Locution or collocation: comparing linguistic and statistical methods for recognising complex prepositions

Claudia Oliveira[1], **Cícero Nogueira**[1], **Milena Garrão**[2]

[1]Departamento de Engenharia de Sistemas
Instituto Militar de Engenharia, Rio de Janeiro

[2]Departamento de Letras
Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro

{cmaria, nogueira}@de9.ime.eb.br, migarrao@uol.com.br

Abstract. *Multi–Word Expressions (MWE) include a large range of linguistic phenomena, such as nominal compounds and institutionalized phrases. These expressions, which can be syntactically and/or semantically idiosyncratic in nature, are frequently used in everyday language as well as in formal contexts. In this work we investigate a type of MWE, the complex preposition (CP). The purpose is to establish the relationship between the notions of locution - the linguistic view of CPs - and collocation - the statistical view - when we look into the corpus, and to consider how these notions can be applied to the delimitation of CPs.*

Resumo. *Expressões Multivocabulares (EMV) englobam uma grande coleção de fenômenos lingüísticos, tais como compostos nominais e frases institucionalizadas. Essas expressões, que podem ser sintaticamente e/ou semanticamente idiossincráticas, são freqüentes na linguagem oral assim como em contextos formais. Neste trabalho nós investigamos um tipo de EMV, a locução prepositiva (CP). O objetivo é estabelecer o relacionamento entre as noções de locução - a visão lingüística de CPs - e colocação - a visão estatística - quando analisamos o corpus, e consideramos como essas noções podem ser usadas na delimitação de CPs.*

1. Introduction

In recent years, there has been a growing interest in the issues involved in dealing with Multi–Word Expressions (MWEs) in most areas of Natural Language Processing (NLP). MWEs include a large range of linguistic phenomena, such as nominal compounds (e.g. “table cloth”), and institutionalized phrases (e.g. “fish and chips”). These expressions, which can be syntactically and/or semantically idiosyncratic in nature, are very frequent in everyday language as well as in formal contexts. Applications that require some degree of semantic interpretation (e.g. machine translation, question-answering, summarisation, generation) and require tasks such as parsing and word sense disambiguation are particularly sensitive to MWEs’ delimitation problems.

Brazilian grammarian Mattoso Câmara Jr [Mattoso Câmara Jr, 1984] considers a locution to be “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 and with a distinguished part of speech. He emphasizes that the nouns forming the prepositional locution have already undergone grammaticalisation; in other words, they have been through a "process that consists of turning simple, lexical semantically full words into grammatical words". What should be considered, therefore, is the signifying block and not the meaning of each of the items belonging to the locution. Even though it is not explicitly discriminated in [Mattoso Câmara Jr, 1984], we notice that locution is mostly used with respect to functional classes such as adverbial, conjunctional or prepositional locutions. In a descriptive grammar one would find the following definition for an adverbial locution: "... two or more words working as an adverb".

Collocations, on the other hand, are defined in terms of frequency of co–occurrence. Manning and Schütze [Manning and Schütze, 1999] define collocations as "two or more words that correspond to some conventional way of saying things", which highlights the habitual place of a word in relation to another.

There are some important common points between the notions of locution and collocation. First of all, both present limited compositionality of meaning. This is the most distinguishable feature of MWEs. Another characteristic of MWEs present in both notions are non–substitutability of its components by near synonyms and non–modifiability of the phrase as a whole, for instance the impossibility of insertion of other lexical items.

The purpose of this work is to establish the relationship between the notions of locution and collocation when we look into the corpus, and to consider how they can be applied to the delimitation of Complex Prepositions (CPs). There are some recent works with similar aims and different target languages. In [Trawinski, 2003], the focus is on the representation of the syntax of German CPs in HPSG. More closely related to our work is [Bouma and Villada, 2002], in which a list of Dutch collocations of the form "Prep_1 NP Prep_2" is extracted from a corpus and analysed by human judges to determine which ones are CPs and therefore to establish the effectiveness of statistical methods.

The remainder of this paper is organised as follows. In section 2 we review the grammatical notion of CPs and a summary of the operational criteria for the delimitation of CPs. In section 3 we describe the statistical approaches to extracting collocations from a corpus. In section 4 we describe the data and the statistical experiments that we carried out, comparing the results with a list of well established CPs. Final comments and conclusions are presented in section 5.

