the grammaticalization level of its parts.

Dias (Dias, 2002) comes to a conclusion that the CPs are a subgroup of prepositions, for they present more similarities than differences when compared to simple prepositions. “Their semantic content is the same, they have similar functional role and similar complements”[...]“in order to better evaluate the classification of a CP, we must consider its function, which is to introduce prepositional phrases, specially those which convey circumstance. Such role has, in fact, a simple preposition function [...] CPs are considered as unfolded prepositions, carrying the same syntactic role (heading prepositional phrases) and discourse function (connectors), being the two groups interchangeable”.

The fact that Quirk et al. (Quirk,1978) adopts the terms *simple* and *complex prepositions* for the English language confirms the level of generalization regarding this functional perspective. Likewise, Perini (Perini, 1996) considers a CP as a single lexical item, though it is formally a complex sequence.

Summing up, the terminology chosen to consider the phenomenon is not the defining factor for its delimitation; one may call it a CP, a composed preposition or even a prepositional locution. However, by considering it as a subgroup of simple preposition, it is implied that: there is only one class for what otherwise was thought of as two; the syntactic-semantic aspect or the functional role is hegemonic.

Therefore, the focus on the phenomenon delimitation is the prepositional function and no longer its internal components arrangement. This way, it is considered the

productive character of the CP instead of its structure description, what seems to be a very pertinent solution if we take into account the difficulty in formalizing the phenomenon.

3. Operational definition of CP

The main focus of our investigation is the formulation of systematic criteria for recognizing CPs. Considering that the class of CPs is open and productive in Portuguese, the task of characterising it goes beyond the trivial enumeration of expressions. It is important to keep in mind that the resulting criteria is to be employed in spotting multi-lexemic prepositions in a sequence of words that includes a noun, in order to rule out the detection of noun phrases containing that noun. In other words, we are interested in distinguishing sequences introduced by a CP followed by a NP, with the structure

i. *CP(Prep₁ N₁ Prep₂) NP*

from prepositional phrases with the structure

ii. *Prep₁ NP(N₁ Prep₂ NP)*

The speaker's choice between i. and ii. is based on the underlying principle of semantic incompatibility possibly found in the sequence *NP(N₁ Prep₂ NP)*. It is very difficult to judge whether *N₁ Prep₂ NP* works as a phrase based on singled out semantic features. Mostly it is just an oddity in the sequence that makes the speaker dismiss it. For example, in

1. Por isso, em vez da compreensão, eu procuro estimular a sensação do público **em torno de** um ritual, que celebra um momento de harmonia entre homens e anjos.

the choices are

iii. *CP(Prep(em) N(torno) Prep(de)) NP(um ritual)*

which translates as “Therefore, instead of understanding, I try to encourage the sensation of the public **around a ritual** that celebrates a moment of harmony between men and angels”, or

iv. *Prep(em) NP(N(torno) Prep(de) NP(um ritual))*

which translates as “[...] I try to encourage the sensation of the public **in lathe of a ritual** that celebrates [...]”. For semantical reasons, it is very hard to construe a meaning for the noun phrase **torno de um ritual** (lathe of a ritual), therefore parsing of structure iii. (around a ritual) is strongly favoured. In many instances, the semantic incompatibility occurs with respect to the whole sentence. In the following example, if preference is given to ii. it is impossible to construe the meaning of the sentence

2. O preço do café no mercado interno deve ficar estável até dezembro **em [função de um acordo]** firmado entre governo, produtores e industriais do setor. (“The price of coffee in the internal market should be stable until December **in [function of a deal]** between government, producers and industry”).

Again, the only alternative is to consider the sequence **em função de** (due to) as a CP.

The above examples illustrate how hard it is to give a computational treatment to the problem based on the speaker's intuitive mechanisms. Instead, we are looking for systematic tests that yield the same results.

The order in which the criteria are presented is not incidental, but rather it reflects the decisiveness of the corresponding testing mechanism in spotting the CP. On the

other hand, it is not the case that a single test, or even the combination of all the tests, will result in a foolproof interpretation of the expression. We apply them to gather evidence in favour of structure i. or ii. At the same time, it is possible that a CP will test positive to some criteria and not to others. In summary, the tests or any groups of tests, are neither necessary nor sufficient to categorically determine the interpretation of a given expression.

