Natural Language Processing System Applied in Public Health for Assessment of an Automatic Analysis of Patterns Generator

Anabel Fraga, Juan Llorens, Valeria Rodríguez and Valentin Moreno
Computer Science Department, Carlos III of Madrid University, Av. Universidad 30, Leganés, Madrid, Spain

Keywords: Indexing, Ontologies, Knowledge, Patterns, Reuse, Retrieval, Public Health, Oncology, Genetics.

Abstract: Nowadays, there are many scientific articles referring to any topic like medicine, technology, economics, finance, and so on. These articles are better known as papers, they represent the evaluation and interpretation of different arguments, showing results of scientific interest. At the end, most of these are published in magazines, books, journals, etc. Due to the fact that these papers are created with a higher frequency it is feasible to analyse how people write in the same domain. At the level of structure and with the help of graphs some of the results that can be found are: groups of words that are used (to determine if they come from a specific vocabulary), most common grammatical categories, most repeated words in a domain, patterns found, and frequency of patterns found. This research has been created to fulfil these needs. A domain of public health has been selected and it is composed of 800 papers about different topics referring to genetics such as mutations, genetic deafness, DNA, trinucleotide, suppressor genes, among others; and an ontology of public health has been used to provide the basis of the study.

1 INTRODUCTION

There are many scientific articles referring to any topic like medicine, technology, economics, and finance, among others. Christopher Manning states in his book that: “People write and say lots of different things, but the way people say things - even in drunken casual conversation - has some structure and regularity.”(Manning, 1999)

The important aspect in here is to ask ourselves: how do people write? Nowadays, researchers conduct investigations using natural language processing tools, generating indexing and semantic patterns that help to understand the structure and relation of how writers communicate through their papers.

This project will use a natural language processing system which will analyze a corpus of papers acquired by Ramon and Cajal Hospital from Madrid. The documents will be processed by the system and will generate simple and composed patterns. These patterns will give us different results which we can analyze and conclude the common aspects the documents have even though they are created by different authors but are related to the same topic. (Alonso et al., 2005) The study uses as center of the study an ontology created in a national founded project for Oncology and it has been extended with general terms of public health.

The reminder of this paper is as follows: a state of the art of the main topics to deal with in the paper, section 3 includes the summary of the information processing; section 4 summarizes the results, and finally conclusions.

2 STATE OF THE ART

A. Information Reuse

Reuse in software engineering is present throughout the project life cycle, from the conceptual level to the definition and coding requirements. This concept is feasible to improve the quality and optimization of the project development, but it has difficulties in standardization of components and combination of features. Also, the software engineering discipline is constantly changing and updating, which quickly turns obsolete the reusable components (Llorens, 1996).

At the stage of system requirements reuse is implemented in templates to manage knowledge in a higher level of abstraction, providing advantages

Fraga, A., Llorens, J., Rodríguez, V. and Moreno, V..
Natural Language Processing System Applied in Public Health for Assessment of an Automatic Analysis of Patterns Generator.
In *Proceedings of the 6th International Workshop on Software Knowledge (SKY 2015)*, pages 31-38
ISBN: 978-989-758-162-5

over lower levels and improving the quality of the project development. The patterns are fundamental reuse components that identify common characteristics between elements of a domain and can be incorporated into models or defined structures that can represent the knowledge in a better way.

B. Natural Language Processing

The need for implementing Natural Language Processing techniques arises in the field of the human-machine interaction through many cases such as text mining, information extraction, language recognition, language translation, and text generation, fields that requires a lexical, syntactic and semantic analysis to be recognized by a computer (Cowie et al., 2000). The natural language processing consists of several stages which take into account the different techniques of analysis and classification supported by the current computer systems (Dale, 2000).

