Metalinguistic Information Extraction from Specialized Texts to Enrich Computational Lexicons

Ph.D. dissertation

Carlos Rodríguez Penagos

Tesi Doctoral

Departament de Traducció i Filologia

Universitat Pompeu Fabra

Sota la direcció de:

Toni Badia i Enric Vallduví

Barcelona, 2004

Dipòsit legal: B.14491-2005
ISBN: 84-689-1258-1

To Roser, with deep love and admiration

I wish to thank, in no particular order, the following friends and colleagues for their support, contributions, useful references and criticism, that helped me bring this work to fruition:

T. Badia, E. Vallduví, G. Sierra, H. H. Clark, S. Garrod, L. McNally, the IULATERM research group, J. Pustejovsky, J. Castaño, L. F. Lara and J.M. Fontana, as well as the anonymous reviewers of the various publications derived from this research.

TABLE OF CONTENTS

I STUDYING KNOWLEDGE THROUGH KNOWLEDGE OF LANGUAGE

I.1 Presentation

This is a dissertation about knowledge and language, about the ways they sustain and enable one another, and about their interplay in the generation, evolution and expression of Science and technical meaning. It is also about the understanding of how groups interact, and how they build, communicate and negotiate knowledge, makes possible the design of computer applications that allow better use of our own brainpower; This is vital when the combined efforts of many individuals throughout centuries have produced vast and complex theories and conceptual systems that aim at comprehension, representation and manipulation of our changing reality and the objects in our world. We want to show how language can be decisive in creating and sharing knowledge about reality as is jointly put forward by a group of experts in a domain.[1]

The ever-increasing size of the on-line production of modern scientific research is getting very difficult to manage and use efficiently. Databases and specialized dictionaries representing compilations of highly-structured conceptual systems are fairly efficient knowledge sources, but ongoing research, in the form of free-text technical papers and lively discussions among scientists, constitutes the true, real-time state of the art of a discipline. The wide adoption and availability of electronic text in academic and technical domains presents both a challenge and an opportunity: a challenge because, unless computational tools and resources are used, the sheer mass of information can overwhelm any analytical attempt; an opportunity, because intelligent gathering and processing of information from large amounts of text can be of great value in the examination, discovery, interpretation and creation of new and original knowledge.

Although many academic and commercial projects for the computational processing of natural language have been, for some years now, laboriously creating the resources (like lexicons, ontologies and computational grammars) that are needed for these important tasks, development of optimal domain-specific resources that best reflect a specific field still has to be

[1] Although this statement might bring to mind the endlessly debated Sapir-Whorf hypothesis, we want to restrict our arguments and evidence to the role that agreeing on what Clark (1996, 1998) has called a Communal Lexicon plays on the advancement of the consensus-based general knowledge of an

undertaken manually in a case by case basis, consuming a considerable amount of time, money and effort. This seems to be unavoidable, since it is accepted that technical sublanguages, although sharing most of their lexicon and grammar with common, everyday language, have very specific terminological lexicons and grammatical and stylistic quirks that need to be taken into account for the proper interpretation of the complex processes and conceptual structures they convey (Kittredge, 1982).

For a long time the specificity of technical knowledge has not received, in our view, a satisfactory theoretical account that might properly ground its computational processing, especially in the field of domain-specific lexicography, or Terminology. Until recently, the standard description of terms was a mentalist and conceptualist one best represented by the work of E. Wüster and the International Standards Association, of neopositivist inspiration. Although the Vienna Circle advocated an empiricist epistemology, it also gave much weight to a logicist viewpoint when concerned with scientific knowledge. As a result, modern Terminology and the resources and applications produced by it were very much static and concept-driven, and terms considered as the linguistic expression of domain-specific conceptual systems were not understood in a cognitive and communicative framework that could do justice to their true complexity and richness. Onomasiology had been the preferred methodology for terminological work, although this paradigm is slowly and cautiously being challenged. Scientific knowledge, under this view, was a platonic world of rational ideas or concepts somehow shared by all scientists, and the sublanguages and the terms where it became alive did not seem to be much more that pale shadows projected over the rugged walls of a dark cave. The representation of the dynamic nature of technical concepts and knowledge required a data-driven approach similar to the one that the use of corpora has offered to the field of modern-day Linguistics, which brought it closer to the truly scientific status of other empirical disciplines like Physics or Biology. We would like this dissertation to be an important step in this direction.

In this work we will study, using various corpora from highly specialized domains, a very specific aspect of specialized discourse in order to apply Corpus and Computational Linguistic techniques to the machine extraction of sublanguage information from unstructured text. This work studies metalinguistic predication in text, as well as demonstrate how to represent and manipulate it computationally. This is important since Metalinguistic predication is present in all textual domains. We have attempted a corpus-based description of what we have termed Explicit Metalinguistic Operations (EMOs):[2] sentences where discourse reflects upon itself,

---

academic field, as analyzed in corpora of leading-edge papers from peer-review journals and other highly-specialized texts.

[2] C. Rodríguez, 1999a, 1999b, 2000, 2002, 2003 & 2004

where language is itself the subject, where language is creating and manipulating the elements and rules that make it possible. What follows is an example from our research corpus of the kind of definition-like metalinguistic utterances to be examined and processed here, although, as in this case, only part of the sentential construction can be considered to be truly about some aspect of language:

> Integral power results in a fundamental type of social classification which, adapting Bernstein's terminology, I shall call "frame" (Bernstein 1971).

In this case, the author is explicitly specifying what will be its usage of a term (“frame”), what if will refer to, and what is the inspiration or source for this terminological definition.

We will be concerned here with a relatively small but vital sector of technical communicative exchanges, where discourse takes a brief detour from its usual domain subject matter to reformulate the very meaning conditions that enables it to be a medium for the exchange of ideas and information. Metalinguistic exchanges are fundamental discourse operations with powerful and lasting consequences not only for the successful continuation of scientific dialogue, but for the very constitution of: (a) technical knowledge as a collective enterprise and (b) social groups defined by a domain expertise and a shared lexicon. The basic configuration of an epistemic common ground through negotiation of a shared terminology is possible because metalinguistic exchanges have cognitive and formal characteristics that include, among other things, powerful lexical markers that make them prominent in discourse and guide their interpretation.

The core claims of this dissertation are: (1) that those interactive discourse processes are foundational actions in the intersubjective construction of scientific knowledge (especially theories and explanatory models); (2) that metalinguistic predication, by virtue of its formal and pragmatic properties, can be processed in a robust and theoretically-motivated manner by computer applications that obtain domain-specific terminological and linguistic information; (3) that the resources obtained with these methods can be useful for empirical studies of scientific knowledge and expert communities, as well as for Natural Language Processing applications and specialized lexicography. Vital to these general goals is an adequate theoretical account of metalanguage’s role in the buildup of knowledge (presented in Chapter II), as well as a descriptive model that properly analyzes those metalinguistic sentences in order to understand how they function in discourse (Chapter III). An important part of the foundations for the applications described and evaluated in later chapters is the identification of the basic components of the definition-like sentences we have termed EMOs. In general, Explicit Metalinguistic Operations like the following one from a Biomedicine paper:

> The characteristic syndrome associated with lumbar stenosis is termed neurogenic intermittent claudication.

can be decomposed into three constitutive and distinct element that have particular semantic and cognitive features:

1.- A term that is being defined or for which some information is being provided, in this case: "*neurogenic intermittent claudication*".

2.- The actual information provided or put forward, that defines for the first time or modifies the standing semantic content for the term, here: "*The characteristic syndrome associated with lumbar stenosis*"

3.- Some lexical or paralinguistic elements that allow the adequate processing and interpretation of a metalinguistic sentence as such, in this case the verbal item "*termed*" that predicatively connects the two previous elements and flags the metalinguistic nature of the whole utterance.

Mining terminological information from free or semi-structured text in large-scale technical corpora is slowly becoming a reasonably mature technology, with term extraction systems leading the way. These systems are capable of locating and extracting terminology by exploiting the syntactic and statistical regularities of terms, but most of the work involved in defining them and informing them into terminological databases still has to be left to human experts. Compiling the extensive resources needed by modern scientific and technical disciplines to manage the explosive growth of their knowledge has become a necessity. A good example of the NLP-based processing need driving these efforts is the MedLine abstract database maintained by the National Library of Medicine[3] (NLM), which incorporates around 40,000 Health Sciences papers each month. Researchers depend on these electronic resources to keep abreast of their rapidly changing field. In order to maintain and update vital indexing references such as the Unified Medical Language System (UMLS) resources, the MeSH and SPECIALIST vocabularies, the NLM staff needs to review 400,000 highly-technical papers each year (Powell et al., 2002). Most of these terminological knowledge sources have been compiled from existing glossaries and vocabularies that can become dated fairly quickly, and eliciting this enormous amount of data from experts in the field is not a practical option. Neology detection, glossary update and other tasks can benefit from automatic search of semantic and pragmatic information in highly technical texts, e.g. when new information about sublanguage usage is being put forward.

[3] http://www.nlm.nih.gov/

In the work described here, proven Information Extraction techniques combined with novel strategies for locating knowledge-rich contexts have allowed us to mine vast textual corpora to obtain valuable lexical resources that can drive exploration, description and representation of the dynamic evolution of scientific enterprise. The core contribution of this dissertation is a proof-of-concept system, the Metalinguistic Operation Processor (MOP, described fully in Chapter IV), that exploits our theoretical models and empirical findings to automatically generate, from unstructured documents of a highly technical nature, special lexical-knowledge bases we have termed Metalinguistic Information Databases (or MIDs). These databases are not full-fledged terminological knowledge bases, but they do constitute richly-textured knowledge sources about a discipline's configuration and sublanguage.

The MOP system described here applies standard pre-processing techniques such as tokenization, POS tagging and partial parsing on domain-specific textual corpora. It then locates and extracts metalinguistic fragments employing lexical and punctuation indicators obtained from analyses of an extensive corpus of such operations, and disambiguates them using a two-pronged approach involving machine-learning and finite-state techniques. After obtaining an accurate set of EMO sentences to work with, the MOP system parses them into a database structure by using heuristic rules derived from manual analysis of such sentences and of their lexical markers, as well as from relevant semantic frames represented in resources such as the FrameNet project (Baker et al., 1998). The MOP system is to a large degree independent of the domain, and it strives to maintain simplicity in its processing mechanisms, avoiding complex or resource-costly techniques such as full syntactic parsing or full semantic interpretation. Although we have opted for low-level processing, further improvements of the present system undoubtedly can be expected if more sophisticated NLP machinery is implemented in the future, like an anaphora resolution module.

Although we will discuss the MOP system and MIDs in depth in latter chapters, we can briefly offer herean overview of what we believe to be their relevance for the representation and processing of specialized language. Conventional resources like lexicons and dictionaries are considered stable references for the sublanguage shared by a domain-defined community. They can be seen as static repositories of the default, core lexical information of terms used by a research community (that is, the information available to an average, idealized speaker in that domain- and linguistically-defined community). A limitation of lexical databases understood as "holding devices" for lexical data is their failure to properly represent a language's productivity and its open-endness (Boguraev and Levin, 1993). A Metalinguistic Information Database might contain the multi-textured real-time data embedded in research papers and technical documents, and in this sense could be viewed as something completely different from a

lexicographic artifact: it can be seen as a listing of exceptions, special contexts and specific usages where meaning, value or pragmatic conditions have been spotlighted by discourse for cognitive reasons. Terminological data in MIDs can be more specific and might be better suited for the interpretation of certain texts or utterances. To rely solely on lookup of manually-compiled resources might miss some of the data provided by metalinguistic speech acts, where the term is put forward for the first time, or where important, context-sensitive information about the term is offered. Another difference of MOP & MIDs with regard to conventional terminology extraction systems based on syntactic regularities of term formation is that our context-centered approach offers better precision than them, although its recall is much lower because its goal in not to locate all terminology, but only those terms for which something is being stated.

In addition, the MOP-created Metalinguistic Information Databases might more accurately mirror the dynamically changing nature of technical knowledge, which is always subjected to the unavoidable social influence of groups of experts and scientists interacting as peers. Although there has been recent interest in applying NLP techniques to terminological analysis in order to capture the implicit systematicity of conceptual (Kageura, 2002) or linguistic (Jacquemin, 2001) term formation, the MOP system's exploitation of the metalinguistic dimension of text allows to go beyond term and neology detection through formal and quantitative means; Using explicit statements expressed in discourse, the MOP system can locate the non-predictable and idiosyncratic ways in which specialized knowledge evolves and expresses itself. These diverging approaches reflect the options of symbolic or statistic approaches to general language computation.

Even with the obvious limitations of the semi-structured lexical resources that are MIDs (to be discussed in later sections) we believe that our analyses and their application in the MOP system can prove to be useful tools for empirical research into the nature of expert knowledge, as is attested in the interaction of scientists and scholars that struggle to convey theoretical explanations and descriptions to their colleagues. As we have stated above, early approaches to the study of scientific knowledge and discourse involved historical, logical, sociological or speculative lines of thought, or were carried out by elucidating goals and beliefs from members of the scientific community. Although locating and processing metalinguistic repair in scientific communication involves a sparse portion of textual segments, the information obtained with our techniques could be combined with computational analysis of rhetorical structures (as in Teufel and Moens, 2002) or text summarization techniques (Fuji and Ishikawa, 2004) to provide a more detailed look at how knowledge-building in science functions from a discourse perspective. We believe that using computational and Corpus Linguistics techniques for this

goal can provide an alternate methodology to speculative and inductive studies about the nature of scientific thought and the interaction of scientists. Up until now, enormous amounts of information in digital format have been underutilized to research interactions among scientists. In contrast, modern linguistic research has received an enormous boost in its consolidation as an empirical science by the development of Corpus Linguistics techniques that ensure access to actual, hard data about real linguistic performance, in contrast to previous methodologies that relied on intuitions and introspective judgments of personal competence and grammaticality. It is now possible to evaluate linguistic theories or compare descriptive approaches by employing all manner of corpora and lexical resources, with standards in the hundred of millions of words that ensure good coverage and representation of a wide array of phenomena. Using techniques employed to increase our knowledge of language in order to study the linguistic realization of scientific and technical knowledge seems a logical step. Our data-driven approach to terminological compilation would allow direct empirical observation of the evolution of scientific language in the ongoing debate that reflects a changing conceptual state of the art in academic disciplines. By observing actual scientific exchanges as they focus on its own conditions for lexical and linguistic expression, we should be able to observe intersubjective knowledge-construction processes in the context of peer-to-peer communication. Quantitative as well as qualitative data can be expected from such an approach, and no recourse to purely mentalist explanations would be called for.

After presenting an overview of the kinds of issues this dissertation addresses, discussing some methodological considerations and describing our research corpora in this first introductory chapter, we start by reviewing the general concept and function of metalanguage in Chapter II. We present an elaborated theoretical account of the conditions under which metalanguage functions in one kind of learned exchange: printed scientific discussion mediated by peer-review. We show how such exchanges allow the systemic construction of a shared epistemic common ground, and how they help further the ultimate goal of scientific pursuit: the theories and models that help explain and describe reality through consensus, empirical evidence and rational interpretation. This chapter is heavily oriented to theoretical issues from the perspective of philosophical and epistemological considerations, and some application-oriented readers might either want to leave them for later or skip them altogether to go to the next two chapters, which contain the core contributions of this dissertation. Chapter III presents a more empirically-rooted speech act analysis of metalinguistic sentences, based on specialized corpora and previous literature on the subject, followed by a simple descriptive model of metalinguistic predication in general, in which lexical markers constitute the axis that signals, articulates and enacts metalinguistic operations of an explicit nature. In this chapter we give an

account of the basic elements and functions of the EMOs we have already mentioned before, and point at how this knowledge can be used in a terminology-informing application, in effect grounding the implementation of an Information Extraction system geared towards metalinguistic data. Chapters IV and V describe and evaluate a computational tool, the MOP system, devised to exploit the pragmatic and semantic regularities described by our model, in order to automate the creation of the non-standard terminological resources we have called Metalinguistic Information Databases. We use and refine finite-state and stochastic techniques commonly employed in state-of-the-art Natural Language Processing such as Information Extraction and text classification, and implement them in novel ways. Finally, in chapter VI, we discuss and describe the unorthodox repositories of lexical and terminological information that constitute the end result of MOP processing, the Metalinguistic Information Databases. We also discuss here some of the limitations of the approaches we have chosen and of the tools we have developed, as well as the potential uses and applications envisioned for them, such as Question & Answering systems or tracking and representation of the evolution of scientific thought as mirrored by the dynamically-changing terminological systems that sustain it. An enclosed CD-ROM contains the application code, data files and module documentation, as well as electronic copies of this and other publications that refer to this research.

I.2 Formulating the general problem

It is an obvious fact that the retrieval of any kind of useful knowledge from text has to proceed from a framework of linguistic analysis and interpretation. What is more difficult to say is at what level of complexity this analysis should be attempted, and where we should best invest our processing resources in order to detect those sections of text that are relevant for our purposes. Until fairly recently, Natural Language Processing (NLP) research developed approaches that privileged those aspects of language that more easily adapted to logic and formal constraints, and so were better suited for algorithmic treatment. Syntactic patterns or grammatical classes could, for example, be used to provide candidates for term detection. On the semantic side, observed regularities of meaning in terms prove to be more difficult to formalize though one could claim they are the actual repositories of knowledge about the world. In some fairly recent applications, semantics was sometimes restricted to queries sent to machine-readable dictionaries (MRD) or other implicit means of semantic representation, such as conceptual hierarchies or ontologies which are ruled by meronymic and hyperonymic relationships, as in WordNet (Miller et al., 1990), although more sophisticated theoretical frameworks like the Generative Lexicon (Pustejovsky, 1995) and LKB (Copestake, 1993) have contributed to a more robust representation of linguistic meaning. In recent theoretical frameworks (Generalized

Phrase Structure Grammar, Head-driven Phrase Structure Grammar or Lexical Functional Grammar) the lexicon has become a central part of the linguistic apparatus, including syntax and grammar, and the need to acquire a systematic and exhaustive representation of it is now part of mainstream research in linguistics.

With the advent of full-fledged textual corpora the need for sophisticated retrieval tools has increased significantly. The quality and quantity of linguistic data potentially available for actually using and understanding knowledge inherent in written documents has grown commensurately with the growth of the world-wide web. But we already have a powerful device for storing and transmitting human knowledge. Already in Aristotelian times definitions were identified as discourse devices able to transmit and create knowledge, both linguistic knowledge (when we define a word) and encyclopedic knowledge about the word when we define a concept meant to reflect how reality is organized. Traditional dictionaries and specialized vocabularies are one obvious source of semantic information, but the incredible amount of available digital text about all kinds of topics, contexts and fields constitutes an inexhaustible reservoir of meaning. There is a need to learn how to efficiently locate relevant fragments of text in data not yet lexicographically structured. This dissertation shows how this information can be located, processed and exploited computationally.

As mentioned earlier, another discipline that can benefit from a data-driven approach to specialized meaning is Terminology, especially now that it is slowly starting to depart from its Wüsterian origins and move into a less conceptual-based definition of terms. Heir to the Vienna Circle's philosophy of Science, the previous onomasiologic conventional wisdom dictated that linguistic terms were only labels to be applied to technical concepts, the latter supposedly being clear mental ideas shared by domain experts. Although this narrow conception still dominates ISO methodologies, a fresher, more cognition-oriented and communicational view of terminology (Hermas, 1991; Ahmed, 1996; Temmerman, 1997; Zawada & Swanepoel, 1997; Cabré, 1998 & 1999; Rodríguez 2001b) now recognizes that both concepts and terms are constructed in the midst of personal, social and group conditions that shape all linguistic interchange. This line of research parallels a historicist view of science inaugurated by names such as Kuhn (1962), Achinstein (1962, 1968), and Feyerabend (1965).

Recent access to domain-specific corpora of sufficient size shows the high fluidity of scientific concepts, and the non-referential, non-denotative nature of many of the important terms used in technical sublanguages, contradicting one of the central tenets of terminological work.[4] The purported univocity of terms was just a convenient way to organize them, a

[4] "As items of natural language discourse, terms are elements of language and could in principle be described purely linguistically by means of the sense relations they form in discourse. Their meaning

lexicographic convention that is still valid only if we recognize it as such. The primacy of concepts is a methodological illusion.[5] The MOP system we will describe, and the MIDs that it can generate, show terminologies to be metastable systems in constant change, with new paradigms constantly being born and old ones forgotten.[6] Terms and concepts are dynamic entities which are fixed neither mentally nor linguistically, except in the artificial snapshot of a particular point in time, and even then many of them are conceived and expressed differently, depending on context, theories and even specific textual instances. Modern lexicographic reference has to reflect this fluidity, and it can only do so by combining an onomasiological approach with a semasiological, corpus-based study of terms as eminently linguistic entities.

Any serious attempt to consider the extensive problematic fields of language and knowledge (and their mutual interactions) needs to take into account many related issues that have been explored from the perspective of many disciplines throughout the history of human thought. Many of them are ancient problems first tackled by Greek philosophy, or at the dawn of scientific inquiry. At one point or another, our study will have to deal with some of these questions, or variations thereof. Even if we attempt to consider them systematically, most of them will have to wait for other research in order to get partially satisfactory answers, or alternatively they will end up uncovering in turn new problems and issues that this dissertation is not committed to solve, as they pertain properly to other disciplines or extend beyond our present goals. We will merely mention some of these relevant issues, as they can be formulated from different theoretical and descriptive perspectives:

From an epistemological perspective:

How is new knowledge created, validated and circulated among groups of experts in the field? How do group dynamics in expert communities help shape that knowledge? What is the linguistic nature of theories? Is scientific advance also a consensual matter, or is it just a matter of logic, evidence and rationality? What role does persuasion play in science? How is discovery

---

would thus, with Wittgenstein, reside exclusively in their use in the context of other words. But in a theory of terminology, the nature of the special reference (...) leads to a different approach in the method of definition. By convention special lexical items are considered to be devoid of other than referential meaning within their area of usage, i.e. special subject communication. Because they occur in a limited range of collocations only, for the purpose of definition terms are also considered to be context-free." (Sager, 1990)

[5] "The primary object of terminology, the terms, are perceived as symbols which represent concepts. Concepts must therefore be created and come to exist before terms can be formed to represent them. In fact, the naming of a concept may be considered the first step in the consolidation of concepts as socially useful or usable entities." (Sager, *ibid*)

[6] Besides the classic description of revolutionary change in science by Khun (1962), see Gentilhomme (1994) for the idea of metastability (a concept originally from chemistry) in terminology, and Beaugrande & Dressler (1981) for similar ideas in discourse studies.

turned into new knowledge? Is it possible to study these epistemic processes empirically using text collections of a highly specialized nature?

From a linguistic perspective:

From a strictly linguistic point of view, what is expert knowledge? What is the nature of complex terminological systems and technical sublanguages? How are terms different from the items of everyday use? How is everyday language different from domain-dependent sublanguages? Are these learned differently? How are new meanings and usage conditions in technical sublanguages established and manipulated by their main users? What are the basic characteristics of metalinguistic statements? Can a theoretically-motivated taxonomy of metalinguistic information account for its variety and efficacy? Can we study specialized discourse to explore the crossroad issues of semantics and pragmatics?

From a cognitive perspective:

Do linguistic factors influence knowledge-building processes? How are these operations interpreted by readers of highly specialized texts? What are the features that make them prominent, and flag their special nature? Is there a single property that explains their cognitive importance? Is there any functional redundancy in this cognitive processing?

From a computational approach:

Is it possible to formalize such high-level discourse processes? Are they regular enough to allow for algorithmic treatment and exploitation? Is an Information Extraction system that obtains purely linguistic knowledge feasible? Can this data be suitable to formalization into a lexical knowledge base without losing too much valuable information? How can multidimensional and heterogeneous language-related information be incorporated into a useful data structure? How can we complement regular lexical databases and computational lexicons with non-standard lexical information? What would be the usefulness of such applications for Artificial Intelligence, Knowledge Engineering and specialized lexicography?

I.3 Claims of novelty

All of the previously-mentioned issues are important, in one way or another, as general problems related to the research we present here, but most of them are not part of the core contributions of this dissertation. We do not attempt or claim to bring a definite answer to any of them. The novelty of this proposal should not be expected to come from wholly original techniques for understanding, manipulating or representing meaning in domain-specific texts, although some of these techniques, like the learning-based identification of metalinguistic activity described in chapter IV, have not been reported in the literature before. There have been

many efforts at automatic extraction of terminology and definitional information from text,[7] but our work extends and refines those experiences to include metalinguistic information of a wider spectrum than classic definitional structures, and which up to now had not been compiled because it was not deemed structured enough or valuable enough to be suitable for computational processing.

Our central claims to novelty lie, then, in presenting an adequate theoretical object (EMOs) on which to build an application that can retrieve metalinguistic information in general, including definitions similar to the ones encountered in lexicographic entries. Even though some of the methods employed here have proven to be successful in Information Extraction tasks to obtain the extra-linguistic data of events and situations, our use of some of those computational techniques on the metalinguistic dimension of free-form text is certainly novel. Even without the theoretical apparatus we present (in chapters II and III) for the phenomena over which our MOP system operates, we believe that its implementation and the Metalinguistic Information Databases produced constitute a clear advancement of the state of the art in natural language technologies, especially with regard to the processing and study of domain-specific sublanguages.

The MOP system is a work in progress, and should be conceived as a proof-of-concept implementation aimed at showing the basic soundness of the theoretical description of Explicit Metalinguistic Operations, and of how some of their cognitive and linguistic features allow for computational manipulation. The MOP system can be a useful source for the compilation of more sophisticated NLP resources from MIDs, and any further valid uses for them others might come up with, like neology detection or taxonomy bootstrapping, are indeed welcome, but they lie beyond the of our contribution.

But our proposal is just one example of the use of knowledge about how metalinguistic activity proceeds cognitively and linguistically. We are sure other applications that benefit from them could be envisaged. Here we claim only that the indicators we have found point to conceptually important nodes of text, and that they are more than just collocation data of no epistemic importance. Even though we do not claim to have provided "the definitive application", we do suggest the need for applications that can "mimic" our impressive human competence as efficient readers of technical subjects, as incredibly good lexical-data processors that constantly update and construct our own special purpose vocabularies. Our suggested application may not be the most computationally suitable to the task or the most efficient, and some problems with how to organize and classify the resulting data remain unsolved. At its

[7] The British National Corpus manual has a chapter on how to find definitions using lexical makers.

most basic level, this dissertation is an invitation to look beyond the notion of conventional definitions and exploit the more general and powerful dimension of metalinguistic activity in the search for better and smarter processing of language and knowledge.

I.4 Methodological issues and data sources

Corpus Linguistics techniques (McEnery and Wilson, 1996) now allow study of linguistic phenomena that is data-driven, and it grounds research to linguistic occurrences actually attested in various corpora that provide adequate representation of a subject, field or genre. Properly marked-up textual collections also serve as training datasets for NLP applications. Corpus Linguistics constitutes the empirical foundation of modern linguistics. In following these methodological principles, we have accessed resources, preexisting or compiled by us, that can accurately reflect how technical or scientific language is produced and used in various fields; in doing so we can reduce to a minimum the umber of made-up and altered example sentences used to present specific aspects of metalinguistic predication. We have not studied oral or conversational communications, and we have centered on written text to have a more homogeneous data set, and, although we do not dismiss the importance of face-to-face exchanges (such as academic congresses and classroom conversation) for the advancement of a discipline we do believe that written text has a more profound impact on mid- and long-term terminological constitution. Scientific texts have as their main function to enhance and transmit a widely-accepted knowledge about a "real world", through the analysis of empirical or documental evidence (de Beaugrande & Dressler, 1981).[8]

In what follows, we describe the text collections we have employed in our study of metalinguistic phenomena in highly technical documents. All corpora are available in the enclosed CD-ROM.

I.4.1 Description of specialized corpora used in our research

I.4.1.1 Sociological research corpus

Our first exploratory corpus was obtained from 19 sociology articles published during a five year period (1995-1998) in various British, American and Canadian academic journals (both in print and on line) with strict peer-review policies. The selection of these articles attempts to cover a wide range of subjects within the discipline, from computer-simulation of social processes to analysis of advertising images. We also made sure that the articles and papers have a theoretical component in the discussions, and represent samples from different regional varieties of English, so as to dilute bias by these factors.

[8] J-P. Bronckart (1985) also emphasizes their theoretical component.

Another factor influencing our decision to start with a sociology corpus was the existence of previous studies about the discipline, like Lachenmayer's (1971) essay on the methodological and terminological status of the discipline that could be useful points of contrast. The complete list of text for this Sociology corpus in presented next.

Breen, R. (1997)	Risk, Recommodification and Stratification; **Sociology** Vol. 31 No. 3, August
Campbell, C. (1996)	On the concept of "motive" in sociology; **Sociology** Vol. 30 No. 1 February
Delanty, G. (1996)	'Beyond the Nation-State: National Identity and Citizenship in a Multicultural Society - A Response to Rex', **Sociological Research Online**, vol. 1, no. 3, http://www.socresonline.org.uk/socresonline/1/3/1.html
Delgado-Moreira, J. (1997)	Cultural Citizenship and the Creation of European Identity. **Electronic Journal of Sociology**: 2, 3. http://www.sociology.org/content/vol002.003/delgado.html
Fahey, T. (1995)	Privacy And The Family: Conceptual And Empirical Reflections **Sociology** Vol. 29 No. 4
Gilbert, N. (1997)	'A Simulation of the Structure of Academic Science', **Sociological Research Online**, vol. 2, no. 2, http://www.socresonline.org.uk/socresonline/2/2/3.html
Jargowsky, P. (1996)	Take the money and run: economic segregation in u.s. metropolitan areas. **American Sociological Review.** Vol. 61 (pp.: 984-998)
Jewitt, C. (1997)	Images of Men: Male Sexuality in Sexual Health Leaflets and Posters for Young People; **Sociological Research Online,** vol. 2, no. 2, http://www.socresonline.org.uk/socresonline/2/2/6.html
Lee, D. (1994)	Class as a social fact ; **Sociology** Vol. 28 No. 2 May 1994 p. 397-415
Lehmann, J. (1995)	The question of caste in modern society: Durkheim's contradictory theories of race, class, and sex. **American Sociological Review.** Vol. 60 No. 4; p. 566
Mainprize, S. (1996).	Elective Affinities in the Engineering of Social Control: The Evolution of Electronic Monitoring. **Electronic Journal of Sociology**: 2, 2. http://www.sociology.org/content/vol002.002/mainprize.html
McKie, J. (1996)	'Is Democracy at the Heart of IT? Commercial Perceptions of Technology', **Sociological Research Online**, vol. 1, no. 4, http://www.socresonline.org.uk/socresonline/1/4/1.html
Orbell, J., Zeng, L. & Mulford, M. (1996)	Individual experience and the fragmentation of societies. **American Sociological Review.** Vol. 61 (pp.: 984-998)
Payne, G., Payne, J. & Hyde, M. (1996)	"Refuse of All Classes"? Social Indicators And Social Deprivation, **Sociological Research Online**, vol. 1, no. 1, http://www.socresonline.org.uk/socresonline/1/1/3.html
Rex, J. (1996a)	'National Identity in the Democratic Multi-Cultural State', **Sociological Research Online**, vol. 1, no. 2, http://www.socresonline.org.uk/socresonline/1/2/1.html
Rex, J. (1996b)	'Contemporary Nationalism, Its Causes and Consequences for Europe - A Reply to Delanty' **Sociological Research Online,** vol. 1, no. 4, http://www.socresonline.org.uk/socresonline/1/4/rex.html

Smith, R. D. (1998) Social Structures and Chaos Theory; **Sociological Research Online**, vol. 3, no. 1, http://www.socresonline.org.uk/socresonline/3/1/11.html

Thomas, R. (1996) 'Statistics as Organizational Products', **Sociological Research Online,** vol. 1, no. 3, http://www.socresonline.org.uk/socresonline/1/3/5.html

Treanor, P. (1997) Structures of Nationalism; **Sociological Research Online,** vol. 2, no. 1, http://www.socresonline.org.uk/socresonline/2/1/8.html

The complete Sociology corpus has 138,183 words, with 5,583 lines identified by our tokenizer and normalization apparatus via punctuation clues as clauses and/or complete sentences. Using criteria described in Section III.1 (Listing I), we manually identified 243 sentences as providing metalinguistic information, a 4.35% of total. We flattened all texts to ASCII format for processing.

