

# Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area

DOI: 10.3395/reciis.v3i1.244en



*Lucelene Lopes*  
Catholic University of Rio Grande do Sul, Porto Alegre, Brazil  
lucelene.lopes@pucrs.br



*Renata Vieira*  
Catholic University of Rio Grande do Sul, Porto Alegre, Brazil  
renata.vieira@gmail.com

## *Maria José Finatto*

Federal University of Rio Grande do Sul, Porto Alegre, Brazil  
mfinatto@pq.cnpq.br

## *Adriano Zanette*

Federal University of Rio Grande do Sul, Porto Alegre, Brazil

## *Daniel Martins*

Catholic University of Rio Grande do Sul, Porto Alegre, Brazil

## *Luiz Carlos Ribeiro Jr.*

Catholic University of Rio Grande do Sul, Porto Alegre, Brazil

## Abstract

In this article we demonstrate the use of the tool OntoLP in the ontology construction process in an experiment in the health care area. Specifically, terms based on a corpus in the pediatrics area are extracted. We compare the result obtained by the tool with the reference results of a list of terms obtained manually. In this comparison, bigrams and trigrams obtained through different methods are analyzed. We conclude the work by observing the advantages of processing by including complex linguistic information such as syntactical and semantic analysis.

## Keywords

processing of natural language; ontologies; construction of ontologies for the health care area; semi-automatic extraction of terms

## Introduction

The development of computerization and also the constant evolution of the means for storing great masses of data, in the most diverse sectors of society, gave origin to a significant quantity of digital bases and data sources. Most of these are available on the internet, in the most varied forms, such as texts, images, videos, services, hypertexts, etc. In this sense, it is necessary to perfect descriptive models of knowledge provided by these resources so that they are recovered when necessary. Lack of standardization in knowledge representation may make it difficult to understand the content of the various bases, consequently making their use impracticable. As an alternative to solve this problem, the information systems and computer science area has been adopting the use of knowledge representation with ontologies (Gomez-Perez et al. 2004).

Ontologies are being used as a way of conceptualizing, structuring and representing, in a document, the knowledge of a domain in such a way that it can be shared. This practice has been adopted in various domains, and especially in biology, bioinformatics, biomedicine and medicine, which is the domain of knowledge explored in this article. However, it is known that ontologies have an arduous construction process, which demands much time and effort, especially in their large-scale use (Brewster et al. 2003). One solution for this is to invest in research to be able to automate the task of constructing ontologies for specific domains (Buitelaar et al. 2003). This research often considers text bases as knowledge sources. These sources, for their part, are expressed in different languages, causing methods based on the use of linguistic information to be developed for the different languages.

In this context, we present a study directed specifically at the Portuguese language and considering the domain of medicine, in particular the pediatrics area. In this study we focus on comparing alternative approaches for identifying composite terms (concepts expressed in more than one word), considering that this is simply the initial stage in the ontology construction process.

In the medicine area, more specifically in the theme of prevention and promotion of health, for example, automated treatment of text information tends to help researchers and managers of information policies to recognize the best ways to present the most relevant data based on the aims and the existing communicative situation. Here the research area known as e-health, already recognized outside Brazil, is outlined. The idea of compacting and representing extracts of information in health sciences (and its challenges) has had space for discussion not only with respect to language recognition and scientific terminology that needs to be “facilitated” for the layman, but also in medical publications themselves. For this purpose, information processing systems must be developed and this development must be made by multidisciplinary teams made up of health professionals, scholars of language and communication in the sciences and engineers who build computerized systems with practical and simple interfaces.

The Journal of Medical Internet Research – JMIR <<http://www.jmir.org/>>, for example, contemplates e-health themes. The abstracts, text extracts of content representations are generated with the support of software of programs and have been appearing as:

- 1) synthetic text cut out of a source text or a group of texts;
- 2) conceptual map schemes based on a text or groups of source texts;
- 3) hierarchical relations schemes of conceptual nodules in ontologies based on one or more texts;

In the Brazilian setting, we can take an illustrative example of the usefulness of e-health systems. In the case of the Health Department, looking at the institutional advertising from the end of 2008 and the start of 2009, we know, for example, that hanseniasis is a disease that still has a considerable impact in Brazil. However, in spite of the Department and civil society’s actions, the population still seems to resist the informative campaigns that advertise preventative measures. Among various resources, the articles on hanseniasis prevention transmitted on television and print media do not seem to have the desired effect.

In this particular case, mapping the available information, especially in advertising media dealing with health themes, based on great masses of data from online text, scientific texts and texts aimed at the lay public, could demonstrate, for example, that the word “leprosy”, negative and stigmatizing as it may be, is rarely mentioned in the advertising texts that deal with hanseniasis. Maybe this interrelation gap explains the lay public’s delay in understanding the messages. A similar mapping could also show situations with the use of connected terms and their more or less popular equivalents in texts directed at specialists or semi-specialists and if the scientific community itself pondered the use of this terminology, in its different matrixes, more or less strongly.

Thus, it would be possible for the manager to see, based on a statistical collection of language configuration in texts that seek to inform laymen, that the language informed has to favor notional or logical links between a conceptual nodule X and a conceptual nodule Y (with the due information about the character of the different denominations). Once the lack has been identified and the information has been supplied, a lay citizen, as in the case above, could relate notions and, activating his cultural memory about a certain disease or risk, could respond pro-actively when faced with the information he is receiving.

Another example of the usefulness of these systems, but on another dimension, can reveal how a term such as “prevalence” is being employed by the doctor-researcher community in scientific publications in an interval, for example, of two years of publication and a corpus of over a million words. In a collection of pediatrics texts, published by the *Jornal de Pediatria* <<http://www.jpmed.com.br/>>, it can be seen how use of this expression can generate misunderstanding when it tends to extrapolate

the field of the technical term and fuse with the word in common language, which means “that which predominates”, while in medical terminology, it is a term that describes a statistical measurement in epidemiology. In these texts, the presence of a construction like “prevalence of comforter use” may signal an important piece of information for the magazine’s publishers. Apart from specific cases about the employment of technical terms, there are other cases sparsely indicated by the medical community itself. They are situations that extrapolate terminologies and reach the statute of recurring constructions in the texts, as we see in the work *Expressões médicas: falhas e acertos* (Bacelar et al. 2003).

Beyond the observation of language uses, prospecting scientific information available in collections of scientific periodicals is also important because “the increase in knowledge production – and, therefore, in the number of periodicals – in the second half of the last century led the community of professionals and researchers to confront the challenge of developing quality criteria that could guide readers when slecting the best scientific evidence” (Blank et al. 2006, p.97). This type of data, when it is extensively obtained and studied in great text bases available on the internet, offer valuable information for the managers and editors of specialized periodicals as they allow themes and notions of a particular community to recur. This community would then have, based on computerized exploration and representation of knowledge techniques, access to a broad range of information about its own communication practice, which could be useful to reflect on the production and circulation of knowledge.

