PAST – A text analysis tool "to discover" Distinctions

Germán Bravo
José Bermeo
Sebastían Contreras
Los Andes University – Bogotá, COLOMBIA
{gbravo, j-bermeo, seb-cont}@uniandes.edu.co

ABSTRACT

In the course of Cybernetics of Cybernetics (CC), arises the problem of analyzing, for each student, three written texts from different moments during the course, searching for the distinctions, and their evolution throughout the course. PAST is a text analysis tool to help the professor in this "distinction discovering" in an impersonal and repeatable way. The process, based on professor-defined rules, consist in realizing successive text transformations that lead to a Conceptual Proximity Graph (CPG) reflecting the distinctions used in the text. This work describes the problem, the solution proposed by PAST, and its effectiveness by means of some simple examples and analyzing its use in the course of Cybernetics of Cybernetics.

Keywords:

Text analysis, distinction, dependency grammar, conceptual proximity graph, text transformation

1. INTRODUCTION

The interpretation of a text is a cognitive process involving, among others, the author's knowledge, experience and expectations, as well as of the reader's. The ideal case is when the reader understands the text as the author expects, but this is not always successful. An inappropriate interpretation of a text may be the author's responsibility, but also of the reader's. If it is the author's responsibility, there may be a lack of clarity in the sent message, deficiency in the writing, language misuse, and bad text organization. If it is the reader´s responsibility, there may be lack of knowledge in the subject and language withdraws, but also his physical and emotional morn affects the interpretation.

There are two kinds of author/reader relationship: the first one is when there is an author and many readers, like novels, books, and articles; the second one is when there many authors and one reader. The latter occurs mainly in academic environments, which is the case of students' assignments, where the reading, interpretation, and evaluation of tests must be the more "objective" possible; although this "objectivity" is usually lost because of the quantity of tests, the texts' quality, the fatigue, and the distractions to which the reviewer is exposed.

It is then necessary a tool to make the objectivity of the interpretation easier, without involving the cognitive process of the reader. This transformation must be uniform for all texts and reflect in some way the subject treated in the text. This requirement is accomplished by extracting from the text the words that are more frequently used and the relationships amongst them. When this transformation is applied to the whole set of course texts the professor may identify the words and relationships used by the students, and thus the conceptualization that they all have about the subject.

This article presents PAST (Platform for Text Analysis) a software tool that applying transformation rules to a text, allows a reader to perform a "subject analysis" of the text. First, it describes the course of Cybernetics of Cybernetics, as a case study where the need of PAST is evident. Then, it presents the theoretical concepts supporting the proposed text transformations. Next, it describes the proposed solution and the developed software tool. Then, it presents the application to the case study, and finally some conclusions and outcomes for this work.

Disclaimer: The analyzed texts were written in Spanish, so the analysis rules defined in PAST. Because of the differences between English grammar and Spanish grammar, the examples and results shown in this paper are in Spanish. The authors apologize for an eventual lack of comprehension due to this fact.

2. THE PROBLEM

PAST appears in the context of the course of Cybernetics of Cybernetics (CC) [1], part of the MSc. program of Industrial Engineering in Los Andes University, Bogota, Colombia. The leading edge of this course is the cybernetics of second order, having as a thematic "The Observer of the Observer". Having a system to be observed (a game for instance), an observer observing this system, and another observer observing how the first observer observes the system, the students take successively the two observer roles and then study, make questions, think about this situation, and finally rebuild their conceptions in a written text. In other words, the central question of the course is "What kind of observer of the observer am I?"

The epistemological concept of distinction[1] is the foundation of course's development. Some "master distinctions", including the language (connotative and non-denotative), the observer, the observer of the observer and the dispositions, amongst others, articulate, throughout the language, the fields of action in the course. By means of a recurrent exercise of reading, gaming and writing, the course proposes the student to analyze his writings, to observe the used concepts and the relations among them, and finally to extract the distinctions he made.