I.4.1.2 The expanded EMO corpus from the BNC

From our initial exploratory corpus described above, we selected lexical patterns that could be indicators of metalinguistic activity in text, and expanded the list to 116 different patterns using other plausible verbal forms, as well as lexical items and nominal modifiers such as *term, word, phrase, vocabulary, terminology,* etc., that could indicate that the sentence was metalinguistic in nature. Our observed markers broadly overlap inventories done by Pearson (1998), Meyer (2001) and others that have described technical definitions in context. These patterns were sent as queries to the British National Corpus[9] written portions of Social Science, Applied Science, Natural and Pure Sciences, Belief and Thought domains. The query results were retrieved in SGML and transformed into XML format for later markup and classification. A total of 10,837 hits (641,214 tokens including words and punctuation) were compiled, with a variable amount of context allowed. The list of queries performed is shown next:[10]

1. ('a term')	8. ('applied to')	15. ('called')
2. ('a term')	9. ('applies to')	16. ('called')(''=PUQ)
3. ('a term'|'the term'|'this term')	10. ('apply the term')	17. ('calls')
4. ('adjective')	11. ('apply the word')	18. ('calls')(''=PUQ)
5. ('adverb')	12. ('apply to')	19. ('christened')
6. ('ambiguity')	13. ('as `')	20. ('coin')
7. ('ambiguous')	14. ('call')	21. ('coined')

[9] The BNC is accessible at http://www.natcorp.ox.ac.uk/.

[10] Some queries were repeated to obtain sufficient examples in light of either their relevance or the ambiguity of the lexical item.

22. ('coins')
23. (', a term')
24. (', known as')
25. (``'=PUQ)*(', known as')
26. (``'=PUQ)*(', known as')
27. (``'=PUQ)*(', the term')
28. (','=PUN) ('or')
29. ('concept')
30. ('connotation')
31. ('context')
32. ('could be called')
33. ('define')
34. ('defined')
35. ('defines')
36. ('definition')
37. ('definition of')
38. ('definition of')(``'=PUQ)
39. ('designate')
40. ('designates')
41. ('dub')
42. ('dubbed')
43. ('dubs')
44. ('euphemistic')
45. ('expressed')
46. ('expresses')
47. ('expression')
48. ('heading')
49. ('i.e.')
50. ('implies')
51. ('imply')
52. ('in terms of')
53. ('in terms of')(``'=PUQ)
54. ('is called')
55. ('is called a')(``'=PUQ)
56. ('is called')(``'=PUQ)
57. ('known as')
58. ('known as')(``'=PUQ)
59. ('knowns as')
60. ('label')
61. ('labeled')
62. ('labelled')
63. ('labels')
64. ('mean')
65. ('meaning')
66. ('means')
67. ('meant')
68. ('might be called')
69. ('namely')
70. ('oxymoron')
71. ('paraphrasing')
72. ('refer to')
73. ('reference')
74. ('referent')
75. ('referred to')
76. ('refers to')
77. ('sense')
78. ('sign')
79. ('so-called')(``'=PUQ)
80. ('so called')
81. ('so called')(``'=PUQ)
82. ('stand for')
83. ('stands for')
84. ('synonym')
85. ('synonymous')
86. ('talks of')
87. (``'=PUQ)*('talks of')
88. ('term')
89. ('term `')
90. ('termed')
91. ('terminology')
92. ('terms')
93. ('use'|'uses')*('the phrase')
94. ('the term')
95. ('the term`')
96. ('to refer to')
97. ('usage')
98. ('use')
99. ('use of the term')
100. ('use the term')
101. ('use'|'uses') ('a term')
102. ('use'|'uses')('that term')
103. ('use'|'uses')('the term')
104. ('use the term') (``'=PUQ)
105. ('use'|'uses')('the word')
106. ('use'|'uses')('this term')
107. ('used to refer to')
108. ('uses')
109. ('verb')
110. ('vocabulary')
111. ('what') (_) ('called')
112. ('what') (_) ('calls')
113. ('where') (_) ('is')
114. ('where') (_) ('refers')
115. ('word')
116. ('word') (``'=PUQ)

Hits were later manually reviewed and marked up as metalinguistic or not using the criteria defined in Section III.1. The statistics were calculated automatically and incorporated into the xml document. Overall, 5,430 of those sentences were found to be true EMOs, while 5,407 were not, for a 49.6 % rate that is close to a 0.5 ratio of chance distribution. This benchmark corpus, henceforth referred to as the EMO corpus, allowed us to experiment with machine-learning techniques for the task of EMO identification, and also served as a test ground for our final IE system. A sample of one of the resulting xml files as visualized with a web browser is shown next.

Pattern: **('stands for') *****

Of 85 hits in the BNC query, there are 45 that are EMOs. **There are 40 negative hits.** Percentage: 52.9411764706 %

Hit	*EMO?*	Ref.
It is, in fact, what the brand **stands for** .	**N**	685
This sum is impossible if you do not know what x **stands for** .	**N**	1608
We then collect together all the remaining processor flags and registers into what is known as a process state word or PSW (some computer manufacturers say that the P **stands for** "processor", or, incorrectly, for "program").	**Y**	1401
The change in the orbit of the earth is too slow to be observed, but this same effect has been observed over the past few years occurring in the system called PSR 1913+16 (PSR **stands for** "pulsar," a special type of neutron star that emits regular pulses of radio waves).	**Y**	713
2 If you are told that 1 **stands for** 8, that is 1 8, find the value of the following: ***formula***	**N**	1613
What will be the new proper numbers of the agents at the side of the page when now **stands for** 8?	**N**	1536
The linguistic expression ` **stands for"** , `is a substitute for", the natural expression.	**Y**	1580
(…)		
The `p and q" **stands for** `peace and quiet", and it meant that she could be totally on her own.	**Y**	1450
ROM **stands for** `read only" memory.	**Y**	121
A diagram of its rhythmical structure can be made, where s **stands for** `strong" and w stands for `weak".	**Y**	814
7.3 If they touch on the point, discussions of restriction invariably proceed in a way which implies that the property of a restrictive adjective **stands for** a property which is to be ascribed to the entity of the noun phrase.	**N**	875
When numbers are written as figures a given digit **stands for** a word.	**N**	1245
The abbreviation for computer display is surely wrong; VGA **stands for** video, not virtual, graphics array.	**Y**	2373
The Mountain View, California-based Unicode Consortium has merged its multi-lingual encoding standard with the recently approved ISO 10646, developed by the International Standards Organisation: the idea is that computers all over the world should agree on which number **stands for** which character so they will be able to communicate other in any language.	**N**	407

I.4.1.3 Other corpora

Another document set used as corpus for our work is the on-line Histology manual[11] written by Dr. William A Beresford, professor of the Anatomy department, at West Virginia University. It has a total of 98,915 words (including punctuation) and 9,368 lines. The text presented 69 metalinguistic sentences.

The MedLine biomedical sciences abstract database provided another document set for exploration of metalinguistic activity in text. MedLine is an indexed resource containing references for 4,600 biomedical journals and more than 14 million documents, and is part of what has become known as the biobibliome, the ever-growing literature on Biomedicine. Our sample contains abstracts from 400 articles, with a total of 1,043 normalized lines, 43,943 words and 10 EMOs located. Although abstracts are not prime textual sections for

[11] http://wberesford.hsc.wvu.edu/histol.htm

metalinguistic definition, as they constitute condensations of the main lines of arguments in the article, we believe that a search over full-text papers would present considerable material for metalinguistic information. Such full-text corpus was not available for this dissertation.

Other textual corpora were explored for this work, like the online edition (1992) of the reference Merck Medicine manual,[12] but were not used for evaluating the MOP system. Other resources used for analysis, research and development purposes were:

Corpus textual especialitzat plurilingüe / IULA	http://www.iula.upf.es/corpus/corpus.htm
Corpus Crea / Real Academia de la Lengua Española	http://www.rae.es/NIVEL2/recursos.htm
Brown Corpus	http://corp.hum.ou.dk/corpsearch/search/
WordNet 1.6	http://www.cogsci.princeton.edu/cgi-bin/webwn/

[12] http://www.merckmanualhomeedition.com

II A THEORETICAL FRAMEWORK FOR METADISCOURSE

Summary

From a formal point of view, all symbolic systems like language are founded and enacted through a metalanguage of a higher semiotic order that describes the elements and rules of an object language. The autonymical state of regular lexical items occurs when self-referential words are present as signs for themselves, when they are being mentioned instead of being used normally. This constitutive distinction and its foundational role are enacted through various linguistic and paralinguistic devices in language, like quotation marks and metalinguistic verbs and descriptors. As metalanguage has to be marked somehow, from a cognitive point of view, to obtain special processing conditions that allow it to be interpreted correctly, these previously mentioned devices play that role also. This saliency and regularity is one of the factors allowing for automating their processing. Metalinguistic information, in order to be relevant and informative, cannot be inferable from previous knowledge or from regular language competence. Regular words and technical terms differ, among other things, in that terms require

volitional creation and introduction into generalized use for specific purposes, unlike the everyday lexicon, and in that they are modified quite rapidly by conscious agreement and specification through metalanguage between its users.

Metalinguistic sentences create and materialize specialized knowledge through linguistic means. Conventional definitions that follow an Aristotelian scheme are some of the textual acts that perform those functions in text, but are by no means the only possible ones, or even the most common. The information supplied can be non-complete and have a wide variety of forms. Conventional dictionary definitions represent highly processed compilations of generalist meaning that can't realize fully the spectrum of possible information about a sublanguage's denotation, connotation, usage conditions and meaning that metalanguage can convey.

From the point of view of pragmatics, metalinguistic sentences as performatives not only create meaning, but function as interactional grounds for expert peers in their consensus-based activity of jointly constructing a shared knowledge space and a communal lexicon. Felicitous metalinguistic operations depend on previous context, but create new interpretative contexts for future discourse (II.3.1). These highly complex processes can't be automatic or implicit. They rely on explicitness, although subsequent usage might signal implicit acceptance. There is an undeniable rhetorical and interactive component in negotiating meaning through metalanguage use in the specialized papers and domain-specific texts. Terminology is a common ground for knowledge as much as a battle ground. In fact, scientific advancement can be reflected in the metastable nature of theoretical-conceptual systems and terminological networks that end up changing systemically through slight and disperse modification of their elements and relationships.

Before describing the computational system for automating the identification and processing of metalinguistic activity in text, which constitutes the core contribution of this dissertation, we will discuss in detail the metalinguistic phenomenon from formal, cognitive and communicative perspectives, in order to fully understand its properties and linguistic realizations. We will present (in Section II.1) a generalist formal analysis of metalanguage, and proceed (in Section II.2) to discuss the pragmatic and cognitive aspects of metalinguistic interaction in domain-specific texts, especially insofar as Metalinguistic predication influences the collective construction of knowledge within expert communities (Section II.3).

II.1 Formal languages and metalanguages

II.1.1 Bridging semiotic levels: Autonymy, reflexivity and isomorphism in symbolic systems

The concept of metalanguages originates with Hilbert and Gödel in the traditions of logic and mathematics and is the cornerstone of modern formalization of abstract systems,[13] especially symbolic systems like language. It originally addresses the need to have a formal language of a higher order with which to describe and establish the truth-conditions of another one. In fact, Gödel showed that no formal language is autonomous, or capable of being its own foundation. A metalanguage thus supplies the "conditions of the possibility" (to use Kantian terminology) of the language to which it refers, defining on the one hand the signifying elements of it, and on the other the combinatorial rules necessary for the creation of propositional sense. Its predication builds its "content plane" (as Hjelmslev would call it) though semiotic elements, elements of a code that acquire their semic value by virtue of membership in a system of symbolic interrelationships. Of course, a *metalanguage_1* can always be described through another *metalanguage_2* (of which it becomes an object-language), and so on. This formal infinite regression can be troublesome only if we require that there be an absolute reality or system underlying and providing foundation to everything else, but for our present purposes this is neither logically nor ontologically necessary.

[13] See also Hjelmslev 1943; Tarski 1944; Carnap 1934; Jakobson 1963; Rey-Debove 1978, to name but a few.

In order to function at a very basic semiotic level, natural language has to be split (at least methodologically) into these two distinct systems that share the same rules and elements: a metalanguage, and an object language which in turn can refer to and describe objects in the mind or in the physical world. Metalanguages and object languages thus belong to two different semiotic levels, even though they may share the same surface elements and be isomorphic.[14]

An interesting aspect of the coexistence of two distinctive systems in the same representational space is that a discontinuity in semiotic levels occurs when linguistic items are *mentioned* instead of being *used* normally in an utterance.[15] Other formalisms, including those used in linguistics, should in principle be able to mirror such discontinuity in semiotic levels without introducing ambiguity, and some researchers have urged to break this isomorphism when doing linguistic analysis.[16] As we will discuss in Chapter III, natural language employs when necessary avariety of resources (quotation marks, typography, lexical items) to functionally mark this distinction in discourse. Logic and mathematics use special symbols and conventions that are distinct from their object language elements, though both of them can coexist in formulas.

The *Use/Mention* dichotomy can explain what happens when the word "Socrates" does not refer to a person, but to itself as linguistic sign, e.g. in the classic example from Lyons (1977):

> "Socrates" is an eight-lettered word.

or account for a change in the grammatical properties of a verb such as "moving" when used in a metalinguistic context, where its normal interpretation would have to be blocked:

> The word *moving* means changing position in space through time.

Reflexivity (the property of referring to itself) has been ascribed to language as one of its most important features, and one that sets it apart from other semiotic systems. Coseriu (1986) stated that any element of linguistic code can become a name unto itself and automatically acquire nominal features, in what Rey-Debove (1978) terms a “metalinguistic rewriting rule“. The fact that two identical lexical items with different interpretative conditions are isomorphic requires us to consider its reflexivity, as when linguistic items are mentioned instead of being used normally in an utterance. Rey-Debove, following Carnap (1934), calls this condition *autonymy*. A metalinguistic element embedded inside the object language stands in autonymical condition, as a sign for itself. It is self-referential and reflexive. This implies certain semantic,

---

[14] These formal considerations can also be found at the birth of modern computer science, for example in von Neumann’s (Burks, et al., 1963) specification for a computational architecture, that instructions must have the same nature as the information over which they operate.

[15] This distinction is usually credited in the philosophical literature to Quine (1960).

pragmatic and logical conditions, and has some theoretical consequences when studying metalinguistic sentences.

II.1.2 Metalinguistic predication as a foundational operation

Statements about language, whether in a formal system or in natural language, are vital nodes in the interaction between content and form, between knowledge and expression, between personal intuition and intersubjective scientific understanding. Metalinguistic statements establish the formal and semiotic systems that constitute the basis of communicative codes. Metalanguages are foundational in nature (Lara, 1989), since they have the power not only of *describing* a communicative code, but of directly *enacting* and creating it. A metalanguage supplies the framework where a linguistic code can actually mean anything at all, defining on the one hand the formal elements that belong to it and on the other the combinatory rules allowed for in the construction of meaning and sense in a well-formed sentence. A metalinguistic predication thus establishes the conventionality of meaning (first posited by Saussure) that enacts a linguistic code system, and establishes specific signs as elements capable of conveying significance or sense. Metalanguage sets the elements, structures and rules of the sentences in an object language.

Zelig Harris (1991), in his attempts at mathematical representation of linguistic phenomena, suggested that metalinguistic representation underlies all linguistic utterances, but these formal devices are elided for various reasons from the actual sentential surface. Deep structural specification of meaning is implicit for all sentences, but it can be retrieved in a complete representation of text. In Harris' view, any sentence is a reduction,[17] from another one where those metalinguistic directives would be expressed explicitly. But those structures can also appear at the sentence's surface (Harris, *ibid*.): "Certain features of language structure create a family of metalinguistic devices within language, i.e. of sentences talking about sentences and their parts." Among these devices Harris lists: explicitly metalinguistic sentences (5.2, 9.2) constituting a sublanguage proper (10.2), pronouns and other cross-references (5.3), the use-mention dichotomy, certain performatives (5.5), and "meta-scientific operators" in scientific writing.

An important point to make here is that to distinguish between these two semiotic levels is not just a practical convenience to dispel confusion, but a theoretical prerequisite for its logical workings. Metalanguage must in principle be distinguishable from its object language in order to fulfill not only a communicative goal but the fundamental act of code enactment, as it must

[16] Nirenburg & Levin (1991) specify that in order to avoid ambiguity a metalanguage for the description of natural language should not consist of lexical units of the same language. See also Lyons, 1980.

be implemented from a distinct and higher semiotic level. Otherwise, it would be impossible for one language to provide the logical leverage to be the foundation of another one. From a formal perspective, even with isomorphism, marking these distinct levels is a fundamental property of the metalinguistic aspect of all languages. Metalinguistic discontinues are not just hypothetical parentheses, but are really temporary detours into another (meta)language altogether. As such, making them prominent in discourse and being able to recognize them, as readers or hearers of technical discourse, is a fundamental trait of them.[18]

Another important precision concerning this issue, and one that is not in contradiction with this last point, is that metalinguistic operations are not always explicit. In categorization and sortal operations represented by copulative sentences like "The coelomates are monophyletic", there can always be a metalinguistic interpretation inferable from it, in which we can state that all entities called "coelomates" can also be described as "monophyletic". In conceptual definitions where the metalinguistic dimension is not prominent (e.g. *A triangle is a 3-sided figure*) there is an implication that allows for a statement such as: "the term *triangle* can be applied to any 3-sided figure". When allowing for these interpretations, the decoder provides the framework for metalanguage and inserts the required differentiation between semiotic levels, even if they are not apparent in the original surface sentence. As noted by Darien (1981) for the case of definitions, many kinds of levels are involved in the assignation of meaning in discourse, "interlocking systems" at various levels. It is the whole articulation of a pragmatic structure, a semantic contribution, and typographical, layout and orthographic clues that jointly perform the complex semiotic textual act that allows for code modification. This multiplicity of levels makes it more difficult to define criteria for automatic term or information extraction, though a combination of two or more items functioning as markers increases the chance of successful automatic processing, as we will argue in later chapters.

Everyday language acquisition, as well as learning how to use more formal and technical sublanguages, implies a process of assimilation of the rules and lexical components of a communicative code. Although there is (especially in the first case) an emulation of competent speakers, deliberate metalinguistic directions play a very important role in both of them. Both children and students of a technical domain need instructions about how to use a sublanguage to manipulate knowledge about the world. Specialization is the acquisition of a particular domain

---

[17] "zeroable" is the term used by Harris.

[18] The "jump" in semiotic levels might also be described in terms of what Clark (1996) calls "layers", for example when a conversation introduces a narrative of a joke that must be processed as fictional, but these discourse layers are not as different from normal descriptions as embedded metalinguistic utterances are, in terms of long term consequences for subsequent understanding, nor they need special devices as autonymy to parse them.

of knowledge (of a special, thematically-restricted slice of knowledge about something); it is a gradual and conscious effort developed throughout a considerable amount of time. Knowledge of language (or a sublanguage) is also a special kind of specialized knowledge. Directive utterances concerning language use are vital in those learning processes, be it for acquiring a language as a whole or a sublanguage restricted to a domain.

In the case of learning natural language, lexical items and grammar rules are part of a collective heritage that historically and culturally is relatively stable, and can be retrieved from a repository where source and origins are lost and there is no personal attribution for meaning. But terms and sublanguages are another matter. They are consciously created, evolve rapidly and are linked to specific theories, names and texts, to events and to specific functions. They were introduced somewhere, sometime by someone, and then adopted (or rejected) volitionally by a community of speakers that is a subset of all the speakers of a language. From the viewpoint of our analysis, the validity of a concrete lexical choice (or the objective existence of one or another conceptual referent, for example) matters less than the actual fact of that choice being made and accepted by a big enough number of users of the sublanguage. It could be argued that all specialized language elements could theoretically be traced to original baptismal instances, though a corpus to corroborate this claim would have to be so exhaustive as to be virtually impossible. Nevertheless, at least in the context of acquisition of personal domain competence in expert communities this baptismal hypothesis can be considered plausible.

II.1.3 Definitions and metalanguage

Terminological control is enacted in well-bounded textual fragments that usually serve to state something about the value, meaning and/or usage conditions of the lexical items that are focused in a metalinguistic statement. Definitions are the kind of metalinguistic statements that more easily come to mind, but they are not the only possible ones. Reformulation and paraphrases also help define lexical meaning, as well as restrictions on usage or semantic extensions otherwise allowable by grammar or custom.

In general, definitions and other metalinguistic operations can be viewed (Jakobson, 1963; Bierwisch & Kiefer, 1969; Riegel, 1987; Kleiber, 1990) as deep-structure equations that relate a term with its semantic content, or as answers to a lexical question such as: "¿what is the <meaning | usage conditions | referent | value> of the linguistic sign X?", that carry out the transformation of language by explicit modification of the lexicon. For Sager (1990), a definition adopts “the form of a simple predication about a word or expression and has also been described as an equation of an unknown term and the sum of its constituent meaning elements.”

Bierwisch & Kiefer (1969) present the following logical form for definitions:

"(5) If M(x) then R(x)

(…) (5) amounts to saying that every time an object is referred to by the hitherto unknown NP_1 it has the properties *R*, i.e. it might be referred to by NP_2 as well. In other words, (5) is an instruction to enter the item [X] into the dictionary and to associate with it the reading *R*."

Bierwisch & Kiefer (ibid.) suggest that semantic characterization of a lexical entry can be done through core and peripherical information, the former being the characterizations that establish the place of entry E within a lexical system as a whole, while the latter refers to characterizations that contribute to the meaning of E without supplying the specificity of E with regard to other entries in the system. Core information affects categorization, while peripherical information does not. They also suggest that linguistic knowledge cannot be of an altogether different nature than extra-linguistic (or encyclopedic) knowledge, and that there is no way to set a precise boundary between them that is language-independent.

Reichenbach (1947: 20-21) described definitions as highly abstract equisignificance equations (with system-wide consequences) between two terms. Strictly speaking, it is not a product of logical derivation or inference, but an axiomatic foundation of a system:

The relation of equality by definition can be considered a special case of the relation of equisignificance, i.e. of having the same meaning. It constitutes the case where the equality of meaning is not derived from other statements but is introduced by a volitional decision with reference to the introduction of a new sign. The question of whether equisignificance is demonstrable or a matter of definition will therefore be answered differently according as the system of language is constructed.

The classic view of a definition, and one that traditional lexicography employs even today, is Aristotelian: specification of a *genus*, followed by a statement of *differentia* that allows unique identification of a referent. By virtue of a substitutability principle, the predicates of a definition could substitute for the defined item in sentences, without changing truth-conditions.

Flowerdew (1991) presents the following schematization of classic definitions, with optional elements parenthesized.

(an) X	is (a) Y	which has characteristic(s) Z
definiendum	+ *genus*	+*differentia*

In the linguistic realization of definitions, the *genus* generally corresponds to the syntactic kernel and the *differentia* to the modifiers (Vossen and Copestake, 1993). This definitional

schema has been exhaustively studied and a complex typology has been developed to describe it, but certainly is not the only possible one, or even the most common one outside of reference and lexicographic documents. This emphasis on the identification of reference in definitions can also be found in Ogden & Richards (1929), but, as we will discuss later, unique identification of a conceptual or physical object for a denomination is not the only kind of metalinguistic information that is possible in discourse. Good reviews of terminological definition, including an overview of the extensive typology that has been developed for them and an adequate record of the state of the art, can be found in Pearson (1998) and in Auger (1997).

Structure and form in sentences supplying information about language is much more heterogeneous than the classic *word = genus + differentia* schema, as evidenced in the following fragment from a scientific text:

> By the late 1970's it was clear that the protein-coding sequences in a eukaryotic gene do not necessarily consists of a single continuous stretch of DNA, as they do in a bacterial gene. Instead the coding region is often discontinuous, being interrupted by stretches of noncoding DNA; such noncoding DNA segments are called intervening sequences or introns, and the coding segments of the genes -those that generally direct polypeptide synthesis - are referred to as exons. [19]

or the following example from our sociology corpus:

> So called Realism models a world of aggressively competitive states —sometimes identified with mediaeval Europe.

What is needed is a different textual object that can account for all the varieties of metalinguistic activity actually encountered, and not just for formal definitions that follow a classic pattern along the following lines:

> "An intron, also known as intervening sequence, *is_a* noncoding DNA segment ".

Definitions of the kind employed by dictionary makers, where an hyperonymic *genus* is linked up with specifying *differentiae* in order to conceptually establish a word within the framework of language (or to constraint within a technical domain) are just some of the ways in which metalinguistic operations can materialize in texts. Another usual device is to provide a referent that anchors the meaning of the lexical item through ostension and referring expressions. Nevertheless, on many occasions the kind and scope of the information retrieved does not reflect an inflexible paradigm in which reference, hyponymy, meronymy and conceptual completeness rule. For example, Pearson focuses its corpus work on formal and semi-formal defining expositives (her label for "classic" definitional structures), and although she also reviews other structures that involve synonymy, substitution and paraphrasing she does

[19] In Berg & Singer: *The language of Heredity*. Blackwell Scientific Publications, California. 1992 p. 126 [taken from Temmerman (1997:59)]

not consider all the variety of metalinguistic information possible in discourse. Very often what is really being provided is partial information; for instance, just an additional semantic feature for an item already described, the modification of a pragmatic restriction or maybe the writer is merely suggesting synonym for a given lexical item, or re-evaluating it in a restricted context.

In previous work (Rodríguez, 1999a; 1999b; 2000; 2001a & 2002) we have introduced the theoretical notion of Explicit Metalinguistic Operations (or EMOs) to account for all varieties of metalinguistic modification. Under this framework, classic definitions are just one of the several concrete instantiations of EMOs. We will describe and formalize this notion in later chapters, but for the time being we can describe an Explicit Metalinguistic Operation as a discourse act where specific knowledge about the sublanguage being used is explicitly provided, a knowledge that can be about the items or rules of usage of the technical sublanguage, be it of a semantic or a pragmatic nature. This last sentence is in itself one instance of an EMO, as it has served to introduce this very term into our common lexicon and our shared knowledge space.

EMOs are discourse operations, a certain kind of discursive action, and should not be considered a sentence typology, as they might be expressed in a wide array of linguistic structures and forms, with similar but diverging functions. They might be realized as complete grammatical sentences, or they might be embedded inside textual segments that are non-metalinguistic as a whole. In that case, a logical decomposition into two or more statements is possible. Borrowing a distinction by Leech (1980), a metalinguistic sentence can be embedded as a GUEST inside another sentence that acts as HOST. In the following sentence, from our Sociology corpus, the main topic refers to certain demographic reviews that were performed first in a text identified as "Marsh, 1993", but at the same time a denomination (*The Missing Million*) is provided for the phenomenon, and linked to a referent represented by the WH-pronoun *what*.

> The preliminary demographic checks on what have become known as 'The Missing Million' are described by Marsh, 1993.

We could illustrate this embedding with the following figure:

Host ….	**Guest**	**….. Host**
demographic checks on …	*X(what)* has become known as ⇨ THE MISSING MILLION	are described by…

An example from the *Nature* corpus described in (Pearson, 1998) shows another such embedding through apposition.

Surface uplift	(the term is used to mean that the average elevation of the ground increases)	on a regional scale is difficult to demonstrate.
	Guest	
Host		

A purely formal examination of definitional sentences can advance real knowledge of how they work only so far. Modern linguistics has been shaped by functional analyses pioneered by Wittgenstein's (1953) use of the "language games" analogy, and extended and adapted by theoretical frameworks like Halliday's Systemic-Functional Grammar (1985) and work on pragmatics by Austin, Searle and others. Within this tradition, language has to be studied in the context of its use, as a media of social interaction as well as a communication tool. Socio-cultural functions are also vectors of meaning, and a purely formal analysis of syntax misses a big dimension of how meaning comes about. This is especially true in the case of terms, linguistic artifacts created for the specific goal of transmitting and creating special domain knowledge, where form follows function much more closely than in everyday conversation, and where the mechanics of consensus within an expert group play a vital role in the buildup of conceptual and terminological systems reflected in sublanguages. Modern philosophies of language and science, pragmatics, sociolinguistics and discourse theories cannot be understood without examination of the dynamic interactions of the participants in a linguistic exchange.

> "In addition to sentence structure itself, language users need cognitive structures that permit them to understand the goals of communication and to attach significance to the associated metalinguistic signals. (…) Many aspects of this communicative competence are subsumed under a larger theory of how people manage to carry out any set of cooperative activity." (Jackendoff, 2002)

This is why the purely conceptualist and onomasiological approach to terminology of such authors as Wüster (1959) or Lachenmayer (1971) was at loss to fully describe the phenomenon of technical denomination from linguistic or cognitive standpoints.[20] External constraints like standardization efforts or univocity of meaning for terms might help teach and structure technical knowledge, but in the highly creative and fluid framework where Science advances by constructing and rejecting epistemic paradigms, they constitutes more a prejudice than true theoretical prerequisites. Traditional theoretical frameworks and methodologies used by terminologists tend to present technical knowledge as a fairly static edifice defined by concepts and the relations between them. What we call classic definitional frameworks, following the Aristotelian or lexicographic models tend to reflect a concept-centered view of specialized knowledge, which our more empiricist consideration of metalinguistic exchanges is committed to avoid. Mainstream General Terminology Theory considers reference as the most usual

[20] For a more thorough critique of conceptualist positions in Terminology, see Zawada & Swanepoel (1997), Temmerman (1997), Rodriguez (1999b) and Cabré (1999)

function of specialized texts,[21] denies terminological units the possibility of connotation, and explicitly prohibits for terms the denominative variation that they would naturally posses if they were to be analogous to normal lexical units of natural language. At least in the case of technical sublanguages, a purely analytical view of language, as proposed by some members of the Vienna Circle, is bound to failure; how those languages operate in the real-world environment of a group's interaction thus becomes a central question:

> " (…) a new term or grammatical construction does not become a true part of a sublanguage until its use has been conventionalized by the community of specialists." (Kittredge, 1982)

A metalanguage might not be limited to formal definition of all possible items of a propositional calculus so that their combination under certain rules is allowed. It might also specify the conditions for a successful enunciation, or the presuppositions or intentions derived from the text. We can then widen the notion of metalanguage and suggest that there are such things as a meta-pragmatics and a meta-discourse (Gülich & Kotschi, 1995; Hyland, 1998) that are the foundations of language, understood as a paradigmatic model of communicative interaction. These new perspectives on the nature of language and terminologies have been driven by the use of text corpora, as in Pearson (1998) study of term definitions.