2. Complex prepositions: linguistic facts

A Complex Preposition is a type of MWE that functions as one preposition. In Portuguese and other romance languages, a CP can have a variety of internal structures, such as: "Adverb Prep" (**dentro de**), "Prep Prep" (**por sobre**), "Prep Prep Prep" (**por trás de**), "Prep V Prep (**a partir de**) and "Prep N Prep" (**de acordo com**). We address the delimitation of the latter type of CPs for two reasons. First, because they are more numerous and more frequent. Secondly, given the utmost importance of spotting noun phrases (NP) in text systems, parsing prepositional structures such as "Prep_1 N Prep_2 X" prevents the fragment "N Prep_2 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.

Prescriptive grammarians of the Portuguese language have not treated the concept of CPs 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.

The grammarians' 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 that the CP is applied as a preposition. One may say, however, that the definitions do not exhaustively describe the phenomenon because they fail to provide consensual and operational criteria to identify the CP.

From a functional perspective, Dias [Dias, 2002] comes to a conclusion that the CPs are a subgroup of prepositions, since they present more similarities than differences when compared to simple prepositions. She considers a CP to be an unfolded preposition, carrying the same syntactic role (i.e. heading prepositional phrases) and the same discourse function (i.e. connectors). The fact that Quirk et al. [Quirk et al., 1978] adopts the terms simple and complex prepositions for the English language confirms the level of generalization regarding this functional perspective.

One 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–word 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(Prep*$_1$ *N*$_1$ *Prep*$_2$*) NP* from prepositional phrases with the structure ii. *Prep*$_1$ *NP(N*$_1$ *Prep*$_2$ *NP)*.

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.

Criterion 0: A priori lexicalisation 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**. 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 a computational standpoint, since it does not require syntactic parsing.

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. For instance, the sequence [**em virtude de**] ("in virtue of") in 2.1 can be replaced by the CP [**por causa de**] ("because"). In example 2.2, the sequence [**em companhia de**] ("in the company of") can be replaced by the preposition [**com**] ("with").

2.1 *Senna morreu* **em virtude de** *uma falha mecânica da Williams ... (por causa de)*

2.2 *Ele vai passar os próximos meses* **em companhia de** *outros dois cosmonautas que chegaram à Mir no mês passado. (com)*

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 in 2.3.

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

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 2.4, 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))".

2.4 *Esta multa só se [aplica em] processo de execução; não cabe em procedimento de jurisdição voluntária (JTJ 151/90).*

On the other hand, in example 2.5, 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** ("in") (criterion 1).

2.5 *Algumas empresas [***em processo de***] finalização de balanço anual estõ 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 2.6.

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

In other cases we verified a significant semantic change in the resulting expression, as in the case of the definite article **a**, in the contraction [**em + a = na**], in example 2.7.

2.7 *A maioria dos reféns foi libertada na quinta-feira* **em troca de/ na troca de** *armas e drogas.*

There are also cases in which the semantic impact of the determinant is very slight, as in 2.8, which leads us to recommend the use of this criterion to corroborate others, rather than to be used on its own.

2.8 *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/ 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 2.9 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 2.10 shows a simple plural inflexion of **em forma de**.

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

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

This criterion could be generalised to cover the insertion of any lexical material such as pronouns or adjectives.

3. Collocations: statistical tests

The most straightforward method for finding collocations in a corpus is computing the frequency of word pairings (bigrams). Frequent bigrams have a good chance of being collocations. The problem with this approach is that the most frequent words of the language will tend to combine more. In our particular case, the part–of–speech (POS) pattern of the expressions we are investigating contains at least two function words: the prepositions. They are so frequent that their combination with frequent nouns will be wrongly analysed as collocations. The statistical tests we selected balance the effect of individual frequencies, and measure whether a sequence of words occurs more often than would be expected on the basis of individual word frequencies. Therefore, these tests are often used to determine whether two co-occurring words are potential collocations.

Log-likelihood score. A typical problem of statistics is determining whether something is a chance event or linked to another. We want to know whether a bigram is a collocation or whether it appears together by chance. This type of hypothesis testing requires two hypotheses to be formulated (see [Manning and Schütze, 1999]):

$$H_1 : P(w_1|w_2) = P(w_1|\neg w_2)$$
$$H_2 : P(w_1|w_2) \neq P(w_1|\neg w_2)$$

H_1 assumes that the two words are independent and H_2 assumes otherwise. H_2 is the collocation hypothesis and the log-likelihood test measures how much more likely H_2 is than H_1.