In order to observe the distinctions uniformly, PAST transforms the students' texts in conceptual graphs by extracting the most frequently used concepts and the relations among them. In the resulting structure, the extracted words (nodes) are interpreted as the distinctions used by the author to construct his text. This

[1] Spencer-Browns [2] consider "the departure point of all knowledge requires to invent and to draw up a distinction: The distinction restores an observant act that constitutes a border, dividing the space in two sub spaces, two continents complementarily delimited. The border is the first passage in the production of a world: it organizes all a topology of the perception from an ontology of the cut ".

analysis allows the professor to make an alternate observation of a student's writing. Also, when applied to the set of students' texts, the professor (the observer of the observers of the observer…) might observe the evolution of the main concepts in the course throughout its development.

The formalization principles include:

- The author (observant) can reconstruct his discourse in a written text.
- From this text, based on the linguistic context of the course and on a set of translation rules, it is possible to extract a network of interrelated words.
- The user-defined translation rules allow the identification of related words and the customization of the resulting structures. The union of these structures constitutes, as a non-directed graph, the searched network of words.
- The analysis of these graphs permits to identify the distinctions used by the author in his written discourse. This analysis may be realized by any other person (observer), by the author itself (observer of the observer), or by the professor (observer of the observes of the observer)

The students write their texts in three occasions using iterative and recurrent operators. The iterative operator implies the writing of several versions of the text until the author finds it satisfactory. The recurrent operator implies the observing of the observer, because the conceptual graphs generated by PAST are used to feedback the author, showing him the distinctions he had used in the preceding texts.

3. THE PROPOSED SOLUTION

The process associated to the transformation of the text suggests a sequence of structures representing its content in different forms, each one demonstrating some of the characteristics in the original text. Having in mind the considered case study, this section shows, very roughly, the transformations obtained in each phase of the process, as shown in Figure 1. The linguistic background section gives a one more detailed explanation of some needed linguistic elements. Some details of the proposed solution are found in the detailed description of PAST.

The process starts with the original text of the author, written using any text editor. The text is cleaned up by removing styles, figures, tables, indexes and references, leaving only the text contents and obtaining a plane text file (1).

The first transformation (T1) creates a representation of the text preserving the relationships among the words, not introducing additional linguistic structures to those existing in the text, and that is workable by software. By means of a syntactic parser a set of dependency trees, one by each phrase of the text (2). Each node of any of these trees contains a word, its associated motto and its morphologic category, represented by means of EAGLES labels [3]; the arcs represent syntactic dependencies, like "noun - adjective", "verb - noun", etc.

The second transformation (T2) joins all these dependency trees, in order to get a unique structure of the whole text. This transformation applies a set of morphologic composition and transformation rules (R1) to every dependency tree and generates a graph of morphologic structures (3).

The rules of morphologic composition and transformation (R1) define the translation of the relations found in the dependency trees, hierarchic by nature, into nodes and arcs of a graph. These rules have two components: a syntactic dependency, represented like a tree, and its transformation to a graph.

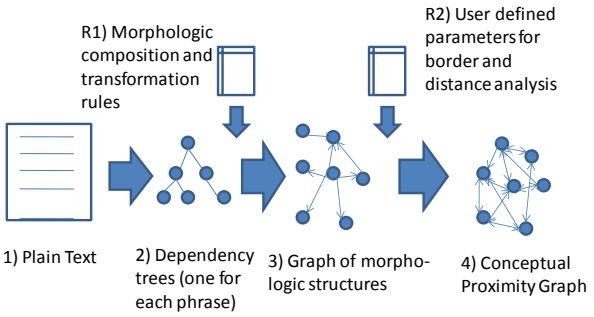

Figure 1. Transformation process of the text

The process consist of the traversal of the dependency trees of every phrase in the text, looking for occurrences, as sub trees, of the syntactic dependencies defined in the rules. For every occurrence of a syntactic dependency, the corresponding transformation is added to the resulting graph, following a strategy that puts in evidence the concepts used by the author in the text. Assuming that the frequency in the use of nouns and their (strong) relationships reflects what the author wants to talk about, i.e. the concepts, the construction of the graph centralize the use of nouns, making them to appear only once in the resulting graph. The first occurrence of a noun creates a node in the graph; further occurrences of the noun add weight to it. In a similar way, the first occurrence of a dependency creates an arc in the graph and further occurrences add weight to it.