The empiricist turn in Terminology we pursue here by exploiting the wider metalinguistic dimension of discourse, has consequences not only for terminological theory and for the practice of terminography, but for the design and development of computer applications that aid domain-specific lexicographers. One of the constraints on recent lines of research in computational terminology (Pascual & Pery-Woodley, 1997; Pearson, 1998; Klavans & Muresan, 2001) is their focus on definition-like sentences, a theoretical object that, although undoubtedly useful and extensively described, presents by its very nature certain limitations when studying expert-domain peer-to-peer communication.[22] The meaning normalization process inherent in compiling definitions may be desirable when creating human-readable reference sources, but might lead to a loss of valuable information for specific contexts where the term appears. Pragmatic information (valid usage conditions or contextual restriction for the terms), or purely evaluative statements (usefulness or validity of a certain term for its intended purpose), might not be found in *classic* definitional contexts. Metalinguistic information in texts can provide us with information not only about what terms mean, but also how they are actually

---

[21] See, for example, Rey (1995): "Terminologists are interested in signs, i.e. words and units larger than the word, only to the extent that they function as nouns, denoting objects and as indicators of concepts."

[22] In more recent approaches, Meyer (2001) and Condamines & Rebeyrolles (2001) exploit wider lexico-conceptual relations in free-text that can be difficult to model and locate accurately.

used by domain experts, how they are valued, ranked or modified. A wider spectrum of sentential realizations of these kinds of information obtained from corpus analysis has been reported by Meyer (2001), Klavans & Muresan (2001), Condamines & Rebeyrolle (2001) and Rodríguez (2001), but organizing it to provide fully useful terminological resources might still be left for manual review by human lexicographers. We believe that using the more general concept of metalanguage (in tandem with present trends applying stochastic models to language processing) can automate as much as possible the extraction of fine-grained knowledge about terms, as well as better capture the dynamic nature of the evolution of scientific and technical knowledge created through the interaction of expert-domain groups. Instead of relying on artificial and abstract conceptual systems posited by terminologists or domain experts, such a data-driven methodology would open the doors to an empirical study of what and how technical terms mean in the context of their actual use by a community of peers, as well as to theoretically sound technological implementations. The metaphysics implicit in Wüsterian terminology would not interfere with the functionality of term-processing applications, and theoretical prejudices would not get in the way of studying the nature of scientific discourse and knowledge. A multidisciplinary model that is data-driven, as well as term and discourse-centered (instead of mentalist and concept-centered) could go a long way in improving understanding of conceptual and terminological systems, and in providing the foundations for sophisticated and practical computer applications to manage and process them.

In Chapter III we will present a theoretical model of metadiscourse that can sustain a more complex picture than the one presented by the classic description of definitions, but before doing so we have to discuss in more detail the conditions under which some relevant discourse actions are performed, and what functions or roles they might enact.

II.2 The pragmatics of metadiscourse

II.2.1 Metalinguistic operations as textual acts

Metalinguistic statements are special kinds of speech or textual acts, and involve a definite performative dimension. Using some of the theoretical tools employed by modern pragmatics to analyze dialogues and conversations might help clarify their interactional nature (see Section II.3.1), and help obtain a more dynamic view of knowledge constitution in highly complex disciplines.

In the terminology of speech act theory, "executive nomination" (Austin 1958, Lyons 1977) includes definition of terms, and is clearly a performative in which the utterance itself enacts the "baptismal" action. Any follow-up semantic or pragmatic reformulation for the same term in other statements could be considered an "executive re-nomination". In contrast with

constatives, executive acts lack truth value. Its truth is established precisely by its own enunciation.

For Searle (1969), to speak a language is to perform actions according to rules, and those rules can be constitutive (creating or defining new norms of conduct) or regulative (ruling over preexisting ones). Metalinguistic statements would supply both kinds of rules to apply over a concrete linguistic context. For Searle the semantic structure of a language "is the conventional realization of a set of underlying constitutive rules". A definition, then, in this line of thought, would constitute the proposal of a constitutive rule if it generates new linguistic behavior through the introduction of a new lexical item or concept that is to be used in a specific way, or a regulative rule in case new rules for semantic interpretation are being introduced for a preexisting sign. These are rules that can regulate linguistic performance in all its dimensions and that will lead to the semantic and pragmatic coding that systematizes meaning, both with regard to the intersubjectivity of semantic interpretation of lexical units as well as to the possible combinatory at a propositional level and the pragmatic acceptability of statements. Searle also states that "regulative rules characteristically (…) have the form «do x» or «if Y, do Z». Within systems of constitutive rules some will have this form, but some will have the form «X counts as Y» or «X counts as Y in context C»". An example of a regulative rule in a definition of the word "dog" would perform an enunciative act that could be paraphrased as: «If you want to communicate the conventional semantic content to denote the referent for *dog* or the concept 'dog', always use the sign "dog" as part of the linguistic rules following this enunciation. »

Uttering a definition would also entail the supposition that such an act is truly informative, and (at least within the conventions of scientific discourse) that it is a sincere statement without any other ulterior motive beyond providing precisions or clarifications about an issue. Metalinguistic textual acts can also inform about a linguistic state of affairs, presupposing the "truth" of the statement in the sense of "truth" in the framework of a model. Following Wittgenstein's late functionalist orientation that suggests that using a language is applying rules in an interactive game among interlocutors, statements that make those rules explicit are not in principle neither true nor false, but merely operative.

> "The analysis of metalinguistic phenomena requires reference to speech-situations, and to the different "models of reality" entertained by different users of the language. In other words, an account of metalinguistic phenomena requires a conception of meaning which is model-theoretic in the sense that truth is defined as "truth in a model", but in which models are relativized to users of the language". (Leech, 1980)

Inquiring about correspondence to an "objective", extra-linguistic truth is not, strictly speaking, either a conceptual or a linguistic task. Empirical verification of theories belongs to an extra-linguistic realm of scientific activity. The special reference that Sager et al. (1980) attribute to terms is mediated in theory-driven research by a descriptive, observational language connecting theoretical entities with concrete experiences and phenomena, as noted by Hempel (1958). This kind of logic can be taken to an extreme, as in Lachenmayer's (1971) critique of Sociology's status as a "science", because of the ambiguity, imprecision, opacity and contradiction of the terms it uses. It is interesting that Lachenmeyer's Hempelian position provides here a more language-centered framework for understanding scientific thought: "if terms like "institutions", "social organization", [etc.] have to be part of a scientific terminology, they should never occupy a nominal position that masks their inherent predicative function." For Halliday (1993), this tendency to nominalization would not be a result of chance, but of the natural rhetorical aspects of scientific debate, "since you can argue with a clause, but you can't argue with a nominal group".

Scientific theories as such are hypothetical models of reality that we can approach only through the material medium of terms and texts. Although we can certainly postulate "concepts" that change, store and create knowledge (Budin, 1990), we only know those conceptual entities as incarnated in a technical lexicon and a collection of descriptive and explicative statements. Terms and concepts can be understood and comprehended without scientists having any assurance of objective referents for them. If theories are to be meaningful, that is, capable of conveying meaning, they should be considered so on linguistic grounds only, and we should leave up to scientists themselves the issue of their truthfulness, adequacy or extra-linguistic correspondence to reality. I am not concerned here with truth-value, but with significance. Not with extra-linguistic referents for terms, but with how theories can actually refer to anything. The purposeful assignation of meaning has to be done by making explicit both the rules of use and the semantic content of terms within a theoretical framework that is a linguistic context as much as a tool for thought. Such an approach to theories as artifacts should account for how they come into being and change when they are subjected to the dynamics of social exchange among an expert group, a group that communicates by using words and terms and negotiates meaning through consensus and debate in the highly structured context of academic journals. Although such a study exceeds the bounds set for this work, we will examine these questions insofar they pertain to metalinguistic activity in the texts we have examined.

From a discourse perspective, metalinguistic statements can be considered textual acts that perform two distinct pragmatic functions: they can provide **directions** for generating and interpreting linguistic code, or they might **inform** about the state of an assumed or existing

communicative code. That is, they can be of a descriptive or instructional nature. These two functions can operate simultaneously, and might be ultimately equivalent: directing the workings of a code informs about its system, and vice versa. Pearson (1998) presents the following formulation of this duality of function:

"The performative act of defining can be interpreted either as a defining exercitive or as a defining expositive depending on how the definition is expressed and on who is expressing it. The concept of a defining exercitive will be proposed for situations where new concepts are being described, and definitions are being formulated for the first time. The concept of a defining expositive will be introduced for situations where definitions which already exist are being repeated or rephrased for the purpose of clarification or explanation."

Metalinguistic operations in text can convey different ilocutionary forces through hedging, and they may also reflect different configurations and relationships of participants in them. Scientists as writers position themselves with regard a domain's issue, and describe how their personal contribution to the ongoing debate of science can be inserted in the history and the state of the art of the discipline. Hyland (1998) presents this aspect of scientific writing in these terms: "academic arguments have to be won in both subjective and logical arenas, which means a writer's metadiscourse must work to establish credible propositional connections while also conveying an appropriate interactional stance."

A last couple of points to make here are the following: even though metalinguistic operations do not present independent truth conditions, their felicity conditions as successful textual acts can be defined by the subsequent adoption of the proposed modifications of the standing code by a significant portion of its "speakers". A successful metalinguistic operation can be judged by the actual modification of a linguistic behavior, as evidenced in the textual production of a discipline. As instances of the introduction of interpretative directives and regulations, metalinguistic operations will show their effectiveness in subsequent discourse. Felicitous metalinguistic operations depend on previous context, but create new interpretative contexts for future discourse (II.3.1). These highly complex processes can't be automatic or implicit. They rely on explicitness, although subsequent usage might signal implicit acceptance (II.3.2).

II.2.2 Cognitive processes in metadiscourse

II.2.2.1 Relevance and informativity

Metalinguistic utterances contribute to the relevance, coherence, cohesion and informativity of specialized discourse. Their cognitive prominence contributes to the efficiency and overall effictiveness with which a reader can process the message and acknowledge the writer's intention. The fundamental contribution of Paul Grice (1971) to this issue is the proposition that

textual acts can be understood only in the framework of communicative intentions and implicit presuppositions about the shared knowledge of participants in the exchange, of their veracity, sincerity and coherence. Interpreting those intentions is part of the general linguistic competence of interlocutors. Bach and Harnish's Speech Act theory (1979), Sperber and Wilson's (1986) Relevance Theory and de Beaugrande and Dressler's (1981) Textual Linguistics provide us with further elements for understanding the cognitive prominence of metalanguage embedded in specialized discourse.

Metalinguistic saliency contributes to the processing of the information by helping to choose adequate contexts for the interpretation not only of a present textual fragment, but of subsequent or previously encountered texts. Information about term usage mirrors in many ways the conceptual information of a domain because what is being done when articulating language is bringing about (or staging in a sequential and syntactically articulated manner) a representation of events, entities, processes and relations (that is, a universe of discourse or ontology), even when we are not dealing with sophisticated theoretical modeling.

Relevance, in particular, is ensured by the implicit understanding that when we are "crossing over" semiotic levels (which is a resource-consuming process) to state something about our own linguistic code, we are doing so because we think that there is actually some communicative problem or issue that should be dealt with. From this perspective, metalinguistic information is relevant because code change cannot be assumed to be part of previous subject or lexical competency, nor it can be inferable from a previous one available to both parties, but nevertheless is important enough to warrant the additional processing effort involved. It is very important contextual information, quite distinct from regular subject-matter content. In metalinguistic statements the writer is assuming that there is a problem in the attribution of semantic content or pragmatic value regarding the lexical item that stands as the logical or grammatical subject, either because that information differs from the one expected in general language or because a new referent or a distinct new meaning is being introduced in discourse. In either case, the semantic or pragmatic information being contributed cannot be inferred from previous context or retrieved inductively from a previous encoding. This is the main reason why a metalinguistic operation can stand for an answer to a lexical question, providing information relevant to subsequent usage and interpretation; it is also the justification for the role that its textual prominence plays in the cognitive processing of such sentences, a feature that can help us identify them as such.

Computing what constitutes relevant metalinguistic information can be done dynamically by figuring out which terminological items are assumed to be shared by all, and which are new or have changed. We assume the informativity of a sentence that provides instructions or directives

about how the message conveyed by a sublanguage must be understood. This kind of calculus of basic common ground is seldom done for normal, everyday lexical competence (although normal conversation commonly employs various mechanisms for checking the continuity of dialogue), but this lexical alignment is crucial in special domain communication (see Section II.3.2). Altering the meaning conditions of scientific discourse can only be justified if such operations will have some impact on the common cognitive state of the expert-domain community, if it allows better understanding, structuring or communication of scientific knowledge.

Metalanguage creates and informs a common code needed to talk about a specialized subject. As we have suggested before, the *markedness* that makes metalinguistic discourse explicit is a constitutive, core feature of it. Insofar a speech (textual) act that needs to be interpreted, a metalinguistic statement presupposes, in order to be considered relevant and informative, that 1) an actual problem obtains with regard to the linguistic code itself, that 2) there is some information that is new to whom it is addressed, and 3) that this information is not inferable from other sentences already uttered or from information that is readily available. The very explicitness of metalinguistic discourse is grounded on the fact that relevant information about the communicative code is being provided, in effect making opaque and visible language itself, which is usually "transparent" and invisible to the user. Metalinguistic predication allows grounding of shared knowledge states by establishing conventional meanings for discourse.

We could state this along the lines of a Gricean maxim:

- *When a speaker puts forward a modification of the standing linguistic code, or he informs of a linguistic convention, he does so to ensure felicity of previous or subsequent communicative acts, and to establish successful reference within a shared cognitive state.*

Two more principles of metalinguistic predication can be formulated thusly:

- *When describing or modeling a state of affairs, it is possible to use lexical items with interpretative conditions that are not assumed to be known by the linguistic community to whom the utterances is addressed, provided that the producer makes explicit the pragmatic or semantic rules needed for intersubjective interpretation of those utterances.*
- *Metalinguistic statements imply an invitation to generalize and adopt, from that point forward and within certain contexts and a stated scope, a convention in the standing communicative code, and to incorporate it into lexical and cognitive common grounds valid for the language-defined community.*

II.2.2.2 Lexical, linguistic and paralinguistic markers for metalinguistic interpretation

Besides formal requirements mentioned earlier, metalinguistic discourse also has to be flagged somehow to allow for special processing by participants in an exchange. Introduction of metalinguistic information in discourse is highly regular, regardless of the specific topic or domain. This can be credited to the fact that the writer needs to mark these sentences for special processing by the reader. The successful interpretation of metalinguistic statements requires that the writer provide clues as to its proper cognitive processing, since they operate across two distinct semiotic levels. The decoder has to provide the adequate context for the successful interpretation of the utterance.

Markers of metalinguistic activity become vital components in the interactive alignment process that has far-reaching consequences for the constitution of a common code. A metalinguistic statement embedded in the regular statements of an object language is usually signaled prominently using various expressive means: lexical items that act as descriptors (e.g., *term* or *word*) or metalinguistic verbs (*called, termed, dubbed,* etc.), recurrent syntactic structures, pragmatic or paralinguistic resources (hedging, typographical conventions, layout on the physical page, or punctuation). Certain lexical markers commonly signal metalinguistic activity, and trigger the adequate processing of this new information about the common code. This crucial and complex interaction takes place throughout all levels of linguistic expression,[23] with various resources contributing to the global task of enacting or modifying a specialized lexicon, as well as to the more restricted, but equally important, goal of signaling or marking the uncommon nature of metalinguistic information. Some of the markers identified from our preliminary sociology corpus are presented in Table A, and overlap the inventory presented by Pearson (1998):

A) **Lexical:**

- Descriptors: *term, word, phrase, terminology, vocabulary, name, definition,* etc.
- Metalinguistic verbs: *calls (is called), means, termed, name, refer, use (is used), speak, designates, known as, stands for, defined as, coins, corresponds to, applies to, dubbed, designates, labels, indicate, said (to be),* etc.
- Other lexical indicators: *subtitle, oxymoron, where (in formulas),* etc.

B) **Syntactic: Apposition**, copulative clauses, etc.

C) **Pragmatic:** Informational structure, hedging, etc.

D) **Paralinguistic:**

- Layout: footnotes, highlighted text, tables, etc.
- Typography: bold, italic, different type face, etc.

[23] Halliday (1993) points out that "all the features of a text participate in the creation of meaning".

- Punctuation: quotation marks, parenthesis, etc.

Table A. Common items that mark metalinguistic activity in English

Candel (1993) views these metalinguistic phenomena as reformulative strategies, and presents the following table of linguistic devices for French.

Verbs	*appelé, baptisé, dit, parler, est, désigner, signifier*
Generic names (descriptors)	*nom, terme, en un mot, appellation*
Conjuntions	*"X ou Y", "X, Y"*

Table B. Common items that mark metalinguistic activity in French (Candel 1993).

Similar surveys of definitional markers for Spanish in specialized domains have been reported in Sierra & Alarcón (2002) and Alarcón & Sierra (2002), in the context of an ongoing project to automate acquisition of conceptual information from an Engineering corpus. Pearson (1998) and Sager (1980) use the term "connective verb" to describe markers of a verbal nature,[24] but they restrict them to strictly definitional contexts which, as we have discussed, are too narrow a theoretical model to account for all varieties of metalinguistic phenomena. Pearson uses her *Nature* corpus to provide marker statistics for special domain text (Table C).

Verb	Total	Total *per* 100,000 words
is/are	3906	1700.65
consist(s)	72	31.35
define(s)	18	7.9
comprise(s)	13	5.66
is/are defined as	8	3.48
denote(s)	5	2.18
is/are called	3	1.31
is/are known as	2	0.87
designate(s)	0	0

Table C. Connective verb statistics from the *Nature* corpus (Pearson, 1998)

Markers like quotation marks or lexical descriptors force the interpretation within a framework of metalinguistic activity of the items under their influence, coercing its typing as linguistic signs. This crucial operation takes place at all levels of linguistic expression, with various resources contributing to the global task of enacting or modifying a specialized lexicon. Although language has an impressive variety of ways to express the same content, our corpus analyses and our examination of the literature has shown that the repertoire of metalinguistic devices in text is limited and predictable, and does not change very much across genres,

[24] Pearson also uses "connective phrases" for markers not necessarily of a verbal nature.

languages or domains. Even though there might be very idiosyncratic ways to convey them, metalinguistic operations in general are highly protocolized and regular. Ambiguity in this aspect of linguistic communication would have grave and far-reaching consequences for the very possibility of mutual intelligibility, as well as for subsequent mutual understanding of the parties involved in the exchange. As in the case of normal conversation, where various ways to check understanding of the message are common and very prominent, participants in an exchange where complex terminology is needed will be expected to spend time and effort to ensure that their ideas are coming through.

Our corpus analysis also suggests that metalinguistic signaling is usually carried out concurrently by at least two or more different lexical and paralinguistic elements. This is consistent with results reported by Malaisé et al. (2004) in similar work. Their constitutive markedness means that most of the times these sentences will have at least two indicators present, for example a verb and a descriptor, or quotation marks, or even have preceding sentences that announce them in some way. Redundancy of markers diminishes ambiguity or confusion. Some elements are just markers whose sole function is to flag a given item as metalinguistic, while others also imply some processing directive by virtue of their semantic content. This double aspect of markers and operators in EMOs will be returned to in later sections. As we will see then, the formal and cognitive properties of EMOs facilitate the task of locating them accurately in text.

Before attempting to establish a theoretical model of Explicit Metalinguistic Operations that underlies their computational representation and manipulation, we take a brief detour and discuss how these discourse actions are expressed in special domain text, what functions they fulfill, and how they contribute to the creation, assimilation and representation of the technical knowledge of the community of domain experts.

II.3 Knowledge and terminology control in expert groups

II.3.1 Creating, modifying and controlling expert knowledge

The original concern that drove our research had to do with the fact that, regardless of the exact nature of its mental organization and computational properties, human expert knowledge is undoubtedly an intersubjective, collective enterprise. It is possible only in the interaction of a community of individuals. As such, it needs to be communicated persuasively (usually in the context of linguistic interaction) in order to be validated by a group's consensus.

When experts interact amongst themselves, they must assume that their audience of peers knows most of the Communal Lexicon (Clark, 1996, 1998) that holds for their field. The audience also assume that a speaker will use both a common repository of knowledge accepted

by all and an unambiguous terminology to refer to it. They also expect that new knowledge will be offered by the speaker. Furthermore, they will also assume that if new knowledge or new terminologies are introduced in discourse, then evidence, arguments or interpretative conditions will be provided so that these innovations, terms or ideas can be accepted at large by the community of expert whom the speaker is addressing. Of course, expert communities can also form subcommunities defined by areas of specific knowledge in the field, or by acceptance of diverse and competing explanatory theories of the phenomena. These subcommunities will share some (but maybe not all) of the terminology and common ground of the whole field, and must accordingly adjust their assumptions about what is commonly accepted as far as linguistic conventions, conceptual inferences and empirical evidence is concerned.

Domain-dependent words or terms have low occurrence statistics in the normal interaction of the general population, but not so in the specialized contexts where they have been spawned and where they evolve. They help define the communities that use them as much as these communities define those very terms. Language and knowledge-based communities are defined by their practices, beliefs and interaction, by their common ground. They are constituted against the wider frame of society in general, and establish with other groups and individuals certain links and relationships. Putman's hypothesis of the universality of the division of linguistic labor[25] (1975) suggests that some lexical items require interpretative recourse to (and also some form of interaction with) groups of individuals that in some sense control meaning conditions outside the mainstream of normal linguistic competence.

Linguistic competence as member of an expert community requires assimilation of structured knowledge and of complex theoretical frameworks. Meaning in those settings is assigned to new lexical units or modified for existing ones depending on the state of the art of different (and sometimes conflicting) disciplinary accounts about how or what is constitutive of observable reality. Linguistic issues interact with empirical research and conceptual constraints as theoretical knowledge is being built up or discussed through rational consensus within a scientific community.

A fundamental part of the cycle of scientific advancement consists of refining and establishing a communicative code that is both well suited to reflect whatever conceptual advances the expert group has attained, and at the same time can achieve enough acceptance to be able to be shared by very diverse (and sometimes antagonistic) research groups. We have to consider an additional lexical competence on the part of the specialist (besides his

[25] "Every linguistic community (…) possesses at least some terms whose associated 'criteria' are known only to a subset of the speakers who acquired the terms, and whose use by other speakers depends upon a structured cooperation between them and the speakers in the relevant subsets."

understanding of everyday language) that is topical in nature, and that includes not just knowledge about concrete or abstract entities, but also knowledge about how to describe, express and communicate highly structured data. Acquisition of a technical sublanguage is very different from acquisition of everyday language, as it involves not just learning a pre-existing lexicon, but actually participating in a constant creation of new lexical meaning. But seeing sublanguage conventions just as a passive reflection of what is known about the world would be misleading. Scientific advancement is also conditioned by its expressive means. The relevance of this point can be illustrated with the next quote from an article in a scientific journal (Guterl, 1996):

> "Because the language of physics does not contain a vocabulary for granularity, engineers must treat granular material as either a liquid or a solid. These approximations work most of the time, but occasionally they lead to disaster."

Although these facts seem clear enough, the multidisciplinary effort needed to understand the complex construction of a linguistic common ground for science has made it very difficult to engage in meaningful theoretical debate; this is not surprising, since many very different disciplines claim a stake in this vast uncharted territory: philosophy of science, epistemology, sociology of knowledge, cognitive science, sociolinguistics, discourse analysis, etc. Researching research and the specialized knowledge it creates has usually been subjected to speculative or logicist arguments, instead of benefiting from modern corpus-based empirical methods. At most, surveys trying to elicit from scientists the nature of their work can reflect a conventional wisdom about their research and their methodology. When asked about the importance of the language they use, a common answer from practicing scientists is that terminological issues are secondary both for their methods and results. But contemporary philosophy of science and language would not be so sure of that. Following a number of Wittgensteinian clues, we can apply to any scientific activity, *mutatis mutandis*, what Richard Rorty has stated for philosophy:

> Interesting philosophy is rarely an examination of the pros and cons of a thesis. Usually it is, implicitly or explicitly, a contest between an entrenched vocabulary which has become a nuisance and a half-formed new vocabulary which vaguely promises great things. (Rorty, 1989)

Scientific discourse as a joint activity (Clark, 1996) is at the same time cooperative and adversarial. In order to enact a personal discovery into the social and cultural artifact that is expert knowledge, it has to be communicated effectively to others (Malinowski, 1944). Modern scientific thought can't appeal to belief, to instinct or to an incommunicable *gnosis*: any valid claims or hypothesis have to be describable or explainable through linguistic resources, even

though that discourse can be highly abstract and use other formal expressive means, as is the case in mathematics. The expert community has to establish particular relationships among its members, relationships that have a dual collaborative and confrontational nature. Scientists as speakers have to assume a cooperative stance from their counterparts, though it has a confrontational component as well, as adoption of one or other sublanguage conventions can have profound theoretical and practical consequences.[26]

Scientific interaction through language can be viewed as an extended conversation among expert peers,[27] and the ongoing production of scientific papers as turn-taking in these conversations regarding both the subject matter of the discipline as well as the correct or acceptable way to convey that message employing a sublanguage's conventions. Academic papers in isolation might be viewed as monologues, but in a wider perspective they are part of dialogues with established protocols for communication, with their own specific devices for clarification and reformulation. Scientific communication follows certain basic principles, especially in the case of interaction through sequential publications in public forums such as academic journals. Participants play certain roles, and become at times readers and at times producers of the elaborate turns in the technical dialogue that is at the heart of a discipline. These turns are anything but spontaneous, and are carefully crafted within stylistic, argumentative and thematic constraints. In these contexts there is little room for improvisation, and positions are expected to be backed by empirical or rational arguments that must be accessible, in principle, to all that have similar training and previous knowledge of the general state of the field.

A communicative common ground has to be negotiated between peers, although factors other than rationality or communicative efficacy influence such decisions. From the perspective of the logic of Science, Rorty's (1997) next quote about Thomas Kuhn might help illustrate how linguistic aspects of the scientific debate have a real impact on how science is produced, beyond what, as many scientist would argue, might be "simply" terminological conventions:

> Khun fuzzed up the distinction between logic and rhetoric by showing that revolutionary theory-change is not a matter of following out inferences, but of changing the terminology in which truth-candidates were formulated, and thereby changing criteria of relevance.

---

[26] Proposing a terminology and modifying a sublanguage are also ways in which researchers establish the relevance of their own work, and gain recognition for their theories and for themselves as leaders in their field.

[27] The context of written scientific communication in a journal can be viewed, in Clark's terms, as a non-personal setting, addressed to a community of experts instead of to a concrete person.

Domain experts form communities of peers[28] that not only control high level knowledge but also actually produce new knowledge as one of their main activities and goals. The discovery and systematization of new knowledge is a dialogic, interactive and collaborative enterprise. Scientific activity involves constructing intricate conceptual systems that, by virtue of rational consensus, can become communal knowledge spaces, repositories of accepted facts, theories and terminology. Scientific interlocutors share an enormous amount of assumptions and ontological commitments as domain-specific common ground. In fact, a significant aspect of their own constitution as domain experts consists in assimilating publicly available empirical evidence, explanations, models and theories that are viewed by the community with a lesser or greater degree of acceptance, but that all can refer to and thus become the contexts for the dynamic debate that powers scientific advancement (be it evolutionary or revolutionary in nature).

These knowledge-based communities are also language-based communities, insofar as one of their foundational conditions, along with a common domain subject, is a shared sublanguage. Debate and pragmatic cooperation within a discipline creates an epistemic state-of-the-art that can be followed in the copious scientific and academic literature by noting how researchers modify the lexical items and usage rules of the sublanguage that they employ to put forth and communicate new knowledge. For Clark (1998), a Communal Lexicon (CL) is "the vocabulary associated with a community of people who are distinguished by their common knowledge of a particular field of expertise." These linguistic conventions can arise in what Clark (1996:80) calls *explicit agreement*, when an author stipulates explicitly how he is going to use a term, and what meaning conditions he will attach to them. Clark identifies the following as some of the phrasal and lexical markers used in these operations: "what I shall call", "let us call this", "hereafter", "for short", "termed", "named" and "abbreviated", but states that more elaborate formulas are possible.

If we extend the notion of "conversational record" (Thomason, 1992) to include the multi-party subject-specific exchanges of scientific debate, we can imagine that participants keep in mind such a registry of meaning-in-context when they write or read scientific papers, in order to establish the conceptual common ground of the expert community they belong to. Such "score-keeping" would of course extend into conceptual positions and established empirical facts, but it would also crucially include a semantic and pragmatic dimension that ensures close alignment of linguistic expression and intended meaning. It is unlikely that either in scientific settings or in

[28] To call the participants of a scientific debate "peers" can be misleading. As in some conversations, certain speakers' utterances carry more weight than others, and the influence of a *pecking order* in learned discussion is hard to dismiss.

everyday conversation the entire aggregate of all possible data constituting a common ground between two interlocutors is activated and presented when processing language. A more realistic model of common sense and shared technical knowledge would involve selective activation of inferential chains as discourse develops and memory brings into focus, as required, one or another sector of a systematic network of concepts. A situation model (especially one as complex as a scientific theory) presupposes an enormous amount of interlinked facts, events and actors, but does not require such a lattice to actually be present in the mind for adequate interpretation of language. It only needs to be available for lookup when needed, and for some mental and linguistic generative devices to be used in a way that is shared by all.

In contrast to the relatively stable lexicon of everyday language, specialized terms are continually being created, put forward or modified purposefully by an expert community in order to accomplish their communicational and representational goals. Language conventions are enacted through what Lewis (1969) calls *signaling doublets*, in which producers and interpreters of utterances link meaning with linguistic representations. As coordinating devices, they allow modeling of intentions, actions and contents, and can involve the lexicon, grammar, language use and perspectives. Considered from the point of view of speech (or textual) acts, they allow alignment of production and interpretation. At the speaker's side, the intention: "when I use X to refer to Y" will be matched by the hearer with "when I encounter X I should take it to refer to Y". For Lewis, these conventions are valid language-wide (language as a whole is a system of signaling conventions), and this dovetails with Harris' (1991) metalinguistic hypothesis that implicit linguistic specification devices underlie systemically for all normal utterances.[29] In any case, as we have stated in previous sections, the information that a metalinguistic statement provides is not inferentially derivable from either the context or a previously standing code, or from information publicly available at that point in discourse. This fact ensures its relevance and informativity (Sperber & Wilson, 1986; de Beaugrande & Dressler, 1981). The kind and quantity of information that is deemed relevant in each metalinguistic exchange depends on the context in which it appears (Thoiron & Béjoint, 1991), the computed linguistic common ground that holds for the targeted audience, and the methodological protocols that ensure successful reference in scientific communication. Metalinguistic exchanges provide information that might go beyond the normalized default information contained by dictionaries. When interpreting text, regular lexical information is applied by default under normal conditions, but more specific pragmatic or discursive

[29] "[...] all sentences can be thought of as originally carrying metalinguistic adjunctions which state all the structural relations and word meanings necessary for understanding the sentence, these being zeroed if presumed known to the hearer."

information can override it if necessary, or if context demands so (Lascarides & Copestake, 1995). The contextual "effect" (to use Sperber & Wilson's expression) of metalinguistic exchanges for ongoing communication can be evaluated if we take into account the dynamic nature of scientific debate, where new knowledge is constantly being created and new linguistic expressions have to be constantly pressed into circulation.

