Building Domain Specific Parsed *Corpora* in Portuguese Language

Lucelene Lopes, Renata Vieira

PPGCC – FACIN – PUCRS
Av. Ipiranga 6681 – 90.619-900 – Porto Alegre – Brazil

{lucelene.lopes, renata.vieira}@pucrs.br

Abstract. *This paper discusses the process of building domain specific parsed* corpora*. The whole process is detailed on how the texts were chosen, validated, converted to a common format and, particularly, how they have been subject to a careful refinement in order to keep only relevant and well-formed sentences for parsing. The resulting* corpora *are described by its numerical characteristics and practical applications are mentioned. Finally, these* corpora *are made available to the research community and brief examples of use are presented.*

Resumo. *Esse artigo discute o processo de construção de* corpora *de domínios específicos. Todo o processo é detalhado descrevendo como os textos foram escolhidos, convertidos para um formato comum e, particularmente, como foram submetidos a um cuidadoso processo de refinamento buscando manter somente frases relevantes e bem formadas para serem submetidas a um* parser*. Os* corpora *resultantes são descritos por suas características numéricas e aplicações práticas são mencionadas. Finalmente, esses* corpora *são disponibilizados à comunidade de pesquisa e pequenos exemplos de utilização são apresentados.*

1. Introduction

Domain specific parsed *corpora* are relevant resources for trendy Natural Language Processing (NLP) tasks, such as ontology learning. The process of achieving such resources for academic works from their usual PDF format is quite laborious. Besides converting to plain text formats, these documents have to be well selected, converted to the right encoding format, and learned from a multitude of non textual elements so that they can be fit as input for currently available parsers.

The use of domain specific *corpora* is quite common in many languages. One example is the work of Bourigault *et al.* [Bourigault et al. 2005] which uses a Law *corpus* in French language to extract noun phrases in order to build an ontology. Another examples of *corpus*-based term extraction are the works of Kietz *et al.* [Kietz et al. 2000] which presents an ontology extraction from a German language *corpus* composed by texts from the intranet of an insurance company, and of Kilgarriff *et al.* [Kilgarriff et al. 2006] which describes the creation of an huge 55 million words bilingual (Irish and English) *corpus*.

The work of Lopes *et al.* [Lopes et al. 2009b] presents the automatic extraction of relevant terms from a Pediatrics *corpus* in brazilian Portuguese builded by Coulthard [Coulthard 2005]. This work describes with a relative success that the automatic extraction of terms is comparable to a list of terms manually generated by a group of Linguistic and Pediatric specialists over a couple of years [TextCC-x 2012].

Applications as those demonstrate that *corpora* availability allows a considerable save in time and expensive specialist resources. In fact, the availability of *corpora* from different scientific domains represents an important asset in order to identify the relevant terms with a considerable lower cost than the use of specialist of the domain, and probably with more reliable results, since it will furnish the terms that are actually used in the area avoiding possible prejudice from the specialists themselves. According to Perini [Perini 2007], the use of *corpora* in the scientific process becomes relevant because of its impartiality and reliable indication of frequencies of the forms, since it represents the language reality without any theoretical preconceptions.

The construction of automatic tools to extract information from *corpora* also is very popular as the works of Gregoire and Duelme [Gregoire 2009] which describes a term extractor for Dutch language *corpora*, and of Pantel and Lin [Pantel and Lin 2001] which presents an automatic tool for extraction of terms applied to *corpora* in English and Chinese languages.

This paper presents the building process of four original domain specific parsed *corpora* in Portuguese (brazilian, actually) language, plus the refinement and parsing of a pre-existent *corpus*. Additionally, the results, *i.e.*, the *corpora* main characteristics are presented, and some initial application are described. The domain of the four original *corpora* are:

- Geology (Geo);
- Data Mining (DM);
- Stochastic Modeling (SM);
- Parallel Processing (PP).

