Originally published in: Proceedings of the 8th ICCC/IFIP International Conference on Electronic Publishing. ELPUB. - Brasília: Universidade de Brasília, 2004. Pp. 111-120.

ONTOLOGY EXTRACTION FOR INDEX GENERATION

CLÁUDIO GOTTSCHALG-DUQUE[1]; HENNING LOBIN[2]

[1] Justus-Liebig-Universität Gießen
Angewandte Sprachwissenschaft und Computerlinguistik

[2] Justus-Liebig-Universität Gießen
Angewandte Sprachwissenschaft und Computerlinguistik

The administration of electronic publication in the Information Era congregates old and new problems, especially those related with Information Retrieval and Automatic Knowledge Extraction. This article presents an Information Retrieval System that uses Natural Language Processing and Ontology to index collection's texts. We describe a system that constructs a domain specific ontology, starting from the syntactic and semantic analyses of the texts that compose the collection. First the texts are tokenized, then a robust syntactic analysis is made, subsequently the semantic analysis is accomplished in conformity with a metalanguage of knowledge representation, based on a basic ontology composed of 47 classes. The ontology, automatically extracted, generates richer domain specific knowledge. It propitiates, through its semantic net, the right conditions for the user to find with larger efficiency and agility the terms adapted for the consultation to the texts. A prototype of this system was built and used for the indexation of a collection of 221 electronic texts of Information Science written in Portuguese from Brazil. Instead of being based in statistical theories, we propose a robust Information Retrieval System that uses cognitive theories, allowing a larger efficiency in the answer to the users' queries.

Keywords. electronic publication, information retrieval system, computational linguistics, ontology, index generation

1 INTRODUCTION

This paper presents an approach based on Natural Language Processing and Ontologies to construct an Information Retrieval System. An Information Retrieval System is, here, understood as a system that can treat the classical problem of effective and efficient retrieval of relevant documents out of a collection in response to a certain information need of a user. Basically, Information Retrieval is composed of three processes, collecting, indexing and ordering the documents. Now, particularly due to the increase of the information accessible in digital format [1], [2], especially because the Web, the biggest human information repository, studies about Information Retrieval Systems and these processes cited above grew vastly in importance, [3], [4], [5], [6], [7]. One relevant problem is that, despite the evolution of the hardware systems used to collect, index and order digital documents, nowadays an Information Retrieval System can not incorporate proficiently the entire information content of a collection of documents. To try to solve this problem and keep the efficiency of the system, the index process in an Information Retrieval System usually deals with a manipulated representation of the document's information content,

normally the determination of the document representation using key words extracted from the text. To order the documents that can possible answer the user's query it is possible to make the same to the query, when is obtained a matching it's seems that both of them are co-related. This way out is very simplistic and don't work for big collections, that have a lot of information that can't be automatically fully extracted. So, the form of representation, the manner the representation is achieved and the way this information is showed are extremely important to all Information Retrieval System.

Our goal is construct an Information Retrieval System from the beginning using basic linguistic and computer linguistic theories by means of syntactic parser [8], semantic parser [9] and a knowledge representation technique [10] used in a central ontology, allowing to create lightweight ontologies [11], [12], [13]. We believe that this is the better approach instead of making some adaptations in a finished Information Retrieval System [14], [15]. To validate our ideas we developed a prototype and we are evaluating our theories using it to index a little collection (221 papers) of electronic Brazilian documents about Information Science ("Revista Ciência da Informação", Information Science Journal, http://www.ibict.br). Preliminaries results validate our ideas and make subsides to optimize the prototype to be full automatic.

This paper is organized as follows. All over the sections we show examples of the process using some real text sample extracted from the collection. We started describing in section 2 the use of concepts and ontologies in our prototype. Why index by concepts, the importance of developing new approaches using Natural Language Processing to treat the classical Information Retrieval problem. Then, in section 3, we explain the use of syntactic and semantic parsers. Section 4 is dedicated to the use of the ontology. After, section 5, we make a brief description of the index module. Finally, in section 6, we discuss our theories and future works.