Articles and papers of a theoretical nature propose representational frameworks to explain and describe their domain's problems. They constitute leading-edge research where entire conceptual systems are created and communicated. Scientific theories are thus conceptual constructs that introduce new entities, processes and states into the shared discourse space. The link between theories as purely intellectual productions and their terminological realization is emphasized in the following quote from Halliday's landmark study of scientific writing (1993):

> "The language of science is, by its nature, a language in which theories are constructed; its special features are exactly those which make theoretical discourse possible. But this clearly means that the language is not passively reflecting some pre-existing conceptual structure; on the contrary, it is actively engaged in bringing such structures into being. They are, in fact, structures of language (...) A scientific theory is a linguistic construal of experience." (p. 12)

Denomination is a fundamental scientific operation (Benveniste, 1969, Sager & Kageura, 1994), very close to the discriminatory core of rational thinking and conceptualization, but its real complexity as a textual and discourse action has not been fully described or accounted for. Moreover, denomination as referential attribution is not the only way in which we can talk about our own words. Metalinguistic propositions don't always provide complete information on lexical meaning, like definitions usually do. We can also state that a term is valueless, or ambiguous, without providing any concrete semantic trait.

Metalinguistic statements in highly specialized texts thus enact the sense or usage specificity (with regards to a posited general lexical competence) which constitutes the core technical nature of terms, while at the same time facilitating the interpretation and cognitive processing of scientific discourse. Through a conventional link between things, processes, states and events, and the lexical items that convey them, technical knowledge that goes beyond personal, individual perception, intuition and experience can collectively be built up by an expert group.[30]

---

[30] See Putnam's previously-quoted "division of linguistic labor" hypothesis (in *The Meaning of 'Meaning'*, 1975:227) for a philosophical discussion of this semantic phenomenon. Note, however, that referential mechanisms of ostensive definitions in Kripke, Quine and Putman's discussions are neither relevant nor useful for our discussion of strictly textual devices.

An important aspect of peer-to-peer discussion oriented towards laying the ground for new knowledge involves introducing or modifying the semantics, the pragmatic conditions or even specific parts of the grammar[31] of a sublanguage in use. The very arduous learning process that allows someone to be admitted as a peer in a research community involves learning the terminology, the pragmatic quirks and the style of academic publications. Whenever scientists and scholars advance the state of the art of a discipline, the language they use has to evolve and change, and this build-up is carried out under metalinguistic control (Jakobson, 1957). That is why when we want to structure and acquire new knowledge we have to go through a resource-costly cognitive process that integrates, within coherent conceptual structures, a considerable amount of new and very complex lexical items and terms.

> "In those cases that conceptualizations deviate from the knowledge we have, expert knowledge on the object obviously has to use different words, constructions or specialized senses to be able to refer to this knowledge." (Meijs and Vossen, 1991)

The *technification* of meaning can be understood as an abstract re-presentation of empirical phenomena (Wignell, 1998), which in the realm of lexical knowledge results in the need for putting forward terms that have very specific meanings and usage conditions constantly negotiated and dependent on the consensus of an expert community or an academic group. Thus, *technicality* is here understood as a controlled and consensual deviance from the point of view of a general linguistic competence (from language understood as a common shared linguistic code), in order to open up cognitive space and expand knowledge. Terms are the means to objectify reality in order to allow for its cognitive manipulation. As in the case of the *Language Hypothesis* of child development in psycholinguistics (Xu, 2002), terminological labels in highly-technical discourse might allow for mental representation and cognitive manipulation of non-sensory data, such as abstract categories and property-kind information. Parallel to the introduction or modification of conceptual structures (and intimately linked to them), sublanguages grow and evolve in a metastable process[32] in which insidious global change slowly but surely transforms the theoretical landscape, and brings about revolutionary transformation of conceptual frameworks and theories.

Figure 1 illustrates some aspects of the dynamics of terminological evolution: one author might introduce term T_1 in a publication to refer to a new theoretical entity; another one might suggest modifying that term's semantics slightly to better adapt it to new observations (T_2); other papers might adopt the new proposal, but change its relationships (T_3) with other terms (t_a

[31] Like the part-of-speech of a lexical item, or its combinatorial properties.
[32] See end of this section for the concept of metastability.

& t_b) in the domain in order to reflect a new theoretical framework, or they may attach some usage restriction, or even offer a completely different interpretation (T_4) that is far removed from the original proposal.

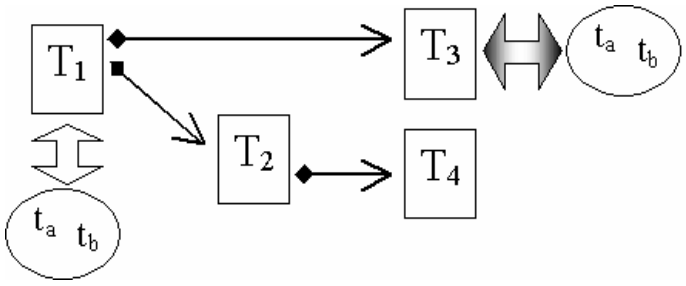


Figure 1. Terminological evolution

The evolution of a proposal to create or modify a terminology across different papers and authors reflects both the dual confrontational-cooperative debate between rival theories, as well as the logical or empirical refinement of scientific explanation. A terminological proposal can be said to be successful if the term is subsequently adopted by at least some of the speakers of the domain community. Previous knowledge is thus transformed into new scientific common ground and ontological commitments are introduced and defended as semantic reference is established.

We have pointed out that non-specialized language is not abundant in metalinguistic exchanges because (except in the context of language acquisition) because we usually rely on a lexical competence that, although subsequently modified and enhanced, reaches the plateau of a generalized lexicon relatively early in our adult life. Unlike the largely stable vocabulary of everyday language, specialized terms evolve rapidly in order to accomplish their communicational and representational goals. Regular words and technical terms differ, among other things, in that terms require volitional creation and introduction into generalized use for specific purposes, unlike the everyday lexicon, and in that they are modified quite rapidly by conscious agreement and specification through metalanguage between its users. Technical terms can be thought of as semantic anomalies, in the sense that they are *ad hoc* constructs strongly bound to a model, a domain or a context, and are not, by definition, part of the far larger linguistic competence in a first native language. Unlike everyday words, terms are the result of an agreement within a specific group, even if they have derived from a word that originally belonged to a more general collective competence, as is the case with, for instance, in Spanish 'gato' (an animal) and 'gato' (a hydraulic or mechanical jack). This argument reinforces the idea introduced earlier that *all* technical terms owe their existence as such to a baptismal speech act, and that given a big enough sample (an impossibly exhaustive corpus of all expert language

exchanges), an initial metalinguistic sentence could be located that constitutes an original, foundational source of meaning. Even if this is difficult to grasp for a whole discipline or field, it is undoubtedly true in the case of the individual competence as is accomplished through training and experience.[33]

Theories and terminological systems constitute networks of concepts linked by logic and conceptual relations, but also by semantics and pragmatics. As with all such highly interdependent and structured formal systems, change in one of the components involves a systemic change, though this can be expressed very subtly. Change in one linguistic item might involve rearranging a whole set of relationships and properties in others. As with chemical reactions, we can posit that, as a whole, these systems can undergo global reorganization with very small local changes. They are what have been termed metastable systems (Gentilhomme, 1994). As stated in previous sections, terms have a somewhat different semantic nature than lexical items from a general lexicon, in the sense that they involve going beyond a basic assumed competence of an idealized speaker of a language, and require an additional learning effort in very specific settings. They can be said to be anomalous with regard to everyday words, since they are *ad hoc* constructs established through a conscious and temporal agreement, and are strongly bound to a specific representational model or theory. As textual acts, EMOs enact the sense or usage specificity with regard to general language that constitutes the actual technical nature of terms, while at the same time facilitate the interpretation and cognitive processing of scientific discourse (See section II.3.1). An interesting difference that can be suggested between words and terms is that a precise textual origin cannot be attributed to everyday lexical items, while terms, at least in theory, always have a concrete (often textual) first appearance in an specialized sense, in which they are introduced and defined more or less explicitly. Being terms, they are not a core part of a common linguistic competence; if they become part of a general lexicon, then they are not terms anymore, or keep just one of its senses as terminological. In that sense, words are akin to natural classes and objects, while terms resemble use-defined artifacts like tools. A conscientious modification of meaning or the establishing of sense peculiarity is required for terminological constitution. This over-specification is the foundational act upon which all terms are constituted. We can conceive terminological systems created by these discourse devices as metastable systems where gradual transformation belies stativity and stability. Such terminological dynamics are “a key determiner of scientific or technological change" (Ahmad 1996).

[33] See systemic functional theory’s (Halliday, 1993) concept of semiogenesis, the creation of meaning in a species-wide phylogenesis, in a personal ontogenesis and in a textual logogenesis.

II.3.2 The mechanics of consensus: the search for a shared code and a common ground

We have already argued that the discovery and systematization of new knowledge is a dialogic, interactive and collaborative enterprise; In theoretical research, knowledge is constantly being created and monitored through metalinguistic statements that allow for negotiation of the conceptual frameworks that materialize in complex terminologies. The background for scientific debate is a standing terminology accepted by a significant portion of the participants in a learned exchange, but it becomes a veritable "common ground" when it can be changed by a community of peers,[34] and in that way is appropriated by them. It is a framework that is, paradoxically, continuously strengthened by its potentiality for change. Terminological alignment is systemic, and it crucially involves modeling the sets of beliefs and ontological commitments of the interlocutors, similar to what Bach and Harnish (1979) would call "mutual contextual beliefs".

Although a terminological proposal is put forward in a textual instance that does not receive immediate feedback or challenge (as in regular face-to-face conversation), a joint process with two or more interlocutors (and thus with a wider scope than just dialogical) is initiated, resulting in either tacit, implicit or explicit acceptance of the new linguistic representation, or in a counterproposal that reinitiates the interactive alignment process. In terms of situation models, EMOs try to align the communicative channel through which a shared world-view can be agreed upon. It's an extended and more complex instance of lexical alignment between interlocutors (Clark and Wilkes-Gibb, 1986; Garrod and Anderson, 1987; Wilkes-Gibb and Clark, 1992, Pickering and Garrod, 2003). Since the "dialogues" inherent in scientific debates are played out in widely separated turns (represented by distinct and consecutive publications) the simple priming mechanisms suggested by Garrod and Anderson are unlikely to be at work here, but nevertheless we can conceive metalinguistic marker-operators as "priming" the reader to process special sentences that deviate from the expected thematic content of the exchange, and require special processing, long-term storage and the full cooperation of all participants.

Metalinguistic lexical markers can provide processing hints just by signaling that the limits of a common ground have been reached, and in that way constitute an invitation to extend or modify that common ground by accepting the next utterance at face value. Lexically marking an utterance as metalinguistic creates special conditions for its interpretation, and signals the need for a repair in the information flow. Metalinguistic verbs and descriptors make clear that there is

[34] To call the participants of a scientific debate "peers" can be misleading. As in conversation, certain speakers' utterances carry more weight than others, and the existence of a *pecking order* in learned discussion is undeniable.

an issue with the shared linguistic conventions that needs to be resolved before discussion of normal domain-centered topics can resume. In the background of these processes, there is the belief or assumption that the speech community shares a more or less established (linguistic) convention (Clark, 1998; Lewis, 1969) that allows them to communicate, and that anyone with the right expertise in the subject matter shares a core vocabulary and rules of usage for the particular sublanguage, even if he doesn't master every possible term in the field. The underlying assumption is that there is a consensus about what is meant when using shared linguistic expressions, a belief that goes beyond the truth of falsity about what is actually being stated. Metalinguistic markers do more than signal a common ground. They also signal something that is not assumed to belong to the common knowledge state of the participants, something that has to be negotiated or reported in order to become part of the lexical common ground needed to ensure mutual understanding. Metalanguage brings a new or modified lexical item to the attention of the expert community in order to specify its meaning conditions. What needs to be addressed in a metalinguistic speech act is the presumed contextual inoperability of what Bach and Harnish (1979) call the *Linguistic Presumption (LP)*: "The mutual belief in linguistic community C$_L$ that:

i. the members of C$_L$ share [language] L, and

ii. that whenever any member S utters any [expression] *e* in L to any other member H, H can identify what S is saying, given that H knows the meaning(s) of *e* in L, and is aware of the appropriate background information."

Metalinguistic repair ensures that expression *e* is clarified and again becomes part of the common communicative code that S and H share, which in an important way defines them as a linguistic community.[35] In the case of technical terminology without wide currency "special beliefs to the effect that H is acquainted with the vocabulary are required by S." Metalanguage, then, not only affects linguistic performance *per se*, but establishes and renews socio-cultural links between individuals.

In metalinguistic exchanges a communicative misalignment needs to be repaired and a coding issue agreed upon, if only for the sake of the arguments that might follow. A clarification, qualification (Gülich & Kotschi, 1995) or reformulative procedure is called for that might resolve communication problems such as ambiguity or reference.[36] The expert

[35] Bach and Harnish state that "as a matter of social fact, the LP in a community is so strong that not to know the language is often a sign of nonmembership in the community."

[36] "(…) expressions to which speakers refer by means of a qualification procedure are usually characterized as a trouble source. This characterization as a trouble source concerns form, content or

community must actively maintain coordination of a usable lexicon that reflects the states, objects, events and causes that their theories, explanations and empirical descriptions refer to as they change with the advancement of their discipline. A cognitive contrast has to be presented between what is assumed to hold for all, and what the speaker will introduce as new information. Marked metalinguistic topics introduce an issue that has to be dealt with on a semiotic level that is distinct from the purely referential level of explanatory and descriptive discourse. It puts over the table the notion that there are alternatives to the assumed meaning conditions of a linguistic item, and that at least one of those alternatives should be considered. Even in the field of linguistics, where language itself is at the center of discussion, this vital distinction between representational conventions and the attributes, states and processes of the objects of study, can and should be made. Explicitness is, again, paramount for correct interpretation of such repair actions.

Fetzer (2003) argues for both explicitness of topic and lexicalized verbs when putting forward a felicitous reformulation strategy. Reformulations are repair strategies concerned with contextual meaning, but metalinguistic operations can in fact be also concerned with "default" meaning beyond the local scope of ongoing discourse, beyond the immediate "conversational record". Modification of lexical meaning can work retroactively or prospectively to change interpretative conditions for the most basic level of terms, in addition to introducing new ones for manipulating new knowledge, with epistemic and ontological consequences for the conceptual systems underlying terminological networks. Also, the specification of scope for a proposed metalinguistic modification or clarification becomes important to establish a wider or more restricted context for a repair procedure.

Negotiating a new terminological usage interactively establishes a new descriptive and theoretical background against which mutual understanding can be achieved. As scientists occupy themselves with their own words, rather than with their usual subject-matter, they are setting up the groundwork for effective communication as much as actually constructing knowledge from the ground up. There is a "mapping" from meaning and conceptualization towards denomination and textuality, and this kind of complex discourse operation can take on many forms and configurations.

If we understand knowledge in the general sense of an enhancement or modification of a previous informational state, linguistic knowledge can be described as the modification of a standing communicative code that allows for some special content or message to be transmitted and materialized adequately. Terminological modification is explicitly negotiated using

---

conditions of use of the respective expressions qualified, and frequently more than one of these aspects at the same time." Gülich & Kotschi (1995: 53)

arguments about communicative efficacy, descriptive power and systematic coherence. In some fields there will be parallel or even conflicting theories that will require alternative conventions or negotiations. From an epistemological point of view, an Explicit Metalinguistic Operation can be viewed as a (proposed) update of the global cognitive state of an expert community, of the common ground for discussion, specifically of the linguistic infrastructure that supports the conceptual configuration of the mental space shared by the whole group. EMOs in theoretical discourse help identify and establish the entities, states, relationships and processes involved in explanatory models; they also provide information about the communicative interaction of the expert speaker group, through the specification of proper usage, values and connotations. In the thin but epistemological vital slice of metalinguistic clarification, expert-domain language users perform work that is of a peculiar linguistic nature:

> "A scientific theory is a specialised, semi-designed sub-system of a natural language; constructing such a theory is an exercise in lexicogrammar. Science and technology are (like other human endeavours) at one and the same time both material and semiotic practices; knowledge advances through the combination of new techniques with new meanings. Thus reconstructing experience is not merely rewording (regrammaticising) it is also resemanticising." Halliday (1993:228)

The 5,400 metalinguistic sentences from our BNC-based corpus[37] (see section I.4.1.2 for its description) reflect an important aspect of scientific sublanguage, and of the scientific enterprise in general. The parallelism between a conceptual knowledge structure and a terminology conveying and modeling it, allowing its cognitive and social manipulation, suggests that the proposed Metalinguistic Information Processor system described in Chapter IV can be used, beyond its more restricted motivation as a compilator of computational resources, as a tool to study the evolution and generation of domain-specific knowledge. It could be employed to trace the introduction or modification of terms and the concepts they embody into the framework of the lexico-conceptual systems that constitute the materialization of theories. Monitoring the "exercises in lexicogrammar" described by Halliday could improve our understanding of the scientific enterprise as a whole, employing empirical data analysis as opposed to the mentalist and speculative methods commonly resorted to by other disciplines. We will return to this matter in Chapter V when we discuss applications of our information extraction system.

[37] Henceforth, the EMO corpus

II.3.3 Science and persuasion

We have argued earlier that expert knowledge has to be communicated effectively to others and that the community has to establish particular relationships among its members that have a dual collaborative and confrontational nature. A communicative common ground has to be negotiated between peers, though factors other than rationality or communicative efficacy influence such decisions. Proposing a terminology and modifying a sublanguage are also ways in which researchers establish the relevance of their own work, and gain recognition for their theories and for themselves as leaders in their field. Terminology is thus as much a battle ground as a common ground.

Beyond the stated rational goal of unveiling "truths" about the world, science evolves under the influence of rhetoric, peer pressure, misconceptions and power struggles in highly structured contexts and institutions. Some scientists are reluctant to accept the reality of this intrusion of irrationality into their activities, and are guided by misconceptions about what they are actually doing. Some of them would prefer to adhere to a mentalist perspective in which personal ideas and insights are elaborated into concepts and then dressed with the materiality of words, and they would flatly reject that knowledge can be highly determined through discourse conditioning in a social context, affected by all factors involved in such exchanges. They may view the debates and arguments that accompany the advancement of science as an unavoidable chore on the crystal-clear path to a rational explanation of the facts of their fields, instead of viewing them as a fundamental part of doing science. Gross (1991), Hyland (1998) and others have shown that when scientists argue their discourse is persuasive and rhetorical, for they have to convince an audience of their right to assert something, while adopting the position that this right should be granted to them not by virtue of who they are as groups or individuals, but, as if were, by letting things to speak for themselves through science's transparent language. To achieve a consensus about what they believe to be true, scientist need to do more than come up individually with ideas or explanatory models.

Consensus is the negotiation of shared cognitive states that underlie the communicative acts of an expert domain community. It is the negotiation of an ontological commitment, a group's decision which also affects terminological knowledge and assumes the referentiality of discourse. As such, consensus is more a process than a state, dynamically changing in response to new data, to better descriptive or explicative models, even to the forceful personalities that from time to time shape a whole discipline by virtue of their powerful argumentative resources. Peer discussion about adequacy, meaning and usage conditions of a term creates a new shared cognitive state about a sublanguage that ensures understanding of the arguments developed in a text, and it employs, even in a self-declared "neutral" language like that of science, complex

persuasive strategies and sophisticated rhetorical devices.[38] The rhetorical aspect of scientific debate, and the role that language plays in competitive interaction, have been studied extensively in recent years, in research lines from critical hermeneutics and social semiotics that examine scientific production as social interaction (Fowler, Hodge et al., 1979; Kotschi, 1986; Gross, 1991; Hermas, 1991; Myers, 1992; Hunston, 1996; Liddicoat, 1997; González & Soláns 1997; Hyland, 1998, among others). Although we will not delve further into this matter in this dissertation we believe that our proposed system can become a very useful tool for studying scientific interaction through these theoretical frameworks.

The complex picture we present here for linguistic scientific interaction involves simultaneous analyses in various dimensions of language, knowledge and group behavior. It is a multidisciplinary approach that relies on corpus potentiality as an empirical instrument that can go beyond more traditional argumentative and speculative traditions of study. Given the quantity and quality of digital text now available for scholarly research, use of computational procedures is vital for these tasks. But this kind of examination of scientific discourse calls for more than just a sophisticated corpus search application. It involves a veritable expert system that can distinguish among intentions of exchange participants, recognize specific and complex verb usages, model communicative settings or consensual processes, and interpret many kinds of formal and structural cues, much in the way our ingrained competence as readers allows us to "read between the lines". Although we do not claim to have attained the level of sophistication that is possible, the chapters that follow present the core contributions of this dissertation along those lines: an adequate theoretical model of metalinguistic predication in specialized text, and the description of a proof-of-concept application (the MOP system) to automatically extract from natural language texts the metalinguistic information contained in them. The adequate treatment of all the issues mentioned above begs for a computational system of lexical acquisition capable of storing and retrieving multidimensional information that is not always semantic, or "lexicographic" in the classic sense of the word. It also calls for a comprehension of scientific theories as semiotic and linguistic artifacts which not only order and systematize knowledge, but actually construct meaning dynamically, opening up its abstract nature with the empirical approach that Corpus Linguistics allows. Only by providing a complete theoretical foundation for such an expert system, Terminology will really overcome its restrictive idealist and normativist origin and become finally the scientific study and engineering of specialized

[38] This "intrusion" of the mechanics of social interaction in the realm of meaning is illustrated by Fowler & Kress' opinion (1979: "Rules and Regulations") that part of the process through which language is used to exercise power implies a "semantic control" of a social group. They introduce the notion of "relexicalization", a process of re-codification of experience by introduction of new lexical items, which is in effect what they themselves are doing in introducing this new term.

knowledge as is materialized in terms, a much more ambitious goal for which acceptance of the interdisciplinary nature of the field is required.

III EMOS: A DESCRIPTIVE MODEL OF METALINGUISTIC PREDICATION

Summary

To construe a more adequate and formalized descriptive object for metalanguage in specialized contexts, we elaborate the notion of Explicit Metalinguistic Operation (EMO). EMOs perform two basic functions in the code specification inherent in all metalinguistic acts, informing on the standing conventions for the system, or directing the interpretation or elaboration of linguistic messages done through that code. EMOs represent textual acts consisting of 3 basic elements that materialize in them: (A) a single or multiple lexical unit that stands in autonymical condition as the subject of the metalinguistic predication; (B) an informational segment in the sentence that provides information or instructions regarding the self-referential unit in A; and (C) lexical, punctuational or paralinguistical elements acting as a predicative articulation of A and B and functioning as flagging devices on the metalinguistic nature of the utterance.

The inventory of elements acting as markers/operators is in principle finite and our corpus analysis has brought forward a number of them (with statistics of use) that have been used to compile our EMO Corpus of metalinguistic speech acts. Realizations and general kinds of possible informational segments are varied, but some common categories have been listed, among them: unexpressed existential variables, full clauses with intensional or denotational references, etc. Other informations that EMOs can carry, but that the MOP system will not attempt to extract, are: agentivity, hedging, illocutionary force, participants in the exchange, etc.

In previous chapters we have described the diverse document sets used as corpora to study metalinguistic activity in domain-specific text. In using text from diverse disciplines we have attempted to cast as wide a net as possible in our analysis of EMOs in order to better generalize about their properties, functions and linguistic realizations. We have also stated that the retrieval of any kind of useful knowledge from text has to proceed from a framework of linguistic analysis and interpretation, an analysis we will carry out in this chapter. All examples presented in the following sections (unless noted otherwise) were extracted from our corpora.

In previous chapters we have presented a critique of wusterian General Terminology Theory and have shown its failure to present a more accurate and dynamic model of how specialized knowledge is enacted through theories and terminological systems. In a very real sense, Terminology was too much focused on concepts and terms abstracted from specific realizations, and was not aware of the importance of the knowledge-rich contexts in which they appeared, or of the communicative conditions under which they necessarily function. We have stated repeatedly the need for a better empirical model for metalinguistic predication that goes beyond the exhaustively studied example of conventional definitions, both as a foundation for better research as well as for grounding improved computational lexicography applications. We have hinted at how the formal, pragmatic and cognitive properties of metalinguistic exchanges can help understand how they perform the vital foundational and interactional roles they play within knowledge- and language-based communities. In this chapter, we will present a descriptive model of Explicit Metalinguistic Operations that provide a theoretical framework for our MOP application, to be discussed in Chapter IV.

III.1 Explicit Metalinguistic Operations: a discourse model for studying specialized linguistic exchanges

As argued previously, the definitional schemata used to study the phenomenon of metalinguistic predication are too narrow a framework to capture the multidimensional workings of such textual instances, or to explain fully their vital role and long-term effect over the constitution of scientific and technical knowledge. We will propose a theoretical object and a model of how EMOs function that attempts to describe these discourse operations, while at the same time allowing for an elementary formalization for its computational treatment. Explicit

Metalinguistic Operations (EMOs) have been put forward in previous work (Rodríguez, 1999a&b, 2000, 2001a&b, 2002, 2003 & 2004) as theoretical entities modeling those discourse acts where specific knowledge about the sublanguage being used is explicitly provided, a knowledge that can be about the items or rules of usage of a sublanguage.

We have discussed metalinguistic predication as a deep-structure equation that helps explain its foundational nature. These formal structures do not always strictly imply equality between the two elements, and more complex relationships like equivalence, equisignificance, referentiality or evaluation are involved. EMOs can be logically analyzed as constituting complex formulas where the predicates receive their functional value from the semantics of the verb (among other factors); thus, a metalinguistic verb such as "*X* means Y" will receive a different interpretation from another one like "Y is known as *X*". We suggest that metalinguistic operations are discourse actions involving three distinct constitutive elements (listed below), articulated into a predication where those elements play specific roles that make the sentence or phrase meaningful and effective at the same time: autonyms, markers/operators and informational segments.

In metalanguage an element of a clear semiotic nature (word, sign, term, or however we might refer to it) is specified with regard to a system or code by virtue of a link with some entity or information acting as a predicate for it. This articulation into meaning is done by one kind of element, those items that have a dual role in the predication: on the one hand, ***flagging,*** or ***marking,*** the metalinguistic nature of a sentence as a whole, and the autonymical condition of one of the two items being connected, and on the other hand, as ***operators*** that perform or enact a concrete modification of the code, supplying interpretative cues as to the specific operation being performed over the code. These two different roles can materialize in text through the same surface elements. In short, all metalinguistic operations can be characterized by presenting, either explicitly or implicitly, the following roles and elements: (A) a self-referential item, (B) relevant information about it, and (C) a way to convey that information and attach it to that autonym.

In order to identify EMO elements in natural language text, we specify some minimal requirements that should always appear in the structure of the sentences performing metalinguistic functions. The features in listing I describe the general requirements that a sentence or a phrase should have in order to be considered an EMO.

Listing I. EMO requirements:

- The presence of a linguistic sign that is the subject (either in logic, discourse or grammar) of a predication.

- The utterance should count as a contribution of relevant information about the status, coding or interpretation of the linguistic sign.
- The whole operation should be prominent or conspicuous on account of its metalinguistic nature.

An EMO, then, should always have at least three basic constitutive elements corresponding to these requirements, which we present in listing II. Following each one, we show how each of those elements can be projected over an example sentence:

Listing II. Constitutive elements of EMO

A) a (complex or simple) term or linguistic sign functioning as **a self-referential term** or **autonym**, which stands as the subject of the metalinguistic operation;

KENES

B) a semantic or pragmatic **information** to be linked up with the term or lexical unit;

The bit sequences representing quanta of knowledge
KENES **intentionally similar to "genes"**.

C) a metalinguistic, or "definitional" verb or verbal phrase, a typographical, punctuation or dispositional mark, or a combination of other semiotic resources that connect (and conceptually articulate) element A and element B, while flagging the extraordinary (non-referential) nature of the segment (**markers/operators**).[39]

The bit sequences representing quanta of knowledge **will be called**
" KENES **", a neologism** intentionally similar to "genes".

In order to better illustrate how each element is expressed in two example sentences, we have adopted the following parenthetical markup:

1) / <The bit sequences representing quanta of knowledge> [will be called "]KENES[", a neologism] <{intentionally} similar to 'genes'>. /

2) The new world order could be / [']SYNCRETIC[',] [a term] <from the study of religion> {(see Colpe, 1987).} /

where: < angular brackets > signal the semantic or pragmatic content (the informational segment) of the EMO, that is, the actual information supplied about the autonym, which is

[39] We group these functions together because they often materialize with the same linguistic and orthographic items, though they can be kept separate from a logical point of view.

highlighted by means of UPPERCASE letters. [Square brackets] show the markers or operators that articulate and make salient such operations, and the {curly brackets } identify information we could term "peripheric", or encyclopedic, which do not belong to the core of the semantics of or the lexical knowledge about the autonym (see I.1.3). The slashes / ... / spotlight the boundaries of the operation, spanning the whole sentence in (1), and just a segment of it in (2). Thus, we could paraphrase (1) into a semic equation that "unrolls" an underlying metalinguistic control process:

The neologism KENES ⇨ <The bit sequences representing quanta of knowledge>

Regardless of the possible syntactic representation of an EMO, its deep structure can be considered an articulation of the three theoretical objects described above. We have also suggested conflating the diversity of EMOs to two very basic discourse actions: to **inform** on a (linguistic) state of affairs, e.g., establish and present a standing communicative code, or **direct** either the coding or the decoding of a message. Both actions can be seen as being equivalent at a higher level of abstraction, very much like an informative statement that may also imply that a convention stands but could be interpreted as a hidden directive. We have observed that most of the time EMOs are multifunctional, that is, they, for example, inform on a state of language and at the same time supply a referent through denotation.

III.2 Elements and basic features of EMOs

For the sake of clarity and in order to go in more detail into the components of EMOs, I present some examples in Table D. The first full length row contains the complete sentence where the EMO is present. In the second row, the first column shows the lexical item(s) figuring in it as *autonyms*. The second column contains the lexical, pragmatic or paralinguistic elements that help flag and articulate these discourse operations, while the third column presents the actual informational segments where something is stated about the lexical item.