The examples given, about simple incidences of words in texts, aim to illustrate the possibilities for using advanced systems to deal with text information. These systems, as well as locating information itself, show how the information is linguistically and notionally configured in various situations. They also reveal which other vocabulary units accompany them more frequently or more rarely. They are innovative systems that integrate search and ordering of data in text collections, which integrate lexical statistics, summarizing (context synthesizing) and ontologies. They are information selection tools that need qualified methodologies for treating phenomena of scientific language in use. In these systems, the cooperation between health professionals, linguists and computer scientists is a necessity.

In this article, the tool OntoLP (Ribeiro 2008) is used, which aims to help in a semi-automatic manner the engineers of ontologies in the Portuguese language, whether they be specialists in the domain in question (health professionals) or linguists. The tool shows suggestions of terms, concepts and of organization of ontology hierarchies, based on knowledge recorded in a text base or domain corpus.

More specifically, the article presents a study about the analysis and identification of composite terms, that is, terms which contain two (bigrams) or more words (n-grams). In the context of this work, only bigrams and

trigrams are extracted and this is the initial stage of the complex process of ontology construction. Alternative methods of text processing are compared. The study is developed with application in the health care area, considering a pediatrics corpus and a list of reference terms for method assessment.

## OntoLP tool

Natural language processing (NLP) presents itself in ontology construction through texts. NLP uses linguistic techniques based on syntactical and morphological analysis of texts, representing information on various levels.

As the ontologies created are restricted to a specific idiom, the difficulty increases in the Portuguese domain, a language which has been little compared to English. To answer this difficulty, new methodologies for the semi-automatic construction of ontologies are being created along with the tools to aid this construction.

OntoLP is a tool, actually a plug-in, for the ontologies editor Protégé (Gennari et al. 2002), a widely used editor in the scientific community and which gives support to the construction of ontologies, following the Semantic Web techniques, as for example the construction of OWL Web Ontology Language ontologies, as defined by the World Wide Web Consortium (W3C) (Mcguinness et al. 2004).

The process of automatic ontology construction is divided into five basic areas (Buitelaar 2005): extraction of terms that could be domain concepts; identification of hierarchical relation between the terms; identification of non-hierarchical relations; identification of instances and rule extraction (axioms). This process can be represented in layers as shown in Figure 1.

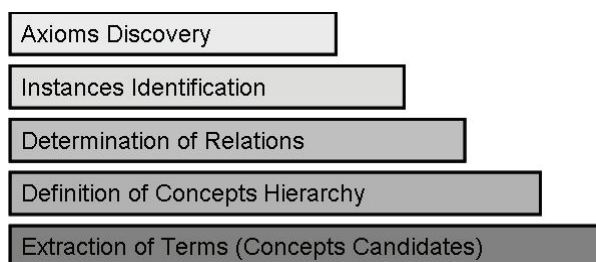


Figure 1 - Basic stages of ontology construction.

From the point of view of this work, which aims in the long term to achieve automatic ontology construction, extraction of terms is the initial and fundamental task, as the extracted terms represent the concepts of a specific area and are the base for performing the other stages.

For extraction of terms, there are three main approaches:

- Statistics – the documents contained in the corpus are seen as a set of terms and their frequency of occurrence is measured;
- Linguistics – the texts are annotated with linguistic information (morphological, syntactic and semantic)

and this information is taken into consideration in the extraction process;

- Hybrid – considers the union of the two approaches (statistical and linguistic).

The OntoLP tool (Ribeiro 2008) is made up of a series of hybrid methods. These methods of term extraction are grouped into two stages:

- CorpusXCES: in this stage, the corpus is annotated with linguistic information by the parser PALAVRAS (Bick 2000). The annotated corpus contains morphological, syntactic and semantic information, represented in the XCES/PLN-BR (Ide et al. 2000) format. The text processing for term extraction is based on this annotated corpus. Through morphosyntactic analysis, information is added to the original text that can allow more or less linguistically-informed methods to be employed. In this work, the linguistic information used by the extraction methods are the grammatical categories of the words (for example, *noun, verbs, adjectives*), prototypic semantic categories (for example, *humans, animals, diseases*) and the identification of nominal grammatical groups (nominal syntagms such as *exclusive maternal breastfeeding*);

- Term extraction: for this stage, different methods are applied which combine statistical measures of frequency with the linguistic information mentioned above, with the finality of extracting simple terms (unigrams) and composite terms (n-grams, where  $n > 1$ ).

In the OntoLP plug-in, the extraction methods have a set of functionalities to help the ontology engineer in the stages that may have human interaction. The extraction interface is divided into three parts: (1) selection of semantic groups; (2) extraction of simple terms; and (3) extraction of composite terms.

Figure 2 shows in a general way how these functionalities appear in the OntoLP tool, where the number 1, 2 and 3 indicate the extraction interfaces mentioned above. In the figure, the selected semantic group is (H – Human). The window on the far right shows the words in the group as tag clouds, where the most frequent terms are highlighted, in this case <patient> and <child> appear most frequently. The ontology engineer can choose to exclude a semantic group or not. In this example, the group (H – Human) would not be excluded because of its relevance in the analyzed corpus.