3.1. Criteria for the recognition of CPs

Criterion 0: A priori lexicalization.

This criterion consists in the recognition of the CP as an unambiguously frozen expression, which means that the sequence of lexemes in question is always interpreted as a CP. The most decisive test for frozen CPs is the inexistence of the noun in any other context in the language, for example **em prol de** and **em cima de**. Even though the nouns **prol** and **cima** still deserve an entry in most major dictionaries, they are not found in corpora of current spoken or written Brazilian Portuguese. The corpus produced the following example.

3. Um crítico compreensivo haverá de relaxar um pouco a sua severidade **em prol de** uma candidata mulher, cujos erros e acertos são os da cabeça mas não os do coração .

According to this criterion, the CP can be unambiguously recognised as a frozen expression, precluding the need for further testing. It precedes any other criterion for it is the most decisive and the cheapest, from the computational standpoint. The delimitation of a frozen expression presumes the non-compositionality between the individual words; therefore criterion 0 does not require syntactic parsing. The most appropriate way to implement it is the inclusion of the expression in the electronic lexicon.

Criterion 1: Substitution.

This criterion is based on the notion presented by some grammarians that a CP can normally be substituted by a simple preposition or by another CP. In example 4, the sequence [**em virtude de**] (“in virtue of”) can be replaced by the CP [**por causa de**] (“because”). In example 5, the sequence [**em companhia de**] (“in the company of”) can be replaced by the preposition [**com**] (“with”).

4. Senna morreu **em virtude de** uma falha mecânica da Williams, «potencializada» pelas falhas encontradas pela polícia na pista de Ímola que até agora está fechada, segundo a Justiça italiana «em processo de seqüestro». (**por causa de**)
5. Ele vai passar os próximos meses **em companhia de** outros dois cosmonautas que chegaram à Mir no mês passado. (**com**)

Together with a list of simple prepositions the lexicon must contain the list of CPs that will be considered as standard, which will be used to replace the tested expression.

This test is very attractive at first glance, but presents a few problems when it comes to its implementation. We encountered several examples in the corpus for which we could not find a suitable substitution, such as 6.

6. Só é legal a doação feita **em troca de** bônus no valor correspondente.

Note that the CP [**em troca de**] (“in exchange for”) is listed in section 4 a borderline case, largely due to its failure with regard to criterion 1, even though we recognise the prepositional value of the whole expression. It is very difficult to create a list of standard CPs, which would cover all possible prepositional meanings, to ensure the completeness of this criterion.

Criterion 2: Valency of the preceding verb.

This criterion uses a very important feature of the syntactic structure of the sentence. If the preposition *Prep*₁ in a sequence “*V Prep*₁ *N Prep*₂ *X*” is part of the valency of the verb *V* then “*Prep*₁ *N Prep*₂ *X*” is the complement in the verb phrase and consequently “*Prep*₁ *N Prep*₂” is not a CP.

The following examples make this statement clearer. In 7, the preposition *em* (“in”) is the head of an essential complement of the verb *aplicar* (“to apply”) therefore the sequence “*em processo de execução*” is interpreted as *PP(em SN(em processo de execução))*.

7. Esta multa só se [**aplica em**] processo de execução; não cabe em procedimento de jurisdição voluntária, como é o caso da alienação de coisa comum (JTJ 151/90).

On the other hand, in example 8, a similar analysis is not possible. In this case, the sequence “*em processo de*” is clearly a CP, which shows semantic non-compositionality and which can be substituted by the simple preposition *em* (criterion 1).

8. Algumas empresas [**em processo de**] finalização de balanço anual estão remetendo recursos para o exterior para fugir de obrigações fiscais.

Criterion 3: Insertion of a determinant

This criterion consists of analysing the consequences of inserting a determinant into the sequence [*Prep*₁ *N Prep*₂] to obtain [*Prep*₁ *Det N Prep*₂]. The idea is to break the integrity of the expression, so that a CP, as a unit, should not allow such insertion. Sometimes the insertion is simply impossible, as in the case of the definite article **o**, in the contraction [**em + o = no**], in example 9.

9. Dizem que o Sr. articula a saída de Bisol **em favor de** Roberto Freire.
*Dizem que o Sr. articula a saída de Bisol **no favor de** Roberto Freire.