1) Tokenization: The tokenization corresponds to a previous step on the analysis of the natural language processing, and its objective is to demarcate words by their sequences of characters grouped by their dependencies, using separators such as spaces and punctuation (Moreno, 2009). Tokens are items that are standardized to improve their analysis and to simplify ambiguities in vocabulary and verbal tenses.
2) Lexical Analysis: Lexical analysis aims to obtain standard tags for each word or token through a study that identifies the turning of vocabulary, such as gender, number and verbal irregularities of the candidate words. An efficient way to perform this analysis is by using a finite automaton that takes a repository of terms, relationships and equivalences between terms to make a conversion of a token to a standard format (Hopcroft et al., 1979). There are several additional approaches that use decision trees and unification of the databases for the lexical analysis but this not covered for this project implementation (Trivino et al., 2000).
3) Syntactic Analysis: The goal of syntactic analysis is to explain the syntactic relations of texts to help a subsequent semantic interpretation (Martí et al., 2002), and thus using the relationships between terms in a proper context for an adequate normalization and standardization of terms. To incorporate lexical and syntactic analysis, in this project were used deductive techniques of standardization of terms that convert texts from a context defined by sentences through a special function or finite automata.
4) Grammatical Tagging: Tagging is the process of assigning grammatical categories to terms of a text or corpus. Tags are defined into a dictionary of standard terms linked to grammatical categories (nouns, verbs, adverb, etc.), so it is important to normalize the terms before the tagging to avoid the use of non-standard terms. The most common issues of this process are about systems' poor performance (based on large corpus size), the identification of unknown terms for the dictionary, and ambiguities of words (same syntax but different meaning) (Weischedel et al., 2006). Grammatical tagging is a key factor in the identification and generation of semantic index patterns, in where the patterns consist of categories not the terms themselves. The accuracy of this technique through the texts depends on the completeness and richness of the dictionary of grammatical tags.
5) Semantic and Pragmatic Analysis: Semantic analysis aims to interpret the meaning of expressions, after on the results of the lexical and syntactic analysis. This analysis not only considers the semantics of the analyzed term, but also considers the semantics of the contiguous terms within the same context. Automatic generation of index patterns at this stage and for this project does not consider the pragmatic analysis.

C. RSHP Model

RSHP is a model of information representation based on relationships that handles all types of artifacts (models, texts, codes, databases, etc.) using a same scheme. This model is used to store and link generated pattern lists to subsequently analyze them using specialized tools for knowledge representation (Llorens et al., 2004). Within the Knowledge Reuse Group at the University Carlos III of Madrid RSHP model is used for projects relevant to natural language processing. (Gomez-Perez et al., 2004); (Thomason, 2012); (Amsler, 1981);(Suarez et al., 2013) The information model is presented in Figure 1. An analysis of sentences and basic patterns are shown in Figure 2.

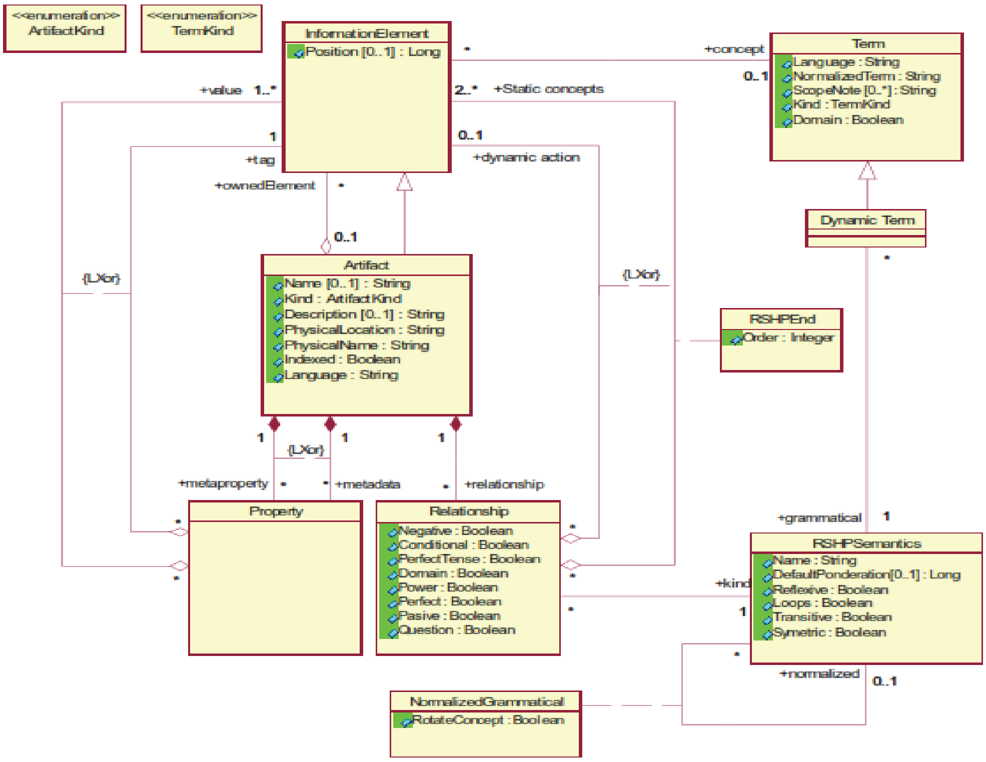


Figure 1: RSHP information representation model. (Alonso et al., 2005).