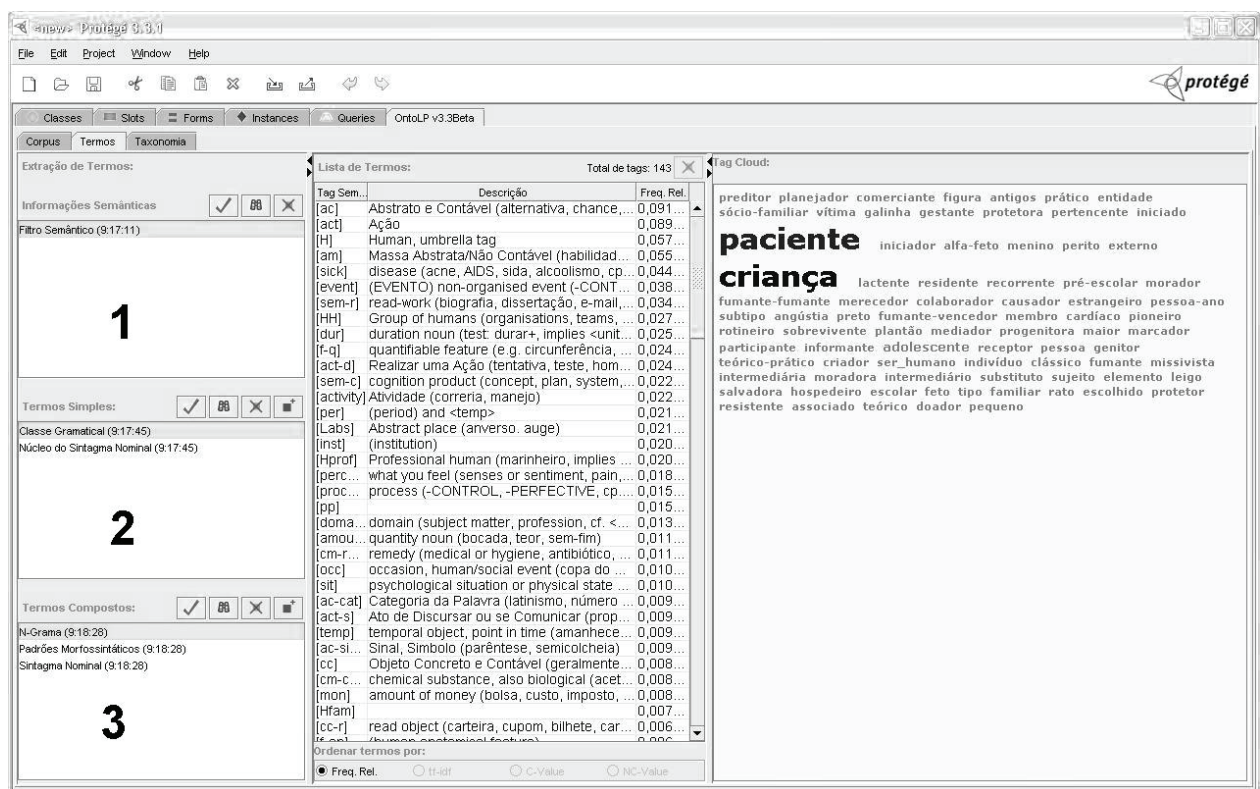


Figure 2 - Extraction of terms interface and its stages proposed for the task.

The Semantic Group Selection stage is optional. The tool shows the user the semantic information that the parser PALAVRAS associates to the words of the corpus. It is prototypic information that classified common nouns into general classes, for example, the tag

<an>, attributed to the noun “muscle”, indicates that the word belongs to the class “Anatomy”. Figure 3 shows some examples of these groups and subgroups that exist in the pediatrics corpus analyzed in this article.

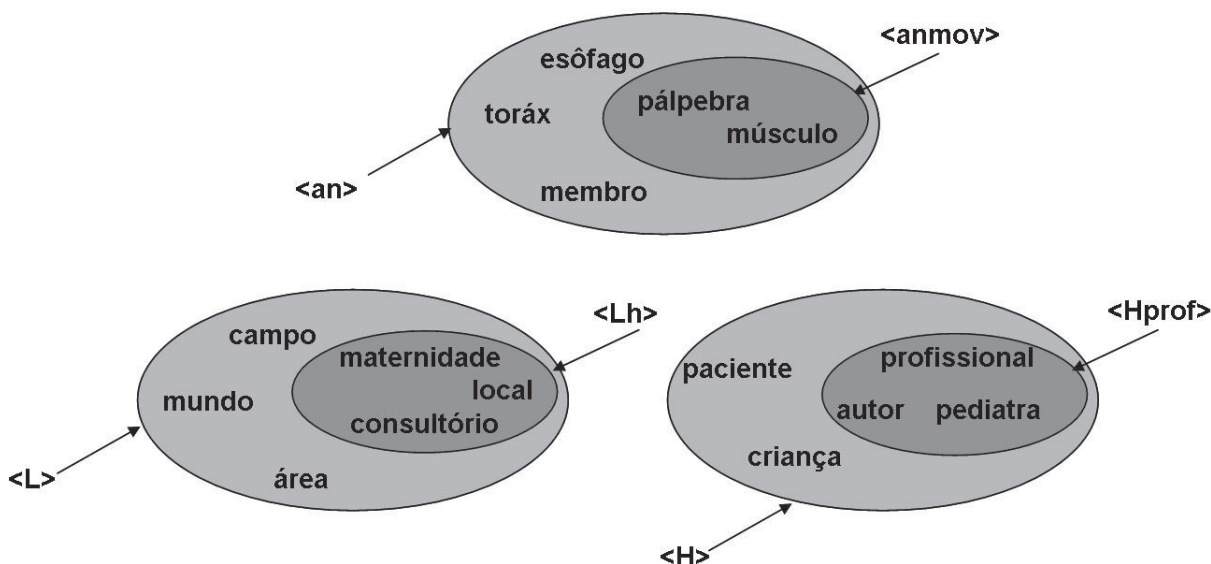


Figure 3 - Examples of semantic groups and subgroups.

This way, the nouns tagged with the same tag are grouped into semantic sets, for example:

- Group <an> (Anatomy): {esophagus, thorax, member, eyelid, muscle}
- Group <L> (Place): {field, world, area, maternity, location, office}
- Group <H> (Human): {patient, child, professional, author, pediatrician}

The semantic groups can also present subdivisions, as can be seen in Figure 3: a) anatomy (<an>) and anatomy of movement (<anmov>); b) place (<L>) and functional places (<Lh>); c) human (<H>) and professional human (<Hprof>).

The OntoLP tool provides the Filter by Semantic Group method for the user. This uses the following steps:

- The semantic tags present in the entry corpus are extracted;
- The calculation of relative frequency (RF) is applied to the list of semantic tags that are presented to the engineer ordered according to this measurement;
- The engineer excludes the semantic groups that he considers as having no relation to the domain represented by the entry corpus.

This method can be considered as the construction of a specific stopwords list for a domain. These stopwords are items not to consider. The correct selection of groups depends on the ontology engineer's knowledge of the area involved. The tool helps the engineer to show the occurrence of each group's terms and their relevance by the tag clouds method, that is, a method which attributes greater sources and text emphasis to more frequent terms in the corpus.

After the selection of semantic groups, the second and third stages of extraction (extraction of simple terms and extraction of composite terms) are performed, implemented by hybrid methods (statistical and linguistic). In the second stage, the extraction of simple terms is performed. The method used is the grammatical classes method, detailed in Ribeiro (2008).

The third stage, extraction of composite terms, the focus of this work, consists of indentifying bigrams and trigrams. In this stage, we use three different methods, with different linguistic complexities. Firstly, the extraction of n-grams is made, by co-occurrence frequency, simply, by applying simple filters such as eliminations of terms with initial or final prepositions. The second method considers the grammatical class of the terms and patterns of extraction, such as:

- adjective noun – maternal breastfeeding
- noun preposition noun – saturation of oxygen

The third method extracts nominal syntagms, as they are recognized by the syntactical analyzer. This is a more complex level of structural linguistic information and its production requires specialized tools.

During the stages of extraction of simple and composite terms, the methods receive the list of semantic groups generated in the first stage and go through the corpus, selecting the terms that are part of at least one group present on the entry list. The tool offers four relevance measurement options: FR, *tf-idf* (Manning et al. 1999), NC-Value and C-Value (Frantzi et al. 1998). The extracted terms are organized in decreasing order based on the results of the application of these measurements, and the ontology engineer can analyze and edit the final list of terms. It must be stressed that the analyses made in this article considered Relative Frequency (FR), which

looks at the number of times a term appears in a document divided by the total words in the document.

## Experiments

The corpus used in the experiments with the tool is made up of 283 texts in Portuguese extracted from the *Jornal de Pediatria*, a total of 785,448 words. On this corpus, the experiments described in Figure 4 were performed. Initially, the corpus was annotated by the PALAVRAS parser, which generated an XML representation, being converted to XCES type archives. This annotated corpus in XCES format was read by the OntoLP plug-in. On this basis, the extraction of composite terms followed the stages:

- extraction of semantic groups – through manual analysis, the specialist excludes the semantic groups that are considered not relevant, the semantic groups are created with the semantic tagging of PALAVRAS and are verified by the user with tag clouds;

- extraction of simple terms; and

- extraction of composite terms where the methods were analyzed: n-grams, morphosyntactic patterns and the nominal syntagm.