2 WHY USE CONCEPTS AND ONTOLOGIES TO INDEX

Index using concepts can make a positive difference. It is possible to combine different techniques from Natural Language Processing and Ontologies for indexing terms, words and phrases. These can permit that an Information Retrieval System create relevant connections between the terminology of users query and related terminology in the document or documents that contains the information that can really correspond to the user need. This approach can dramatically reduce the frustrating experience trying to discover the most exact and specific term or terms that will permit the real matching between the user query and the system answer. Using an automatic semantic parser it is possible to extract from the texts the conceptual structure, descriptions of phrases; the semantic relationships between words and concepts to establish connections between then. This structure, a meta-level description, is a representation that brings order to the collection of documents, so can be understood as Ontology. Ontology, as a formal explicit specification of a shared conceptualization, can help to solve the problem of inefficiency, overloaded "fake information", ambiguity and chaos that exists in our "Information Era". Ontologies can be very useful to an Information Retrieval System because they are structured in a way that goes considerably beyond the possibilities offered by others classification systems like thesauri [16], [17]. The idea is that structured index concepts generated by means of extensive analysis result in a better performance for answer user's queries. Knowledge-based techniques [18] are used to infer that some concepts extracted from texts are co-related.

Our prototype is composed of three modules: A Natural Language Processing Module, an Ontology Module and an Index Module (Fig.1).

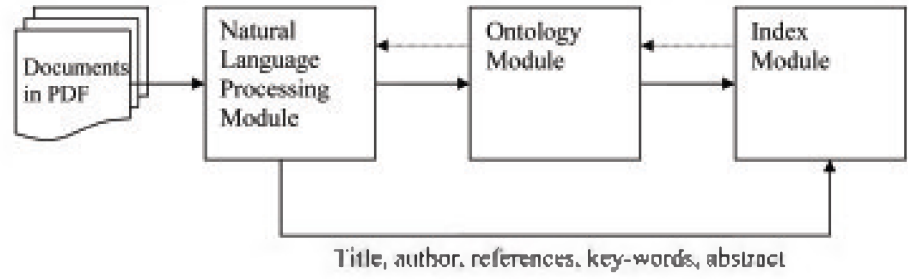

FIG 1. THE PROTOTYPE. DOCUMENTS ARE SENDING IN PDF FORMAT, IN THE FIRST SUB-MODULE OF THE NATURAL LANGUAGE PROCESSING MODULE ARE CONVERTED TO HTML. THE TITLE, AUTHOR, REFERENCES, KEY-WORDS AND ABSTRACT ARE RECOGNIZED AND SENT DIRECT TO THE INDEX MODULE. AFTER THIS, THE SENTENCES ARE EXTRACTED, WORDS SYNTACTIC LABELED AND PROPOSITIONS EXTRACTED AND SENT TO THE ONTOLOGY MODULE. IN THE ONTOLOGY MODULE THE PROPOSITIONS' TERMS ARE CONFIRMED, STORED IN AN ONTOLOGY, LABELED AGAIN AND SENT TO THE INDEX MODULE. SOME STAGES CAN BE RECURSIVE TO GUARANTEE THE ROBUSTNESS OF THE INFORMATION RETRIEVAL SYSTEM.

3 NATURAL LANGUAGE PROCESSING MODULE

The Natural Language Processing Module is used to enhance the indexing process by producing structured concepts. This module analyzes the sentences in the documents identifying concepts. It is composed of three sub-modules: Tokenizer Sub-module, Syntactic Sub-module and Semantic Sub-Module.