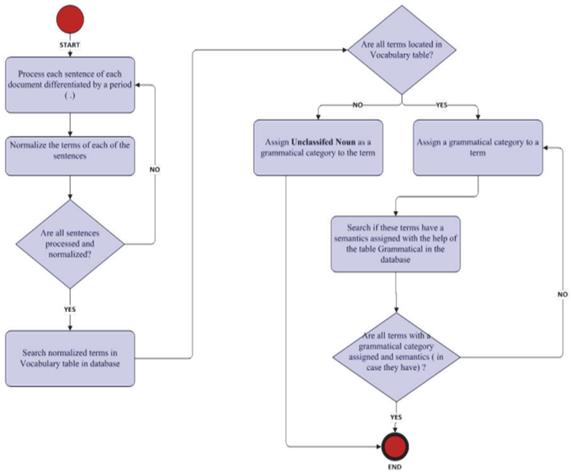

START
Process each sentence of each document differentiated by a period (.)
Normalize the terms of each of the sentences
Are all sentences processed and normalized?
NO
YES
Search normalized terms in Vocabulary table in database
Are all terms located in Vocabulary table?
NO
YES
Assign Unclassifed Noun as a grammatical category to the term
Assign a grammatical category to a term
Search if these terms have a semantics assigned with the help of the table Grammatical in the database
Are all terms with a grammatical category assigned and semantics (in case they have) ?
NO
YES
END

Figure 2: Analysis of sentences and basic patterns.

3 INFORMATION PROCESS FOLLOWED

Several steps have been done to start analyzing the set of documents (known as Corpus). These are mentioned below:

1. Choose the documents to analyze out of the eight hundred. These have been picked randomly.
2. Due to the fact that the system for analyzing text and discovery patterns (Suarez et al., 2013) only analyzes text documents, these papers had to be converted since they were given in a PDF format. The conversion of these documents has been manual. Headers, footers, page numbers, references have been removed because the tool analyzes sentences of each document and the ones that have been avoided are not relevant at the time of the analysis.
3. After the documents have been inserted in the tool, you proceed to create basic patterns and patterns with all the fifty text documents. The minimum of frequency can be changed.
4. Different scenarios will be used in order to analyze the different results and have a final conclusion. These scenarios will be described below.

The scenarios selected for the study are as follows:

Scenario I: Use all grammatical categories with a minimum frequency of 1 and without a difference in semantics

This scenario has the following characteristics:

1. The basic patterns for the text documents in the pattern analysis system (Figure 3) were generated in this scenario.

2. Generate all patterns for these documents using all grammatical categories located in Create patterns tab in pattern analysis system.
3. Minimum of frequency (Figure 4) to create a pattern
4. is 1.

Scenario II: Use all grammatical categories with a minimum frequency of 1 and differ patterns by its semantics

This scenario has the following characteristics:

1. Since basic patterns were already created in scenario one, it is not necessary to create them again because basic patterns analyze each of the documents and with the result of these is how patterns can be created. In this case the steps shown in Scenario I (step 2 and 3) are the same.

Scenario III: Use all grammatical categories with a minimum of frequency of 5 and differ patterns by its semantics

This scenario has the following characteristics

1. In this case the steps shown in Scenario I (step 2 and 3) are the same, but applied to a frequency of 5.

Scenario IV: Use all grammatical categories with a minimum frequency of 10 and differ patterns by its semantics

This scenario has the following characteristics

1. In this case the steps shown in Scenario I (step 2 and 3) are the same, but applied to a frequency of 10.

Scenario V: Use all grammatical categories with a minimum frequency of 20 and differ patterns by its semantics

This scenario has the following characteristics

1. In this case the steps shown in Scenario I (step 2 and 3) are the same, but applied to a frequency of 20.

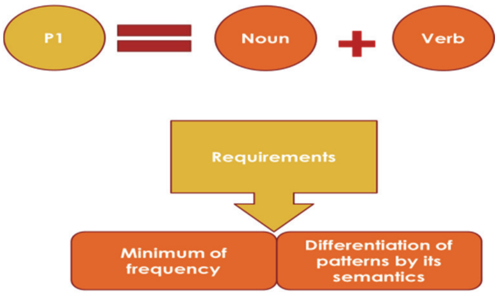


Figure 3: Frequency of patterns.