Thus, the nodes of the graph of morphologic structures are the words in the text matching at least one rule of composition. The arcs, directed, represent the matching of syntactic dependency in the text, according to the translations defined in the morphologic composition and transformation rules.

The visualization of the graph, taking into account the weight of the nodes (color code) and of the arcs (thickness), allows the analyst to identify, as concepts, those nouns having a great weight and many relationships. However, due to the size of the graph, this is not an easy task.

The third transformation (T3) extracts the concepts in the text and puts in relevance the relations among them, allowing a semantic analysis based on the syntactic structure of the phrases of the text. The process contemplates only the nouns in the morphologic structures graph and further analysis over them. A border analysis, to a given distance passed by parameter (R2), generates a sub graph with the nouns whose distance to other nouns is lesser or equal to the given parameter; the arcs in this sub graph conserve the transitivity relationships existing in the original text. A further analysis concerns the importance of the nouns, based on the number of relationships they have, a number also passed as parameter (R2): Only those nouns with more relationships than the parameter passes are preserved in the final sub graph. The result is a Conceptual Proximity Graph (4), non-directed and weighted, where the nodes are those nouns considered important (by the defined parameters), i.e. the distinctions, and the arcs represent the existence of a relationship between two distinctions. The weight of the nodes is the number of relationships they have, representing thus the relevance of the distinction in the author's discourse. The weight of the arcs is the structural distance between two distinctions, representing thus how strong their relationship is.

Depending on the desired analysis, it is possible to define new or alternate transformations, following a similar process: Rule definition, establishment of the parameters of analysis, application to the text and visualization of the result.

4. LINGUISTIC BACKGROUND

There are two main perspectives to study the structural properties of human language. The first one identifies the discrete units conforming the language and classifies them in classes. The second one studies the rules and principles that govern the phrase construction: The syntax [4]. Although numerous syntactic theories exist, they all have some common concepts. From all these concepts, the Generative Grammar provides the foundation to this work.

The Generative Grammar is the linguistic branch supporting the idea that the most important element of study is how to construct the sentences. Noam Chomsky [5] establishes that: (1) what people know is a collection of words and rules to generate chains of words, called sentences in our language. (2) Although there exist a finite number of elements in that collection (some thousands of words and some hundreds of rules), it is possible to generate an infinite number sentences, because of the recurrence of some of these rules. (3) It is not possible to build an infinite sentence. (4) It is not possible to build a sentence having all the words in the collection.

Two approaches exist to describe the structure of a phrase in natural language: Constituent grammars and Dependency grammars. The constituent grammar divides the sentence in several components, and divides these components in smaller ones, until arriving to words. In the dependency grammar, a word is the nucleus of the phrase and the other words of the phrase, either depend syntactically of the nucleus or depend syntactically of another word of the phrase [6] [7]. The dependencies represent the grammar production rules, as for example: "the nucleus is the sentence's main verb", or "an adjective depends on the affected noun". The dependency analysis aims to obtain, for a sentence, a dependency tree respecting a dependency grammar, as shown in Figure 2, for the sentence: "Este es un ejemplo de dependencia gramatical".

One of the most important advantages of dependency analysis, Covington [8], is that dependencies are close to the semantic relationships needed for further text interpretations. Therefore, it is not worth to try to look for relations that the tree does not show directly.