• This means that they ingest oxygen from the air via fine hollow tubes, known as tracheae.		
Term	**Markers/Operators**	**Informational segments**
Tracheae	known as | *Apposition*	fine hollow tubes
• Computational Linguistics could be defined as the study of computer systems for understanding, generating and processing natural language [Grishman, 1986].		
Term	**Markers/Operators**	**Informational segments**
Computational Linguistics	defined as | *Caps*	the study of computer systems for understanding, generating and processing natural language [Grishman, 1986].
• In 1965 the term soliton was coined to describe waves with this remarkable behaviour.		
Term	**Markers/Operators**	**Informational segments**
Soliton	coined | the term	to describe waves with this remarkable behaviour
• Integral power results in a fundamental type of social classification which, adapting Bernstein's terminology, I shall call "frame" (Bernstein 1971).		

Term	Markers/Operators	Informational segments
frame	call | *Quot. marks*	a fundamental type of social classification | adapting Bernstein's terminology

Table D. EMO elements

Other examples taken from Catalan and Spanish corpora are presented next. Identified marker/operators are shown with boldface:

Corpus Crea (RAE):

> Los Ornitisquios fueron todos ellos herbívoros y sus tallas son menores que las de los anteriores. Presentan en la mandíbula inferior un hueso sin dientes, **llamado** predentario.

> En este entramado se puede llegar a distinguir sucesivas expresiones de lo heredado, aunque **el término** fenotipo se suele reservar para la expresión final.

Corpus textual especialitzat plurilingüe IULA/UPF:

> La fàcies, amb l'embotornament causat per edema palpebral, presenta un aspecte clínic característic que **s'ha anomenat "**fàcies xarampionoide**"**.

> **El terme** Air-Mass **expressa** la longitud del camí efectuat a través de l'atmosfera pel raig solar, expressat com a múltiple del camí entre un punt al nivell del mar i amb el sol en el zenit.

Our initial survey of these operations in the sociology corpus presented 240 sentences that complied with our requirements for metalinguistic activity. On the BNC-based EMO corpus, a total of 5,430 sentences were identified as EMOs using these requirements.

III.2.1 Lexical elements in autonymical condition

EMOs present as their core element lexical items that stand in autonymical condition, as linguistic signs, regardless of their usual grammatical properties, and they present a nominal syntactical category. They constitute the subjects about which something is predicated with regard to its value, meaning or interpretation. As to the full realization of autonymical items, sometimes determiners and modifiers need to be considered as part of the full term being focused in the predication.

We have also encountered cases where there is a definitional sentence that does not carry an item explicitly marked as metalinguistic. In those cases, it can be argued that such sentence can receive an auxiliary, potestative metalinguistic interpretation; e.g., in a sentence such as:

> A triangle is a three-sided figure

we could infer that any three-sided figure can be referred to as a "triangle", as it conforms to the conceptual definition for that term. Many cases of copulative sentences can also be given a metalinguistic interpretation in technical corpora, and constitute cases that could be called "Implicit Metalinguistic Operations". We have chosen not to include those cases in our compilation of EMOs from the exploratory corpus, and restrict ourselves to cases where the

autonymical nature of the items is explicitly presented, either through semantic, textual or pragmatic means. Processing those cases would greatly increase our proof-of-concept system's complexity, and require inference mechanisms that are not within the scope of the present work. Disambiguating valid metalinguistic sentences in those cases is, by itself, complex enough a problem to merit a dissertation. A definitional sentence, such as the following example from the MedLine corpus, would not be considered an EMO under our current framework, but could certainly be exploited as an extension of the EMO schema to other copulative sentences that are given a metalinguistic interpretation:

> Amyotrophic lateral sclerosis (ALS) is a neuro-degenerative disorder with both sporadic and familial forms.

III.2.2 Metalinguistic markers/operators

Metalinguistic markers/operators ensure that autonyms are successfully interpreted as such, and they provide further clues as to how the metalinguistic predication must be performed. We have argued that this markedness is a constitutive feature of EMOs, and it allows for its foundational nature. The most common markers/operators in our exploratory corpora were quotation marks[40] and the verbal elements "called" and "known as". Some of these markers were bounded to very specific kinds of EMO: for example, the combination "*where :*" usually signals meaning attribution within a formalizing expression, as when defining the variables in a formula, and must be considered to have a very "local" scope that usually does not affect a global linguistic code (see next section for description of *scope*). Appositional structures are also very common markers in technical texts that usually signal relations of equisignificance, especially when providing synonyms and acronyms. In general, our inventory is in line with previous ones by Pearson (1998) and others.

It should be mentioned here that when normalization and tokenization is performed, textual corpora generally lose some typographical and dispositional characteristics that could function as autonymical markup. Italics and definitional tables that might perform the function of flagging the metalinguistic nature of sentences are not considered in general relevant for a purely linguistic interpretation, and are not preserved in corpora that aim at more general phenomena than metalinguistic predication (this is the case, for example, with the British National Corpus). Our suggestion that in general the autonymical nature of terms is done redundantly by two or more markers/operators should take these restrictions of available data

[40] Quotation marks are also used as "scare quotes" to indicate tentativity in the use of a certain item, or to indicate reported speech within a sentence, especially in spoken text. Autonymical use must be distinguished from those uses, which is not always an easy task for computer applications.

into account. Our final inventory of markers for the MOP application does not contain dispositional or typographical markers, but does contain punctuation and lexical ones.

In some cases it is very difficult to disambiguate between valid metalinguistic uses of a verb lemma such as "call" and other uses that are not metalinguistic, as we shall discuss when we describe our Information Extraction application that uses the principles and models from these sections. In general, using certain tenses and auxiliary verbs in the co-presence of other indicators like quotation marks will help perform a better job at correctly identifying metalinguistic instances from our corpora.

In the following listing (III) we present an open-ended catalogue of elements that appear in sentences with EMOs. The nature of each element contributes in its own special way to the overall predicative and cognitive processes. Some of the information that we can obtain for terminology or lexicography is encoded in the choice of elements that a writer actually uses (or avoids using). The semantics of each of the verbs, for example, establishes a unique link between the term and the semantic-pragmatic information; "dubbed", for instance, supplies different connotations than "called" when used in a pattern «X *verb* Y», while the use of one descriptor or another, *name* vs. *term,* for example, is actually providing additional information about the proper classification of the linguistic item being considered. As pointed out above, orthographic and dispositional elements introduce important aspects of metalinguistic activity that can thus be easily detected. Our findings suggest that metalinguistic activity is a discourse process in which many linguistic dimensions interact, and in which non-linguistic, semiotic, resources can have an effect on what is being meant, on what is being stated. At the same time, we have argued that metalinguistic predication is a very regular process, and that the actual set of recurring items that operate those processes is limited and can in principle be recorded exhaustively, even if their sentential realizations are very diverse.

Listing III. Common recurring elements in EMOs

A) Lexical**:**

- Descriptors: *term, word, phrase, terminology, vocabulary, name, definition,* etc.
- Metalinguistic verbs: *calls (is called), means, termed, name, refer, use (is used), speak, designates, known as, stands for, defined as, coins, corresponds to, applies to, dubbed, designates, labels, indicate, said (to be),* etc.
- Other lexical indicators: *subtitle, oxymoron, where* (in formulas), etc.

B) Syntactic: Apposition, copulative clauses, etc.

C) Pragmatic: Informational structure, hedging, etc.

D) Paralinguistic:

- Layout: footnotes, highlighted text, tables, etc.

- Typography: bold, italic, different type face, etc.
- Punctuation: quotation marks, parenthesis, etc.

The following table shows some of the most common verbal elements functioning as markers/operators in our Sociology corpus:

Verb or VP	Ocurrences
call	36
Is, use	19
refer	17
known as	8
use, refer	5
Imply, mean	4
Apply, define, restrict	3
Designe, use, mean	2
amounts to, coins, conceptualize, consider, correspond, denote, dub, evoke, extend, favor, has, has become, includes, indicates, insist, name, note, reserve, speak, stands for, substitute, strech the meaning, taken to embrace, termed, terms	1

Table E. Verbal markers/operators statistics for Sociology corpus

There are two important observations to make here. First of all, none of these elements by itself is enough for a positive identification of an EMO. It is the whole articulation of items from different dimensions of linguistic and textual structure that bring about a complex, foundational discourse action such as those we have been analyzing. These elements and patterns are not always exclusive to metalinguistic statements, but most often they do signal that such processes are taking place in a text. Unlike the fields of lexical databases or the feature structures we sometimes employ to represent lexical entries, actual language use does not usually segregate semantic, pragmatic or syntactic data for our convenience. Moreover, some of those items are polysemic or can perform different functions in different contexts; for example so-called "scare quotes" can merely indicate tentativity, and some otherwise reliable lexical indexes of metalinguistic activity could be functioning in a completely different way, as in the following example from the *Brown* Corpus, where they have been bold-faced:

> In any case it is by no **means** clear that formally structured organs of participation are what is **called** for at all.

Therefore, the presence of any element as indicator of autonymy should be reinforced by other elements (either formal or semantic) in the EMO. The following table presents some of the most common markers/operators (including descriptors, modifiers and punctuation, and their combinations) found in the exploratory Sociology corpus:

marker_1	marker_2	Total
quot. marks	term	17
term		14
term	quot. marks	10
quot. marks		6
quot. marks	calls	6
known as		5
quot. marks	called	5
quot. marks	call	4
called		4
word	quot. marks	3
referred to as		3
quot. marks	phrase	3
The term		2
quot. marks	[footnote]	2

quot. marks	so-called	2
words		1
word	context	1
the term	refer	1
the term	taken to imply	1
term	aplied to	1
term	mean	1
term	refer	1
term	referent	1
term	refers to	1
term	refer to as	1
term	synonymous	1
term	taken to refer to	1
term	used	1
term	word	1
so-called	quot. marks	1
[Acronym]		1
refers to		1
refers to	[footnote]	1
refer to as		1
refer to	term	1
oxymoron	labels	1
namely	quot. marks	1
[footnote]	definition	1
known as	[abbreviation]	1
[Formula]	Where X is	1
[Lexicographic entry structure]	[Phonetic info]	1
Designates		1
definition		1
defined as	advocates	1
define	[abbreviation]	1
quot. marks	adverb	1
quot. marks	aposición	1
quot. marks	been called	1
quot. marks	Calling	1

quot. marks	coins	1
quot. marks	defined	1
quot. marks	designated	1
quot. marks	[Lexicographic entry structure]	1
quot. marks	is called	1
quot. marks	is known	1
quot. marks	is meant	1
quot. marks	mean	1
quot. marks	meaning	1
quot. marks	meant by	1
quot. marks	name	1
quot. marks	neologism	1
quot. marks	refer to as	1
quot. marks	referred to as	1
quot. marks	refers	1
quot. marks	refers to	1
quot. marks	refers to as	1
quot. marks	terminology	1
quot. marks	the term	1
quot. marks	use	1
quot. marks	word	1
calls		1
calls	quot. marks	1
called	quot. marks	1
called	derisively	1
called	[footnote]	1
call		1
call	refers to	1
call	term	1
[abbreviation]		1
[abbreviation]	where	1

Table F: Combination of markers/operators in the Sociology corpus

III.2.3 Informational segments and predicates

We have argued that the range of metalinguistic information that can be provided in discourse is much wider than what traditional views of the definitional phenomenon have considered. In fact, referring expressions and pronouns representing referents for the autonymic signs are common instances of informational segments identified in our exploratory corpora, but they are by no means the only kind of linguistic realization for these EMO components. Besides providing denotation through a referent or through conceptual data to identify it (that is, referent identification through extension or through intension), EMOs might also present a judgment on the usefulness or validity for a term or its use within a particular context, or a wider characterization of word usage, as in the next example.

Durkheim's usage of the terms "psychic" and "affective" seems euphemistic .

Assignation of synonymy or equivalence between two lexical items is also a common operation that does not involve presenting “additional” information, except in the sense that this stated equivalence was something not known beforehand.

What we call the language of isolation (I-you) corresponds exactly to his I-it .

Sometimes the information provided in metalinguistic contexts is restricted to the fact that some meaning conditions need to be reexamined, without presenting locally a new meaning configuration, as in the next two examples:

At this point we should give some further consideration to what exactly the widely used term identity is supposed to mean.

One suspects that non-sociologists, on learning this, might well be inclined to wonder what, in that case, nearly all sociologists do see as being the referent for this term;

Furthermore, the information provided might amount to a restriction on an item's usage, or a specification about what that item does not mean, rather than specific usage directions or what it actually means in an specific context. We could term "differential" information the features that merely serve to reflect a particular change in the information for a previously-known lexical item:

By primacy we do not mean exclusivity

Intra- and inter-sentential relationships also play a role in identification of the informational segment that contains what is being linked to the autonym. Anaphoric pronouns and other similar lexical items (like WH-pronouns in relative clauses, or determiners) might point to other components of the sentence, or refer to other preceding or subsequent sentences that express reference or specifications for the term. The following example presents an instance where the information about how exactly the term "flux pattern" will be used (more precisely what will be its referent) will have to be retrieved from a previous utterance, connected to the EMO through the use of lexical item "this" (boldface is mine).

This shall be referred to as the flux pattern.

Another special case of informational segments is when there is no explicit or surface textual fragment that can be linked to an autonym, unless we look at a deeper, logical level. Some markers like "so called" flag a denomination in which the referent is elided completely, and has to be inferred from context:

But the so called 'opinion' polls have become more frequent.

What is being stated here is merely the existence of a discourse referent for the autonym, that is, we know that there **is** some entity that is given the name *'opinion ' poll*. In these cases of indirect reference we have to posit a logical form where an existential operator binds a variable, or what we may call an "existential variable", formalized as ∃x.

∃x N(x,y)

where x is a term denoting an entity, y is a linguistic sign, and N is a nominative predicate linking y with term x.

This analysis accomplishes Quine's goal of translating a linguistic expression into a logical form that better evidences its ontological commitments. The hidden reference would be paraphrased as "there is something, and that something is referred to with autonym y". It is a case of the most basic denominative operation restricting itself to the predication that "something is referred to with this term", possibly within the context of a sentence whose main point is not metalinguistic clarification at all, as in the next (truncated) example:

> The so-called "neo-liberals" who detect Marxist bias in contemporary class analysis have often …

But not all cases where this peculiar marker operates will be devoid of retrievable denotational reference. We have identified some instances where a referent precedes the marker, and an automatic extraction system should in principle be able to pick it up, as in the following example where "the Mountain division and the East-Midwest division" would count as referents for the term *Rust Belt*:

> The highest coefficients appeared for the Mountain division and the East-Midwest division, the so-called "Rust Belt".

Finally, information about an autonym might not be localized in one place only, and if so it will have to be retrieved from various sentence segments and constituents, either to be reconstructed as a single piece of information or as two or more distinct pieces of information. The following example shows a WH-pronoun that represents a possible referent for the expression "logic trap" as used by *Craib (1992:12),* and an additional conceptual characterization of that referent as the situation of "dismissing a substantive theory by purely rationalistic means".

> However, there could hardly be a better example of what Craib (1992: 12) calls the "logic trap" of dismissing a substantive theory by purely rationalistic means.

Likewise, a previously quoted example's information might be decomposed into two distinct predications:

> Integral power results in a fundamental type of social classification which, adapting Bernstein's terminology, I shall call "frame" (Bernstein 1971).

One would specify a categorization of "frame" as *a fundamental type of social classification,* while the other would simply qualify the term "frame" used in the local context of this text, as an adaptation of "Bernstein's terminology". Although we have argued against the classic definitional schema that presents terminological information as a hyperonym *genus* qualified by specifying characteristics, it has to be noted that this kind of sortal information occurs quite frequently. Thus, a "frame" in this context can be assumed to belong in an *is_a_kind_of* relationship with the EMO's informational segment *social classification,* so that an application that bootstraps ontologies directly from corpora would use such information to classify that

concept into a hierarchical structure; This kind of processing has been done with machine-readable dictionary entries in the ACQUILEX project (Vossen and Copestake, 1993), and has been proposed recently for semantic re-rendering using the biobibliome (Pustejovsky et al., 2002a). We will examine such uses of automatic EMO processing in Chapter V.

Our last example shows another sentential instance in which we can perform a decomposition of a complex metalinguistic operation into more simple logical predicates:

> Culture is an ambiguous term and often refers to ways of life which have little to do with the market place.
>
> 1.- CULTURE is a term
> 2.- This TERM is ambiguous
> 3.- The TERM often refers to ⇨ *ways of life which have little to do with the market place*

III.2.4 Other functional elements and more information potentially extractable from EMOs

Besides core semantic and pragmatic data, EMOs can also supply other potentially useful pieces of information for the metalinguistic operation as a whole, like scope, ilocutionary force, hedging, participants in the exchange, source attribution, etc. These expressive devices help the speaker position himself with regard the actual knowledge being evaluated, as well as from the point of view of the consensus processes taking place in the framework of the expert group. These kinds of fine-grained data about terms are not commonly found in usual terminological resources like knowledge bases or specialized dictionaries, since this information is either not considered important linguistically or conceptually by terminography, or is dismissed as too context-dependent to be useful for the generalizations expected from lexicography.

As we have argued in Chapter II, metalinguistic exchanges in technical text are subjected to the same complex dynamics that govern dialogic interaction in conversation, and language provides the devices that are used by the participants both to crate meaning as well as to perform discursive actions. These intentions, strategies and inferences, and the whole discourse production process (Gülich & Kotschi, 1995), leave a *trace* in the linguistic materializations of EMOs. In listing IV we present other types of information that can potentially be retrieved from some metalinguistic statements, although at present our implementation of the MOP system does not attempt to extract them in an automatic fashion.

Listing IV. Other pertinent information retrievable from EMOs

- **Extent or scope of the terminological proposal:**[41]
  - Local (applies only to present text)
    "In this papers, X means Y"
  - Regional (for a specific theory, school or problem)
    "In this context, X is known as Y"
  - Global (valid for a whole domain or discipline)
    "X, commonly referred to as Y"
- **Participants** in the communicative exchange
  "In contrast to W's usage, I will refer to…"
- **Locutionary force**
  "X might be called Y" vs. "X should be called Y"
- **Attribution of semantic responsibility.**
  who is putting forward the new term, where and when; e.g. who is the agent of terminological change, what are their sources. This can be done in reference to individuals or to specific papers or documents
- **Attitude of speaker** towards others or towards his own utterances.

III.2.5 Processing EMOs

Our findings from careful analysis of metalanguage in corpora suggest that EMOs can be exploited fruitfully as knowledge-rich contexts for extracting information that can be used by lexicographers and terminographers, as well as for other applications such as text indexing for Information Retrieval and knowledge management. Although statistically sparse, Explicit Metalinguistic Operations can be located in a straight-forward manner by using some recurring elements in the textual surface as series of indexes or indicators that allow for its recognition by finite-state devices in NLP applications that do not need large lexicons or high-level processing. In contrast with terminological methods that use pre-compiled term lists or syntactic groupings to detect new terminology, focusing textual search on EMOs would present a two-fold superiority (at least for neology detection): on the one hand, it would retrieve knowledge-rich contexts that can provide multidimensional terminological information, and on the other hand would ensure that the terms obtained were deemed important enough by the writer to merit metalinguistic specification. Term discovery using EMOs would not eliminate other

[41] Extending Pearson's (1998) initial idea: "In the corpora, metalanguage statements fall into two categories, specific and generic. Specific metalanguage statements are statements which are qualified in some way by the author. The author may wish to restrict the scope of the statement to the particular text segment in which the statement appear and will use hedges such as in this context, here to stipulate that the scope of the metalanguage statement is restricted. Generic metalanguage statements, on the other hand, are, as the name suggests, statements which have general applicability."

terminological methodologies, like following collocates or term expansions, but would feed them with reliable term candidates and would update term knowledge bases.

Modern term extraction systems in general use techniques based either on statistical or linguistic information, or a combination of both (Estopà et al. 1998, Cabré et al., 2001). In this, they follow the same general trends of all modern NLP techniques, which can be categorized as either rule-based or as stochastic, or a combination of both. Some of these techniques, for example, use syntactic regularities of term formation to identify the nominal groups in texts that are most likely forming terminological compounds. Others use collocations and mutual information statistics to guess groupings of two or more lexical units into complex terminological units. What most term extraction systems share is their focus on terms and their contexts, while on an EMO-based application terms would be just one of several items to retrieve from a complex and richly-textured speech act. Processing terminology through EMOs conveys much more adequately the interactive and communicative aspects of specialized knowledge associated with terminological systems, which we have attempted to showcase in contrast to conventional terminology theories of Wüsterian inspiration. Employing the Use/Mention dichotomy presented earlier to describe metalinguistic contexts, a term extraction system exploiting EMOs works over mention-based contexts, as opposed to the use-based contexts of mainstream terminological extraction technology, which can review, albeit more superficially, all term occurrences of a text. These extraction systems are semi-automatic, and present to a human lexicographer a list of term candidates for their manual validation. Although they might be useful for compiling dictionaries and glossaries, term-location applications can seldom provide semantic or pragmatic information for terms without costly human intervention. Strictly speaking, an EMO-based application would not have as its main purpose to extract all possible terms from a text, but to obtain the explicit information provided for terms in those texts. In contrast with non-stochastic term-centered approaches to terminological extraction (for example, Meyer, 2001) locating terms by focusing on EMOs would not require lookup on previous resources or lexicons (and thus no bootstrapping using a precompiled glossary), and might involve actual discovery of new lexical units and the meaning conditions associated with them. Used as a term identification tool, the EMO-centered MOP system would generate less spurious terms (or "noise") than a statistical term extraction application, for which overgeneration of term candidates is a major hurdle. EMOs as contexts for terminological work allow for better precision and more richly textured information about terms. Also, the material provided is more precisely centered in language, and no complex inferences are required to extract useful data about them.

Another advantage of exploiting metalinguistic contexts when processing texts for terminological work would be to provide a more dynamic view of the evolution, consolidation and change of terminological and their parallel conceptual systems. Metalinguistic Information Databases that are dynamically constructed by processing specialized texts would maintain glossaries and ontologies up-to-date, but would also constitute a detailed record of each concept's evolution since its first introduction, and would allow for an empirical study of some aspects of scientific activity. In this sense, and in contrast with the resources usually obtainable with conventional computational terminology systems, the EMO-centered MOP system would create simultaneously in a fully automatic way practical reference sources and machine-readable data for empirical study of the evolution of scientific thought and language. Using stochastic methods and heuristic rules, the MOP system is not intended to compete with conventional term extractors, but to provide access to a whole new dimension of information useful for managing and studying empirically specialized knowledge in corpora of technical text, and which conventional systems can process only incidentally. By processing in a theoretically-motivated manner metalinguistic exchanges we can also attempt to know more about the communicative context of highly structured sublanguages and terminologies, about domain-specific linguistic resources, and about the interaction of scientists and technicians as speech and knowledge-centered communities, among other important issues. Besides the more practical and applied side of an computational system such as MOP, we think that researching research in the systematic and data-driven manner made possible by the Metalinguistic Information Databases is a definite improvement over speculative and conceptualist approaches to epistemology and terminology.

In the two previous chapters we have discussed the formal and cognitive properties of metalinguistic statements, which allow them to be the foundational cornerstone of symbolic systems. We have also presented a complex perspective on the metalinguistic exchanges in scientific and technical sublanguages that constitute the communicative context within the community of users of those sublanguages, with an extended description of what role metalanguage plays in the creation and structuring of knowledge—the main declared activity of these expert-domain groups. Finally, we have elaborated a descriptive model of Explicit Metalinguistic Operations (EMOs) that allows for the identification of their constitutive elements and the function each one contributes to the global role of EMOs in discourse. The analysis, grounded on the observation of thousands of actual EMOs in our various corpora, leads us to propose the feasibility of a computer application tfor the extraction and parsing of EMOs from free-text sources. The predictability and regularity of metalinguistic predication across domains, along with the constitutive markedness of such sentences, suggests that use of

Information Extraction techniques and heuristics derived from our corpus study will be enough to create a program designed to compile the metalinguistic information that all specialized text contains alongside its regular subject-matter information. The proof-of-concept Metalinguistic Operation Processor system we describe in the next chapter constitutes the core computational contribution of this dissertation, and it shows how EMO extraction and processing can be achieved without highly-complex NLP machinery, such as deep syntactic parsers or laboriously-built specialized lexicons. The workings of the MOP system are theoretically motivated by the cognitive, pragmatic and formal features of metalinguistic exchanges as have been described in previous chapters of this dissertation. Although few of the computational techniques used by the system are novel, their use and adaptation to metalinguistic information extraction has not, to the best of my knowledge, been reported before.

IV METALINGUISTIC OPERATION PROCESSOR (MOP):

AN APPLICATION FOR THE AUTOMATIC EXTRACTION OF METALINGUISTIC INFORMATION FROM NATURAL LANGUAGE TEXT

Summary

The MOP system described in this chapter applies standard pre-processing techniques (tokenization, POS tagging and partial parsing) on specialized, natural language texts. It first locates and extracts candidate metalinguistic fragments employing patterns of lexis and punctuation compiled from extensive analyses of corpora of Explicit Metalinguistic Operations; The MOP system selects metalinguistic sentence candidates using two different approaches, one involving machine-learning algorithms and another one using pattern-matching over collocations obtained from corpora. After retrieving EMO sentences, the MOP system performs semantic labeling of the chunks, and parses the ones that might fulfill desirable semantic roles into a database structure by using heuristic rules derived from manual analysis of such sentences and of their lexical markers, as well as from relevant semantic frames from the FrameNet project. The MOP system carries out this processing avoiding overly complex and sophisticated NLP machinery, and consequently presents limited parsing and co-reference resolution capabilities. Final output for this processing is a Metalinguistic Information Database containing a three-entry record of infomational segments, markers/operators and autonyms.

By reviewing the formal and cognitive properties of metalinguistic exchanges we have previewed the empirical and theoretical foundations of our proposed computational system for locating and processing metalinguistic information embedded in text. Our use of Explicit Metalinguistic Operations (EMOs) as a descriptive model has hinted at how their unique cognitive and linguistic characteristics make it feasible to implement an Information Extraction system that can create the lexical knowledge databases we call Metalinguistic Information Databases. In this chapter we will describe in detail a proof-of-concept implementation of such a system, the Metalinguistic Operation Processor (MOP). First (IV.1 and IV.2), we will present its goals and the previous state of the art of comparable efforts in Information Extraction, lexical knowledge acquisition and terminology processing. The EMO system technical specifications and the architecture of its two main stages (localization of candidate sentences and predicate processing) will be described in sections IV.3 and IV.4. The next chapter (V) presents a full evaluation of the current version of the system using standard metrics for the tasks, with golden standards and answer keys derived from three different datasets. The full code of the application, output files and golden standards are included in the enclosed CD-ROM for full evaluation.

IV.1 What are the goals of the MOP system?

The Metalinguistic Operation Processor (MOP) is basically an Information Extraction system that locates and processes the metalinguistic information of an explicit nature that resides in domain-specific texts. As discussed in previous chapters, it is language-specific information about the domain that provides the reader with knowledge about the meaning and usage of lexical items, as well as pragmatic and general coding instructions for interpreting and producing meaningful messages. In that sense it is information about the object language, instead of about the entities, relations and processes of the domain. MOP is a special subclass of lexical-knowledge acquisition system that searches text both for terms and lexical items, as well as for the information that is provided about them by the special kind of speech acts or discourse operations we have called Explicit Metalinguistic Operations (EMOs). The final output of the MOP system are Metalinguistic Information Databases that contain a record of sublanguage-

specific information contained in processed text. A complete description of these data structures and their characteristics is presented in chapter VI.

The application inputs full text from scientific papers and technical documents, and after tokenization and general linguistic preprocessing proceeds to locate candidate sentences that carry metalinguistic information. The predicate processing phase then identifies and labels the sentence segments that are potential carriers of relevant information, and organizes them into a database structure corresponding to the EMO components described earlier in section III.2. Through this general processing architecture it automatically analyzes metalinguistic predications that might be embedded in sentences that are globally non-metalinguistic. Consequently, an application that successfully tackles this general task will:

A. Identify accurately the sentences where metalinguistic information resides, while ignoring or filtering out lexically or structurally similar instances of no metalinguistic value.
B. Identify accurately, using frame case analysis, the sentence constituents or segments and the roles they play in metalinguistic predication, performing typing operations over them for a set of custom labels that are specific to the application's goals.
C. Create as output of the application a database record for each EMO instance in which identified constituents or textual fragments constitute an entry for Autonyms, Marker-Operator(s) and Informational Segments, with additional records that register the full context of the analysis (the candidate sentence) and an unique ID string that shows provenance and record number. This database structure constitutes the Metalinguistic Information Database (MID) we will discuss in detail in section VI.

IV.2 Previous work: Information Extraction and Lexical Acquisition

IV.2.1 Information Extraction techniques

The quantity of specialized knowledge available to scientists and technicians has surpassed their ability to process information without computer-based tools. Availability of large-scale corpora has made it possible to mine specific knowledge from free or semi-structured text, resulting in what many consider by now a reasonably mature NLP technology. Information Extraction (IE) systems process documents to extract information about events, entities or relationships.[42] A pioneer implementation was done by Sager et al.. (1982; 1987).

[42] Although an Information Retrieval application ascertains that relevant information is present in a set of documents and presents them to the user, Information Extraction systems go the extra mile and actually go in and get that information, structuring it further for ease of processing.

IE structures knowledge embedded in text into a relatively fixed template frame, and focuses on a small part of the information carried by the text. What is attempted is, in effect, "selective concept extraction" (Riloff, 1993), a technique which sidesteps irrelevant surrounding text while focusing on the relevant one for the task or the domain. Although these techniques generally do not require a processing complexity comparable to that of some dialogue systems, where full semantic interpretation might be necessary, and in contrast with extracting information from fully structured resources, like databases, or even semi-structured ones like classified job announcements, they do require some level of linguistic preprocessing and limited analysis, like chunking, or partial parsing. Although full understanding of text is still a long way from the present state-of-the-art, extraction of conceptual relations on a domain-restricted basis is becoming more and more common. Among other applications that can benefit from these data are Q&A, Word Sense Disambiguation, indexing or summarization systems.[43]

Extensive research in IE techniques, especially with the successful series of Message Understanding Conferences (MUC) in the nineties (sponsored by the U.S. Defense department's DARPA) centered on tasks such as creating and updating databases of corporate joint ventures or terrorist and guerrilla attacks. The next figure (taken from Grisham, 1997) shows a text fragment from the MUC-3 competition dataset, along with the final expected template to be extracted automatically.