In this article, we compare the results obtained with and without extraction of semantic groups. In extraction without exclusion of semantic groups, all terms are considered in the computing of later stages.

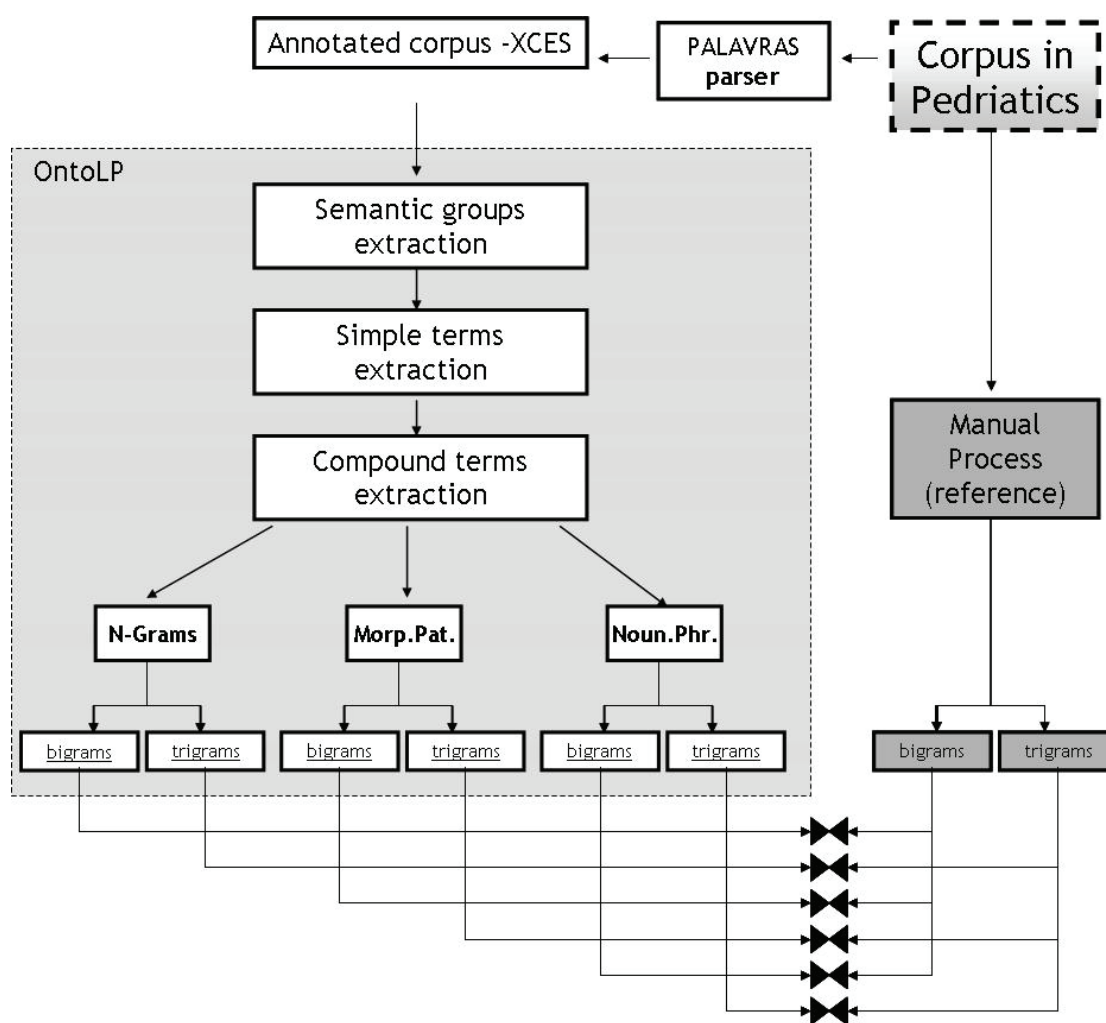


Figure 4 - Methodology used in the experiments.

The stages of extraction of simple and composite terms are reproduced for the two groups of identified terms (with or without exclusion of semantic groups) in identical manner, with the process ending in six bigram lists and six trigram lists. Each of these was compared with the reference lists of bigrams and trigrams. The reference lists were constructed by a process strongly

supported by manual tasks, performed by the TEXTQUIM/TERMISSUL group at the Federal University of Rio Grande do Sul (TEXTQUIM/UFRGS, <<http://www.ufrgs.br/textquim>>). The work of extracting terms present in the pediatrics corpus aimed to elaborate a glossary for support to translation students. This material also furnishes the items of a Recurring Expressions Catalogue

in pediatrics. The glossary and catalogue, designed as online resources for long-distance education, aim to help qualify translators and reviewers of pediatrics texts.

In generating the reference lists, initially n-grams with more than 5 occurrences in the corpus, extracted automatically, were considered. Based on this list of 36,741 n-grams, an automatic filtering process based on heuristics was begun. For example, terms that began or ended with prepositions were transformed by excluding these prepositions; n-grams contained in larger n-grams were excluded. In this way, a bigram that appeared in a trigram was discarded, because, for the purposes of translation learning, terms with a greater number of words are preferable to smaller terms. For example, “exclusive maternal breastfeeding” appears as a tri-gram, therefore “maternal breastfeeding” does not appear in the list of bigrams. The process resulted in a list with 3,645 n-grams. This list was manually checked by translation students, resulting in a final list of 2,407 terms, with 1,293 bigrams, 775 trigrams and 339 terms made up of more than 3 words.

The lists obtained with the OntoLP tool were compared with the reference lists using the following metrics: precision (P), scope (S) and f-measure (F). Precision (P) indicates the method’s capacity to identify the correct terms, considering the reference list, and is calculated by formula (1).

$$P = (\text{Reference Terms} \cap \text{Extracted Terms}) / \text{Extracted Terms} \quad (1)$$

Scope (assesses the quantity of correct terms extracted by the method and is calculated through formula (2)).

$$S = (\text{Reference Terms} \cap \text{Extracted Terms}) / \text{Reference Terms} \quad (2)$$

The f-measure (F) is the harmonic measurement between precision and scope, and is given by formula (3).

$$F = (2 * P * S) / (P + S) \quad (3)$$

Table 1 presents the total number of terms found in the experiments for each analysis made. In addition, it shows how many of these terms are present in the reference list that has a total of 1,293 bigrams and 775 trigrams. The number of recovered terms is much higher than the number of reference terms, as all terms extracted from the corpus are considered, without using a frequency cut point. Obviously, this number decreases as semantic groups are excluded. In this case, as the proportion of excluded not relevant terms is greater than that of relevant terms, an increase in precision is seen.