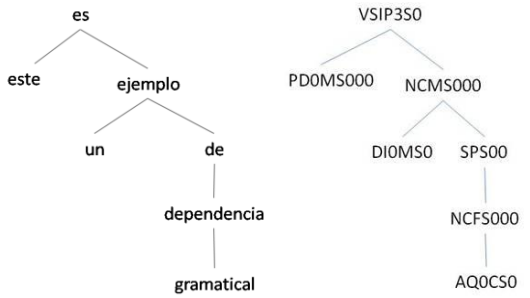

Figure 2. Dependency and morphologic categories trees

In order to accomplish the second text transformation, the dependency tree also includes a morpho-syntactical annotation for each word in the sentence, using the labels proposed by the Expert Advisory Group on Language Engineering Standards (EAGLES [3]) for all the European languages. These labels define a morphologic hierarchy in three levels: (1) the first level contains the main morphologic categories (verb (A), adjective (A), noun (N), etc.). (2) The second level contains the categories' attributes (for the adjectives, the attributes are Type, degree, sort, number and function). (3) The third level contains the possible values for each attribute (for example, the genre of the adjectives can be masculine, feminine or common).

For example, the word "alegres" ("glad") has the label "AQ0CP0", indicating that it is an adjective (A), qualifying (Q), without degree (0), without genre (C), plural (P) and without function (0).

Figure 2 also shows the EAGLES' tree for the sentence "Este es un ejemplo de dependencia grammatical".

From the software point of view, there are many available software tools for (Spanish) text analysis. Among them, there are WorldNet [9], Conexor [10] and Freeling [11]. For this work, Freeling was chosen to realize the syntactic analysis of sentences, mainly because of the features it provides, including the dependency analysis, because it can be used as a software library, and because the GNU/GPL licensing scheme.

However, for the other text transformations proposed in this work, it was necessary to develop PAST.

5. PAST' S DETAILED DESCRIPTION

PAST (Platform for Syntactic Text Analysis) is a software tool implementing the construction of conceptual proximity graphs from a text, based on the syntactic relationships existing in the text. This section describes first the user roles supported by PAST and then its five composing modules. The operation of these modules are explained using as example the sentence "El juego es una herramienta para incorporar la metodología"

The users

The process associated to text transformations requires of the definition of three user roles: the professor, the analyst and author, respectively.

The professor is the person who proposes and structures the thematic about the text to be developed. In consequence, he is the most appropriate to define the morphologic composition rules, which are the base for the analysis. The professor must be aware of grammatical concepts used (dependency grammar, EAGLES labels, etc) and have a clear idea of the desired results. The professor may use PAST as test bed with some known scenarios, in order to define (and redefine) the rules and parameters of analysis.

The analyst is in charge of the analysis of texts, aided by the trees and graphs generated by PAST. For the case study, the analysis includes the comparison of different graphs, the definition of guidelines to interpret the graphs and the generation of conclusions, and, maybe the most important, the feedback to the authors. Other analysis may include the study of pertinent relations to one or several nodes in a same graph, and the analysis of common nodes to one or several versions of the same text. In a wider context, PAST may receive as input a joined text form several authors, in order to detect common opinions, or make analysis based on a particular subject (What a community think about a subject), among others.

The author writes the text. In cooperation with the analyst, the author receives the feedback based on PAST graphs and rewrites the text. This process can occur many times, if the purpose is to observe how some concepts evolve in different versions of a given text.

Depending on the context, a same person can play several of these roles. The roles of professor and analyst are compatible, as well as the roles of author and analyst. It is complicated and no desirable that a same person plays the three rolls, because of a risk of lack of "objectivity".

Module of Syntactic and morphological Analysis

This module is responsible of the first text transformation. In this module, the user imports a plain text file, and use Freeling to process it in order to obtain the dependency trees for each

sentence. Finally, PAST shows the sentences and their associated dependency trees. Figure 3 shows the results of the module for the sentence example. As shown in the example, Freeling associates the EAGLES' labels to every word in the tree.