19 March — A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb — allegedly detonated by urban guerrilla commandos — blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador: San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

Figure 2. News report and extracted template

One of the features that sets IE apart from other NLP tasks is the possibility of performing accurate evaluations of similar systems, due to the end result being a database that can be used

[43] "This area addresses information processing needs associated with large volumes of text containing information from some domain of interest. For example, a stock analyst might want to track news stories about corporate mergers; an intelligence analyst might need to track descriptions of terrorist events in a geographic region; an insurance adjuster might want to compile data from text-based hospital records. In general, information extraction refers to the problem of finding useful information in a collection of texts, and encoding that information in a format suitable for incorporation into a database." (Lehnert et al., 1994)

as a benchmark. In the initial stages of these early efforts, gathering knowledge from text often required manually crafting knowledge-engineering rules that were both complex and deeply dependent of the domain at hand. Extraction rules combine lexical patterns with frame semantics, heuristics or a mixed bag of computational techniques to decompose linguistic structures into basic semantic components and their predicative relations. In general, these patterns roughly reflect semantic and subcategorization frames of verbal items and structures involved in expressing the events and data of interest for the extraction task. A good survey of the variety of the patterns used in various projects is Muslea (1999). A typical IE system has two stages: in one (Named Entity Recognition), facts and entities are identified in text; in a later stage, relationships holding between mentioned entities are resolved and a database template is filled with that information. These processes do not typically involve discourse processing, and are more focused on local text analysis, but might involve co-reference resolution and performing inferences of one kind or another.

In synch with the general trend in NLP towards using statistical techniques, some successful experiences using learning algorithms to create extraction rules have been reported (Riloff, 1993; Fisher et al., 1995; Ratnaparkhi, 1997; Riloff and Jones, 1999; Califf and Mooney, 1999; Chieu et al., 2003). CIRCUS (Lehnert, 1991), Autoslog (Riloff and Lehnert,1994) and CRYSTAL (Soderland, 1997), for instance, used iterative machine-learning algorithms trained on marked-up corpus to create IE pattern dictionaries that attained good metrics. These systems (either fully automatic or semi-automatic) were comparable with others compiled with hand-crafted rules, providing a good alternative to laborious manual analysis of domain-dependent text. Adaptation to new domains (or scenarios, as extraction situations are called) has been one of the bottlenecks that have hindered IE systems' abilities to evolve into commercial applications, and using stochastic analyses of text to solve this problem is one of the most active areas in recent research. Gildea and Jurafsky (2002), for example, applied statistical classifiers trained on annotated data to automatically label a constituent's semantic role. Using full parsing and sophisticated learning algorithms, Chieu et al., (2003) reported competitive performance in scenario template fill-up tasks, as compared with manually engineered systems of MUC-4.

With regard to Information Extraction techniques, the MOP system we present here has two particularities: first, and unlike many other systems, it is only moderately domain-dependent, since it can be used to process corpora from various disciplines with little or no customization. It alternatively uses collocation-based manually-compiled rules and Machine Learning classifiers to choose its target sentences (see IV.4.3.2), in what is basically a categorization problem. The reason MOP is moderately domain-dependent system has to do with the second peculiarity: the information which MOP is designed to retrieve is present in all kinds of texts

and in all disciplines: given its systemic role in knowledge-building, metalinguistic information resides in documents of all fields and presents a high regularity in its expression due to cognitive and formal reasons. Even though there have been other experiences extracting linguistic information from text (either implicit or explicitly expressed, see next section), the MOP system is unique in its goal of working with metalanguage in general, as opposed to, for example, focusing on specific aspects of language, such as conceptual or structural relations or conventional definitions.

IV.2.2 Extraction of linguistic knowledge from text

Lexical knowledge acquisition is, and has been for some years now, an active and exciting field within Computational Linguistics.[44] A consensus has emerged that the main obstacle to improve performance and coverage of current NLP systems is the exhaustive lexical and linguistic resources that they need. For Boguraev and Pustejovsky (1996) the core issue of lexical acquisition is "how to provide, fully and adequately, the systems with the lexical knowledge they need to operate with the proper degree of efficiency. The answer is (…) to extract the lexicon from the texts themselves." The kind of information gathered can be as sparse as synonyms or as richly textured as lexical entries including morphology, glosses, POS, complements, phonology, corpus statistics, etc. Although some resources, like semantic networks and ontologies, restrict themselves to very specific aspects of lexical knowledge, like hiperonymy or synonymy, and other lexical or terminological knowledge bases represent fully structured databases using typed features and inheritance mechanisms where multidimensional linguistic information is converted into machine-readable data ready to be interpreted by computer applications, the Metalinguistic Information Databases created with the MOP system contain a wide variety of semi-structured data that retains some of its linguistic organization, and needs to be interpreted by humans or further processed to be made useful by incorporating its information into more conventional computational resources. They are midway between highly structured computational terminological knowledge bases and the raw corpora from where those KBs are compiled.

Whereas mining specific semantic relations and subcategorization information from unconstrained free-text has been successfully carried out before (Manning, 1993, Hearst, 1999), automatically extracting lexical resources (including terminological definitions) from text in special domains has been coming of age rather slowly. Creating a corpus-driven database of

[44] "The need for wide ranging lexical-semantic knowledge to support NLP, commonly referred to as the Acquisition Problem, has generated a great deal of research investigating automatic means of acquiring such knowledge." (Lauer, 1994)

domain-specific lexical knowledge using Information Extraction techniques has been shown to be both feasible and practical by recent research, an automation process which is especially important considering the modern lexicon might contain more than 400,000 entries (Pustejovsky, 1999). When focused on language information, IE techniques should be called "Linguistic Information Extraction." These techniques include purely mathematical modeling of texts (Harris, 1968;1991), lexical statistics as text structuring factors (Phillips, 1985), lexical clustering methods (Basili et al. 1996) or a variety of terminology extraction systems using corpus statistics (Cabré et al., 2001). A distinctive feature that has proven useful to classify linguistic IE projects is the kind of information they aim at: in some cases, explicit information like glosses or definitions might be what we are after, while in other cases we might want to retrieve data that is either structurally or semantically implicit in the text, like POS category, subcategorization frames or semantic typing. Recent examples of the later trend are Kageura (2002) or Jacquemin (2001), in which morpho-syntactic and semantic term-formation regularities are exploited using computational techniques.

The last decade of the XXth century saw a flurry of activity in this area. It started with the pioneering work of Amsler and White (1979) which extracted IS_A links from the relationship between a defined term and the syntactic head of dictionary definitions in order to create taxonomical structures; it continued both in the extensive literature of the ACQUILEX projects and in other similar efforts where computational methods were used to create lexical databases, thesauri and ontologies using the highly structured environment of machine-readable dictionary entries and other reference resources like specialty corpora (Chodorow et al. 1985; Cuouto, et al. 1999; Desclés, J-P. et al. 1997; Dolan et al. 1993; Hearst and Schütze 1993). These approaches contrast with traditional terminology identification methods that are term-centered. They focus more on knowledge-rich contexts where terms may reside, instead of attempting to extract and organize all terminology in a document. Some efforts (Kruijff and Schaake, 1995; Pearson, 1996 & 1998; Pascual & Péry-Woodley, 1997; Cartier, 1998; Rebeyrolle and Péry-Woodley, 1998; Cartier, 1998; Fuji and Ishikawa, 2001; Meyer, 2001; Malaisé et al., 2004) have attempted automatic extraction of knowledge from specialized texts and Internet sources, such as software manuals or textbooks, by analyzing relatively fixed and stable definitional patterns that are delimited (sometimes excessively) by formal and structural constraints.[45] What many of these approaches lacked was a way to extract pertinent semantic data from sentences that did not always adhere to the usual formalism of didactic or lexicographic definitions. Also, the diversity of linguistic data that they were capable of extracting was limited by design or

[45] For example, Kruijff & Schaake (1995) studied ways to establish relevance using informational structure in order to help extract definitions from text.

interest. None, to my knowledge, attempts to exploit metalanguage as a whole in quite the way the MOP system does. In these approaches what was needed was a different theoretical textual object that accounted for all the varieties of metalinguistic activity actually encountered, and not just for formal definitions that follow a classic pattern. An extraction method that could accounted only for clear-cut definitions would miss many other fragments of text with relevant information, information about the way language expresses the knowledge we have about the world.

Automatically extracting lexical resources (including terminological definitions) from free text in special domains has been a field less explored, but recent experiences (Klavans & Muresan, 2001; Rodríguez, 2001, 2003b, 2004; Pustejovsky et al., 2002a and 2002b) show that compiling the extensive resources that modern scientific and technical disciplines need in order to manage the explosive growth of their knowledge, is both a feasible and practical endeavor.[46] In fact, computational terminology has followed the rising trend in NLP of successfully applying either statistical or mixed techniques to various problems, complementing or replacing entirely rule-based algorithms that encode linguistic knowledge. A good overview of this trend can be found in a multi-paper volume on *Recent Advances in Computational Terminology (2001).*

We have already mentioned (I.1) the NLM Knowledge Sources as an example of the very large resources that any language technology supporting scientific activity needs to maintain and create. But the usefulness of robust NLP applications for domain-specific text goes beyond glossary updates. The kind of categorization information implicit in many definitions can help improve anaphora resolution, semantic typing or acronym identification in these corpora, as well as allow "semantic rerendering" of domain-specific ontologies and thesaurii (Pustejovsky et al., 2002a). The extensive resources needed for practical, real life applications in Natural Language Processing require automatic compilation of sublanguage knowledge bases. Robust lexical acquisition systems that use both Language Engineering and statistical techniques must be capable of handling the vast digital resources available today to researchers and lay-men alike, helping to make them more accessible and useful. In doing so, they are also fulfilling the promise of NLP techniques as mature and practical technologies. The MOP system described in the next sections attempts to be a step in this direction.

---

[46] The DEFINDER project at Columbia University (Klavans et al, 2001) examines user-oriented medical documents to extract fully-developed definitions for use by laymen. The MedStract project, a joint effort by Brandeis University and Tufts University researchers (Pustejovsky et al., 2002a & 2002b) mines biomedical abstracts to create specialized resources, like the AcroMed acronym database, and perform semantic "rerendering" of the UMLS ontology.

IV.3 MOP specifications and structure

IV.3.1 Development resources

IV.3.1.1 File format standards

We have described (Section I.4) the corpus work that allowed for an empirically-motivated study of metalinguistic sentences. The EMO Corpus and other files related to that part of our research serve now both as a reference source and as training files for some of the stochastic classifiers we will describe further in later sections. These files and all output files have been encoded in XML standards (and related technologies like RDF, DOM and XSL) referenced in the W3 consortium (W3C) website.[47] We have decided to use these standards for a number of reasons, including:

- Data in XML is highly transportable, independent of the specific operating system, platform or client application.
- As in its SGML parent standard, markup entities maintain independence from the final viewing template or output format, but allow quick transformation both in form and structure using XSL and XSLT. XML markup allows for complex structure in data, without being obtrusive to processing applications.
- XML files are fairly transparent, well-suited for limited human reading as well as efficient machine-readable resources.
- Increasingly, linguistic resources are benefiting from these de facto international standards for data interchangeability that promote comparative evaluation of systems and resource reuse (lexicons, corpora, ontologies).
- Error detection in XML is facilitated by use of Document Type Definitions (DTDs), and by the many parsers developed for XML.
- Many programming languages and software packages include now, either at their core or as add-on modules, tools for handling such files.
- XML is license-free.

Our output files also make use of these flexible technologies for the same reasons. We have included *.xsl* files that transform them into html for easy viewing of application files using Microsoft or Netscape browsers. Samples of these files and formats as implemented can be found the enclosed CD-ROM. Fragments of them will be presented in these sections as their content and formatting is discussed.

---

[47] http://www.w3.org/XML/; Resource Description Framework, Document Object Model and Extensible Stylesheet Language , respectively.

IV.3.1.2 Programming language and development platform

IV.3.1.2.1 Core programming language

Python,[48] a freely-distributed programming language, is our core platform for development of MOP. As described in the reference site FAQs,

> "Python is an interpreted, interactive, object-oriented programming language. It incorporates modules, exceptions, dynamic typing, very high level dynamic data types, and classes. Python combines remarkable power with very clear syntax. It has interfaces to many system calls and libraries, as well as to various window systems, and is extensible in C or C++. It is also usable as an extension language for applications that need a programmable interface. Finally, Python is portable: it runs on many Unix variants, on the Mac, and on PCs under MS-DOS, Windows, Windows NT, and OS/2."

Created by *Centrum voor Wiskunde en Informatica* computer scientist Guido van Rossum in 1990, Python is easy to learn and its code easy to read. Its basic functionality and many libraries and modules make it an ideal choice for string and character processing of the kind the MOP system makes extensive use of. Current versions are very stable, and the great numbers of conferences, documentation sites and projects dealing with Python attest to its success among the developer community worldwide. Nowadays it is the language of choice in many Computer Science departments to introduce students into programming methods. For the non-programmer from other technical fields, like linguistics, it is ideal for prototyping Natural Language Processing systems, especially with the functionality provided by modules like its Regular Expression implementation or the NL Toolkit described in the next section. The use of objects and classes in Python is conducive to extensive code reuse, reducing development and debugging time, and improving code readability.

Even though there are very sophisticated freely-available NLP modules coded in other programming languages, (and Python has good calls and methods for using components from non-native code), for this implementation I have opted to use only Python; The main reasons for this decision are that I wanted code to be perfectly transparent and homogeneous, and that it was important to retain complete control over all subprocesses and their interactions. No software that required paid licenses has been used in order to preserve the GNU or Open Source nature of the MOP system, which will be released under these conditions.

[48] www.python.org is the main reference site for this language.

IV.3.1.2.2 The Natural Language Toolkit

Our development platform is the Natural Language Toolkit (NLTK) made available by E. Loper and S. Byrd (Loper and Byrd, 2002).[49] NLTK is open software and MOP uses release 1.2 of this software.[50] Based on Python's core functionalities, it is conceived both for rapid prototyping of NLP systems and as didactic material for Computational Linguistics coursework. It consists of task-specific data and processing modules that can be built up to a complete linguistic processing system via standard interfaces between them. NLTK modules implement techniques and algorithms from the two main approaches used in NLP, symbolic and statistical treatment of linguistic structures. Its object-oriented approach allows encapsulating theoretically-motivated linguistic objects (word types, trees, constituents) into class-based data structures that can be efficiently processed and have a transparent structure. It has very strict type-checking and is exhaustively documented. As more and more CL courses are using it, they end up integrating into it new processing modules developed as course projects, like a Porter-based lemmatizer or a WordNet query module.

The main modules from NLTK used by MOP are:

- **Data classes:**
  - **Tokens:** used to handle word or phrasal tokens; can store location, POS category, etc.
  - **Trees:** used to represent hierarchical structures, like syntax trees.
- **Processing modules:**
  - **Tokenizers:** for text normalization and tokenization.
  - **Taggers:** to implement stochastic POS tagging n-grams.
  - **Partial parsers**, or "chunkers": for non-arboreal representation of syntactic constituents
  - **Parsers:** implements different kinds of parsing algorithms, including chart parsers and probabilistic CFG parsers.
  - Feature-based **text classification algorithms**, implementing naive Bayes and Maximum Entropy algorithms.

NLTK also comes with a sampler of corpora and data from the Linguistic Data Consortium[51] datasets (including parts of the Brown corpus and the Penn Treebank) to train and evaluate systems. Some of the modules, including the evaluation functions, were adapted to our specifications during the development of the MOP system.

---

[49] Project downloads and documentation at nltk.sourceforge.net. Source files of the version of NLTK used with MOP, as well as source for the Python distribution, are included in the enclosed CD-ROM for reference purposes.

[50] Later releases modify token representation and would not be suitable for use with MOP system as-is.

[51] www.ldc.upenn.edu

IV.3.1.2.3 Other third-party code

Although MOP initially used stochastic POS taggers trained on the Brown, DOE and other freely-available corpora, for our present version we have adapted an implementation of the Brill algorithm done by Hugo Liu at MIT,[52] as part of his Monty Lingua NLP system (v1.3.1). It offered better accuracy and speed than the cascading N-gram taggers used at the outset, and allowed for better customization of the tagging lexicon if needed, while maintaining the system 100% Python.

IV.3.1.2.4 Technical documentation

We have used the EpyDoc[53] tool to create browsable documentation for all classes and modules in the MOP system. This tool extracts the documentation strings that describe the function, restrictions and workings of the code, and helps create an html document tree that shows module hierarchies and facilitates navigation of all classes, modules, functions and variables.

We have attempted to be as thorough as possible in the technical documentation of the MOP-specific code, but will defer language- or platform-specific documentation to the Python and NLTK reference sites.

IV.4 The MOP architecture

In its present form the Metalinguistic Operation Processor has three main modules that have to be called separately, and that perform a variety of different tasks.

normalize.py	A text normalization and tokenizing module used to prepare free-text documents for extraction and predicate processing. This stage is an adaptation of a tokenizer originally written by Oliver Steele and modified by Roser Saurí at Brandeis University. It creates a flat ascii text file where each full sentence resides in a single line. It recognizes abbreviations and uses as end-of-sentence punctuation the following set: [? ! . ;]. It performs other general cleanup and ordering tasks before writing a normalized text file.
extract.py	The candidate extraction module performs a search and filtering process to locate and select from the normalized file the metalinguistic sentences that will merit further processing. It writes an extraction file that will be used by the next module. Two versions of this module were written, using different filtering strategies. They will be described further on in this chapter.

[52] http://web.media.mit.edu/~hugo/montytagger/

[53] epydoc.sourceforge.net

exec_mop.py	The main processing module takes as input an extraction file with candidate sentences and outputs a database structure we have called MID. It imports many sub-modules to perform the preprocessing and the main predicate processing stages of the application. In its development form, the user interface presents 3 choices: the input file can be processed in its entirety to be written to an output file, it can process each metalinguistic sentence one by one for debugging purposes, or it can ask the user to type a single sentence for processing. The last two choices output only to the screen, without a new file being written.

Besides the first text preparation phase, the MOP system has two main, distinct phases: selection of candidate metalinguistic sentences (Candidate Extraction Phase), and final extraction of selected information from them (Predicate Processing Phase). This is reflected in Figure 2, which show the general architecture of the system.

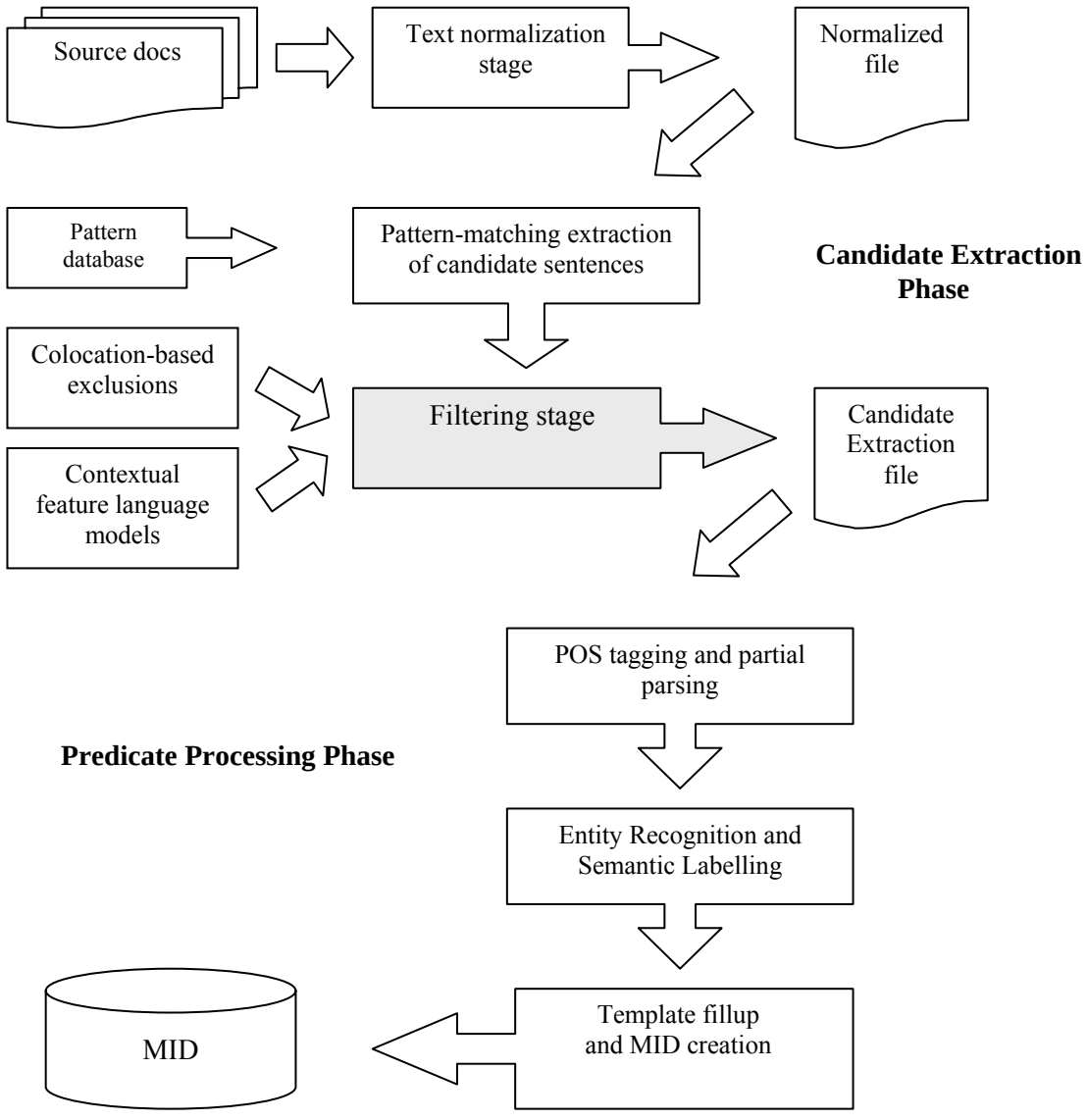


Figure 2: MOP architecture

In addition to these main modules, we have included, in separate directories, file-preparation and evaluation modules that will be described elsewhere, but that are not form part of the general system architecture.

IV.4.1 Text normalization

Input text from specialized discourse can present many styles and formats. It may include tables and graphics, and be in MS Word, HTML, PS or PDF formats. Before processing, these documents have to be flattened into a simple ascii text file. Various file conversion programs or application facilities are used before text can enter program flow. This results sometimes in a loss of possible flagging of metalinguistic conditions through text formatting and layout, but the cognitive redundancy in EMOs marker-operators discussed in previous chapters ensures that our lexically-based detection strategy does not lose much accuracy. To incorporate into our processing important clues of a non-textual nature would require complex modules that are beyond the scope of the current proof-of-concept system described here.

MOP processes text one sentence at a time, and some steps require that various linguistic elements that are normally adjacent in regular text be separated by a text normalization process, like separating parentheses from enclosed words, or separating possessive suffixes from preceding words. Sentence boundary and word form normalization are carried out at the outset, by searching for abbreviations through list lookup and text heuristics. Unknown abbreviations and format conversion errors can induce a modest amount of mistakes than need to be checked manually to ensure accuracy in subsequent stages. All markup enclosed in angular brackets is removed so it will not interfere with XML coding of output files. Output is a simple-text file with a tokenized sentence in each line.

To better illustrate how the MOP system works, in the following sections we will use boxes that exemplify the treatment of a sample sentence. These boxes will have a double-lined frame, as in the following initial instance, which shows the input and output from the text normalization and tokenization process.

Pre-processed text:

The bit sequences representing quanta of knowledge will be called "**KENES**", a neologism intentionally similar to "genes".

Normalized text:

The bit sequences representing quanta of knowledge will be called " KENES " , a neologism intentionally similar to " genes " .

Box A: text normalization module

IV.4.2 Locating EMOs in free-text: two strategies for candidate sentence selection

Our corpus-based analysis of metalinguistic predication (Section I.4.1.2) examined 116 possible lexical patterns that might signal EMO activity in text. These patterns were queried to the British National Corpus and the output files were manually marked-up as metalinguistic or not. The enclosed CD-ROM contains files with the full statistics for each of the patterns tested, as well as the complete EMO corpus. A sample of upper-range statistics is shown in Table E.

Query Pattern	# of Hits	EMOs	Non-EMOs	Percentage of total
('where') (_) ('refers')	3	3	0	100.0
('use'|'uses')('this term')	4	4	0	100.0
('use the term') (``=PUQ)	37	37	0	100.0
('use'|'uses')('the term')	102	102	0	100.0
('the term ``)	100	100	0	100.0
('term ``)	100	100	0	100.0
('so-called') (``=PUQ)	186	186	0	100.0
('so called') (``=PUQ)	27	27	0	100.0
('known as') (``=PUQ)	100	100	0	100.0
('is called') (``=PUQ)	78	78	0	100.0
('dubs')	5	5	0	100.0
(``=PUQ)*(', known as')	100	100	0	100.0
(', known as')	100	100	0	100.0
('calls (``=PUQ)	100	100	0	100.0
('called') (``=PUQ)	200	200	0	100.0
('apply the term')	3	3	0	100.0
('so called')	111	110	1	99.0990990991
('termed')	100	99	1	99.0
(``=PUQ)*(', the term')	89	88	1	98.8764044944
('use the term')	77	76	1	98.7012987013
('use'|'uses')('the word')	66	65	1	98.4848484848
('known as')	100	98	2	98.0
('definition of')(``=PUQ)	79	77	2	97.4683544304
('what') (_) ('calls')	198	191	7	96.4646464646
('what') (_) ('called')	200	191	9	95.5
('might be called')	84	80	4	95.2380952381
('use of the term')	62	59	3	95.1612903226
('word') (``=PUQ)	198	182	16	91.9191919192
('is called')	100	89	11	89.0
('dubbed')	99	88	11	88.8888888889
('coined')	91	79	12	86.8131868132
(', a term')	37	32	5	86.4864864865
('the term')	100	85	15	85.0
('could be called')	51	41	10	80.3921568627
('euphemistic')	5	4	1	80.0
('a term'|'the term'|'this term')	100	80	20	80.0
('called')	200	154	46	77.0
('christened')	29	22	7	75.8620689655
('definition of')	200	150	50	75.0
('in terms of') (``=PUQ)	85	63	22	74.1176470588
('used to refer to')	57	41	16	71.9298245614
('use'|'uses')*('the phrase')	200	136	64	68.0
('connotation')	62	38	24	61.2903225806
('use'|'uses') ('a term')	5	3	2	60.0
('as ``)	100	55	45	55.0
('stands for')	85	45	40	52.9411764706

('refers to')	100	52	48	52.0
('dub')	10	5	5	50.0
('a term')	100	49	51	49.0
....	...	...	...	...

Table E: Pattern reliability statistics from EMO corpus
(PUQ means quotation marks; (_) means any intervening element)

In the case of the pattern ('what') (_) ('calls'), it represents instances where a relative clause introduces an intervening referent and is followed by a term, as in:

Some authors follow Weber himself and assert that what he **calls** " social classes " …

For this specific pattern, 198 similar cases were retrieved from the British National Corpus. Of them, 191 were classified as true EMOs, and 7 were deemed spurious, for a 96.46 % accuracy rate.

From these statistics extracted from the EMO corpus and careful observation of the pattern occurrences, 44 of those patterns were selected to act as lexical triggers for the MOP system. Their selection was based on a high degree of reliability as indicators of EMOs in text (with a lower threshold of around 75% EMOs of all sentences), as well as other considerations; for example: (A) Although some patterns presented a high percentage of EMOs, their number of hits was too low for inclusion since they were too infrequent; (B) In cases where there was a lower but significant percentage of EMOs, if reliable collocation-based exclusion rules could be formulated, they were selected, since metalinguistic instances would be identified with a tolerable margin of error. An XML file, as shown in figure 3, was created with these patterns for use by the search algorithm.

```
<?xml version='1.0' encoding='ISO-8859-1'?> <!DOCTYPE patterns SYSTEM 'patterns.dtd' []>
<patterns>
<pattern pat=' known as ' type='B'/>
<pattern pat=' known \w+ as ' type='B'/>
<pattern pat='So(-| )called ' type='D'/>
<pattern pat=' (S|s)o(-| )called ' type='D'/>
<pattern pat=' a.k.a ' type='B'/>
<pattern pat=' call ' type='B'/>
<pattern pat=' calls ' type='B'/>
<pattern pat=' coins ' type='B'/>
<pattern pat=' coined ' type='F'/>
<pattern pat=' coin ' type='B'/>
<pattern pat=' christened ' type='F'/>
<pattern pat=' christen ' type='F'/>
<pattern pat=' christens ' type='F'/>
<pattern pat=' define as ' type='F'/>
<pattern pat=' defined as ' type='F'/>
<pattern pat=' defined (\"|\') ' type='F'/>
<pattern pat=' defines as ' type='F'/>
<pattern pat=' defines (\"|\')' type='F'/>
<pattern pat=' denote ' type='F'/>
<pattern pat=' denotes ' type='F'/>
<pattern pat=' denoted ' type='B'/>
<pattern pat=' designate (\"|\')' type='B'/>
<pattern pat=' designated (\"|\')' type='B'/>
```

```
<pattern pat=' designates (\"|\')' type='B'/>
<pattern pat=' (\"|\') designates ' type='B'/>
<pattern pat=' dubbed ' type='F'/>
<pattern pat=' dubs ' type='C'/>
<pattern pat=' dub ' type='C'/>
<pattern pat=' labeled (\"|\')' type='B'/>
<pattern pat=' labels (\"|\')' type='B'/>
<pattern pat=' (\"|\') means ' type='F'/>
<pattern pat=' named (\"|\')' type='B'/>
<pattern pat=' us(e|es) (a|the) (phrase|word|term|expression) (\"|\')' type='B'/>
<pattern pat=' refer to as (\"|\')' type='F'/>
<pattern pat=' refers to (\"|\') ' type='B'/>
<pattern pat=' (\"|\') refers to ' type='F'/>
<pattern pat=' refered to as (\"|\')' type='F'/>
<pattern pat=' refered to (\"|\')' type='F'/>
<pattern pat='(\"|\') stand(|s) for ' type='F'/>
<pattern pat=' called ' type='B'/>
<pattern pat=' term "' type='B'/>
<pattern pat='" term ' " type='B'/>
<pattern pat=' terms (\"|\')' type='B'/>
<pattern pat=' termed ' type='B'/>
</patterns>
```

Figure 3: Pattern identification XML file

Each pattern is associated with the "type" attribute to a one-letter code that represents a distinct processing route inside the Predicate Processing stage that we will describe below (IV.4.3.2.1). These processing routes aid the semantic labeling of autonyms and informational segments by indicating a preferred directionality around the axis of the pattern that has been located as the marker/operator in that specific textual instance. The segments before and after the marker/operator (signaled with "kw" tags) are identified in the xml file. The following box shows the extracted entry from our walk-through sentence, which in this case shows a preference for finding the informational segment (bit sequences representing quanta of knowledge) "backwards" starting from the position of the marker "called".

```
<line number='71' pat=' called ' type='B'><bf> The bit sequences representing quanta of knowledge will be </bf><kw> called </kw><af> " kenes " , a neologism intentionally similar to " genes " . </af> </line>
```

Box B: Entry for example sentence in extraction file

IV.4.2.1 Finite-state candidate sentence extraction

With this restricted set of triggering patterns, it is obvious that the MOP system will not attempt to locate every possible instance of a metalinguistic sentence. For example, the pattern comma plus the lexical item *where* (", where") followed by a semicolon (:) is used in local specification of certain meaning conditions for items in formulas or argumentation, as in:

Now we shall assume that there is a linear relationship between the vector potential and the current density <formula> **, where** y is the new macroscopic constant .