Besides the importance of the choice of which texts to include, the most important issue in the *corpus* building process proposed in this paper is the careful refinement of the texts in order to produce a reliable language resources. It is vital to keep in mind that the generated *corpora* will be the input of parsing software tool to perform linguistic annotation. Therefore, it is quite interesting to generate texts with well-formed sentences, *i.e.*, texts as free as possible from pitfalls to the next NLP tools to be employed.

For this reason, in this paper the previously existing *corpus* on Pediatrics (Ped) [Coulthard 2005] was submitted to the same careful refinements as the other four original *corpora*. To illustrate the benefit of such procedure, traditional Information Retrieval metrics (precision, recall and f-measure) are taken from a term extraction procedure applied to the *corpus* as proposed by Coulthard and after the application of the refinements. It is shown that the refinements improve the quality of the *corpus* since the precision of the extracted terms clearly increases as the texts are subject to the refinements.

This paper is organized as follows. Section 2 describes the steps in the construction of the four original *corpora*. Section 3 presents the application of the refinements to the previously existing *corpus* on Pediatrics and the gains achieved in term extraction according to numerical metrics. Section 4 presents the main characteristics of the five *corpora* and the results obtained with automatic term extraction. Finally, the conclusion suggest some future work to be developed using the present four *corpora* and summarizes the contribution of the presented *corpus* building process.

2. The building process

In order to build the four original *corpora*, and re-process the pre-existent one, a five step process was performed (Figure 1). Although quite intuitive, this process was organized to minimize the involvement of specialists from the domain, even though it represents an increase of the work of less specialized people, *i.e.*, in our experiments, computer science and linguistic students. The four original *corpora* were submitted to all five steps, while the pre-existent *corpus* were submitted only to the last two steps.

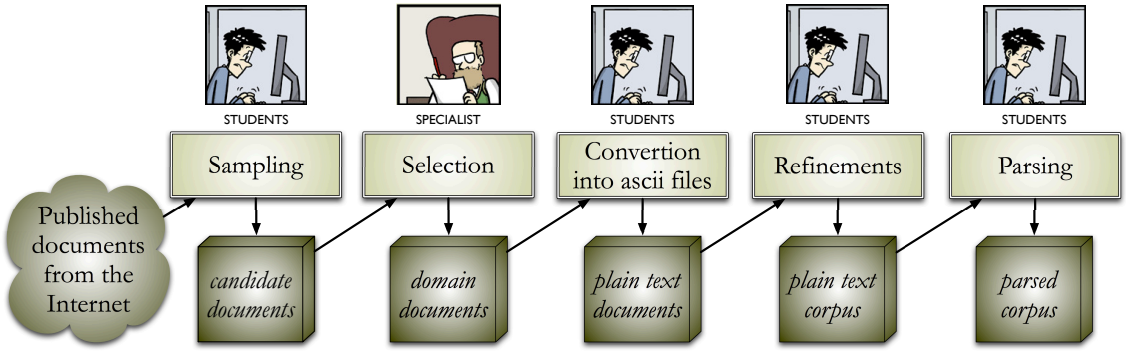

Figure 1. Five steps process to build scientific parsed *corpora*.

The first step consists in collect a considerable number of scientific texts from the Internet in various electronic formats, *e.g.*, `.pdf`, `.ps`, `.doc`, `.tex`, *etc.*. This step is done by non-specialized students searching public databases of theses, dissertations, technical reports and conference and journal papers with keywords or titles or abstracts containing the words of the domain name. Specifically, the Brazilian Digital Library of Thesis and Dissertations (BDTD) [BDTD-x 2012] and the Google Scholar [Google Scholar 2012] were the basic sources for the search performed.

The second step was the only step in which domain specialists were involved. In this step a very shallow analysis of the texts was made to select which texts were actually relevant to the domain. It is important to notice that the specialists were not required to fully read the texts, but only to consider, according to its experience, if the text could be considered as belonging to the domain or not. Nevertheless, a considerable number of texts were discarded during the selection step (see Table 2).