The Tokenizer Sub-module essentially separates and labels the parts of the texts to permit the other sub-modules to process them. First of all the Tokenizer receives the texts in HTML format. Title, authors, keywords and references are identified using heuristics. They are extracted and sent directly to the Ontology Module System. The introduction of the text is also identified and extracted. At this moment, only to test the prototype, the other parts of the text are ignored. This doesn't depreciate our experiment because normally the most relevant words in a text are in the introduction [19], [20] and in the next step of development our prototype will process the entire text. The introduction text is separated in sentences. The Tokenizer covers the text initiating in the first word of the first paragraph, from the left to the right, from top to down. When it finds a specific punctuation signal (. ! ? ;), the end of the sentence is determinate and the same is extracted. All parentheses found in the middle of the texts are eliminated because we discovery empirically that more than 95% of then, in this collection, contents only reference's authors. After this the sentences are sent to the Syntactic Sub-module (Fig. 2).

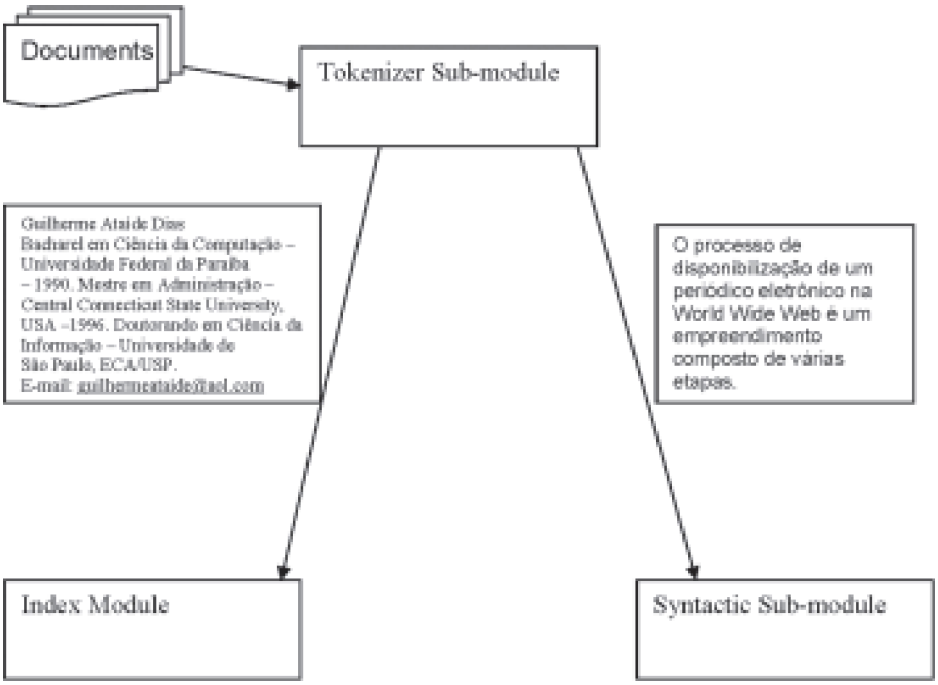

FIG. 2. TOKENIZER SUB-MODULE AND ITS OUTPUTS. THE AUTHOR(S) AND HIS/HER CHARACTERISTICS ARE SENT DIRECT TO THE INDEX MODULE. THE SENTENCES EXTRACTED FROM THE TEXTS ARE SENT TO THE SYNTACTIC SUB-MODULE.

The Syntactic Sub-module in effect is a robust syntactic parser that processes the sentences generating input to the Semantic Sub-module (Fig. 3). It is responsible for the content words (verbs, nouns, etc) and the functional words (prepositions, conjunctions, etc) identification. In the prototype is a Portuguese syntactic parser that uses grammatical rules formulated in the Constraint Grammar formalism [8], [21], [22], [23]. This syntactic parser processes the sentences and the results of the analysis are sentences with tags for syntactical form and functions. After the extraction of the syntactic context of each word this output is sent to the next sub-module.