Table 1: Patterns created in all scenarios.

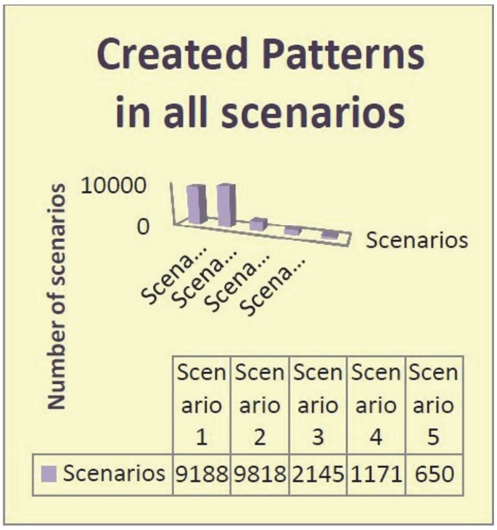
Number of patterns created				
S 1	S 2	S 3	S 4	S 5
9188	9818	2145	1171	650

The frequency pattern creation is showed in Figure 5. The objective of this paper is to discuss the results comparing the scenarios, but in an extended version a detailed presentation of each scenario might be of interest to the audience.

4 RESULTS

Basic Patterns: After the basic patterns were created, all the sentences from the text documents were analyzed and to each of the words (known in the database as token text) a termtag or syntactic tag was assigned with the help of the tables Rules Families and Vocabulary in the Requirements Classification database.

You may find the most repeated words in the domain of documents in the Basic patterns table. The most repeated words in grammatical categories such as nouns, verbs and nouns coming from the ontology we used.



	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
▨ Scenarios	9188	9818	2145	1171	650

Figure 4: Created patterns in all scenarios.

Patterns: The name of patterns is different depending on the requirements you use when you generate them. In this case there are 5 scenarios which have different results because each of them had different characteristics. When the minimum frequency to create patterns is higher the patterns will be less.

Scenario 1 and 2 have the same minimum of frequency (1) but the difference is that scenario 2 has the differentiation of patterns by its semantics activated; the result of differentiating is that more patterns are created due to the fact that there are patterns with distinct identifiers. There has been the decision to use this option for the rest of scenarios (scenarios 3, 4, 5), this way we can analyze the maximum number of patterns created and also the semantics to each of them can be easier to understand.

Patterns Created with Same Termtags: Among all the scenarios there has been one pattern in common. This is composed of two same termtags on the left and right side. This pattern has the same identifier and name for all scenarios it is pattern P1.

Table 2: Patterns with same termtag.

Pattern Name	**Term Tag Left**	**Term Tag Right**
P1	Unclassified noun	Unclassified noun

Unclassified nouns are the most common termtags in all the text documents used. Some words that are unclassified nouns are abbreviations, some words in another language, slang language, uncommon symbols, and scientific terms.

Patterns Created with Two Different Termtags: There are patterns which have a different termtags on the left and right side of a pattern. After comparing the results of the five scenarios it has been observed that the termtags with higher frequency are common between all scenarios for each side.

An example of a pattern generated presented as a binary tree is shown in Figure 5.

Table 3: Most common termtags.

Most common termtags in all scenarios	
Left Side	**Right Side**
Unclassified noun	Unclassified noun
Adverb	Adverb
Noun(ontology oncology)	Noun (ontology oncology)
Noun	Noun
Verb	Verb
Adverb	Adverb
Preposition	Preposition

For all scenarios the 20 patterns with higher frequency as termtags on each side have been shown in graphs. Also, it has been shown how these patterns are composed.

The composition of these patterns is common between all scenarios. The termtags for these patterns are unclassified nouns, noun, prepositions, verbs, nouns (oncology ontology), adverbs, adjectives. These termtags are the ones that also are the most common in all fifty text documents.