The third step consists in transforming the electronic format of the selected texts into a plain textual format using extended `ascii` representation[1]. For most of the selected texts, an automatic converter called Entrelinhas [Silveira 2008] was used, but already textual formats, as LaTeX files, were transformed by the exclusion of embedded commands.

Probably the most important and more laborious step was the fourth one, in which the texts were subject to a semi-automatic application of a set of refinements in order to keep only valid and coherent Portuguese language sentences fit for parsing. The next step in our process is a linguistic procedure (parsing), so valid and meaningful Portuguese sentences are much more likely to be correctly recognized. Titles, keywords, abstracts in

[1]The use of an extended `ascii` was necessary, since Portuguese texts always have non-standard `ascii` characters for diacritics (á, é, í, ó, ú, â, ê, ô, ü, à, ã, õ, ç) in their lower and upper case versions.

other languages, addresses, figures, tables and captions were removed, not by having low relevance, but by not usually being complete sentences. Acknowledgements, dedicatory and other non technical section were also removed since they are not likely to represent the domain of the selected documents.

Unfortunately, the use of such text structures is far from uniform in scientific texts. Despite of that, in order to reduce the manual effort, some automatic refinements were applied through the use of scripts and regular expressions search and replace options of word processors, as Emacs [Emacs 2012] and Notepad++ [Notepad++ 2012]. Table 1 presents the regular expressions employed, however, it is important to stress the fact that this step step was semi-automated, *i.e.*, these expressions were applied with human supervision.

Expression	Aim	Example
“0-9.0-9” change to “0-90-9”	remove period signs inside number digits	1.927 becomes 1927
“0-9,0-9” change to “0-90-9”	remove comma signs inside number digits	3,1415 becomes 31415
“(0-90-90-90-9)” change to “”	remove references years	(2012) is deleted
“[0-90-90-90-9]” change to “”	remove references years	[2012] is deleted
“(” “A-Z a-z”* “,0-90-90-90-9)” change to “”	remove references	(Lopes e Vieira, 2012) is deleted
“[” “A-Z a-z”* “,0-90-90-90-9]” change to “”	remove references	(Lopes et al., 2012) is deleted
“-A-z” change to “- A-z”	insert a blank space for itemizes with hyphen	-backup becomes - backup
“*A-z” change to “* A-z”	insert a blank space for itemizes with asterisk	*backup becomes * backup
“.A-z” change to “* A-z”	insert a blank space for itemizes with period	.backup becomes . backu
“A-z. a-z” change to “A-z a-z”	remove period as abbreviation	Dra. Lopes becomes Dra Lopes

Table 1. Regular expressions semi-automatically employed in the refinement step.

The last step of the proposed process is the submission of the refined texts, *i.e.*, the *corpus* in plain text format to a parsing tool. The chosen parsing tool was the software PALAVRAS [Bick 2000], one of the finest parsers available for Portuguese language. The result of this final step is a parsed *corpus* with part-of-speech tags, semantic tags, and identification of multiple linguistic structures, such as noun phrases and clauses.

One important point in this methodology is that the more tedious steps were performed by non-specialized students and only the Selection step was performed by specialists. Applying the process to the building of the four original *corpora*, a considerable amount of information was processed. Table 2 summarizes the number of texts considered in the Text Sampling step (Before Selection) and the actual number of texts considered in the following steps (After Selection). The texts were divided in three groups: the Ph.D. thesis (**T**), the M.Sc. dissertations (**D**) and the technical reports and conference and journal papers (**P**).

Corpus	Before Selection			After Selection			Final Number of Texts
	T	**D**	**P**	**T**	**D**	**P**	
Geo	55	76	339	17	32	185	234
DM	30	97	51	8	32	13	53
SM	31	70	90	6	33	49	88
PP	43	114	78	9	27	26	62

Table 2. Number of texts processed during the *corpora* construction.

The first observation of data presented in Table 2 is the large amount of documents considered and kept in the Geology domain in comparison with the number of documents for the other domains (DD, SM and PP). This fact is mostly explained by the decision to consider Geology as a single domain, while the computer science related domains were split into three different *corpora*.