The Semantic Sub-module identifies primitive concepts, relevant semantic issues. It is a semantic parser that processes the Syntactic Sub-module output, extracting semantic case roles on the basis of words' syntactic labels. The idea is that the lexicon contains a conceptual structure constituted by formation rules. This conceptual structure permits the combination of primitive categories resulting in more complex ones. The thematic roles are derived from these concepts [24], [25]. The Noun Phrases (NPs) are recognized and extracted using a theory based on NP structure for identifying NP grammar's rules.

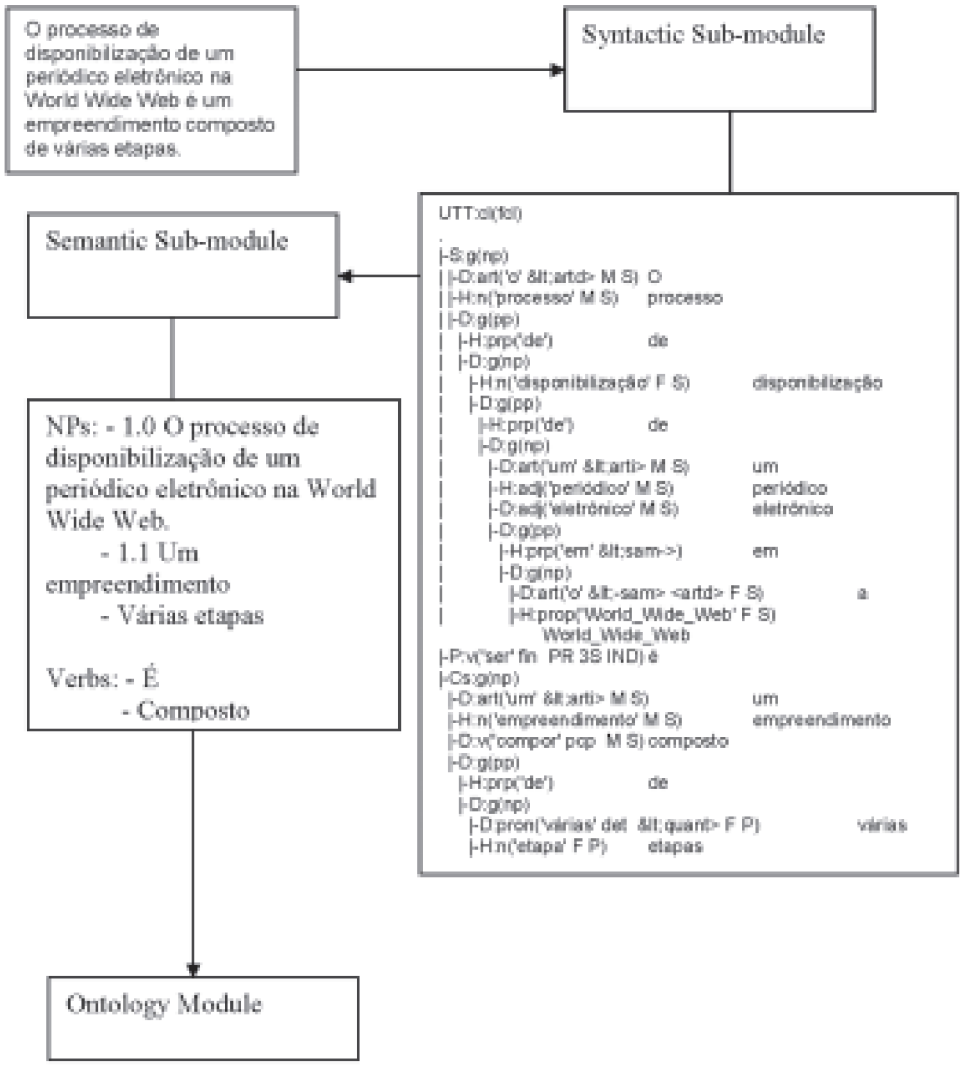

FIG. 3. SYNTACTIC SUB-MODULE, SEMANTIC SUB-MODULE AND ITS OUTPUTS. THE SENTENCE IS PARSED BY THE SYNTACTIC SUB-MODULE, THE WORDS LABELED WITH ITS SYNTACTIC CATEGORIES ARE SENT TO THE SEMANTIC SUB-MODULE. IN THE SEMANTIC SUB-MODULE RULES BASED IN FREDERIKSEN'S MODEL AND IN BRAZILIAN PORTUGUESE GRAMMAR RULES ARE APPLIED AND PROPOSITIONS WITH ITS CLASSES ARE SENT TO THE ONTOLOGY MODULE.