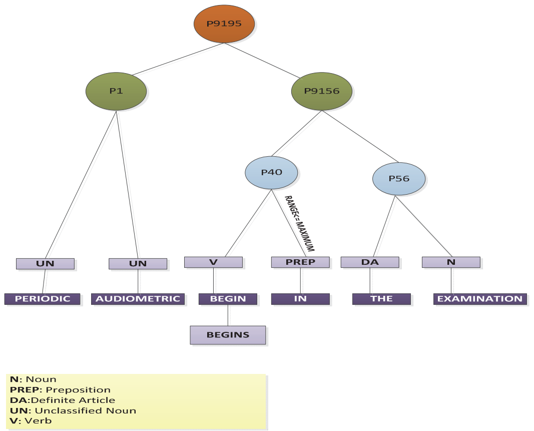


Figure 5: Binary tree representation.

Semantic Patterns: Some of the created patterns in all scenarios have semantics assigned on the left or right side, one side, or some patterns do not have.

In scenario one, there are patterns with different semantics but they have the same identifier and name, this is because the option to differentiate them by semantics was inactive. The rest of patterns for the other scenarios are not repeated and they are unique. Observing the results, the semantics for patterns in different scenarios is similar.

The change is noticeable in the pattern name and ids. In this case, they are different because in every scenario the number of patterns was different due to the minimum of frequency used. While the minimum of frequency is higher, the number of created patterns is less.

- The higher minimum of frequency used for this project has been 20.
- With the help of BoilerPlates the creation of patterns has been successful. Processing fifty documents at a time for basic patterns was not issue for the tool.
- 11% of the term names found in the fifty documents were from the ontology added.
- Writers about genetic deafness use a similar vocabulary and appropriate terms
- Studying patterns will facilitate the search of documents in search engines or databases.
- It can also assist the writing for any user that is not a researcher or scientist. With the help of patterns documents can be written.

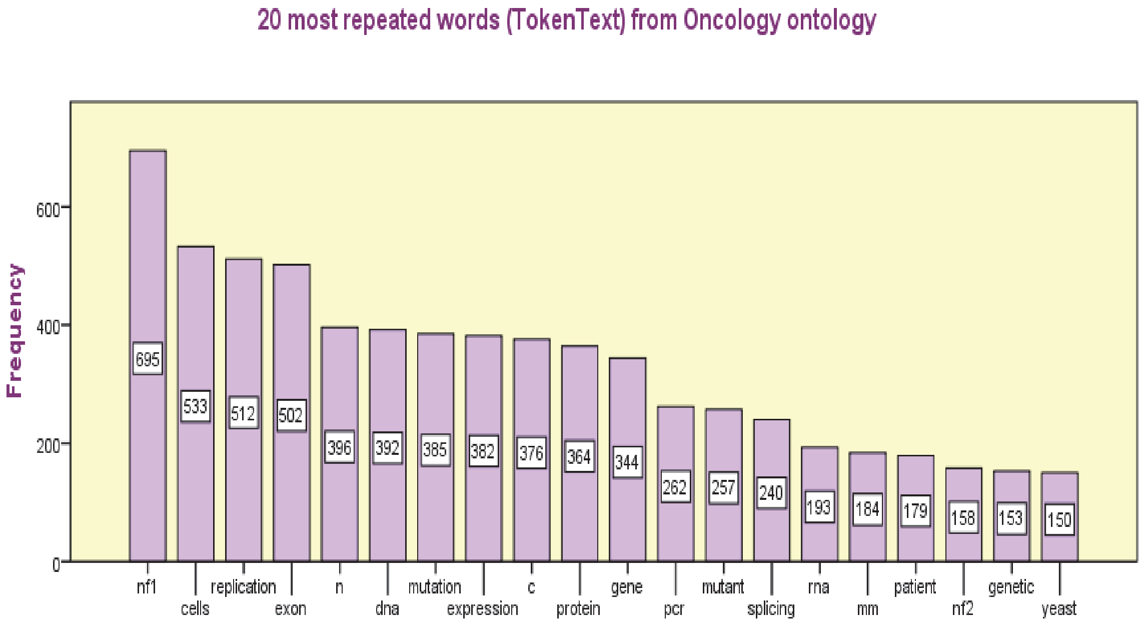


Figure 6: Most repeated words.