**Table 1 - Extracted terms and reference lists**

With the exclusion of semantic groups	n-grams		Morphosyntactic patterns		Nominal syntagm	
	Bi-grams <i>bi nG</i>	Tri-grams <i>tri nG</i>	Bi-grams <i>bi PM</i>	Tri-grams <i>tri PM</i>	Bi-grams <i>bi SN</i>	Tri-grams <i>tri SN</i>
Total Terms	13,115	18,554	27,763	27,322	8,926	5,959
Terms present in the reference	610	407	636	406	588	283
Without exclusion of semantic groups	n-grams		Morphosyntactic patterns		Nominal syntagm	
	Bi-grams <i>bi nG</i>	Tri-grams <i>tri nG</i>	Bi-grams <i>bi PM</i>	Tri-grams <i>tri PM</i>	Bi-grams <i>bi SN</i>	Tri-grams <i>tri SN</i>
Total Terms	18,325	23,588	33,276	30,497	12,691	7,509
Terms present in the reference	769	451	780	441	740	311

The final assessment metrics are presented in Figures 5 and 6. The method that presents the best balance between precision and scope (both for bigrams and trigrams) is extraction by nominal syntagms, with exclusions of terms by semantic groups (f-measure of 11.51% and 8.41% for bigrams and trigrams, respectively). These are the methods that employ a greater level of linguistic processing, both for syntactic

processing and semantic processing. This observation indicates that linguistic pre-processing of the corpus tends to positively contribute to the extraction of composite terms. However, we stress that semantic information is employed in a semi-automatic way; the groups are presented to the domain specialist, who indicates the semantic groups not to be considered in the process.

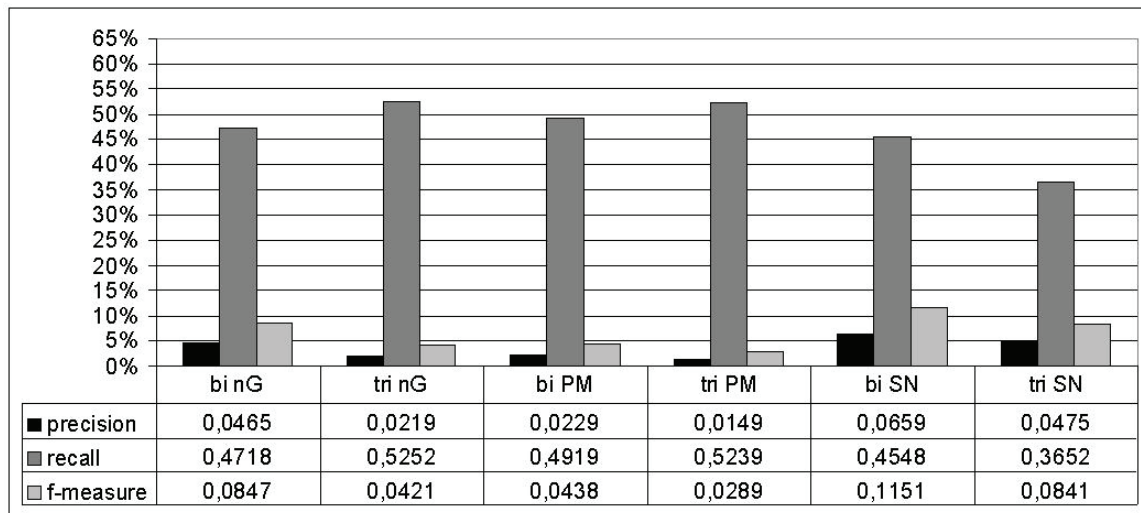


Figure 5 - Graph of metrics with exclusion of semantic groups.

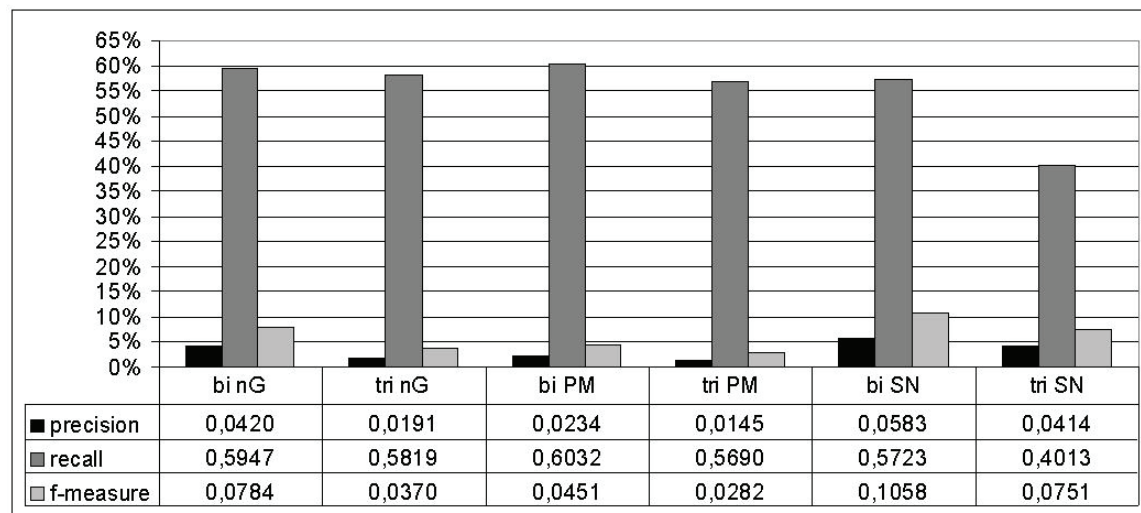


Figure 6 - Graph of metrics without exclusion of semantic groups.

The scope reached only 61% of the reference terms, that is, a significant number of the reference terms were not found by any of the methods employed. This aspect must still be investigated.

In the appendixes (1-4) are listed the first terms of the extracted lists, ordered by frequency. The terms that appear in the reference list have been highlighted in bold. It can be seen that some apparently relevant terms, found by the methods, do not appear in the reference, for example, “gestation age”, “risk factor” and “maternal breastfeeding”. In some cases, relevant bigrams were not found in the reference, for this does not include sub-terms. The bigram “maternal breastfeeding”, for example, is absent from the reference list, as it is a sub-term of the trigram “exclusive maternal breastfeeding”. Facts like these suggest the need for reviewing the assessment method, by refining the reference, as the purposes of this work differ from those for translation

aid. On the other hand, we stress the importance of this resource, as the area of ontology learning, being very recent, still faces serious problems due to a shortage of resources for assessment. Assessments, while difficult, are crucial for developing techniques. Assessments in this area, as well as dealing with a shortage of resources, face problems related to subjectivity. It is common for different specialists to have different judgments about the relevance of terms.

## Related works

In Cimiano (2006) the problem of extracting text ontologies, which has many research questions still open, is dealt with fairly comprehensively.

Suchanek (2006) discusses in a general manner the use of linguistic analysis in the extraction of information from text bases. In particular, the extraction of composite



terms is quite investigated. Ramisch et al. (2008) is an example of recent work in this line. The purpose for investigating composite terms, however, varies. They are not always considered as part of the collection of concepts in a domain.

There is little earlier work relative to Portuguese. Baségio (2006), for example, presents a first approach for the problem of extracting text ontologies based on a domain corpus. Baségio's approach was assessed for the tourism domain by specialists, without a reference list. In this form of assessment, however, it is not possible to calculate metrics such as scope and f-measure.