4 ONTOLOGY MODULE

Ontologies are explicit formal specifications of the terms in the domain and relations among them [26] [27]. Ontologies have become common on the Web; they range from large taxonomies categorizing Web sites to categorizations of concepts in theorical systems of Artificial Intelligent. Ontology defines a common vocabulary for researchers who need to share specific information in a specific domain. Based on text terms and possible relationships between terms, we create a lightweight ontology of Information Science concepts. The Ontology Module has one sub-module named Basic Ontology Sub-module and one (or more, the collection used in the prototype is

very homogeny) named Generated Ontology Sub-module. The first one was created manually, has only 47 classes, one superclass, three parents, 11 children, and is "static". This Basic Ontology (Fig. 4) is based on Frederiksen's (1975) knowledge representation technique. A meta-language used in numerous experiments for the study of textual memory, understanding and composition, besides researches in linguistics and artificial intelligence [28], [29], [30]. It is used mainly to identify explicit elements of the meaning of a text, to determine the density of information of the text (in propositions for sentence) and the degree of the information that is wanted to transmit, as well as for the text analysis in general. An application of private interest in the linguistics is to approach the projection problem, in a systematic way, in other words, to specify the correspondence rules between syntactic structures and semantic interpretations represented with the meta-language used in this project.

Class Hierarchy for *AnáliseProposicional* Project

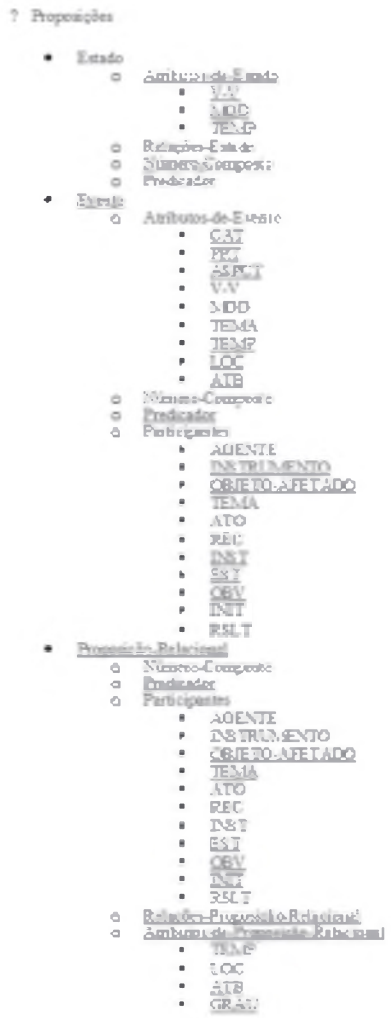

FIG. 4. THE BASIC ONTOLOGY IN AN HIERARCHIC REPRESENTATION IN HTML.

5 INDEX MODULE

Two Sub-modules are part of The Index Module, the Index Rules Sub-module and the Index Structured Sub-module. This module takes as its input the set of ontological information produced by the Ontology Module and some Tokenizer Module outputs (references, author(s), etc.). This module checks several times the Ontology Module output to confirm the validation of all concepts proposals against the ontology and subsequently produces valid index concepts. When an Ontology Module output is considered valid (in accordance with the rules in the Index Rules Sub-module) (Fig. 5).