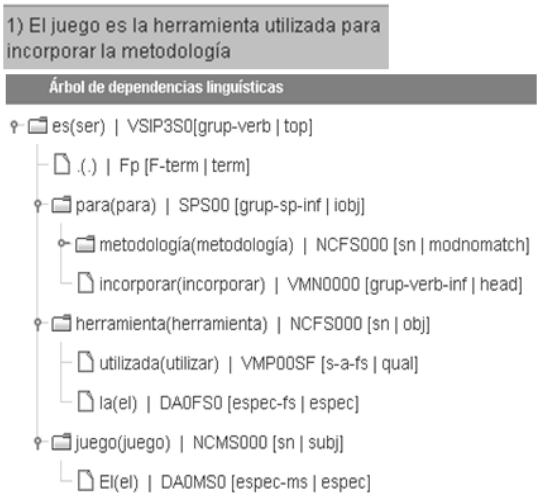

Figure 3. Dependency tree generated by PAST, including the motto and words' morphologic categories

Module of Construction of Morphologic Composition and Transformation Rules

This module helps the professor to construct the morphologic composition and transformation rules. In PAST, it is possible to define all the rules that the professor considers necessary, as complex and numerous as required by the analysis.

The morphologic part of this rules consist of a subtree representing a syntactic dependency, based on the morphological category of words. The nodes in this subtree have an, eventually partial, EAGLES' label and the arcs are the searched dependency, as shown in Figure 4. Every node has also an index, used by the transformation part of the rule. When the analysis searches for generic structures, the label in a node is partial. For example, a label "V" matches all verbs, or the label "VI" matches all verbs in infinitive. PAST adds also a label "ANY" to match any word.

The transformation part of the rule defines how to translate the dependency in a graph. It represents which nodes in the subtree pass to the graph and how they will be related.

Figure 4 shows an example of a rule looking for a syntactic structure having a verb and two nouns (V (N) (N)).

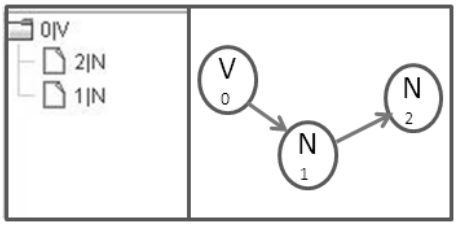

Figure 4. Example of a Morphological and of transformation rule

The construction of this graph consists of generating three nodes and then relating the verb to the first noun and the latter to the second noun.

Module of Morphologic Relationships Graph

This module builds the Morphologic Relationship Graph by joining all the dependency trees in a single structure, using the defined set of morphological and transformation rules. The process looks for matches of every morphological dependency in the rules in every dependency tree of the text. When a match occurs, the process applies the transformation defined in the rule. As stated above, the nouns guide the analysis and hence they appears only once in the result.

The resulting structure is a weighted directed graph, in which the nodes represent the words of the text, with only one occurrence of the nouns. The arcs represent the structural relations defined by the rules. Figure 5 shows the resulting graphs from to two sets of rules for the sentence: "El juego es la herramienta para incorporar la metodología". The two set of rules look for the same syntactic dependencies, but have different transformations rules.

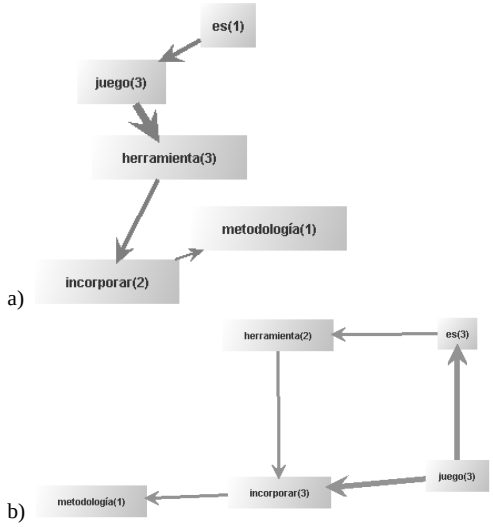

Figure 5 a and b. Graphs of morphological structures

The weight associated to each word corresponds to the number of times that word matched a rule. The arcs' thickness represents the average of weights of linked nodes. The graph also remembers to which sentence the words belong and this information is reflected in the color of the links; in the example, being a single phrase, all the arcs have the same color.

The color of the nodes is function of the frequency of appearance of the word in the text, red for the most used, yellow for the average and grey for the least used, reflecting the relative weight of words with respect to whole graph.