5 CONCLUSION

Using a domain of documents related to genetic engineering and making different scenarios to create patterns it has been possible to conclude the following:

1. While the minimum of frequency is higher, the number of created patterns is less.
2. The higher minimum of frequency used for this project has been 20.
3. With the help of BoilerPlates the creation of patterns has been successful. Processing fifty documents at a time for basic patterns was not issue for the tool.
4. 11% of the term names found in the fifty documents were from the ontology added.
5. Writers about genetic deafness use a similar vocabulary and appropriate terms
6. Studying patterns will facilitate the search of documents in search engines or databases.
   - It can also assist the writing for any user that is not a researcher or scientist. With the help of patterns documents can be written.

After ending all scenarios and analyzing results, we suggest some recommendations for this study:

- Expand the existing vocabulary table in the database. More ontologies can be included, symbols, slang languages, words in different language, among others.
- Differentiate patterns by its semantics to maximize the creation and different compositions of them.
- Using different minimum of frequencies at the moment of creating patterns will help to compare and analyze results

- Expand the existing vocabulary table in the database. More ontologies can be included, symbols, slang languages, words in different language, among others.
- Differentiate patterns by its semantics to maximize the creation and different compositions of them.
- Using different minimum of frequencies at the moment of creating patterns will help to compare and analyze results
- Expand the existing vocabulary table in the database. More ontologies can be included, symbols, slang languages, words in different language, among others. Differentiate patterns by its semantics to maximize the creation and different compositions of them.

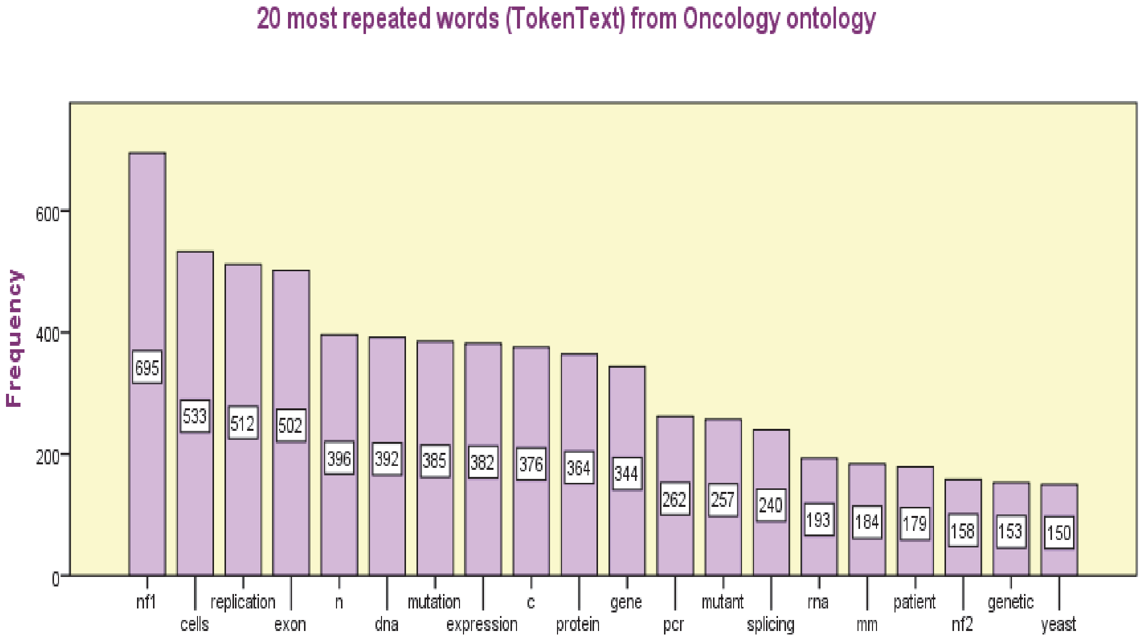


Figure 7: Most repeated words.

ACKNOWLEDGEMENTS

After ending all scenarios and analyzing results, we can conclude that authors writing papers about a same topic (in this case, genetic engineering) have similarity in how they write. They use a similar vocabulary and appropriate terms which makes the reading easier. Some additional enterprises need observation and intelligence.