Other interesting metalinguistic sentences that cannot be selected reliably using this approach include examples like the following two from our sociology corpus:

3) Durkheim is often accused of ambiguity, on the grounds that both "externality" and "constraint" can mean many different things (e.g. Lukes 1975).

4) Finally, the term business elite should be restricted to any exclusive group of individuals and families whose corporate leadership is attributable to contingent exclusion strategies or inheritance.

Example 4 shows an instance where some additional type formatting (like italics in the autonym) might have been lost in the preprocessing, so that it makes it very difficult to identify this instance as a metalinguistic one with just the descriptor "term". In the MOP Corpus, "term" by itself only returns a 28% of true EMOs. When we find quotation marks adjacent to it, its identification is much more reliable, and we do incorporate those candidates into our extraction file, as in:

5) The term "private" is thus a symbolic flag which draws on shared cultural meanings to give day-to-day effect to myriad zones of exclusion and inclusion in day-today life.

At this phase the list of extraction patterns is converted into regular expression patterns and tested over each of the lines from the normalized file. When a match is found, the sentence candidate is subjected to a filtering process to determine accurately if it is or not a true EMO. The problem can be illustrated as the challenge of distinguishing between useful query results such as (6) from non-metalinguistic sentences like (7), with both sentences from the same sociology corpus document and found by searching for the lexical marker "called":

6) Since the shame that was elicited by the coding procedure was seldom explicitly mentioned by the patient or the therapist, Lewis **called** it unacknowledged shame.

7) It was Lewis (1971;1976) who **called** attention to emotional elements in what until then had been construed as a perceptual phenomenon .

The general problem is to find which features in these sentences mark their metalinguistic function, or conversely, which features make them non-metalinguistic. It could be viewed, from the perspective of similar NLP challenges, as a classification problem (akin to Word Sense Disambiguation, POS tagging or document categorization).

In the following table we exemplify this task by presenting a fragment of text where lines 4, 6 and 7 would be selected as metalinguistic candidate sentences by regular expressions capturing the highlighted sections, but sentence 6 would be spurious because the pattern "called" is part of a phrasal verb that includes the collocation "for", and does not have a metalinguistic component.

1.- The first assumption we make is that the simulation may proceed without reference to any external " objective reality " .

2.- We shall simulate scientific papers each of which will capture some quantum of " knowledge " , but the constraints on this knowledge will be entirely internal to the model .

3.- To represent a quantum of knowledge , we shall use a sequence of bits .

4.- The bit sequences representing quanta of knowledge **will be called** " kenes " , a neologism intentionally similar to " genes " .

5.- Kenes could in principal consist of arbitrary bit sequences of indefinite length .

6.- This position is echoed by Lyman and Scott who declare that an account is " a linguistic device employed whenever an action is subjected to valuative inquiry " (1970:112) , and is thus used to explain unanticipated or untoward behaviour , with the consequence that accounts are not **called** for when people engage in routine common-sense behaviour.

7.- In this discussion **the term "** account " is taken to embrace all self-referential reporting by actors , whether expressed in speech or in writing ; specifically those which report on the initiation , monitoring and completion of actions .

Box C: Example extraction task

The simple strategy of obtaining candidate sentences for further processing requires a more sophisticated disambiguation technique that allows non-metalinguistic sentences to be excluded from the extraction files, in order to enhance precision for this task. Two different strategies for filtering non-EMOs were experimented with. These methods are representative of wider paradigmatic approaches to NLP: symbolic and statistic techniques, each with their own advantages and limitations. The first one used corpus-based collocations in the immediate co-text of the markers-operators to reject spurious sentences, using, for example, the evidence of a verbal phrase like "called for" that is not metalinguistic. The second approach to filtering non-EMOs from our initial set of candidate sentences was different, and used machine-learning algorithms trained on our EMO corpus to classify and label the candidate either as EMO or non-EMO. The features used by the classification experiments were either POS tags or word forms

surrounding the sentence marker, identified during the pattern matching stage. Both of these approaches, in fact, aim at representing known regularities in observed data, either by hand-coding rules that mimic them or by constructing stochastic representations, or language models, of those localized regularities. Statistically, the hand-coded approach could be viewed as retrieving the "absolute" probabilities that a certain lexical item, when in co-presence with specific markers, will indicate that the sentence is non-metalinguistic in nature. On the other hand, the statistical language models that reflect more complex algorithms introduce a better and finely-grained representation of those lexical contexts, as attested by a limited, and sometimes sparse, corpus.

IV.4.2.2 Manually-crafted collocations as filtering criteria

Many of the mainstream problems of NLP can be conceived computationally as search and classification tasks, and both hand-coded and machine-inducted rules have proven to be effective for many tasks. Our corpus analysis suggested that for sentences with fairly reliable patterns, the immediate textual items to the left and right of the patterns of markers/operators provided important clues about its metalinguistic nature. Through standard corpus analytics, we were able to compile a list of word forms that blocked the metalinguistic interpretation of such contexts. Another way to put this is it to state that certain collocations helped perform a limited word sense disambiguation task that was geared towards identifying the metalinguistic semantic content of the verb or descriptor involved.

Table F shows a sample of such collocations for some of the metalinguistic verbs used by the pattern-matching extraction.[54]

Preceding	Subsequent
for **calls**	
in, duty, personal, conference, local, per, next, The, the, a, their, his, her, its, house, anonymous, of, phone, telephone, one, counseling, service...	out, anyone, someone, charges, before, forth, charge, and, back, contact, on, off, in, for, for, upon, to, into, off, 911, by...
for **coin**	
pound, small, pence, in, toss, the, this, a, that, one, gold, silver, metal, esophageal, different, same ...	toss
for **terms**	
search, equal, unequal	of, under

Table F. Sample of filtering collocations

Filters based on collocation heuristics were then applied; for example, if we searched for the metalinguistic pattern *called*, we would have to filter out the co-occurrences *called for* and

[54] The complete collocation file (collocations.xml) can be examined in the enclosed CD-ROM.

called upon, which are clearly not relevant for our purposes. We implement other rules that filter out instances where *call* (and other patterns) is followed and/or preceded by numbers or digits, as in imperative sentences that refer to telephone calls:

8) Just call 1-800-607-6872 to get started.

Unfortunately, this step would incorrectly filter out sentential fragments like " … thrown up with the dense lamina propria into projections called 3 papillae of various kinds …", which are not infrequent in biomedical literature, and this mechanism should be refined further in production versions of the MOP system. The algorithm stores filtered sentences in an "exclusions" file for later review of the process accuracy. Sentences that are not filtered-out are written into an extraction file that will serve as input for follow-up processing, and the routing labels, as well as document and sentence IDs, are added to each record, as shown in box B. The sentence is marked-up into 3 elements: the marker pattern items, and the preceding and subsequent text as articulated by the former element. Global and pattern statistics are compiled and written into the XML file. A small sample of an actual extraction file is shown in Figure 4.

<?xml version='1.0' encoding='ISO-8859-1'?> <!DOCTYPE extraction SYSTEM 'extract.dtd' []>

<extraction>

<line number='1' pat=' called ' type='B'><bf> The detailed morphology revealed by EM may be </bf><kw> called </kw><af> fine or submicroscopic structure/ultrastructure . </af> </line>

<line number='3' pat=' so called ' type='D'><bf> Cytoplasm : the </bf><kw> so called </kw><af> soluble phase of the cell , consisting mostly of water , dissolved solutes , and larger molecules in suspension tending to link repetitively with covalent bonds giving the cytoplasm a dense , viscous colloidal sol or gel consistency . </af> </line>

<line number='4' pat=' called ' type='B'><bf> the inner membrane projects inwards as plates or tubules </bf><kw> called </kw><af> cristae , studded with small 9 nm wide elementary particles - rounded bodies on stalks . </af> </line>

<line number='5' pat=' called ' type='B'><bf> (b) A similar cylindrical structure is seen at the base of each cilium and is </bf><kw> called </kw><af> a basal body/kinetosome . </af> </line>

.....

<line number='36' pat=' termed ' type='B'><bf> (i) On the endosteal bone , and in the marrow cavity , as a bony layer and as trabeculae , together </bf><kw> termed </kw><af> the internal/endosteal callus . </af> </line>

<line number='37' pat=' called ' type='B'><bf> This procedure is </bf><kw> called </kw><af> reduction of the fracture . </af> </line>

<line number='38' pat=' termed ' type='B'><bf>the sequences are </bf><kw> termed </kw><af> enhancers , or silencers and repressors , respectively . </af> </line>

<line number='39' pat=' so called ' type='D'><bf> A </bf><kw> so called </kw><af> cell-type-specific TF can be used by closely related cells , e.g. , in erythrocytes and megakaryocytes . </af> </line>

<stats name='HistologyNumberedJP.nor' total='5146' extracted='39' filtered='5' percnt_extrac='0.757870190439' percnt_filtered='12.8205128205'/>

<patstat p=" called " howmany="17"/>

<patstat p=" (S|s)o(-|)called " howmany="6"/>

<patstat p=" known as " howmany="2"/>

<patstat p=" termed " howmany="10"/>

<patstat p=" defines " howmany="2"/>

</extraction>

Figure 4: Sample of candidate sentence extraction file from a Histology manual

Overall, this filtering strategy turned out to be very reliable for our test corpora (see evaluations in Chapter V), and was the one used for the global tests of the MOP system. The collocation information implemented involved 9 of the 44 extraction patterns, as most of the others are reliable enough not to merit any filtering-out mechanism.

One problem with this approach (and one that similar Knowledge-Engineering approaches share) is that some of the hand-compiled rules are domain-specific, and customization of the systems for other domains is very labor-intensive. In our tests, although most of the collocations (phrasal verbs or prepositions) worked language-wide, some of them were very specific to the domain: ***esophageal coins***, for instance, will be quite unusual outside of medical documents. Our collocation exclusions are thus geared towards our main Sociology and Health Sciences evaluation corpora, and would need review if MOP is to be applied to other domains and document sets. Although collocation-based filtering will result in a working system, such customization is an error-prone and laborious task. This was one of our rationalizations for testing machine-learning approaches for the disambiguation task.

IV.4.2.3 Stochastic classification and the filtering task: using contextual feature language models

Another filtering strategy we experimented with involved using statistical algorithms for discriminating between metalinguistic and non-metalinguistic sentences. With the increasing availability of specifically marked-up training corpora, statistical techniques using conditional probability and machine learning algorithms (Riloff and Jones, 1999; Nigam et al, 1999) hold promise in automating and fine tuning some NLP processes, as well as in ensuring quick customization and portability of systems to new domains. Assuming that local co-text around the markers would provide good indication of metalinguistic function, we focused on that context for the task. If that assumption bore out, we could perform the classification of candidate sentences without having to look elsewhere or perform complex processing for interpretation. We would also avoid having to perform laborious corpus analyses to find collocations that could disambiguate EMOs. But, how much context was really necessary? Are word forms sufficient (as in the collocation-driven disambiguation explained earlier), or should we also use the morpho-syntactic context for the classification task?

To test our assumption that the context of metalinguistic markers is important for this problem, we targeted it to obtain relevant features for classifiers based on well-known naive Bayes and Maximum Entropy algorithms that work with sparse data.[55] We used as features either the POS tags or the word form tokens immediately adjacent in 1, 2 and 3 positions before and after our triggering markers. Testing all these possible combinations can help us find out empirically the ideal mix of algorithm, feature type and coverage that would insure best accuracy in locating EMOs. This technique focuses on "local" language models, that is, limits observations to a highly restricted portion of the linguistic structure that we have assumed to be central to our specification. We have called this technique "contextual feature language models" (Rodríguez, 2004a), and is similar to what Berger et al. (1996) call "context-dependent word models".

The naive Bayes algorithm estimates the conditional probability of a set of features given a label, using the product of the probabilities of the individual features given that label. It assumes that the feature distributions are independent, but it has been shown to work well in cases with a high degree of feature dependencies (Rish, 2001). The Maximum Entropy model, on the other hand, establishes a probability distribution that favors entropy, or uniformity, subjected to the constraints encoded in the feature-label correlation that is known. When training our classifiers, Generalized and Improved Iterative Scaling algorithms are used to estimate the optimal maximum entropy of a feature set, given a corpus. In other words, given known data statistics, we construct a model that best represents them, but is otherwise as uniform as possible and does not assume anything else that is not known (e.g. is not attested in the training data).

For the filtering task, two labels "YES" or "NO" were assigned to each sentence. We compiled from our labeled BNC sentences containing our selected patterns for training the classifiers, and converted them into labeled vectors.

After locating the triggering pattern, we constructed a three-part Python tuple and created labeled tokens. The following example used 3 positions before and after the marker with POS tags:[56]

('VB WP NNP', 'calls', 'DT NN NN')/'YES'@[102].

This represents the grammatical context of the marker with which this specific metalinguistic instance was found, that is, the lexical item 'calls'. Similarly, a feature constructed from a

---

[55] We will not describe these algorithms in full in this dissertation, since advancing new stochastic techniques is not part of its core claims. *See* Rish, (2001), Ratnaparkhi (1997) and Berger et al (1996) for a formal description of these algorithms, and Loper (2003) for the implementation specifications of these algorithms within the NLTK platform.

[56] The last number represents location of linguistic unit. The token representation under NLTK is currently under review and probably will change in future versions of the platform.

similar text segment "... *creates what Croft calls a description constraint* ...", using word forms and two adjacent positions would result in:

('what Croft', 'calls', 'a description')/'YES'@[102].

After extraction of candidate sentences using pattern matching, and conversion of the relevant context to unlabeled feature tokens, the application presented the output for the classifiers to decide if they were metalinguistic instances or not. In our tests, the resulting labeled file was compared against a golden standard to obtain metrics to evaluate the accuracy of all the classifiers and training strategies.

The different number of positions considered to the left and right of our training corpus, as well as the nature of the features selected (there are many more word-forms than POS tags) ensured that our 3-part vector introduced a wide range of features against our 2 labels for processing by our algorithms (e.g., 136 for POS tags vs. 1,252 for words in one Bayesian network, with one position before and after the markers).

IV.4.3 Predicate Processing stage

Once correctly identified, EMOs need to be parsed so as to obtain their predicative structure. The Predicate Processing stage decomposes the predication articulated by the markers-operators lexical items in order to find the linguistic constituents that better correspond to the autonyms and the informational segments in the EMO structure. This is done so that grammatical and cognitive properties specific to each structure are considered when incorporating them into a database entry.

As mentioned earlier, during the extraction phase processing route labels were assigned to each candidate sentence. These labels, in combination with heuristic rules and linguistic tests, determine the treatment the MOP system will apply to each one. The following sections explain these pattern-specific processing methods.

IV.4.3.1 NLP pre-processing: tagging, partial parsing and autonym recognition

During the pre-processing phase the application reads a sentence from the extraction file and tags it with POS information. It then inserts a special tag (MKR) for all lexical items considered marker-operators, except for quotation marks, which are tagged 'QUOTs'. Some prepositions, like 'of' receive a special tag so they can be recognized when performing later prepositional attachment. As we have stated before, although initial versions of the MOP system used a stochastic tagger based on the functionalities of the NLTK module, we substituted it in our present version for a rule-based tagger written by Hugo Liu at MIT (Monty Tagger) with a reported accuracy of 95%. No major customization has been performed, except for changing the quotation tags (") for 'QUOT' and adding alternative tags for two terms. Our system uses the

widespread Penn Treebank tagset (Marcus et al. 1993), and defaults all unknown items to nominal categories. Box D shows the result of this initial preprocessing

The/DT bit/NN sequences/NNS representing/VBG quanta/NN of/OF knowledge/NN will/MD be/VB called/MKR "/QUOT kenes/NN "/QUOT ,/, a/DT neologism/NN intentionally/RB similar/JJ to/TO "/QUOT genes/NNS "/QUOT ./.

Box D: Example sentence with POS and customized tags

Since marker-operators in the sentence constitute the processing axis for proper EMO parsing, they are located and isolated before any processing. Next, a module that attempts identification of the elements in autonymical condition is applied. This algorithm implements a cascade of simple rules that check if there already is an item suitably flagged by quotation marks in an expected position according to their processing labels. For example, label 'B' will expect at later stages that an informational segment has to be searched for backwards, starting from the marker-operator and up to the start of the sentence, as in:

9) There 's a long-established and widespread dislike of what is **known as** " miscegenation " .

The marker-operator in bold establishes boundaries for the textual segments where the application will search for relevant constituents that might fit the roles of the EMO components we are looking for. In this case, MOP will look forward of the marker "known as" for any quoted segment that might be an autonym (in this case *miscegenation*) and after attempting to attach pre- or post- nominal modifiers and complements, will store a variable with those segments.

In other cases, when it does not find any quoted element, it will proceed to do a partial parsing of the sentence, starting from the marker-operator boundary and using the predicted directionality, in order to retrieve the most probable chunk that might be a candidate autonym. The following sentence illustrates such a case:

10) These radially-directed nuees are an example of a phenomenon **known as** the base surge or ground surge, first recognized in studies of test explosions of nuclear weapons.

In this case, and starting forward from the bolded marker "known as", the algorithm will select the segment *the base surge or ground surge* as autonymical items that represent the term(s) that the sentence is supplying metalinguistic information for. The process uses

capitalization and punctuation heuristics to attain better accuracy. If the relevant segment is not yet quoted, it will insert quotation marks to help retrieve it at later stages. In rare cases where there is no potential autonym, it will reverse directionality of its search.

The partial parsing module is then applied to the whole sentence, in order to identify noun phrase constituents, leaving as unattached tokens all other items, such as non-modifier verbal and prepositional phrases. Partial parsing (or chunking) does not create complete phrase structure trees, but looks for "flat" syntactic structures in a sentence. It uses only sentence surface information without needing a grammar that depends on abstract categories, and is computationally much easier to implement than a full syntactic parser. Undoubtedly, a full parse of our candidate sentence would improve accuracy in identification of potential segments for filling autonym and informational slots, but the complexity of the system would increase substantially. In most cases, we believe that a full parse is not necessary for this local scope phenomena; we are interested in well-bounded grammatical configurations for a limited set of lexical items, and we have decided to keep linguistic preprocessing as simple as possible to maintain overall robustness of the MOP system. Nevertheless, a future version of the system could rely on a dependency-based parser to better identify syntactic constituents and their semantic roles.

Our chunking apparatus is an ordered cascade of rules using regular expressions over POS tags that isolates quoted elements and marker-operators, and proceeds on to aggregate nominal phrase elements. Our 41-rule implementation uses, among others, limited PP-attachment rules that considers only cases of high probability, for example NP-OF-NP structures. Verbal, adverbial, adjectival and prepositional modification of probable noun phrases is attempted, and merging with possessive suffixes and coordinating conjunctions is controlled at this stage.

The NLTK chunking module provides 5 kinds of rules to do a cascading integration of lexical elements: a) chunking rules that aggregate POS tags, b) chinking rules to break-up previous chunks at predicted tags, c) unchunking rules to break-up a whole sequence of tags, d) merge rules to combine two chunks via a terminal and an initial POS tag, and e) splitting rules to split a previously constructed chunks into two smaller ones. A typical NP chunk representation in NLTK is:

('NP': 'the' 'big' 'dog')@[0w:2w]

At this point MOP will have a sentence structure with two distinct segments articulated by a cluster of marker-operators, and organized into nominal chunks and isolated tokens. Box E shows the chunking results over the POS tags of our example sentence (chunks are shown by sets grouped with curly brackets).

{<DT> <NN> <NNS> <VBG> <NN>} <OF> {<NN>} <MD> <VB> <MKR> {<QUOT> <NN> <QUOT>} <,> {<DT> <NN>} <RB> <JJ> <TO> {<QUOT> <NNS> <QUOT>} <.>

Box E: Chunked example sentence with POS and customized tags

Potential autonyms and nominal constituents have thus been identified. From this input, semantic labeling of these constituents can proceed for predicate processing of the EMO.

IV.4.3.2 Information extraction via heuristic rules: pattern-specific processing routes

IV.4.3.2.1 General semantic labeling process

The next MOP step involves shallow semantic labeling of the sentence segments in order to identify possible roles they play in the metalinguistic predication. This step is similar to what Information Extraction systems do when they use predicative patterns such as the following, customized for an executive database update:[57]

1. "<Person> WAS NAMED <Corporate Post> OF <Organization>"
2. "<Person> SUCCEEDS <Person>"
3. "<Person> FORMERLY <Corporate Post>"
4. "<Person> WHO WAS NAMED <Corporate Post>"

The labeling module uses a data structure for each element in the sentence in which a list is built where the last element is always a string label. The lists contain a single token or a single chunk, plus a string label. The representation of each element is:

[Linguistic Item(s) @ [Range or Location] , LABEL]

The labels applied by MOP at this stage are the following:

LABEL	DESCRIPTION
N-CHUNK	Identifies noun phrases that are yet to be specified for semantic role
AUTO	Identifies potential autonym chunks
ANAPH	Identifies anaphoric elements like pronouns that could point to an entity or referent expressed elsewhere in the text. These elements are potential informational segments, as they stand for the referent expressed elsewhere. Some of the items labeled thusly are: 'them', 'this', 'those', 'these', 'they', 'it', 'what'.
AGENT	Labels personal pronouns that might represent person names, textual references or other entities. In metalinguistic predication they can be credited for the information (original creation of term, modification, reported

[57] Taken from Soderland et al. (1997)

	speech, etc.)
TOKEN	All single lexical items that are not markers, or are not attached to a noun phrase chunk, are labeled with this tag
MARKER	Special label to identify marker-operators constituting the predicative axis

```
[('NP': 'The'/'DT' 'bit'/'NN' 'sequences'/'NNS' 'representing'/'VBG' 'quanta'/'NN')@[0w:5w], 'N-CHUNK']
['of'/'OF'@[5w], 'TOKEN']
[('NP': 'knowledge'/'NN')@[6w], 'N-CHUNK']
['will'/'MD'@[7w], 'TOKEN']
['be'/'VB'@[8w], 'TOKEN']
['called'/'MKR'@[9w], 'MARKER']
[('NP': '"'/'QUOT' 'kenes'/'NN' '"'/'QUOT')@[10w:13w], 'AUTO']
[','/','@[13w], 'TOKEN']
[('NP': 'a'/'DT' 'neologism'/'NN')@[14w:16w], 'N-CHUNK']
['intentionally'/'RB'@[16w], 'TOKEN']
['similar'/'JJ'@[17w], 'TOKEN']
['to'/'TO'@[18w], 'TOKEN']
[('NP': '"'/'QUOT' 'genes'/'NNS' '"'/'QUOT')@[19w:22w], 'AUTO']
['.'/'.'@[22w], 'TOKEN']
```

Box F: Chunked example sentence with structural labels

Our Python processing class constructs a list of labeled segments, shown in table F, over which the heuristics for role identification will be performed. It combines semantic labeling methods with techniques similar to the ones used for Named Entity Recognition in Information Extraction systems. Nevertheless, it does not employ any lookup on Gazetteers, glossaries or other known-entity lists, since it is assumed that new, previously-unencountered terms might exist in the sentential fragments, and that new information might being provided for some of the words of the assigned EMO roles (notably, when identifying neological autonyms). We do not want to introduce any assumptions about the classification, nature or role of entities, other than those introduced by our own heuristics or from explicit mention in text. Search in domain-specific lexicons or ontologies could improve the entity labeling task, but would hurt coverage and portability of the system across domains. We assume that this and other efforts at customization of the system for a specific domain could improve overall performance, but we have not dwelt on this issue here since comparative evaluation of the system for different corpora and domains is one of our secondary goals.

Two normalized examples sentences and their chunked and labeled representation are provided next for illustration of the described processes. NLTK representation of POS-tagged tokens is word form and tag separated by a backslash:

Sentence 1:

Stevenson christened them " persuasive definitions " .

[('NP': 'Stevenson'/'NNP')@[0w], 'AGENT']
['christened'/'MKR'@[1w], 'MARKER']
['them'/'PRP'@[2w], 'ANAPH']
[('NP': ''/'QUOT' 'persuasive'/'JJ' 'definitions'/'NNS' ''/'QUOT')@[3w:7w], 'AUTO']
['.'/'.'@[7w], 'TOKEN']

Sentence 2:

To some the sole alternative to this is to embrace what many contemporary theorists have called " positivism " .

['To'/'TO'@[0w], 'TOKEN']
['some'/'DT'@[1w], 'TOKEN']
[('NP': 'the'/'DT' 'sole'/'JJ' 'alternative'/'NN')@[2w:5w], 'N-CHUNK']
['to'/'TO'@[5w], 'TOKEN']
['this'/'DT'@[6w], 'ANAPH']
['is'/'VBZ'@[7w], 'TOKEN']
['to'/'TO'@[8w], 'TOKEN']
['embrace'/'VB'@[9w], 'TOKEN']
['what'/'WP'@[10w], 'ANAPH']
[('NP': 'many'/'JJ' 'contemporary'/'JJ' 'theorists'/'NNS')@[11w:14w], 'N-CHUNK']
['have'/'VBP'@[14w], 'TOKEN']
['called'/'MKR'@[15w], 'MARKER']
[('NP': ''/'QUOT' 'positivism'/'NN' ''/'QUOT')@[16w:19w], 'AUTO']
['.'/'.'@[19w], 'TOKEN']

Depending on the metalinguistic pattern involved and the general linguistic structure of the sentence being processed, these labels can be ordered in a variety of ways. Although there is not what we could call a “canonical structure” for EMO predication, some forms are certainly more common than others. Argument structure and verb valence dictate some of the initial configurations for processing, but linguistic tests performed at critical points guide the search and selection of segments. The processing algorithms in effect explore case frames for each of our markers to identify entities and relations between them in the predicative structure of the sentence.

Table F shows some of the forms that predicate processing must deal with. It shows examples sentences from our sociology corpus along with labeled bracketing resulting from an analysis of its relevant segments. The markers/operators are in boldface.

11) **[**$_{A}$ Rosaldo**] [**$_{M}$ **calls] [**$_{I}$ this logic**] [**$_{AU}$ " the ethic of the pie " or " hydraulic model "**]** .
12) However , there could hardly be a better example of **[**$_{AN}$ what**] [**$_{A}$ Craib (1992 : 12) **] [**$_{M}$ **calls]** the **[**$_{AU}$ " logic trap " **]** of **[**$_{I}$ dismissing a substantive theory by purely rationalistic means **] .**
13) **[**$_{A}$ He **]** uses **[**$_{I}$ a technique **] [**$_{M}$ **called] [**$_{AU}$ frame flipping **] [**$_{I}$ in which a series of still images are changed slightly each time a frame is taken **]** .
14) The Jovian magnetic field exerts **[**$_{I}$ an influence out to near a surface **] [**$_{M}$ **called] [**$_{AU}$ the " magnetopause " **]** .
15) **[**$_{I}$ Ideas **]** (or as **[**$_{A}$ Dawkins **] [**$_{M}$ **calls] [**$_{AN}$ them **]** , **[**$_{AU}$ memes **]** (Lynch , 1996)) can flow through time ....
16) **[**$_{A}$ I **] [**$_{M}$ **use the term] [**$_{AU}$ " realize "] **[**$_{I}$ here in a deliberate double sense **]** .

Table F. Selected examples of predicative structure for EMOs.
Key: [$_{A}$ AGENT] [$_{M}$ MARKERS-OPERATORS] [$_{I}$ INFORMATIONAL SEGMENTS] [$_{AU}$ AUTONYM] [$_{AN}$ ANAPHORIC ELEMENT]

IV.4.3.2.2 Linguistic realization of informational segments in EMOs

As mentioned earlier (Section 2.1.3), informational segments present many realizations that distance them from the clarity, completeness and conciseness of lexicographic entries. In fact, they may show up as full-fledged clauses (17), as inter- or intra-sentential anaphoric elements (18 and 19, the first one a relative clause), supply a categorization descriptor (20), or even restrict themselves semantically to what we could call a sententially-unrealized "existential variable" (with logical form ∃x) indicating only that certain discourse entity is being introduced (21):

17) In 1965 the term soliton was coined **[**$_{I}$ to describe waves with this remarkable behaviour**]** .

18) This leap brings cultural citizenship in line with **[**$_{I}$ what **]** has been called the politics of citizenship.

19) **[**$_{I}$ They **]** are called "endothermic compounds."

20) One of the most enduring aspects of all social theories are **[**$_{I}$ those conceptual entities **]** known as structures or groups.

21) A **[**$_{I\ \exists x}$ **]** so called cell-type-specific TF can be used by closely related cells, e.g., in erythrocytes and megakaryocytes.

To complement the careful analysis of how the metalinguistic verbs used in corpus create a link between the autonym and the information provided for it, we resorted to the FrameNet (FN) project data[58] to exploit knowledge about the conceptual structure each of these Metalinguistic verbs represent. The FrameNet data shows two relevant frames for metalinguistic predication: the *Name_conferral* and *Name_bearing* frames, which include most of the lexical items in our extraction patterns. To create the processing routes we used corpus analysis and the FN data to establish the following: which arguments (with their required roles to populate our MID templates) adopted which linguistic realization and valency patterns, which arguments were needed, and which were either optional or irrelevant. For example, the verb "call" in its *Name_bearing* frame requires two core elements, *Entity* and *Name*, which are realized through combinations of NP, pronouns, prepositional phrases and quoted material (QUO, in FN terminology). *Name* corresponds to the EMO autonym and *Entity* to the informational segment as a referent. FN also assigns peripheral roles to *Name_source* and *Speaker* that could be retrieved as agents and other roles in the analysis performed by MOP. A marked-up example of this verb's semantic frame in FN, using our customized labels, is:

22) **[**$_{A}$ I **]** **[**$_{M}$ CALL **]** **[**$_{I}$ it **]** **[**$_{AU}$ curry cheese **]** as it is very heavily flavoured with cumin seeds .

A variation of this frame could present a different ordering of elements:

23) **[**$_{A}$ The General **]** would **[**$_{M}$ CALL **]** **[**$_{AU}$ "collateral damage" **]** **[**$_{I}$ the many civilian casualties resulting from the attack **]**.

Other patterns will present different linguistic realizations for an informational segment or no surface realization at all, as we will see in the next sections. For example, the pattern "defined as" might realize the information supplied for the autonym as a full clause or clauses, as a complex noun phrase, or as a pronoun.

Even though the processing labels express the predicted directionality of the main arguments of the metalinguistic predication, in effect assuming that an EMO component will be realized to the right or left of the marker-operators, other modules conduct tests that might reverse that initial assumption, and could retrieve the pronoun "it" in example 22 even though the pattern has a preference for looking backward from the marker to find an informational segment. Thus, in our predicate processing stage default assumptions can be defeated by specific tests or

conditions in order to locate the best candidate segment in a sentence for the structural roles we have identified as relevant. If an autonym has been already located in the expected position of an informational segment, the directionality for the latter's search could again be reversed, subjected to structural and lexical constraints.