As observed, the resulting graphs depend on the set of rules used. This fact constitutes simultaneously a PAST's feature and a challenge to the professor to define the appropriate set of rules for the desired analysis.

Construction of conceptual proximity Graph

This module builds the Conceptual Proximity Graph, receiving as parameters the distance to the border for a node and the desired relevance (measured as the number of incident arcs) of words.

The process of the construction of this graph, starts by getting a non directed graph, then extract the nouns and define their border and finally filter them for their relevance.

A wise choosing of this module's parameters facilitates the visualization of the resulting graph and then the discovery of distinctions: the subject of this work.

6. RESULTS AND APPLICATION TO CASO OF STUDY

Figures 6 and 7 show the graphs of morphologic structures and conceptual proximity, respectively, for a little more complex case. The analysis uses the same set of rules defined for distinctions analysis for the case study of Cybernetics of Cybernetics [1].

The analysis corresponds to the following paragraph:

> *"En el curso de CC consideramos que el aprendizaje es la adquisición y la conexión de diversos conceptos y que los conceptos se entrañan corporalmente. Para el caso del curso, el proceso de aprendizaje y de entrañamiento se realiza en el ejercicio recurrente de hacer lecturas y ensayos relacionados con las lecturas, así como en el diseño de juegos y en el juego de juegos."*[2]

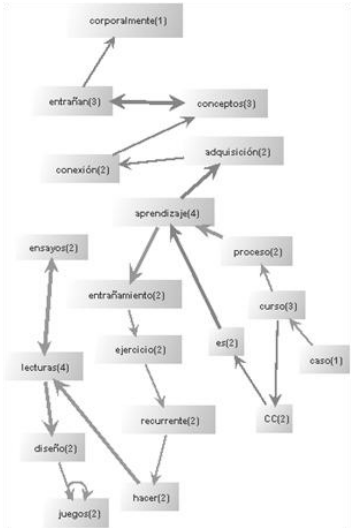

Figure 6. Morphologic dependency Graph

According to the process described, the observation of the graph of morphologic dependencies and the original text, shows that `aprendizaje' appears twice in the text, in both sentences, but only once in the graph. However, it has multiple edges incident arcs, corresponding to the relations of the word in different moments of the text.

On the other hand, the word `concepts' appears twice in the text, but in the same sentence. According with the strategy of graph generation, the word `concepts' appears as a unique node in the graph and its relations, corresponding to each occurrence of the word in the text, appear like arcs.

The conceptual proximity graph represents the connections, direct and indirect, amongst the nouns. In this case, a structural analysis suggests the study of words with greater number of relations and its relevance in context of the paragraph, as for example the words `learning' and `course'.

Application to the case study

For the case study [1], the course professor, José Bermeo, played the roles of analyst and professor. The students played the role of author.

[2] "In the course of CC we considered that the learning is the acquisition and connection of diverse concepts and that the concepts are involved corporally. For the course, the entailment and learning process is accomplished by a recurrent exercise of reading and tests about the readings, as well as in the design of games and the game of games."

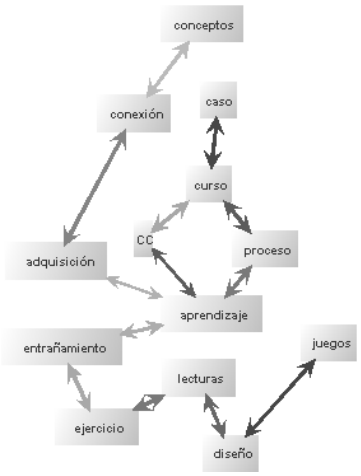

Figure 7. Conceptual proximity graph

The texts analyzed are the essays written by the students in the course. These essays are reviewed and returned to the students in three occasions (T1, T2, T3), as shown in figure 8.