Authors thank AGO2 project, founded by the Ministry of Education of Spain for aiding the author in the research and production of this paper.

The research leading to these results has received funding from the European Union's seventh framework program (FP7/2007-2013) for crystal – critical system engineering acceleration joint undertaking under grant agreement n° 332830 and from specific national programs and / or funding authorities.

REFERENCES

Abney, Steven. Part-of-Speech Tagging and Partial Parsing, S. young and G. Bloothooft (eds.) Corpus-Based Methods in Language and Speech Processing. An ELSNET book. Bluwey Academic Publishers, Dordrecht. 1997.

Alonso, Laura. Herramientas Libres para Procesamiento del Lenguaje Natural. Facultad de Matemática, Astronomía y Física. UNC, Córdoba, Argentina. 5tas Jornadas Regionales de Software Libre. 20 de noviembre de 2005. Available in: http://www.cs.famaf.unc.edu.ar/~laura/freeNLP

Amsler, R. A. A taxonomy for English nouns and verbs. Proceedings of the 19th annual Meeting of the Association for Computational Linguistic. Stanford, California, 1981. Pp. 133-138.

Carreras, Xavier. Márquez, Luis. Phrase recognition by filtering and ranking with perceptrons. En Proceedings of the 4th RANLP Conference, Borovets, Bulgaria, September 2003.

Cowie, Jim. Wilks, Yorick. Information Extraction. En DALE, R. (ed). Handbook of Natural Language Processing. New York: Marcel Dekker, 2000. Pp.241-260.

Dale, R. Symbolic Approaches to Natural Language Processing. En DALE, R (ed). Handbook of Natural Language Processing. New York: Marcel Dekker, 2000.

Gómez-Pérez, Asunción. Fernando-López, Mariano. Corcho, Oscar. Ontological Engineering. London: Springer, 2004.

Hopcroft, J. E. Ullman, J. D. Introduction to automata theory, languages and computations. Addison-Wesley, Reading, MA, United States. 1979.

Llorens, J., Morato, J., Genova, G. RSHP: An Information Representation Model Based on Relationships. In Ernesto Damiani, Lakhmi C. Jain, Mauro Madravio (Eds.), Soft Computing in Software Engineering (Studies in Fuzziness and Soft Computing Series, Vol. 159), Springer 2004, pp. 221-253.

Llorens, Juan. Definición de una Metodología y una Estructura de Repositorio orientadas a la

Reutilización: el Tesauro de Software. Universidad Carlos III. 1996.

Manning Christopher, "Foundations of Statistic Natural Language Processing", Cambridge University, England, 1999, 81

Martí, M. A. Llisterri, J. Tratamiento del lenguaje natural. Barcelona: Universitat de Barcelona, 2002. p. 207.

Moreno, Valentín. Representación del conocimiento de proyectos de software mediante técnicas automatizadas. Anteproyecto de Tesis Doctoral. Universidad Carlos III de Madrid. Marzo 2009.

Poesio, M. Semantic Analysis. En DALE, R. (ed). Handbook of Natural Language Processing. New York: Marcel Dekker, 2000.

Rehberg, C. P. Automatic Pattern Generation in Natural Language Processing. United States Patent. US 8,180,629 B2. May 15, 2012. January, 2010.

Riley, M. D. Some applications of tree-based modeling to speech and language indexing. Proceedings of the DARPA Speech and Natural Language Workshop. California: Morgan Kaufmann, 1989. Pp. 339-352.

Suarez, P., Moreno, V., Fraga, A., Llorens, J. Automatic Generation of Semantic Patterns using Techniques of Natural Language Processing. SKY 2013: 34-44

Thomason, Richmond H. What is Semantics? Version 2. March 27, 2012. Available in: http://web.eecs.umich.edu/~rthomaso/documents/general/what-is-semantics.html

Triviño, J. L. Morales Bueno, R. A Spanish POS tagger with variable memory. In Proceedings of the Sixth International Workshop on Parsing Technologies (IWPT-2000). ACL/SIGPARSE, Trento, Italia, 2000. pp. 254-265.

Weischedel, R. Metter, M. Schwartz, R. Ramshaw, L. Palmucci, J. Coping with ambiguity and unknown through probabilistic models. Computational Linguistics, vol. 19, pp. 359-382.