Observing all the documents considered in this *corpora* building effort it was noticeable that usually Ph.D. thesis are larger and richer in definitions than M.Sc. dissertations. Analogously, M.Sc. dissertations are larger and richer in definitions than technical reports and conference and journal papers. Therefore, the number of texts in Table 2 may serve as an indication of the size of the builded *corpora*, but also the quality of each *corpus* according to the desired future application.

3. The refinement of texts

To illustrate the benefits of the refinements of the texts, we conduct an experiment with the Pediatrics *corpus* [Coulthard 2005]. This *corpus* is composed by 283 texts from papers of the Brazilian Journal of Pediatrics, it has 878,522 words it has been created without any particular concern with refinements of the sentences. For this *corpus* a reference list with the more relevant terms with two and three words (bigrams and trigrams, respectively) was generated [TextCC-x 2012].

The reference list was originally manually created with a deep involvement of domain specialists and the ultimate goal of this list was to build a list of compound terms to help human translation. Nevertheless, the resulting list of 1,534 bigrams and 2,661 trigrams can be considered the relevant terms for the Pediatrics domain, at least according to this *corpus*. It is important to keep in mind that such reference list for a *corpus* is a rare resource, since very few *corpora* have such reliable information to evaluate the effectiveness of a term extraction.

Using an annotation tool, the parser PALAVRAS [Bick 2000], and a sophisticated noun phrase extractor, the EχATO$_{LP}$ software tool [Lopes et al. 2009a], 2,228 bigrams and 2,578 trigrams were considered relevant[2] to the domain represented by the *corpus*. It is worth mention that the combined use of PALAVRAS and EχATO$_{LP}$ allows high quality term extraction, since it relies on powerful annotation of Portuguese texts [Lopes et al. 2010], heuristics to achieve better noun phrase detection [Lopes and Vieira 2012], and advanced relevance induex computation [Lopes et al. 2012].

The intersection between the terms extracted manually (reference lists) and the terms extracted by PALAVRAS and EχATO$_{LP}$ was 1,375 bigrams and 1,941 trigrams. The quality of such automatic extraction can be computed using the traditional Information Retrieval metrics [van Rijsbergen 1975]: precision (P), recall (R) and f-measure (F) metrics, *i.e.*:

$$P = \frac{|RL \cap EL|}{|EL|} \qquad R = \frac{|RL \cap EL|}{|RL|} \qquad F = \frac{2 \times P \times R}{P + R}$$

where $|RL|$ is the cardinality of the reference list (the manually extracted one), $|EL|$ is the cardinality of the automatically extracted list (the one extracted by PALAVRAS and EχATO$_{LP}$) and $|RL \cap EL|$ is the cardinality of the intersection between the lists.

Computing these metrics for the experiment with the Pediatrics *corpus* without any refinements in the texts, the result obtained for bigrams and trigrams were:

$$P = 61.7\% \quad R = 89.6\% \quad F = 73.1\% \quad \text{(bigrams)}$$

$$P = 75.3\% \quad R = 72.9\% \quad F = 74.1\% \quad \text{(trigrams)}$$

After applying the refinements to remove invalid sentences, as it was applied in the construction of the four original *corpora*, the Pediatrics *corpus* became smaller. This refined version discarded two entire texts that have no complete sentences, but also other texts were reduced by the elimination of incomplete sentences. The result was a new Pediatrics *corpus* composed by 281 texts, with 835,412 words.