In other cases we verified a significant semantical change in the resulting expression, as in the case of the definite article **a**, in the contraction [**em + a = na**], in example 10.

10. A maioria dos refêns foi libertada na quinta-feira **em troca de** armas e drogas.
A maioria dos refêns foi libertada na quinta-feira **na troca de** armas e drogas.

There are also cases in which the semantical impact of the determinant is very slight, as in 11, which lead us to recommend the use of this criterion to corroborate others, rather than to be used on its own.

11. Os restos de folículos produzem a progesterona, hormônio que, juntamente com os estrógenos, prepara a parede do útero para receber o embrião **em caso de** gravidez.

Os restos de folículos produzem a progesterona, hormônio que, juntamente com os estrógenos, prepara a parede do útero para receber o embrião **no caso de** gravidez.

A variation of the insertion of a determinant is the inflexion of the noun in the sequence [*Prep*₁ *N* *Prep*₂]. In example 12 such transformation would entail not only the plural, but also a change in the meaning and the parsing of the sentence, as **em forma de** is recognised as a CP. On the other hand, example 13 shows a simple plural inflexion of **forma de vida**.

12. Estes elementos existem apenas **em forma de** átomos separados.

13. Cerca de 57% de todas as florestas tropicais, ambientes mais diversos **em formas de** vida em todo o mundo, estão representados na região Neotropical.

4. Analysis of the data

The criteria outlined in the preceding section has been tested in a Brazilian Portuguese corpus, compiled by NILC (Núcleo Interinstitucional de Linguística Computacional) and described in (Aires, 2001), which contains around 37 million words, including corrected, uncorrected and semi-corrected texts. The texts classified as corrected (around 33 million words) were obtained from publications such as books, newspapers and magazines, which are supposedly revised by language specialists. The genre of the texts are diverse, from Brazilian literary prose, to school texts in sciences, history, to encyclopaedia entries. The uncorrected texts (736,000 words) include original essays from secondary schools pupils and undergraduates, as well as publicity material. The semi-corrected texts (1,390,000 words) are taken from contracts, reports, academic dissertations, etc. Access to the corpus is available on the Internet, from the site of the Portuguese project *Linguateca*, where the concordance queries are made through a formal language of pattern specification.

Initially we concentrated our search on the pattern *em N de*, which seems to be the most productive in Brazilian Portuguese. From the annotated NILC corpus there have been 20,229 matches of this pattern, from which 8,000 have been randomly chosen.

Wed May 1 20:53:45 CEST 2002
Procura: "em" [pos="N"] "de". Pedido de uma concordância em contexto Corpus: NILC/São Carlos anotado v. 3.2
Concordância Procura: "em" [pos="N"] "de". Apresenta-se uma amostra aleatória de 8000 das 20229 ocorrências encontradas.

Figure 1: Header of the corpus query result

Other queries, in which we used variants of the pattern *em N de*, such as *na(s) N de*, *no(s) N de*, *num(a)(s) N de*, etc, did not result in significant additions to the list of CPs collected in the basic query that illustrated in figure 1.

4.1. The CPs in the corpus

In order to present the list of CPs that were found in the corpus, we organized table 1 containing three groups of expressions. The first group consists of expressions that clearly satisfy the CP delimitation criteria presented in the previous section. The second group consists of borderline cases, those which satisfy some of the criteria and whose classification as CP should be discussed further. The third group consists of very frequent expressions, composed of words that frequently co-occur.

Clear cases:	em torno de, em vez de, em virtude de, em função de, em cima de, em volta de, em matéria de, em caso de, em nome de, em razão de, em roda de, em meados de, em face de, em prol de, em favor de, em consequência de, em benefício de, em vias de (em via de), em detrimento de, em poder de, em decorrência de, em nível de, em frente de, em mãos de (em mão de), em lugar de, em proveito de
Borderline cases:	em condições de, em forma de (em formato de), em processo de, em fase de, em estágio de, em troca de, em busca de, em sinal de, em regime de, em defesa de, em socorro da, em companhia de, em prejuízo de
Compositional expressions:	em fim de, em tom de, em estado de, em início de, em média de, em princípios de, em clima de, em ponto de, em esquema de, em compasso de, em questão de, em caráter de

Figure 2: Frequent CPs and compositional expressions

The reason for including this last group is to point out that, even though we recognise that frequency is an influential factor in crystallization and grammaticalization, we could not observe these processes acting in the expressions of this group. In fact, we can say that they do not meet any of the delimitation criteria and, furthermore, they maintain a distinct compositional relationship between their elementary words, as shown in example 14.