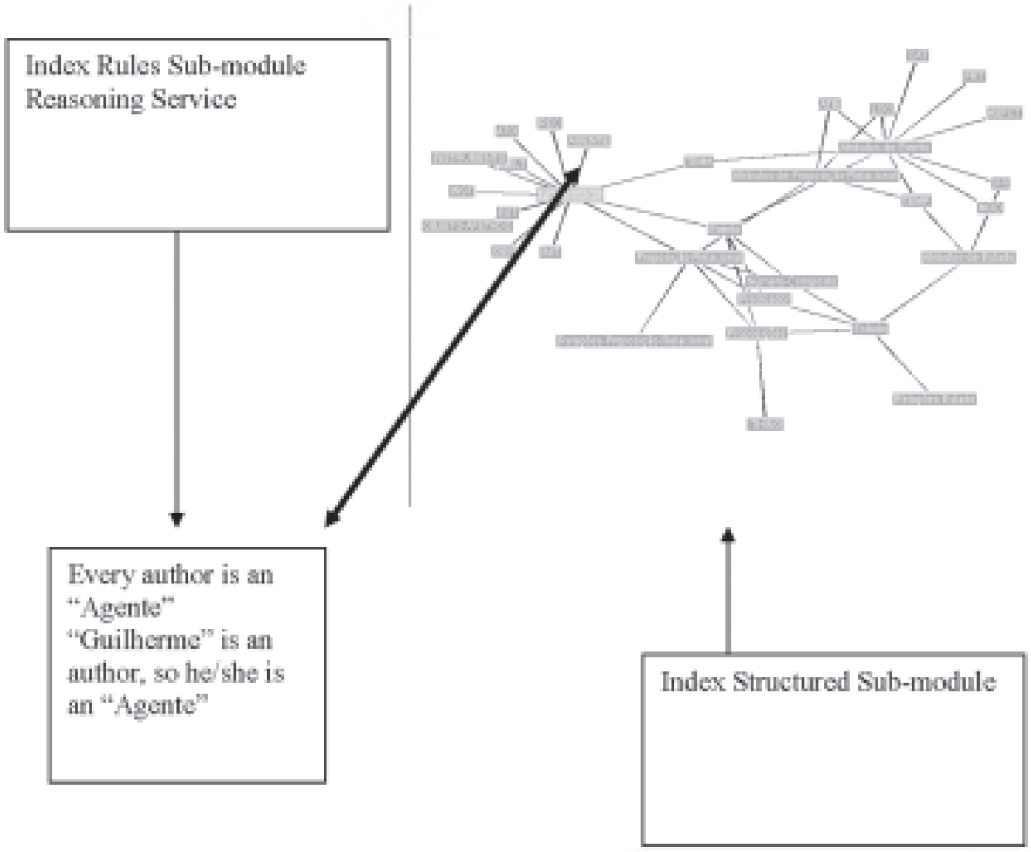

FIG. 5 THE INDEX MODULE. IN THE EXAMPLE THE INDEX RULES SUB-MODULE HAS STORED THAT ALL "AUTHOR" IS AN "AGENTE" AND "GUILHERME" EXTRACTED FROM THE TEXT (THE AUTHOR OF THE TEXT) IS DECLARED AN "AGENTE". THIS "LABEL" IS AUTHENTICATED AND "GUILHERME" IS POSITIONED IN THE INDEX STRUCTURED SUB-MODULE STRUCTURE. SO, HE IS A "PARTICIPANTE" (PARTICIPANT), HE IS PART OF A "EVENTO" (EVENT), AND SO ON, AS SHOWED ABOVE.

6 DISCUSSION

Our prototype works in this way: During first process' steps the texts are converted from the format in which they are, PDF to HTML format. In the next stage the texts will be converted into a canonical format in which XML tags are used to delimit and identify the various parts [31], [32], [33]. Than, the text is tokenized: lexical units are recognized and tagged with the appropriate part-of-speech information (noun, determiner, verb, etc). Now, syntactic structures are identified and a superficial analysis is processed. In the next step semantic elements are identified and a

deep analysis is processed. The goal of the Natural Language Processing Module has been determined in terms of the restrictions the ontology provides, yielding an output specification that consists of linguistic information extracted. Finally, a specific ontology for the collection is created, actualized and the indexing process is ended.