Pearson's χ^2 test. The χ^2 test computes the observed frequencies of the following bigrams: w_1w_2, $\neg w_1w_2$, $w_1\neg w_2$ and $\neg w_1\neg w_2$. If the differences between observed frequencies and expected frequencies for independence are large, then the bigram w_1w_2 is a possible collocation. This computation is done by (see [Manning and Schütze, 1999]):

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where $i \in \{w_1, \neg w_1\}$, $j \in \{w_2, \neg w_2\}$, O_{ij} is the observed frequency of bigram ij and E_{ij} is the expected frequency.

Mutual Information. This test compares the probability of observing two words $w_1 w_2$ together with the probabilities of observing them independently in a given corpus, computed by (see [Manning and Schütze, 1999]):

$$I(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

The collocation tests we selected take bigrams as inputs but the expressions we are looking at consist of 3 or 4 words. In order to apply the bigram tests to our data-set, we transformed the $w_1\ w_2\ w_3\ w_4$ strings into $w_1\ w_2\_w_3\_w_4$ or $w_1\_w_2\_w_3\ w_4$ strings. In our pattern "Prep$_1$ Det N Prep$_2$", we assume that "Det N Prep$_2$" can be seen as a unit or alternatively that "Prep$_1$ Det N" can be seen as a unit.

4. Working with the corpus

We used the corpus CETENFolha (Corpus de Extractos de Textos Eletrônicos NILC/Folha de São Paulo), containing around 24 million words in Brazilian Portuguese, built by the project Computational Processing of Portuguese from the texts of Folha de S. Paulo belonging to the corpus NILC/São Carlos, compiled by Núcleo Interinstitucional de Lingüística Computacional (NILC) [CETENFolha, 2004].

Muitas	[muito] <quant> DET F P @SUBJ>
de	[de] <sam-> PRP @N<
as	[o] <-sam> <artd> DET F P @>N
prioridades	[prioridade] N F P @P<
de	[de] <sam-> PRP @N<
o	[o] <-sam> <artd> DET M S @>N
novo	[novo] ADJ M S @>N
governo	[governo] N M S @P<
coincidem	[coincidir] <fmc> V PR 3P IND VFIN @FMV
com	[com] PRP @<PIV
as	[o] <artd> DET F P @>N
prioridades	[prioridade] N F P @P<
de	[de] <sam-> PRP @N<
o	[o] <-sam> <artd> DET M S @>N
PT	[PT] PROP M S @P<

Table 1: Extract from the corpus CETENFolha

4.1. Extracting the pattern

We extracted the instances of the pattern "PRP (DET)? N PRP" (coded in the CETENFolha tag-set) where "DET" is optional, and arranged the resulting expressions in the format required by the statistical package NSP [Banerjee and Pedersen, 2003].

em_PRP os_DET tempos_N em_PRP
de_PRP as_DET prioridades_N de_PRP
com_PRP as_DET prioridades_N de_PRP
para_PRP a_DET realização_N de_PRP
sob_PRP o_DET comando_N de_PRP
em_PRP a_DET volta_N de_PRP
de_PRP lista_N em_PRP
para_PRP o_DET sucesso_N de_PRP
de_PRP os_DET recursos_N para_PRP
de_PRP o_DET orçamento_N de_PRP

Table 2: The first 10 candidate expressions extracted from the corpus

In the corpus, each line contains a word followed by its morphosyntatic attributes. We extracted sequences of words with POS tags "PRP N PRP" or "PRP DET N PRP". Table 1 shows an extract of the corpus CETENFolha, with the target pattern highlighted.

The extracting program takes as input the corpus and a list of stopwords. This list contains nouns which are not to be allowed as part of CPs, such as months, days of the week and longhand numbers. Upper case nouns and nouns containing numerical digits are also eliminated, as these involve names, acronyms, dates, numbers, etc. which should not be part of potential CPs. Table 2 shows the first 10 candidate expressions extracted from the corpus into a file.

We found a total number of 632,005 matching strings, with 134,803 distinct ones. In the generated file the words are followed by their POS tags in order to enable bigram formation by the package NSP. The bigrams are of the form (PRP_(DET)?_N, PRP) and (PRP, (DET)?_N_PRP).

4.2. Analysing results

For each one of the three tests we selected - Log-likelihood score (LL), Pearson's χ^2 (χ^2) and Mutual Information (MI) tests - and each of the two forms of bigrams, the output of the statistical package is a list of candidate CPs and their ranking according to the test.