Ribeiro and Vieira (2008) assess the Plug-in OntoLP methods in an ecology corpus with respect to the first 1000 terms extracted by each method. The impact of extracting semantic groups is assessed in a set of 150 terms. In this work we present an assessment of the methods based on a more extensive reference list, in another domain, pediatrics, and the assessments consider the total set of extracted terms.

The fact that the ontology learning area is very recent makes it difficult to present a comparative analysis with other works, since there are no standard tests available yet.

## Conclusion

Here we present an initial assessment of Natural Language Processing techniques applied to the problem of ontology construction. The experiments carried out are related to the first stage of the complex process of constructing ontologies, that is, the stage of identifying the terms that could be concepts. The experiments carried out consider a corpus and a reference list of relevant constructions for teaching the translation of medicine/pediatrics texts. This in an area which, among others, can benefit from the development of text processing and information structuring techniques, as there is much acquired specific knowledge registered in text.

Although they are preliminary, the results can be used to observe the behavior of the different methods of extraction employed. The linguistically-informed methods showed advantages in relation to the less informed methods.

In future work, we can refine both the list of terms used as reference and also the term extraction process. As the ontology engineer receives a list of terms ordered by frequency, an important assessment to be made is the analysis of the precision of the most frequent terms; it will also be important to assess how the balance between precision and scope develops according to the cut point. This work is already in progress. It is important to identify a useful balance between precision/scope, that is, one that can contribute positively to the ontology engineer.

In addition, we plan to advance in the other stages of ontology construction. For this, at first we will work with the semantic grouping of expressions, identifying hierarchies and similarities among the terms.

The techniques assessed in this article have been incorporated to the ontology editor Protégé via a plug-in. The plug-in, along with other resources for developing ontology research, is available at <<http://www.inf.pucrs.br/~ontolp>>.

## Acknowledgements

We thank the financing that Capes, CNPq and Sead/UFRGS have granted to the authors of this work.

## Bibliographic references

BACELAR, S.; GALVÃO, C. C.; ALVES, E.; TUBINO, P. Expressões médicas: falhas e acertos. *Revista Brasileira de Cirurgia Cardiovascular*, São Paulo, v. 18 n. 3, Jul/Set, 2003.

BASÉGIO, T. *Uma Abordagem Semi-Automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil*. 2006. Dissertação (Mestrado em Ciência da Computação), Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, Porto Alegre.

BICK, E. *The parsing System "Palavras": Automatic grammatical analysis of portuguese in a constraint grammar framework*. 2000. PhD thesis, Aarhus University.

BLANK, D.; ROSA, L. O.; GURGEL, R. Q.; GOLDANI, M. Z. Produção brasileira de conhecimento no campo da saúde da criança e do adolescente. *Jornal de Pediatria*, Rio de Janeiro, v. 82, n. 2, p. 97-102, 2006.

BREWSTER, C.; CIRAVEGNA, F.; WILKS, Y. Background and foreground knowledge in dynamic ontology construction. In: SIGI, PROCEEDINGS OF THE SEMANTIC WEB WORKSHOP, 2003, Toronto. *Proceedings*. Toronto: August, 2003.

BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: An overview. In: BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. (Ed.). *Ontology learning from text: Methods, evaluation and applications*, v. 123 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2005.

CIMIANO, P. *Ontology learning and population from text: Algorithms, evaluation and applications*. Heidelberg: Springer-Verlag, 2006.

FRANTIZI, K. T.; ANANIADOU, S.; ICHI TSUJII, J. The c-value/nc-value method of automatic recognition for multi-word terms. In: ECDL'98: PROCEEDINGS OF THE SECOND EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, 1998, London. *Proceedings*. Heidelberg: Springer-Verlag, 1998, p. 585-604.

GENNARI, J. et al. *The evolution of protégé: an environment for knowledge-based systems development*. 2002. Technical Report SMI-2002-0943.

GOMEZ-PEREZ, A.; CORCHO, O.; FERNANDEZ-LOPEZ, M. **Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web**. Heidelberg: Springer-Verlag, 2004.

IDE, N.; BONHOMME, P.; ROMARY, L. Xces: An xml-based encoding standart for linguistic corpora. In: PROCEEDINGS OF THE SECOND INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 2000. **Proceedings**. Paris: European Language Resources Association, 2000.

MANNING, C. D. SCHUTZE, H. **Foundations of statistical natural language processing**. Cambridge, Massachusetts: The MIT Press, 1999.

MCGUINNESS, D.L. VAN HARMELEN, F. **OWL web ontology language overview**. World Wide Web Consortium (W3C) recommendation. <<http://www.w3.org/TR/owl-features>>. Acesso em: 01 fev. 2004.

PROTÉGÉ <<http://protege.stanford.edu>> Acesso em: 25 ago. 2008.


RAMISCH, C. ; SCHREINER, P. ; IDIART, M. ; VIL-LAVICENCIO, A. An Evaluation of Methods for the

Extraction of Multiword Expressions. In: LREC 2008 MWE WORKSHOP: TOWARDS A SHARED TASK OR MULTIWORD EXPRESSIONS, Marrakesh, 2008. **Proceedings**. Paris: European Language Resources Association, 2008.

SUCHANEK, F. M.; IFRIM, G.; ANDWEIKUM, G. Leila: Learning to extract information by linguistic

analysis. In: PROCEEDINGS OF THE 2ND WORKSHOP ON ONTOLOGY LEARNING AND POPULATION: BRIDGING THE GAP BETWEEN TEXT AND KNOWLEDGE, Sydney, Australia, 2006. **Proceedings**. Association for Computational Linguistics, 2006.

RIBEIRO, L.C. **OntoLP: Construção semi-automática de ontologias a partir de textos da língua portuguesa**. 2008. Dissertação (Mestrado em Computação Aplicada), Universidade do Vale do Rio dos Sinos - UNISINOS, São Leopoldo.

RIBEIRO, L.C.; VIEIRA, R. OntoLP: Engenharia de Ontologias em Língua Portuguesa. In: ANAIS DO XXVIII CONGRESSO DA SBC - SEMISH - SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE, Belém do Pará, 2008. **Anais**. Porto Alegre: Sociedade Brasileira de Computação, 2008. 

## About the authors

### *Lucelene Lopes*

PhD student from the PUCRS Computer Sciences Course in 2008. She has a Master's in Health Technology from PUCPR (2007). She graduated in Sciences, with Full Qualification in Mathematics, from UNIVALE (2000). She has worked mainly with Artificial Intelligence since her master's and more recently, with the start of her doctorate, she has been working with term extraction within the Natural Language Processing area.

### *Renata Vieira*

Has a PhD in Computer Science from the University of Edinburgh (1998). She is a professor at PUCRS, where she does research and teaches undergraduate and postgraduate courses in artificial intelligence, with an emphasis on natural language processing, knowledge representation, ontologies, agents and the semantic web. She has experience in project coordination and is a member of the scientific committees of the main international conferences in the computing intelligence and intelligent agents area. She participates actively in the development of Brazil's natural language processing area.

## Appendix 1

This Appendix presents the 70 first terms identified with each method (bigrams) **with** exclusion of terms. The extracted terms which appear in the reference have been highlighted.