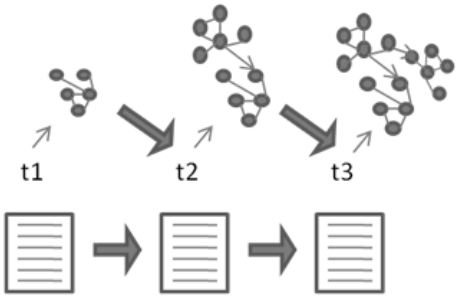

Figure 8. Work schema in the CC course

If the objective of the course is the insight, the reflection and the construction of the observer of the observer, based on commentaries and revisions of the professor, in each opportunity, PAST allowed the student (observant) to find nonlinear relations in his written discourse.

In agreement with the experience of Bermeo [1], the conceptual graphs of students' texts reflect, every time, the expected development of the students' discourse. This development is coherent with the development of the subjects in the course, and with the expected effect of feedback through the generated conceptual graphs: "The number of distinctions follows the chronology of the course. The frequencies and the use of distinctions in the discourse increased, particularly in the transition between the second and third essays", "PAST gives conceptual order to a text, in the dominion of the own text, and it allows to observe diverse relational structures of the distinctions made by an author".

7. CONCLUSIONS AND OUTCOMES

The first conclusion is that PAST fulfilled the objective defined by the case study, the course of Cybernetics of the Cybernetics, and so, it constitutes a valid pedagogical tool, useful to the academic community having the same or similar needs. The flexibility of PAST concerning the definition of the rules and the parameters of analysis facilitates this "customization".

It is also clear that the obtained results depend heavily on morphologic composition and transformation rules used in the analysis. For the case study, the rules were generated in a quite empirical way, mainly because of time restrictions, but certainly, this set of rules deserves a revision from the linguistic point of view.

It is also possible to think about the development of new analyses, based either on the morphologic structures graph or on the conceptual proximity graph. An example would be the analysis of the text centered in a single word (concept), allowing the evaluation of the appropriate use of this concept throughout the text.

On the other hand, although PAST was "born" in the context of CC course, the application of the platform is not restricted only to this course. On the contrary, the strategy and the architecture of PAST is a generic tool with applications in multiple environments, such as:

- Trade: The analysis of open text concerning the consumers' preferences or opinions about a given product, allows the identification of metaphoric or reference relationships.
- Clinical histories: The analysis of the entries of patient clinical history facilitates their "homologation" and eventual importation of this information in health databases or data warehouses.
- Anthropology: The conceptual analysis may identify certain patterns of text repetition from their syntactic structure.

8. REFERENCES

[1] J. Bermeo, G. Bravo, J.S. Contreras, R. Zarama. **Plataforma para el análisis de textos escritos en términos de grafos conceptuales: caso de estudio curso Cibernética de la Cibernética**. WMSCI 2008. June 2008

[2] G. Spencer-Brown. **Laws of form**. Nueva York, Bantam Books. 1973

[3] EAGLES: Expert Advisory Group on Language Engineering Standards. http://www.ilc.cnr.it/EAGLES/intro.html. Last visited 24/03/2008

[4] A. Akmajian, R. Demers, A. Farmer, R. Harnish. **Linguistics: An intro-duction to Language and Communication**. MIT Press. Third edition. 1992.

[5] N. Chomsky. **Syntactic Structures**. The Hague. Mouton & Co. 1957.

[6] C. D. Manning, H. Schütze. **Foundations of Statistical Natural Language Processing**. The MIT Press, Cambridge, Massachusetts, 1999.

[7] J. H. de la Cruz. **Un Modelo Fundamentado en Análisis de Dependencias y WordNet para el Reconocimiento de Implicación Textual**. Universidad Complutense de Madrid. 2005

[8] M. A. Covington. **An empirically motivated reinterpretation of dependency grammars**. Research report AI-1994-01. University of Georgia. 1994.

[9] Wordnet, a lexical database for the English language. http://wordnet.princeton.edu/. Last visited the 24/03/2008

[10] Conexor Natural Knowledge. http://www.connexor.eu/. Last visited the 24/03/2008

[11] Freeling. http://garraf.epsevg.upc.es/freeling/. Last visited the 24/03/2008