We presented the general outline of operation and the basic characteristics of an Information Retrieval System that it uses Natural Language Processing and Ontologies. Syntactic and Semantic parsers are used supplying subsidies to allow the creation of Ontologies. The ontologies can be understood as tools that allow the construction of knowledge bases. In this project a basic ontology is used for the automatic generation of lightweight ontologies, what allows the construction of expressive representations of multiple relationships of conceptual structures of the text and among texts, with the intention of facilitating the answer to the users' queries and the navigation in an Information Retrieval System. We propose the use of syntactic parsers, applications and contributions of an automatic semantic analyzer based in semantic roles detection using meta-language and grammar rules. We believe that the right annotation (tags) of the syntactic context of texts improve the identification of the existent semantic relationship among the words. This allows the identification and creation of knowledge bases, which, together with the inherent characteristics to the ontologies, it will allow the development of a faster and efficient Information Retrieval System. Using our method it is possible to derive a hierarchical organization of concepts from a set of documents, without use of standard techniques based in training data and/or statistical clustering techniques. In the next stage of the project we will develop two others modules, one semantic sub-module to identify and categorize verbs and one to make inferences using computational semantics to solve some ambiguity problems.

ACKNOWLEDGEMENTS

The first author is supported by Conselho Nacional de Pesquisa (CNPq, http://www.cnpq.br).

REFERENCES

1. Lyman, P. and Varian, H. R. How Much Information. Technical Report. UCLA, Berkley. http://www.sims.berkley.edu/how-much-info/ (2000).
2. Dresner E. and Dascal M. Semantics, pragmatics, and the digital information age. Studies in Communication Sciences 1(2): 1-22 (2001).
3. van Rijsberger, C. J. Information Retrieval. http://www.dcs.gla.ac.uk/Keith/Preface.html. (1979).
4. Rowley, J. Organizing Knowledge: An Introduction to Information Retrieval. Vermont Gower Publishing, 2ed. (1996).
5. Baeza-Yates, R. & Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley Pub Co: 1st edition (May 1999).
6. Witten, I. et al. Managing Gigabytes. Morgan Kaufmann Publishers, Inc. Second Edition (1999).
7. Garfield, E. A Retrospective and Prospective View of Information Retrieval and Artificial Intelligence in the 21st Century. Journal of The American Society for Information Science and Technology. 52(1), (2001).
8. Bick, E. The Parsing System Palavras: automatic grammatical analysis of Portuguese in a constraint grammar framework. Arhus University Press. (PhD Dissertation) (2000).
9. Fillmore, C. "The case for case". In Universals in Linguistic Theory, edited by E. Bach and R. Harms. New York: Holt, Rinehart, and Winston. (1968).
10. Frederiksen,C. Representing Logical and Semantic Structure of Knowledge Acquired from Discourse. Cognitive Psychology 7, pp 371-458, (1975).
11. Ding, Y. & Engels, R. IR and AI: Using Co-occurrence Theory to Generate Lightweight Ontologies. DEXA Workshop, pp 961-965. (2001).