Table 3 shows the top–ranked expressions, according to raw frequencies and to tests LL and MI. The data suggests that we are likely to find CPs amongst frequent "PRP (Det)? N PRP" strings. The results from the statistical tests filter the patterns with strong collocational properties, but the improvement is not noticed in the top 20 strings, considering LL and MI. The performance of χ^2 was very poor in this interval, therefore it has not been included in the table.

Given the amount of variation within the class of CPs and the amount of disagreement between linguists and grammarians, one cannot expect to find an exhaustive listing of CPs. This leaves us with the problem of how to validate the statistical results. The best solution is to present a list of statistically discovered collocations to a group of lexicographers, and let them select the CPs, but this is a costly enterprise, which we have not managed to accomplish yet.

raw frequency	LL: PRP_N PRP	LL: PRP N_PRP	MI: PRP_N PRP	MI: PRP N_PRP
em relaç ˜ao a	em relaç ˜ao a	por causa de	em relaç ˜ao a	por causa de
de acordo com	de acordo com	em relaç ˜ao a	de acordo com	em relaç ˜ao a
por causa de	em relaç ˜ao de	de acordo com	em relaç ˜ao de	de acordo com
no final de	de acordo de	ao lado de	de acordo de	ao lado de
em torno de	com base em	em torno de	com base em	em torno de
no o in´ıcio de	em frente a	com base em	em frente a	com base em
no caso de	em direç ˜ao a	ao contr´ario de	em direç ˜ao a	ao contr´ario de
ao lado de	com relaç ˜ao a	ao longo de	com relaç ˜ao a	ao longo de
com base em	com base de	por volta de	com base de	por volta de
ao contr´ario de	em meio a	no caso de	em meio a	no caso de
ao governo de	em entrevista a	por meio de	em entrevista a	por parte de
em vez de	em contato com	por parte de	em contato com	por meio de
ao longo de	por causa de	desde o in´ıcio de	por causa de	desde o in´ıcio de
por volta de	em homenagem a	por falta de	em homenagem a	por falta de
no mercado de	no final de	de causa de	no final de	de causa de
na hora de	de combate a	em vez de	de combate a	em vez de
no centro de	em torno de	no final de	em torno de	uma vez por
na noite de	de volta a	a respeito de	de volta a	no final de
por parte de	no in´ıcio de	em acordo com	no in´ıcio de	a respeito de
em funç ˜ao de	em frente de	ao governo de	no combate a	em acordo com

Table 3: Top–ranked expressions

The alternative was to obtain a good list of CPs, compiled by NILC[1], which is used for tagging the corpus itself. The list was enriched with some "em N de" CPs, obtained from [Oliveira et al., 2003]. This manually compiled list of CPs (henceforth, the CPList) contains 169 CPs, of which 159 occurred in the corpus.

The comparison between CPList and the ranking of the statistical collocations obtained can be seen in table 4. We have found that, in the first hundred best ranked collocations of the log-likelihood test with bigrams "PRP_N PRP", there were 27 CPs from CPList, corresponding to 17.5% of the list. The same test, with bigram "PRP N_PRP", produced a better result of 22%. From this point of view, the best performing test was MI with bigrams "PRP_N PRP" which produced 86% of CPList's CPs, in the range of ten thousand best ranked collocations candidates.

The statistical tests which have been used suggest that there a lot more collocations of the pattern "PRP (Det)? N PRP" than recognisable CPs. If this is really the case then these tests have limited usefulness in building an electronic dictionary.

On the other hand, the tests can be used as facilitators. Given that above 80% of CPList's CPs were spotted within the top 10,000 collocation candidates, it is reasonable to have the 10,000 expressions analysed in the original paragraphs, by a group of human judges, in order to discover more CPs.

We feel that the statistical tests should reflect more closely the linguistic criteria. In particular, the insertion criteria could be used in a rule. For instance, if a potential CP "PRP_1 N PRP_2" occurs also as "PRP_1 Det N PRP_2", it is less likely to be a CP. If the insertion criteria is generalised as a non–modifiability criteria, then the number of variants

[1]The list was obtained from http://www.nilc.icmc.usp.br/nilc/TagSet/locucoesprepositivas.htm, in March 2004.