The combined linguistic based extraction using PALAVRAS and EχATO$_{LP}$ resulted in the extraction of 2,323 bigrams and 2,726 trigrams, were the intersection with the reference list was of 1,440 bigrams and 2,078 trigrams. Such results represent an improvement in the quality of the extraction, since the refinements applied to the *corpus* avoid the extraction of irrelevant terms. In fact, computing the metrics for the experiment with the refined Pediatric *corpus* the following results were obtained:

$$P = 62.0\% \quad R = 93.9\% \quad F = 74.7\% \quad \text{(bigrams)}$$

$$P = 76.2\% \quad R = 78.1\% \quad F = 77.2\% \quad \text{(trigrams)}$$

Such results show a slight increase in the precision of bigrams, that raise from 61.7% to 62.0%. However, the absolute number of correctly extracted bigrams increase

[2]This extraction corresponds to a process proposed in the context of Lucelene Lopes' Ph.D. thesis [Lopes 2012], which recommends to consider the 15% more relevant noun phrases identified by EχATO$_{LP}$, if they have at least two occurrences in the *corpus*.

from 1,375 to 1,440, which corresponds to an increase of the recall from 89.6% to 93.9%. The overall change illustrated by the f-measure shows a slight benficial effect of the refinements in the results for bigrams raising from 73.1% to 74.7%.

The results for the trigrams, on the contrary, show a better result since both precision and recall increase with the refinement of the Pediatrics *corpus*. The number of extracted terms increased from 2,578 to 2,726 trigrams, the number of correct ones increased even more from 1,941 to 2,078. These results indicated an overall improvement from 74.1% to 77.2% for f-measure.

According to this experiment it seems that refinement in the Pediatrics *corpus* is more beneficial to more complex terms, as trigrams, since such terms are harder to correctly detect. In fact, even for bigrams it seems that the detection of terms was not a problem, since the number of correct bigrams increased in 65 terms. An expected effect of the refinement of Pediatrics texts was the increase in the number of extracted terms. However, for bigrams, unlike trigrams, the number of correct terms did not increase as much in order to deliver a more impressive result in terms of the computed metrics.

4. *Corpora* characteristics

The four original *corpora* built were chosen according to the fields of expertise of a multidisciplinary research group in order to make easier the communication between researchers that did not share a common background. The specific areas of the *corpora* were in Earth Sciences (Geology - Geo) and three domains from Computer Science (Data Mining - DM, Stochastic Modeling - SM and Parallel Processing - PP). Table 3 summarizes the characteristics of the four constructed *corpora*, and the two versions of the pre-existent Pediatrics *corpus*, the original (PED) and the refined (Ped) ones.

corpus	Texts	Sentences	Words
Geo	234	69,461	2,020,527
DM	53	42,932	1,127,816
SM	88	44,222	1,173,401
PP	62	40,928	1,086,771
PED	283	30,747	878,522
Ped	281	27,724	835,412

Table 3. *Corpora* characteristics.

After parsing with PALAVRAS [Bick 2000] the five *corpora* that were subject the refinement step, to exemplify their use, all five *corpora* were submitted EχATO$_{LP}$ term extractor [Lopes et al. 2009a] in order to extract their relevant terms. The number of relevant, according to [Lopes 2012], extracted terms to each *corpus* is presented in Table 4 were the terms are classified according to the number of words in each term, *i.e.*, unigrams to quadrigrams and multigrams (ngrams with five or more words).

It is interesting to notice that the total number of extracted terms is proportional to the number of words of each *corpora*. However, while the Geology and Computer Science *corpora* (Geo, DM, SM, PP) are slightly richer in compound terms, the Pediatrics *corpus* has, proportionally, more unigrams. This peculiar behavior may be due to the domain intrinsic writing style.

corpus	unigrams	bigrams	trigrams	quadrigrams	multigrams	total
Geo	1,152	4,616	5,582	4,544	9,281	25,175
DM	630	2,221	2,871	2,104	4,993	12,819
SM	648	2,116	2,831	2,176	4,813	12,584
PP	654	2,145	2,996	2,072	3,724	11,591
Ped	892	2,323	2,726	1,192	1,140	8,273

Table 4. Number of extracted terms to each *corpus* according to Lopes's process [Lopes 2012].