12. Maedche, A. & Staab, S. Mining Ontologies from Text. In: Dieng, R. & Corby, O. (Eds). EKAW-2000 - 12th International Conference on Knowledge Engineering and Knowledge Management. October 2-6, 2000, Juan-les-Pins, France. LNAI, Springer. (2000).
13. Hotho, A.; Staab, S. & Maedche, A. Ontology-based text clustering. In Proceedings of the IJCAI-2001 Workshop "Text Learning: Beyond Supervision", August, Seattle, USA. (2001).
14. Smeaton, A. F. Information Retrieval: Still Butting Heads with Natural Language Processing? In Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology M. T. Pazienza (Ed.), Springer-Verlag Lecture Notes in Computer Science # 1299, pp 115-138 (1997).
15. Lahtinen, T. Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods. Academic Dissertation. University of Helsinki. (2000).
16. Gerda R., 'Automatic detection of thesaurus relations for information retrieval applications', in Foundations of Computer Science: Potential - Theory - Cognition, pp. 499-506, (1997).
17. Brewster, C. Techniques for Automated Taxonomy Building: Towards Ontologies for Knowledge Management . In Proceedings CLUK Research Colloquium, Leeds, UK. (2002).
18. Frederiksen, C. H. Cognitive models and discourse analysis.*In*: COOPER, C. & GREENBAUM, S. (Eds.). *Studying writing: Linguistic approaches*. Beverly Hills,CA: Sage, Pp 227-267. (1986).
19. Larocca Neto, J.; Santos, A. D.; Kaestner, A. A.; Freitas, A. A. Generating Text Summaries through the Relative Importance of Topics. In M.C. Monard And J.S. Sichman (eds.), Lecture Notes in Artificial Intelligence, No. 1952, pp 300-309.Spring-Verlag. (2000).
20. Pereira, M. B.; Souza, C. F. R.; Nunes, M. G. V. 2002. Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português. Revista Eletrônica de Iniciação Científica. SBC. Março de 2002. Ano II, Volume II, Número I.
21. Bick, E. Automatic Parsing of Portuguese. In García, Laura Sánchez (ed.), Anais / II Encontro para o Processamento Computacional de Português Escrito e Falado. Curitiba: CEFET-PR. (1996)
22. _____ Portuguese Syntax (Teaching Manual), epositorio/Bick_Portuguese_Syntax3.doc and http://visl.sdu.dk/visl/pt, (2000-1).
23. Afonso, S.; Bick, E.; Haber, R. & Santos, D. Floresta sintá(c)tica: a treebank for Portuguese, Proceedings of LREC'2002. (2002).
24. Jackendoff, R. *Semantic Structures*. Cambridge: MIT Press, (1990).
25. _____. *Consciousness and the Computational Mind*. Bradford Book. The MIT Press, Cambridge, 1994.
26. Davies J.; Fensel, D. and van Harmelen, F. Towards The Semantic Web. Ontology-Driven Knowledge Management. John Wiley an Sons Ltd. (2003)
27. Erdmann, M.; Maedche, A.; Schnurr, H. – P.; Staab, S. From Manual to Semi-automatic Semantic Annotation: About Ontology-based Text Annotation Tools. In: ETAI Journal – Section on Semantic Web (Linköping Electronic Articles in Computer and Information Science), 6(2001).
28. Frederiksen, C. Just, M. & Carpenter, P. A Capacity Theory of Compreension: Individual Diferences in Working Memory. *Psychology Review* 99 Nº 1, pp 122-149 (1990).
29. Frederiken, C., Bracewell, B. A. & Renaud A. Psycognitive Representation and Processing of Discourse: Function and Dysfunction. *In*: Joanette Y.& Brownell H. (org.). *Discourse Hability and Brain Demade: Theoretical and Empirical Perspective*. N. Y. Springer Verlag. (1990).
30. Gottschalg-Duque, C. 1998. A Leitura em Ambiente Multimidia: A Produção de Inferências por parte do Leitor a partir da Compreensão de Hipertextos (Master Thesis). Programa de Pós-Graduação em Estudos Lingüísticos da FALE-UFMG. 16/11/1998. (1998).
31. Bax, M. P. Introdução às Linguagens de Marcas. Ciência da Informação. Brasília - DF: v.30, Num. 1, p.32 - 38. (2001).

32. Lobin, H. Textauszeichnung und Dokumentgrammatiken. In: Texttechnologie. Lobin, H & Lemnitzer, L. (ed.). Stauffenburg Verlag. (2003).
33. _____. Komplexität und Einfachheit in der Evolution von Dokumentgrammatiken. Zeitschrift für Literaturwissenschaft und Linguistik. Pp.106-122, September (2003).