14. A Vigilância deve aumentar o turno de trabalho, **em esquema de** plantões, durante estes dez dias, para encerrar as análises a tempo.

Some sequences *em N de* occur both as CPs and as part of a larger phrase. In examples 15 and 16, [**em lugar de**] is clearly a CP, which can be substituted by [**ao invés de**] (“instead of”), for instance.

15. As decisões eram tomadas por consenso, **em lugar de** unilateralmente.

16. Algumas empresas, **em lugar de** limitar o valor do reembolso por ato assistencial prestado, adotam um valor anual por empregado, geralmente relacionado com o valor do salário.

On the other hand, in example 17 the phrase [**em lugar de Axel**] can be interpreted both as *CP(em lugar de) NP(Axel)*, translated as “instead of Axel”, or *PP(em NP(lugar de Axel))*, translated as “in Axel’s place”.

17. No segundo tempo, Telê pôs André **em lugar de** Axel, passando Leonardo para o meio.

Similarly, with pattern *por N de*, we recognise a CP [**por força de**], substitutable by [**por causa de**] in example 18. On the other hand, in example 19 there is only one possible interpretation: *PP(por NP(força de trabalho))*.

18. Naquele ano, o Sindicato Solidariedade seria reconhecido **por força de** uma greve geral que abalou o governo comunista.

19. Nos Estados Unidos, segundo o documento, «o crescimento do setor de serviços, dos empregos de meio expediente e a busca **por força de** trabalho mais flexível contribuíram para a expansão do mercado de trabalho infantil».

5. Concluding Remarks

This work began mostly based on a corpus research, where the large variation of sequences *em N de* presenting a prepositional role convinced us that CP is actually an open class and, therefore, it should be formally modelled in automatic text processing systems.

The criteria detailed from the corpus occurrences shall be computationally implemented when formalized. We consider that the next step in the research should be attempting to formally specify the criteria, what will result in an approximation of the presented ideas. The biggest difficulties will be found in the cases where the semantic factors are the most defining ones.

References

1. Aires, R.V.X., Aluísio, S.M, Criação de um corpus com 1.000.000 de palavras etiquetado morfossintaticamente. *Relatórios do NILC*, NILC-TR-01-8, 2001.
2. Said Ali M. *Gramática Secundária da Língua Portuguesa*. Melhoramentos, São Paulo, 1969.
3. Bechara, E., *Moderna Gramática Portuguesa*. Cia. Ed. Nacional, São Paulo, 1980.
4. Bechara, E., *Moderna Gramática Portuguesa*. Ed. Lucerna, Rio de Janeiro, 1999.
5. Bonfim, E., Advérbios, preposições ou conjunções? Fronteiras entre classes de palavras. *Aulas de Português, Perspectivas Inovadoras*, 137-150. André Valente (org.). Editora Vozes, Petrópolis, 1999.
6. Camara Jr., J. M., *Dicionário de Lingüística e Gramática*. Editora Vozes, 1984.
7. Cunha, C. and Cintra, L., *Nova Gramática do Português Contemporâneo*. Nova Fronteira, Rio de Janeiro, 1985.
8. Dias, M. C., Revendo as Locuções Prepositivas. *Flores Verbais*, 193-206. Jürgen Heye (org.). Editora 34, Rio de Janeiro, 1995.
9. Dias, M. C., Locução para quê? *Revista Veredas*, 2002.
10. Jacquemin, C., *Spotting and Discovering Terms Through Natural Language Processing*. The MIT Press, 2001.
11. Kury, A. G. et al., *Gramática Objetiva da Língua Portuguesa*. Editora Rio, Rio de Janeiro, 1981.
12. Perini, M., *Gramática Descritiva do Português*. Editora Ática, São Paulo, 1996.
13. Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J., *A Grammar of Contemporary English*. Longman Group Limited, 7^a. edition, 1978.