### NG Bigrams

maternal breastfeeding, gestation age, mechanical ventilation, age range, intensive therapy, arterial hypertension, low weight, control group, **statistical difference**, cephalic perimeter, large number, neonatal period, high risk, bone mass, **physical examination**, vitamin d, **seric level**, great risk, exclusive breastfeeding, viral load, **urinary infection**, **significant difference**, great frequency, statistical significance, **low stature**, renal scar, adrenal insufficiency, **otitis media**, septic shock, public health, **differential diagnosis**, respiratory insufficiency, logistical regression, plasmatic level, **clinical practice**, neurological evolution, saline solution, **clinical picture**, pulmonary ventilation, by mouth, cystic fibrosis, **lower age**, muscle relaxation, first month, great incidence, great prevalence, **endotracheal tube**, **respiratory frequency**, falciform anemia, free diet, maternal schooling, **clinical assessment**, child obesity, respiratory discomfort, abdominal pain, **z score**, bladder dysfunction, hearing loss, pulmonary hypertension, I degree, **clinical score**, **clinical development**, **pulmonary deposition**, intracranial pressure, hospital discharge, brachial perimeter, acute phase, average time, sixth month, long term

### MP Bigrams

maternal breastfeeding, age range, gestation age, mechanical ventilation, young child, intensive therapy, arterial hypertension, significant difference, cephalic perimeter, renal scar, **otitis media**, **collateral effect**, neonatal period, older child, pediatric patient, **seric level**, **physical examination**, urinary infection, **clinical manifestation**, bone mass, exclusive breastfeeding, **adverse effect**, viral load, statistical significance, **clinical picture**, septic shock, saline solution, clinical test, public health, respiratory insufficiency, physical activity, **chronic disease**, logical regression, plasmatic level, **clinical practice**, neurological development, adverse event, **clinical assessment**, by mouth, **respiratory frequency**, **clinical score**, pulmonary ventilation, **lower age**, I degree, abdominal pain, muscle relaxation, cystic fibrosis, **endotracheal tube**, **diagnostic criteria**, **congenital cardiopathy**, **respiratory infection**, complementary examination, falciform anemia, intracranial pressure, **scientific evidence**, maternal schooling, hearing loss, respiratory discomfort, pulmonary hypertension, laboratory examination, child obesity, **clinical development**, average time, congenital infection, bladder dysfunction, **prognostic factor**, **current volume**, carrier patient, **food allergy**, pulmonary disease

## NS Bigrams

maternal breastfeeding, gestation age, mechanical ventilation, neonatal period, arterial hypertension, cephalic perimeter, renal scar, intensive therapy, high risk, young child, low weight, **physical examination**, bone mass, **significant difference**, statistical significance, exclusive breastfeeding, pediatric patient, public health, **urinary infection**, septic shock, older child, **low stature**, last year, saline solution, by mouth, **endotracheal tube**, cystic fibrosis, falciform anemia, great frequency, **congenital cardiopathy**, child obesity, **clinical practice**, **collateral effect**, **chronic disease**, intracranial pressure, maternal schooling, congenital infection, hospital discharge, **food allergy**, **clinical manifestation**, complementary examination, **otitis media**, **adverse effect**, **cardiac frequency**, neurological development, pulmonary disease, bladder dysfunction, **differential diagnosis**, two patients, **respiratory frequency**, child studied, abdominal pain, **exclusive breastfeeding**, logistical regression, **clinical picture**, pulmonary hypertension, age range, in general, good result, viral load, middle ear, school age, long term, high frequency, adverse event, pulmonary ventilation, short duration, sample size, present work, four patients

## Appendix 2

This Appendix presents the first 70 terms identified with each method (trigram) **with** exclusion of terms. The extracted terms which appear in the reference have been highlighted.\*

### NG Trigrams

year of age, risk factor, month of life, year of life, month of age, weight at birth, day of life, **exclusive maternal breastfeeding**, therapy unit, **confidence interval**, significance level, quality of life, developing country, health care service, week of life, significant statistical difference, hour of life, inclusion criteria, growth rate, **growth curve**, data collection, significant statistical difference, health problem, **type of birth**, age average, **mortality rate**, term of consent, hospitalization time, **risk group**, medication consumption, milk production, 95 confidence, **exclusion diet**, labor, patient number, attention deficit, **oxygen saturation**, lifestyle, prevalence of asthma, need for ventilation, **intensive neonatal therapy**, antibiotic use, bone mineral density, child aged, **infection risk**, patient with disease, **pediatric age range**, free from disease, milk volume, vitamin plasmatic, **regression model**, **generalized anxiety disorder**, use on children, mass index, drug use, **calcium ingestion**, **thorax x-ray**, time period, non-invasive ventilation, **Down's syndrome**, health care team, **cut point**, **hypertonic saline solution**, oxygen use, medication use, variance analysis, medication use, **mechanical pulmonary ventilation**, level of schooling, water seal

## MP Trigrams

year of age, risk factor, month of life, year of life, month of age, weight at birth, health care professional, day of life, patient group, therapy unit, **confidence interval**, quality of life, significance level, developing country, child group, health care service, week of life, hour of life, inclusion criteria, majority of patients, growth rate, health problem, data collection, **growth curve**, sample size, **type of birth**, patient number, average age, **mortality rate**, term of consent, majority of cases, hospitalization time, **risk group**, **cut point**, group patient, medication consumption, milk production, symptom beginning, **exclusion diet**, labor, patient with disease, child's life, attention deficit, antibiotic use, **oxygen saturation**, lifestyle, child aged, increased blood pressure, need for ventilation, breastfeeding time, vaccine application, milk volume, prevalence of asthma, child's health, **infection risk**, majority of children, use in children, **regression model**, **generalized anxiety disorder**, child's age, child's sex, **calcium ingestion**, mass index, **thorax x-ray**, time period, drug use, **complaint time**, oxygen use, medication use, start in childhood

## NS Trigrams

risk factor, **exclusive maternal breastfeeding**, health care professional, weight at birth, inclusion criteria, data collection, health care service, **type of birth**, hospitalization time, developing country, **pediatric age range**, labor, **intensive neonatal therapy**, **mechanical pulmonary ventilation**, antibiotic use, milk production, **confidence interval**, **oxygen saturation**, growth rate, **acute otitis media**, intensive pediatric therapy, quality of life, oxygen use, childbirth room, **exclusion diet**, significance level, non-invasive ventilation, lifestyle, health care team, medication consumption, variance analysis, **thorax x-ray**, transport accident, **risk group**, **bran**, water seal, term of consent, **inadequate muscle relaxation**, medication use, medicine use, **base disease**, bone mineral density, **calcium supplement**, significant statistical difference, **mortality rate**, exclusion criteria, **flu vaccine**, **reference center**, **separation anxiety**, lymphocyte level, vitamin deficiency, older child, **complaint time**, **growth curve**, schizophrenia beginning, air escape, stress situation, **abstinence syndrome**, acute respiratory infection, **blood sample**, **ventilation tube**, **high digestive hemorrhage**, nullity hypothesis, health care center, **control without hepatopathy**, **infection risk**, **comforter use**, **flu vaccine**, **meconium aspiration**, workplace

## Appendix 3

This appendix presents the first 70 terms identified with each method (bigrams) **without** exclusion of terms. The extracted terms which appear in the reference have been highlighted.