	LL		χ^2		MI	
Ranking	P_N P	P N_P	P_N P	P N_P	P_N P	P N_P
up to 100	27 17.5%	37 23.5%	5 3%	5 3%	27 17.5%	37 23.5%
up to 300	46 29%	65 41%	9 5.5%	26 16.5%	46 29%	65 41%
up to 500	58 36.5%	78 49%	16 10%	40 25%	58 36.5%	76 48%
up to 1.000	82 51.5%	97 61%	25 15.5%	67 42%	92 58%	92 58%
up to 10.000	124 78%	130 82%	88 55.5%	117 73%	137 86%	133 83.5%
beyond 10.000	159	159	159	159	159	159

Table 4: CPs from CPList found in the tests.

of the potential CP expressions should count as a negative factor in the statistical scores. Let us consider example 4.1

4.1 *As declarações foram feitas em entrevista* **na varanda da** *casa em que morou o presidente João Goulart.*

The expression **na varanda de** ("in the varanda of") has the desired pattern, but the fact that it has several variants in the corpus, as shown in 4.2 (insertion of the adjective **larga**) and 4.3 (plural inflexion), should be an evidence against CP–hood.

4.2 *Mas teve a compensação de ver, ao lado do seu homólogo,* **na larga varanda da** *pousada, os primeiros veículos não oficiais a atravessarem a nova ponte.*

4.3 *Os "sem convite" permaneceram atrás das cadeiras e* **nas varandas dos** *três pisos do "shopping".*

Another set of linguistic motivated test could be devised to verify whether the structure of the prepositional phrase is *$Prep_1$ NP_1(N $Prep_2$ NP_2)*, rather than *CP($Prep_1$ N_1 $Prep_2$) NP*. If *N $Prep_2$ NP_2* is found to be a collocation then the expression $Prep_1$ N_1 $Prep_2$ is less likely to be a CP. Example 4.4 illustrates this point, if we consider that **bandeira do Brasil** ("Brazilian national flag") is a MWE, which is an indication that **da bandeira de** is not a CP.

4.4 *A Prefeitura de Rio Branco distribuiu 500 kg de cal para os moradores pintarem as ruas com as cores* **da bandeira do Brasil**.

The morphology of the noun can also be used to improve the precision of the statistical methods. Let us consider, for instance, the nominalisation of the verb **extrair** ("to extract") and the impact on its complements in examples 4.5 and 4.6.

4.5 *Família e amigos do maestro comentaram sobre um possível erro de avaliação médica ao submeter Jobim a nova cirurgia, na terça-feira, para retirada do tecido ao redor de onde foi* **extraído o tumor**.

4.6 *Pessoas com câncer de pulmão avançado que são tratadas com drogas antes e depois* **da extração do tumor** *podem viver até seis vezes mais, diz um estudo dos EUA.*

While the verb complement is NP(**o tumor**), the complement of the nominalisation is a PP(**do tumor**), given rise to the potential CP **da extração do**. In summary, if we identify the morphological marks of nominalisation (i.e. derivational suffixes) in the noun then this should be negative evidence of CP–hood.

5. Concluding Remarks

Multi–word expressions have been identified statistically with success, rendering collocation tests a useful tool for building electronic lexicons. Nevertheless, the statistical methods discussed in section 3 have only limited success in identifying complex prepositions.

On the other hand, the linguistic criteria presented in section 2 is not immediately translatable into computer algorithms. They are useful for systematic human evaluation of potential CPs in sentences. Combining both tools provides a feasible method for compiling provisionary list of CPs to be used in computer applications

We suspect that there are ways in which the statistical tests could be improved, by using linguistic knowledge. Some additional filtering of the data involving the insertion criteria seems to be useful. In addition, we observed that each test produces a different ranking, as it should be expected. It would be useful to combine the tests and see if the results would improve.

References

Banerjee, S. and Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City.

Bouma, G. and Villada, B. (2002). Corpus-based acquisition of collocational prepositional phrases. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University.

CETENFolha (2004). In the WWW. http://acdc.linguateca.pt/cetenfolha.

Dias, M. C. (2002). Locução para quê? *Revista Veredas*.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mattoso Câmara Jr, J. (1984). *Dicionário de Lingüística e Gramática*. Editora Vozes.

Oliveira, C., Garrão, M., and Amaral, L. A. (2003). Recognising complex prepositions prep+n+prep as negative patterns in automatic term extraction from texts. In *Anais do I Workshop em Tecnologia da Informação e Linguagem Humana*, São Carlos, SP.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1978). *A Grammar of Contemporary English*. Longman Group Limited.

Trawinski, B. (2003). The syntax of complex prepositions in german: An hpsg approach. In Banski, P. and Przepiorkowski, A., editors, *Generative Linguistics in Poland: Morphosyntactic Investigations*, pages 155–166, Warsaw, Poland. Instytut Podstaw Informatyki Polskiej Akademii Nauk.