In order to have a visual impression about the relevant terms of each *corpus*, Figures 2, 3, 4, 5 and 6 show the relevant terms in a form of concept clouds. These clouds illustrate the topics of all *corpora*, as well as the basic information that can be extracted from the produced resources. For example, it is possible to notice the importance of the terms "arenitos" ("sand stones" in English) for the Geology *corpus* and "crianças" ("children" in English) for Pediatrics *corpus*.

The general observation of the extracted terms from each *corpus* considering their relevance expressed by the font sizes in Figures 2, 3, 4, 5 and 6 shows the importance of unigrams. Such fact is explained by the large number of such terms in comparison with more complex terms. Nevertheless, all *corpora* present fairly relevant bigrams as "events sincronizantes" ("synchronizing events" in English) found in the Stochastic Modeling *corpus*, or even the trigram "regras de associação" ("association rules" in English).

Figure 2. Tag cloud of the relevant terms found in the Geology *corpus* (Geo).

It is also possible to notice that very specific terms, as "class minoritária" ("minority class" in English) for Data Mining *corpus*, are more relevant than other terms employed in Data Mining texts, but not necessarily specific to the domain as "interval de dados" ("data interval" in English). It is also noticeable, the huge importance of the term

Figure 3. Tag cloud of the relevant terms found in the Data Mining *corpus* (DM).

"checkpoint" (an English word) very frequent to the Parallel Processing area documents. Actually, we notice that such English language terms are very relevant to the Parallel Processing *corpus*, in contrast with Geology and Pediatrics *corpora* which do not have almost only Portuguese terms.

Figure 4. Tag cloud of the relevant terms found in the Stochastic Modeling *corpus* (SM).

Figure 5. Tag cloud of the relevant terms found in the Parallel Processing *corpus* (PP).

Figure 6. Tag cloud of the relevant terms found in the refined version of the Pediatrics *corpus* (Ped).

5. Conclusion

This paper presents the effort to build four original *corpora* from the choice of texts, selection by specialists, conversion of electronic format, text refinement, and even some brief applications of the builded *corpora*. In particular, the importance of text refinement step was illustrated with an existing Pediatrics *corpus*.

It is important to call the reader attention that the experiment with the refinement of the Pediatrics *corpus* has shown an increase in the metrics, but the beneficial effect of a "cleaner" *corpus* may be relevant to other applications, than only term extraction. For instance, the availability of texts with well formed sentences may be advantageous to other linguistic applications as concordancing tools or other text analysis tools. Therefore, it is reasonable to believe that the four builded *corpora* must be suitable to further linguistic analysis.

Probably the more important contribution of this paper is the availability of the generated corpora to the research community. The four original *corpora* (Geology, Data

Mining, Stochastic Modeling and Parallel Processing) and the refined version of the Pediatrics *corpus* are available at our research group web page at the address:

`http://www.inf.pucrs.br/~linatural/`

In this web site the reader will find the five *corpora* in a simple text format (.txt), as well as in the annotated format (.xml) as outputted by PALAVRAS parser [Bick 2000]. There the reader will also find more sophisticated linguistic resources as the term clouds presented in this paper, but also term lists and even concept hierarchies for the five domains generated by EχATO$_{LP}$ term extractor [Lopes et al. 2009a]. It is also available the reference lists (bigrams and trigrams) employed for the Pediatrics *corpus* experiment in Section 3.

However, the contribution of this paper is two-fold, since not only the five *corpora* are good language resources to be used by the NLP community, but also the process of *corpus* construction is a valid framework to develop new valid and reliable *corpora*. It is important to notice that the refinement of texts process can be automatized, since it is based on the application of some common replacement options based on textual regular expressions.

As said before, this building effort is inserted in a broader research initiative that congregates researchers from different domain areas. These new *corpora* are already being used to extract relevant terms in each domain in order to build glossaries to help the scientific exchanges among researchers from different domains.

Besides this on going application, these *corpora* can also be employed to other applications, *e.g.*, as source to concept extraction for ontology learning or even more sophisticated tasks as relation extraction. In fact, there is a myriad of potential applications for the new *corpora*.