## NG Bigrams

maternal breastfeeding, gestation age, maternal milk, age range, mechanical ventilation, present study, intensive therapy, arterial hypertension, human milk, low weight, first year, nutritional state, control group, **statistical difference**, cephalic perimeter, one year, bone mass, neonatal period, large number, high risk, 1 year, **z score**, **physical examination**, vitamin d, **seric level**, great risk, statistical analysis, viral load, exclusive breastfeeding, **urinary infection**, **significant difference**, great part, free diet, great frequency, statistical significance, renal scar, adrenal insufficiency, **low stature**, **otitis media**, family history, great time, septic shock, clinical picture, pulmonary lesion, public health, logistical regression, respiratory insufficiency, **differential diagnosis**, plasmatic level, intracranial pressure, **clinical practice**, neurological evolution, t test, populations studied, saline solution, **premature diagnosis**, pulmonary ventilation, by mouth, no medication, first month, **lower age**, cystic fibrosis, antenatal corticosteroid, **clinical assessment**, muscle relaxation, great incidence, great prevalence, falciform anemia, inflammatory response, **endotracheal tube**

## PM Bigrams

maternal breastfeeding, age range, gestation age, maternal milk, mechanical ventilation, young child, intensive therapy, arterial hypertension, human milk, nutritional study, **significant difference**, cephalic perimeter, renal scar, bone mass, **collateral effect**, complementary food, **otitis media**, neonatal period, statistical analysis, **seric level**, pediatric patient, older child, **physical examination**, **clinical manifestation**, **urinary infection**, **adverse effect**, exclusive breastfeeding, viral load, free diet, **clinical picture**, statistical significance, large turn, clinical test, family history, saline solution, septic shock, pulmonary lesion, population studied, intracranial pressure, physical activity, public health, respiratory insufficiency, **clinical practice**, **chronic disease**, logistical regression, plasmatic level, patient majority, cerebral lesion, neurological development, adverse event, **clinical assessment**, **respiratory frequency**, pulmonary ventilation, **premature diagnosis**, **clinical score**, by mouth, **lower age**, muscle relaxation, cystic fibrosis, abdominal pain, I degree, **endotracheal tube**, **diagnosis criteria**, **congenital cardiopathy**, inflammatory response, falciform anemia, **respiratory infection**, complementary examination.

## NS Bigrams

maternal breastfeeding, present study, gestation age, maternal milk, mechanical ventilation, two group, six months, human milk, 2 year, neonatal period, arterial hypertension, renal scar, cephalic perimeter, intensive therapy, complementary food, high risk table 2, bone mass, 5 year, nutritional state, one year, young child, low weight, **physical examination**, 1 year, population studied, **significant difference**, two year, five year, statistical analysis, statistical significance, table 3, exclusive breastfeeding, pediatric patient, last decade, table 1, 24

hour, public health, older child, intracranial pressure, septic shock, **urinary infection**, **low stature**, last year, six month, **endotracheal tube**, saline solution, by mouth, cystic fibrosis, falciform anemia, great frequency, antenatal corticosteroid, 12 month, sample studied, age range, **congenital cardiopathy**, 10 year, **collateral effect**, child obesity, four month, developing country, teenage mother, 6 year, **clinical practice**, **chronic disease**, 30 minute, maternal schooling, congenital infection, **adverse effect**, hospital discharge

## Appendix 4

This appendix presents the first 70 terms identified with each method (trigrams) **without** exclusion of terms. The extracted terms which appear in the reference have been highlighted. \*

### NG Trigrams

year of age, risk factor, month of life, year of life, month of age, **cow milk**, weight at birth, health care professional, day of life, **exclusive maternal breastfeeding**, therapy unit, patient group, **confidence level**, significance level, quality of life, developing country, week of life, health care service, significant statistical difference, hour of life, **cut point**, inclusion criteria, growth rate, soy polysaccharide, data collection, significant statistical difference, **growth curve**, health problem, milk-free, Student t, **type of birth**, child group, **mortality rate**, age average, term of consent, hospitalization time, **risk group**, 95 confidence, medication consumption, milk production, **exclusion diet**, labor, patient number, banana peel, attention deficit, **oxygen saturation**, lifestyle, need for ventilation, bone mineral density, child aged, antibiotic use, milk volume, **intensive neonatal therapy**, **infection risk**, prevalence of asthma, free from disease, patient with disease, **pediatric age range**, use in children, vitamin plasmatic, **generalized anxiety disorder**, **regression model**, **soy formula**, mass index, **thorax x-ray**, drug use, **calcium ingestion**, time period, Down's syndrome, **complaint time**

### MP Trigrams

year of age, risk factor, month of life, year of life, month of age, **cow milk**, weight at birth, health care professional, day of life, patient group, therapy unit,

**confidence level**, significance level, quality of life, child group, developing country, health care service, week of life, **cut point**, hour of life, inclusion criteria, patient majority, growth rate, health problem, **growth curve**, data collection, **type of birth**, sample size, patient number, majority of times, **mortality rate**, age average, majority of cases, term of consent, **risk group**, hospitalization time, child of mother, medication consumption, group patient, milk production, patient with disease, labor, medication consumption, group patient, milk production, sick patient, patient with disease, labor, symptom beginning, **exclusion diet**, child's life, majority of studies, child aged, prevalence of asthma, breastfeeding time, milk volume, vaccine application, need for ventilation, pressure increase, child's health, **infection risk**, majority of children, use in children, **generalized anxiety disorder**, **regression model**, child age, time period, mass index, **calcium ingestion**, child's sex, **thorax x-ray**

### NS Trigrams

risk factor, **exclusive maternal breastfeeding**, health care professional, **cow milk**, weight at birth, inclusion criteria, data collection, health care service, **type of birth**, developing country, hospitalization time, soy polysaccharide, **pediatric age range**, labor, **mechanical pulmonary ventilation**, **neonatal intensive therapy**, banana peel, antibiotic use, milk production, **oxygen saturation**, **confidence interval**, growth rate, quality of life, intensive pediatric therapy, **cut point**, **acute oitis media**, non-invasive ventilation, level of significance, medication consumption, **exclusion diet**, childbirth room, oxygen use, **acute pulmonary lesion**, lifestyle, health care team, transport accident, **thorax x-ray**, **risk group**, **bran**, variance analysis, term of consent, **soy formula**, exclusion criteria, bone mineral density, **calcium supplement**, **mortality rate**, medicine use, significant statistical difference, **base disease**, water seal, **inadequate muscle relaxation**, medication use, **reference center**, **flu vaccine**, 1970s, **separation anxiety**, vitamin deficiency, older child, cut study, air escape, lymphocyte level, schizophrenia beginning, **growth rate**, **complaint time**, **infection risk**, **meconium aspiration**, **comforter use**, **blood sample**, first 24 hours, **systemic inflammatory response**

\* Translator's note: Many of the terms which are trigrams in Portuguese are no longer trigrams once they have been translated into English.