Acknowledgements

The authors would like to express their gratitude to the students that helped in the *corpora* building: Daniel Martins, Kamila Ail da Costa, Guilhereme Rodegheri and Eduardo Schwingel Diederichsen. We thank as well the researchers of the PALEOPROSPEC project at the Computer Science Department (FACIN) of the PUCRS University that contribute in the *corpora* building process as specialist of the four domains.

References

BDTD-x (2012). Biblioteca digital brasileira de teses e dissertações. [accessed online on October 15th, 2012].

Bick, E. (2000). *The parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.d. thesis, Arhus University.

Bourigault, D., Fabre, C., Frérot, C., Jacques, M., and Ozdowska, S. (2005). Syntex: analyseur syntaxique de corpus. In *Actes de la 12ème TALN*, Dourdan. ATALA.

Coulthard, R. J. (2005). The application of corpus methodology to translation: the jped parallel corpus and the pediatrics comparable corpus. M.sc. dissertation, UFSC, Florianópolis, Brazil.

Emacs (2012). Gnu emacs. GNU Emacs - GNU Project - Free Software Foundation (FSF). [accessed online on October 15th, 2012].

Google Scholar (2012). Google scholar. [accessed online on October 15th, 2012].

Gregoire, N. (2009). Duelme: a dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1–2):23–39.

Kietz, J., Volz, R., and Maedche, A. (2000). Extracting a domain-specific ontology from a corporate intranet. In *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning*, volume 7, pages 167–175, Morristown, NJ. Association for Computational Linguistics.

Kilgarriff, A., Rundell, M., and Dhonnchadha, E. U. (2006). Efficient corpus development for lexicography: building the new corpus for ireland. *Language Resources and Evaluation*, 40(7):127–152.

Lopes, L. (2012). *Extração automática de conceitos a partir de textos em língua portuguesa*. Ph.d. thesis, FACIN-PUCRS, Porto Alegre, Brazil.

Lopes, L., de Oliveira, L. H. M., and Vieira, R. (2010). Portuguese term extraction methods: Comparing linguistic and statistical approaches. In *PROPOR 2010 – International Conference on Computational Processing of Portuguese Language*.

Lopes, L., Fernandes, P., and Vieira, R. (2012). Domain term relevance through tf-dcf. In *Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI 2012)*, pages 1001–1007, Las Vegas, USA. CSREA Press.

Lopes, L., Fernandes, P., Vieira, R., and Fedrizzi, G. (2009a). Exatolp - an automatic tool for term extraction from portuguese language corpora. In *LTC'09 - 4th Language & Technology Conference*, pages 167–175, Poznan, Poland. Adam Mickiewicz Univ.

Lopes, L. and Vieira, R. (2012). Heuristics to improve ontology term extraction. In *PROPOR 2012 – International Conference on Computational Processing of Portuguese Language*, LNCS vol. 7243, pages 85–92.

Lopes, L., Vieira, R., Finatto, M. J., Zanette, A., Martins, D., and Ribeiro Jr., L. C. (2009b). Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. *RECIIS*, 3(1):72–84.

Notepad++ (2012). Notepad++. NotePad++ Home. [accessed online on October 15th, 2012].

Pantel, P. and Lin, D. (2001). A statistical corpus-based term extractor. In *Proc. of the 14th Biennial Conf. of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pages 36–46, New York, USA. ACM Press.

Perini, M. A. (2007). *Princípios de linguística descritiva: introdução ao pensamento gramatical*. Parábola, São Paulo, Brazil.

Silveira, F. P. (2008). Entrelinhas - uma ferramenta para processamento e análise de corpus. M.sc. dissertation, FACIN-PUCRS, Porto Alegre, Brazil.

TextCC-x (2012). Textcc – textos técnicos e científicos. [accessed online on October 15th, 2012].

van Rijsbergen, C. J. (1975). *Information Retrieval*. Butterworths, London, UK.