As example sentence 22 illustrates, retrieval of informational segments sometimes encounters entities that either are realized fully elsewhere in text or have not been introduced yet in discourse. In a robust and fully operational system, these anaphors would have to be resolved in order to provide complete semantic information, either with a previously mentioned referent or through a full description of that entity. Example sentences 18, 19 and 21 will display in the MOP output only as unresolved surface element or as existential variable place-holders. The problem of anaphora resolution in discourse is beyond the scope of this dissertation, and our demonstration system will be restricted to identification of the anaphoric element in the predication. We have extended this treatment to cases where relative clauses use a WH-pronoun that serves as a placeholder for another referent (sentence 24, from our sociology corpus), even in cases where they can be resolved within the sentence (sentence 25, modified from the same corpus):

24) This leap brings cultural citizenship in line with **[**$_{\mathbf{I}}$ what **]** has been **[**$_{\mathbf{M}}$ called **]** **[**$_{\mathbf{AU}}$ the politics of citizenship **]** ..

25) **[**$_{\mathbf{I}}$ This change in the Nissl substance **]** , or **[**$_{\mathbf{I}}$ what **]** **[**$_{\mathbf{A}}$ he **]** **[**$_{\mathbf{M}}$ termed **]** **[**$_{\mathbf{AU}}$ chromatolysis **]**, is ….

In (25), even though the segment *This change in the Nissl substance* is more informative that the WH-pronoun that substitutes for it inside the apposition, the MOP system will select the first candidate (the pronoun), and forgo further exploration. Integration of a third-party module for anaphora resolution would solve this problem, but as it is implemented now the MOP system works one sentence at a time and cannot retrieve precedent or subsequent sentences for resolution. Also, a search for such a module written in Python (to preserve code homogeneity) has not presented good candidates. We believe such limitations can be easily overcome in the future without unreasonable effort. If the source document and its lines are properly indexed and retrievable, an anaphora resolution algorithm can be applied after MOP processing, when a client application that encounters anaphoric elements in the informational slot of a MID requires extraction of more complete information about an autonym. Again, we have opted to maintain complexity of the system to the minimum.

---

[58] http://www.icsi.berkeley.edu/~framenet/

Once the limited semantic labeling has taken place, the informational segment selection algorithm will search for the best candidate for that role. In general, it will prefer anaphoric lexical items or noun phrases, but will perform structural and lexical tests to guide both the directionality of the search as well as the final selection. Before selecting a noun phrase, the system will attempt to maximize it by looking for its head noun and all relevant modifiers and complements with an NP processing module. If the algorithm does not find a suitable candidate within its constraints, it will widen its search to include, for example, items labeled as Agent or to isolated determiners that could be playing a subject role in the predication. One pattern ("so-called") contemplates an option for unexpressed arguments, which we have called an "existential variable" for a referent that can be retrieved contextually in discourse.

For certain patterns (for example, "defined as"), the default selection and linguistic tests involves full clauses, instead of nominal sentence constituents. Some of those linguistic tests include limited discourse processing of the sentences, for items such as coordinating conjunctions, appositions or discourse markers. Depending on the pattern-specific restrictions, these tests examine semantic labels, POS tags, word forms, or a combination of them.

Some of the marker patterns that have a default "backward directionality" in search of informational segments are: *known as, call, coins, coin, denoted, labels, labeled, named, terms (verb),termed, etc.* Marker patterns that have a default "forward directionality" in search of informational segments include: *christened, coined, defined as, denote, dubbed,* quotations+*means, etc.*

Other markers, like *so-called,* that might present unexpressed referents use route D. Finally, a small set of patterns such as *dub* follow route C for default structures like [MARKER] [INFO] [AUTO]. Table G presents the algorithms for the processing routes, with a few bolded example sentences included for clarity. **A** stands for autonym, **IS** for informational segment, **Anaph** for anaphoric item and **N-Chunk** for a noun phrase group.

B ROUTE	➲ If markers: *use (the|a) [descriptor]+quotes* or *term+quotes*: • If A in predicted position, look for IS forward of A o Test for possible clause and restrict it by punctuation and discourse markers **→ People use the word " stress " to refer to both the external pressures and demands they are subject to , and the effects that such stressful circumstances have on their performance , feelings and health .** o If not a clause, select first available N-Chunk or Anaph • If no A in predicted position, or marker at end of sentence, reverse search direction for first N-Chunk or Anaph ➲ If NOT *use (the|a) [descriptor]+quotes* or *term+quotes*: • If preceding item is coordinating conjunction, look backward for head noun of previous phrase, instead of first available chunk • If there is a suitable Anaph or N-Chunk candidate between the marker and the forward A, select it.

	→ **"Rosaldo calls this logic " the ethic of the pie " or " hydraulic model " .** o If an apposition is found just before the marker, jump it before selecting a candidate → **Integral power results in a fundamental type of social classification which , adapting Bernstein 's terminology , I shall call ' frame ' (Bernstein 1971) .** • If NP candidate is found, try to maximize noun phrase by finding N-head, PP-attachment resolution, relative clauses and other NP processing rules • If no suitable NP candidate found, search for best Anaph available • If no candidate found, reverse direction of search: o Check for intervening IS o Check for anaphora forward of autonym **O** If all else fails, select any item preceding the marker
F ROUTE	**O** If markers: ' defined as ',' coined ', 'means QUOT': • Do full clause search forward o Check for discourse markers backwards to restrict scope of forward clause → **If " suicide " is defined as the doing of a positive act with the intention of ending life , then ....** o If no clause restriction, use whole segment → **Ergonomics can be defined as systems design with the attributes of people as the frame of reference .** **O** If other markers, search forward for best N-Chunk or Anaph • If A identified BEFORE marker, maintain normal forward search • If A identified AFTER marker o test for IS between marker and A o test for Anaph o select first candidate after A • If no suitable candidate forward, reverse search direction
D ROUTE:	**O** If suitable candidate preceding marker: • Test for apposition, jump it and select best candidate after NP maximization • If no apposition, select best candidate after NP maximization **O** If no suitable candidate preceding marker, use existential variable
C ROUTE:	**O** Search forward if candidate before A → **The initial response of the Left and liberals to this reaction was to dub the conventional wisdom " moral panic "** **O** Reverse search if no intervening info

Table G. Algorithms for the pattern-specific processing routes

In our walk-through example, after a performing prepositional-phrase attachment over the two initial chunks, box G shows the Autonym and Informational segment chunks selected by the application using the B route:

Informational segment: [['The'/'DT'@[0w], 'bit'/'NN'@[1w], 'sequences'/'NNS'@[2w], 'representing'/'VBG'@[3w], 'quanta'/'NN'@[4w], 'of'/'OF'@[5w], 'knowledge'/'NN'@[6w]], 'INFO']

Autonym: [('NP': '"'/'QUOT' 'genes'/'NNS' '"'/'QUOT')@[19w:22w], 'AUTO']

Box G: Selected chunks for Autonym and Informational segment roles

The creation of the Metalinguistic Information Database is done at a final stage, called in IE terminology the template or scenario generation phase. All sentence segments associated with a required role in the EMO predication are extracted and inserted into the relevant slots of the output xml file. For unexpressed entities, a placeholder ("------Existential_Variable-------") is inserted in the required slot.

The following example sentences and their correct database entries are presented next:

26) Here we report the discovery of a soluble decoy receptor, termed decoy receptor 3 (DcR3) ...

Reference:	MedLine sample # 6
Autonym:	decoy receptor 3 (DcR3)
Information	a soluble decoy receptor
Markers/ Operators:	*Termed*

27) The so-called " neo-liberals " who detect Marxist bias in contemporary class analysis have

Reference:	Sociology sample # 3
Autonym:	neo-liberals
Information	------Existential_Variable-------
Markers/ Operators:	*so-called*

28) CFU-S denotes the pluripotent cell in mouse , and forms …

Reference:	Histology sample # 23
Autonym:	CFU-S
Information	the pluripotent cell in mouse
Markers/ Operators:	*denotes*

29) This leap brings cultural citizenship in line with what has been called the politics of citizenship .

Reference:	Sociology sample # 33
Autonym:	the politics of citizenship
Information	what
Markers/ Operators:	*called*

30) Durkheim explicitly and unambiguously advocates class , defined as individual economic inequality--as differential income and private property .

Reference:	Sociology sample # 62
Autonym:	class
Information	individual economic inequality--as differential income and private property
Markers/ Operators:	*defined as*

31) The initial response of the Left and liberals to this reaction was to dub the conventional wisdom " moral panic " .

Reference:	Sociology sample # 19
Autonym:	moral panic
Information	the conventional wisdom
Markers/ Operators:	*dub*

Finally, we present in box H the final xml MID entry for our walk-through example sentence, including records for markers/operators and

```
<OME text="all-soc-numberedJP_Cleaned" n="71">
<Autonym>  kenes </Autonym>
<Info>The bit sequences representing quanta of knowledge</Info>
<Operator> called </Operator>
<line> The bit sequences representing quanta of knowledge will be called " kenes " , a neologism intentionally similar to " genes " .</line>
</OME>
```

Box H: Example sentence MID entry

We have provided a detailed description of the Metalinguistic Operation Processor, as well as a brief example of its operation. The next chapter presents the overall evaluation of the system made with test runs over three different corpora of special domain texts.

V SYSTEM EVALUATIONS OF THE METALINGUISTIC OPERATION PROCESSOR (MOP)

Summary

The evaluation of the MOP system follows standard IE and IR metrics like Precision, Recall and F-measure. We use as Golden standards three corpora of different characteristics from different domains that have been marked-up manually: the Sociology corpus described in previous chapters, an online Histology textbook and a sample of the MedLine abstract database of Bio-Medical papers.

On the Candidate Extraction phase, and using all EMOs identified in our test corpora, whether their distinctive marker/operators pattern were included or not in the extraction list, MOP gave excellent precision rates (P) and but low recall rates (R) in our evaluation runs. We believe a non-exhaustive list of extraction patterns is the cause of such difference in metrics. Using only tagged examples that contained patterns that could actually be recognized, precision was maintained at high levels (0.94 and more), but recall increased substantially (0.79 and more), with F-measures at 0.87 at β of 1 to balance out P & R.

Tests with learning algorithms trained on a subset of the EMO Corpus yielded good metrics also, but were otherwise inconclusive with regard to algorithm and feature set baseline superiorities. With the Sociology corpus, Maximum Entropy using as features a single word form to left and right of markers/operators presented the following numbers: P = 0.9, R = 0.7, while best results for the Histology corpus were attained with a Bayesian network using 3 word forms at each side of the markers (P=0.9, R=0.84). In short, although results were very good for this classification task, they were inconsistent and inconclusive, and so will merit further research. The expectation of improved performance with POS contexts and a wider context did not bore out.

Evaluations of the Predicate Processing task were also very good if compared with similar IE tasks, although these comparisons should be done cautiously for a number of reasons. The global performance of the system ranged from near 0.7 to a perfect score in the autonym identification task in the small MedLine Sample. The best overall F-measure was achieved both in the MedLine sample and in the Sociology corpus, at 0.77. The two different grammatical contexts where entity identification was performed gave also a wide range of metrics. Autonym identification was the most successful (averaging 0.9 P and 0.91 R), while Informational segments averaged 0.8.5 P and 0.8 R. Overall, these are very good numbers (regardless of domain involved), albeit for a very simple extraction with few database slots and a very basic structure (with almost no inference mechanisms). Improvements on coverage and precision are possible, but the system will need to increase its processing complexity significantly, adding deeper parsing and a coreference and anaphora resolution module.

V.1 Evaluation methods and metrics

V.1.1 Comparative evaluation of filtering strategies

Information Extraction and Retrieval tasks use a set of standard metrics for evaluation, which we have adapted for our tests on the MOP system. Recall (R) measures how much of the relevant information in the text (its coverage) has been extracted by the system.

$$\text{Recall} = \frac{\text{Correct answers obtained by the system}}{\text{Correct answers in the text}}$$

Precision (P), or accuracy, shows how well the system performed, in terms of how many of its answers were actually correct.

$$\text{Precision} = \frac{\text{Correct answers obtained by the system}}{\text{Total answers obtained by the system}}$$

Generally speaking, there is always a trade-off involving these two metrics, since increasing the coverage of the system usually results in a greater number of spurious answers. Given our non-exhaustive coverage of potential metalinguistic patterns, this issue is significant in our experiments. A combined measure called F-measure exists that attempts to balance out Precision and Recall ratios using a parameter called β. A β factor of 1 gives P and R equal weight, while a β parameter of less than 1 favors Precision and a greater than 1 favors Recall.

$$\text{F-measure} = \frac{(\beta^2 + 1)\,\text{PR}}{\beta^2\,2\text{P} + \text{R}}$$

Using these metrics, our evaluations of the MOP system are based on test runs over 3 document sets. These files can be viewed in the enclosed CD-ROM:

a) Our original exploratory corpus of sociology research papers, with 5581 sentences and 243 EMOs;

b) An online Histology textbook with 5146 sentences and 69 EMOs;

c) A small sample from the MedLine abstract database, with 1403 sentences and 10 located EMOs.[59]

V.1.1.1 Metrics for collocation-based filtering

Hand-coded collocation rules gave surprisingly adequate precision and recall rates of, respectively, 0.94 and 0.57 for the sociology corpus, 0.9 P and 0.5 R for the Histology manual,

[59] For the MedLine corpus, we did not create a golden standard with all possible EMO patterns, so its Recall at this stage was not measured.

and 0.9 and 0.68 for the more extensive BNC-based EMO corpus. Precision in other corpora, where the total existing numbers of metalinguistic sentences has not been estimated, range from 0.93 (for a sample of MedLine abstracts) to 0.74 for the Susanne corpus[60] (which includes portions of oral text). As we have emphasized before, these low recall numbers reflect the fact that we only selected a subset of the most reliable and common metalinguistic patterns, and our list is by no means exhaustive.

We also tested the extraction process with a run that evaluated against a golden standard where sentences that had patterns that our list wasn't designed to retrieve were removed, such as appositional structures or reformulation cases like:

32) But they tended to see them ethnocentrically, as pre-figuring "pure", i.e. Western capitalistic, market relations.

These runs gave a more realistic picture of how the extraction system was working for the actual dataset it was designed to consider. For the sociology corpus, and a β of 1, P was 0.97 and R 0.79, with an F-measure of 0.87, while for the Histology one, P was measured at 0.94, R. at 0.81 and F-measure at 0.87. Table H presents the results for these extraction runs.

Corpus	Lines extracted	Lines filtered (%)	Precision	Recall	F-Measure
Sociology	143	14 (9.8)	0.97	0.79	0.87
Histology	37	5 (13.5)	0.94	0.81	0.87

Table H. Metrics for collocation-based filtering

V.1.1.1.1 Tests for a single lemma pattern cluster

We also decided to zoom in on a more limited subset of verb forms for extraction (namely: *calls, called, call*) which presented ratios of metalinguistic relevance in our attested corpora, ranging from 100% positives (for the pattern *so called* + quotation marks) to 77% (*called*, by itself) to 31% (*call*). When restricted to these verbs and to our sociology corpus, our metrics shown precision and recall rates of 0.97, and an overall F-measure of 0.97. That is, of 5581 sentences (96 of which were metalinguistic sentences signaled by our cluster of verbs), 83 were extracted, with 13 (or 15.6% of candidates) filtered-out by collocations.

Next, we experimented with classifiers trained on examples from our MOP corpus that contained only patterns based on that lemma. The results for these restricted pattern set are shown in Table I. In this and subsequent tables, GISMax refers to a Maximum Entropy classifier trained using Generalized Iterative Scaling, while IISMax denotes a classifier that uses

[60] Surface and Underlying Structural Analysis of Naturalistic English, available online at the Oxford Text Archive.

the Improved Iterative Scaling algorithm. NB indicates Bayesian network classifiers. Accuracy reported in the fifth column uses NLTK-encoded evaluations of the trained classifiers that employ a small fragment of the MOP corpus training data.

Type	Positions Left/Right	Tags/ Words	Features	Accuracy	Extracted (%)	Filtered (%)	Precision	Recall	Extracted Pattern distribution
GISMax	1	W	1254	0.97	84 (1.51)	12 (14.29)	0.96	0.98	call -> 19 calls -> 11 called -> 54
IISMax	1	T	136	0.95	81 (1.45)	15 (18.52)	0.96	0.94	call -> 16 calls -> 11 called -> 54
IISMax	1	W	1252	0.92	77 (1.38)	19 (24.68)	0.97	0.9	call -> 14 calls -> 10 called -> 53
GISMax	1	T	138	0.91	89 (1.59)	7 (7.87)	0.9	0.96	call -> 18 calls -> 13 called -> 58
GISMax	2	T	796	0.88	82 (1.47)	14 (17.07)	0.93	0.92	call -> 16 calls -> 13 called -> 53
IISMax	2	T	794	0.86	78 (1.40)	18 (23.08)	0.95	0.89	call -> 16 calls -> 12 called -> 50
IISMax	3	W	4290	0.87	95 (1.70)	1 (1.05)	0.85	0.98	call -> 21 calls -> 15 called -> 59
GISMax	3	W	4292	0.87	95 (1.70)	1 (1.05)	0.85	0.98	call -> 21 calls -> 15 called -> 59
IISMax	2	W	3186	0.86	91 (1.63)	5 (5.49)	0.87	0.95	call -> 20 calls -> 15 called -> 56
GISMax	2	W	3188	0.86	91 (1.63)	5 (5.49)	0.87	0.95	call -> 20 calls -> 15 called -> 56
NB	1	T	136	0.88	72 (1.29)	24 (33.33)	0.97	0.84	call -> 6 calls -> 11 called -> 55
NB	2	T	794	0.87	73 (1.31)	23 (31.51)	0.96	0.84	call -> 7 calls -> 10 called -> 56
IISMax	3	T	1910	0.82	82 (1.47)	14 (17.07)	0.89	0.88	call -> 12 calls -> 14 called -> 56
GISMax	3	T	1912	0.82	82 (1.47)	14 (17.07)	0.89	0.88	call -> 12 calls -> 14 called -> 56
NB	1	W	1252	0.85	69 (1.24)	27 (39.13)	0.97	0.81	call -> 3 calls -> 11 called -> 55
NB	3	T	1910	0.8	75 (1.34)	21 (28.00)	0.91	0.82	call -> 5 calls -> 14 called -> 56
NB	2	W	3186	0.74	73 (1.31)	23 (31.51)	0.88	0.77	calls -> 15 called -> 58
NB	3	W	4290	0.73	74 (1.33)	22 (29.73)	0.86	0.77	calls -> 15 called -> 59

Table I. Metrics for restricted lemma test runs

The best classifier, GISMax with one word form before and after, gave a F-Measure of 0.97 with β factor set at 1, a number commensurable with our collocation-based test runs. Maximum Entropy algorithms trained with Improved Iterative Scaling (IISMax) and Generalized Iterative Scaling (GISMax) over a single item surrounding the marker (either POS tags or the orthographic form of the word), provided the best accuracy rates and F-measures, reflecting good precision and recall ratios. These compare very well with the accuracy (78.8%) reported in Maximum Entropy classification experiments using word counts in Nigam et al (1999), or the systems evaluated by Yang (1997), as well as the superiority of the technique over naive Bayes for certain tasks. We have to take into account, though, that classification over a bigger set of labels than our simple "YES" or "NO" options is an altogether different problem that defies direct comparison.

The number of different features used by the stochastic classifiers proved not to be good accuracy predictors, although, as shown in the extracted pattern distribution, *call* as triggering verb suffered from its low relevance in the training corpus and was sometimes filtered out completely. Although some of the naive Bayes classifiers showed remarkable good precision rates with one and two positions, their recall metrics were unimpressive. Maximum Entropy

models that used 3 positions to the left and right of markers maintained good recall, but lowered their precision and overall accuracy.

V.1.1.1.2 Tests for all relevant patterns used in the extraction phase

Next, we tested the classification algorithms using the full set of patterns compiled from the EMO corpus for their reliability and ease of disambiguation. Again, after pattern-matching and feature generation, the filtering code accepted the candidate sentences that were labeled as EMOs, and rejected those indecisive or labeled as non-EMO. Once an extraction file was completed, it was compared against a golden standard that contained only those sentences susceptible to extraction by our complete pattern set. F-measure was again calculated using a β parameter of 1. Tables J and K present metrics in the stochastic filtering experiments for the complete set of patterns used in the collocation approach, over two of our evaluation corpora. Figures 5 and 6 present a comparison of best results for each classifier algorithm over these corpora. The training dataset now included all sentences with patterns related to the 44 used in the hand-coded rules. The complete evaluation metrics for all classifiers can be found in the enclosed CD-ROM.

Although our test runs using only collocations showed initially that structural regularities would perform well, both with our restricted lemma cluster and with our wider set of markers, our expectations about improvement with more features (more positions to the right of left of the markers) or a more controlled and grammatically restricted environment (a finite set of surrounding POS tags), turned out to be overly optimistic. Nevertheless stochastic approaches that used short range features did perform very well, in line with the hand-coded approach.

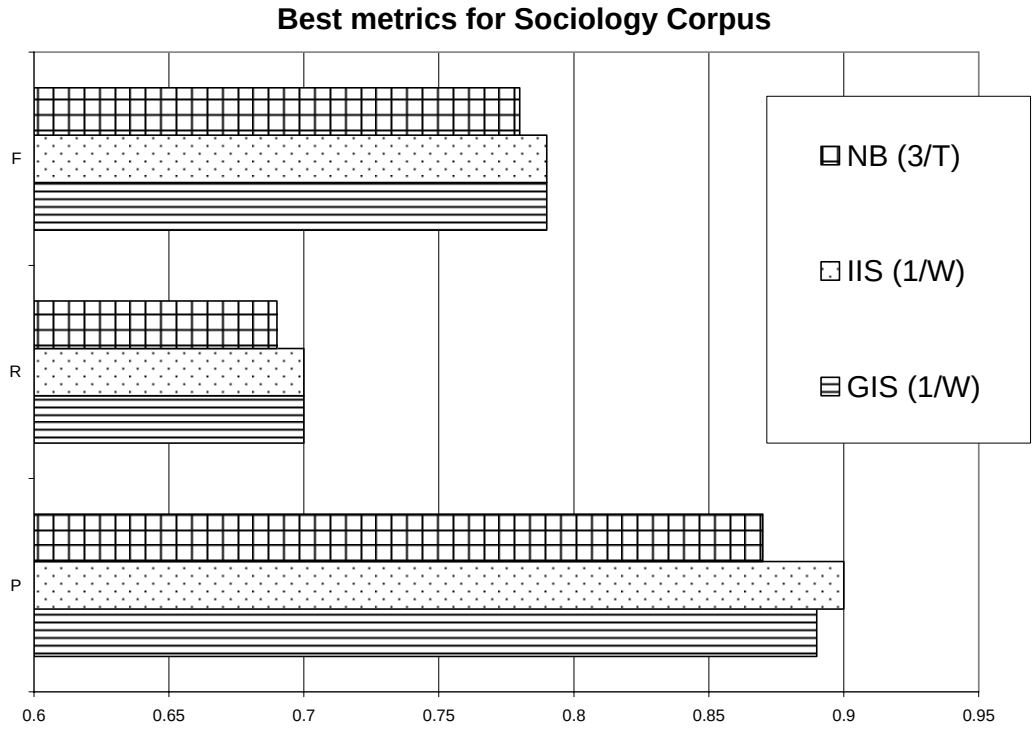
In short, both Knowledge-Engineering and supervised learning approaches are adequate for initial extraction of metalinguistic sentences, although the advantages of learning algorithms over hand-crafted rules are obvious: they allow easier and more accurate transport of systems to new thematic domains. The best results overall were obtained by using the Maximum Entropy algorithm trained with Generalized Iterative Scaling.

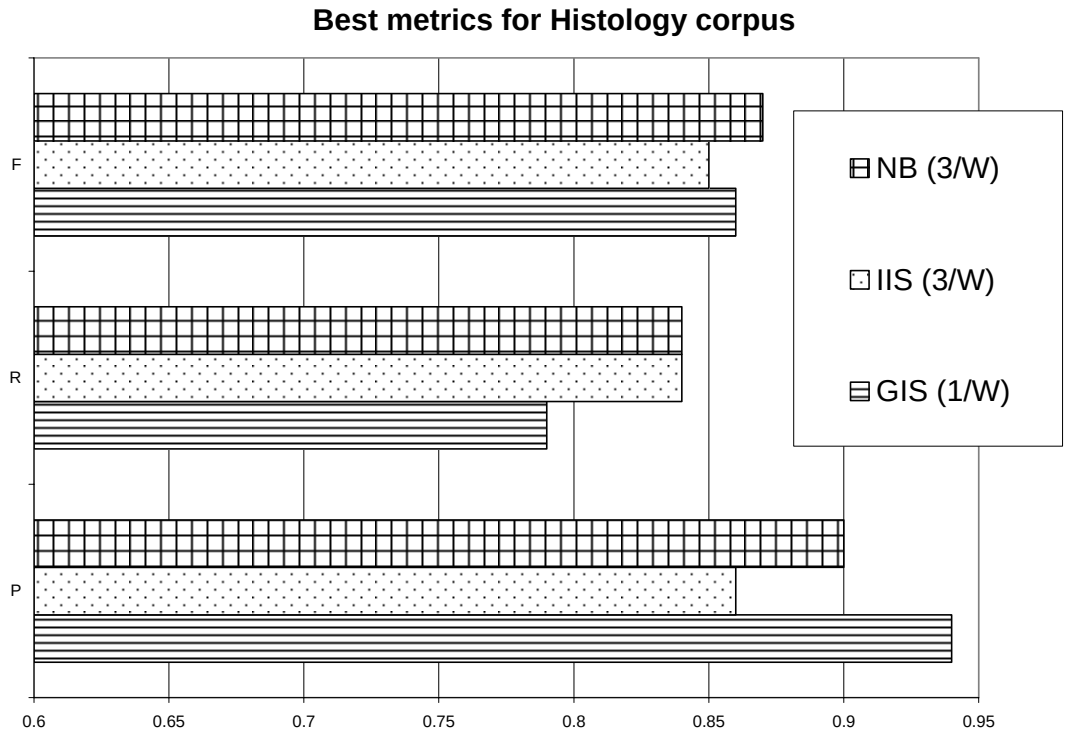
TYPE:	POSITIONS (TAGS/ WORDS)	# OF FEATURES	ACC.	EXTRACTED (%)	FILTERED (%)	P	R	F-M	PATTERN DISTRIBUTION (PATTERN,# FOUND)
IIS	1 (W)	4450	0.81	136 (2.44)	21 (15.44)	0.90	0.70	0.79	(‘called’, 40), (‘defined as’, 6), (‘call’, 19), (‘known as’, 13), (‘calls’, 10), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 6), ('So(-|)called’, 10), (‘QUOTs refers to’, 4)
GIS	1 (W)	4452	0.86	138 (2.47)	19 (13.77)	0.89	0.70	0.79	(‘called’, 40), (‘defined as’, 6), (‘call’, 19), (‘known as’, 13), (‘calls’, 10), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 5), ('So(-|)called’, 11), (‘QUOTs refers to’, 4)
IIS	2 (T)	2012	0.76	132 (2.37)	23 (17.42)	0.90	0.68	0.78	(‘called’, 40), (‘defined as’, 5), (‘call’, 17), (‘known as’, 13), (‘calls’, 10), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 5), ('So(-|)called’, 11), (‘QUOTs refers to’, 3)
GIS	2 (T)	2014	0.83	137 (2.46)	18 (13.14)	0.88	0.69	0.78	(‘called’, 41), (‘defined as’, 5), (‘call’, 17), (‘known as’, 13), (‘calls’, 12), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 5), ('So(-|)called’, 11), (‘QUOTs refers to’, 4)
Nai	3 (T)	6562	0.82	138 (2.47)	19 (13.77)	0.87	0.69	0.77	(‘called’, 42), (‘defined as’, 6), (‘call’, 14), (‘known as’, 13), (‘calls’, 11), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 5), ('So(-|)called’, 13), (‘QUOTs refers to’, 4)
IIS	3 (W)	17554	0.87	155 (2.78)	2 (1.29)	0.82	0.73	0.77	(‘called’, 44), (‘defined as’, 6), (‘call’, 24), (‘known as’, 13), (‘calls’, 14), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 6), (‘QUOTs means’, 2), ('So(-|)called’, 13), (‘QUOTs refers to’, 4)
GIS	3 (W)	17556	0.87	155 (2.78)	2 (1.29)	0.82	0.73	0.77	(‘called’, 44), (‘defined as’, 6), (‘call’, 24), (‘known as’, 13), (‘calls’, 14), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 6), (‘QUOTs means’, 2), ('So(-|)called’, 13), (‘QUOTs refers to’, 4)
IIS	1 (T)	236	0.76	132 (2.37)	25 (18.94)	0.89	0.67	0.77	(‘called’, 42), (‘defined as’, 6), (‘call’, 14), (‘known as’, 13), (‘calls’, 10), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 5), ('So(-|)called’, 10), (‘QUOTs refers to’, 4)
Nai	2 (T)	2012	0.82	135 (2.42)	22 (16.30)	0.87	0.67	0.76	(‘called’, 42), (‘defined as’, 5), (‘call’, 12), (‘known as’, 13), (‘calls’, 12), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 6), ('So(-|)called’, 12), (‘QUOTs refers to’, 4)
GIS	1 (T)	238	0.82	139 (2.49)	18 (12.95)	0.86	0.69	0.76	(‘called’, 43), (‘defined as’, 6), (‘call’, 15), (‘known as’, 13), (‘calls’, 12), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 6), ('So(-|)called’, 11), (‘QUOTs refers to’, 4)
IIS	2 (W)	13226	0.82	144 (2.58)	11 (7.64)	0.83	0.69	0.75	(‘called’, 40), (‘defined as’, 6), (‘call’, 22), (‘known as’, 13), (‘calls’, 14), (‘term QUOTs', 19), (‘us(e|es) (a|the) (*) QUOTs', 5), (‘QUOTs means’, 2), ('So(-|)called’, 12), (‘QUOTs refers to’, 4)
GIS	3 (T)	6564	0.76	129 (2.31)	22 (17.05)	0.88	0.65	0.74	(‘called’, 39), (‘defined as’, 5), (‘call’, 17), (‘known as’, 13), (‘calls’, 9), (‘term QUOTs', 19), (‘us(e|es) (a|the) (*) QUOTs', 5), ('So(-|)called’, 10), (‘QUOTs refers to’, 4)
GIS	2 (W)	13228	0.81	141 (2.53)	14 (9.93)	0.83	0.67	0.74	(‘called’, 41), (‘defined as’, 6), (‘call’, 22), (‘known as’, 13), (‘calls’, 14), (‘term QUOTs', 16), (‘us(e|es) (a|the) (*) QUOTs', 5), (‘QUOTs means’, 2), ('So(-|)called’, 12), (‘QUOTs refers to’, 4)
Nai	1 (T)	236	0.77	132 (2.37)	25 (18.94)	0.86	0.65	0.74	(‘called’, 43), (‘defined as’, 6), (‘call’, 8), (‘known as’, 13), (‘calls’, 12), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 6), ('So(-|)called’, 12), (‘QUOTs refers to’, 4)
Nai	1 (W)	4450	0.78	127 (2.28)	30 (23.62)	0.88	0.64	0.74	(‘called’, 40), (‘defined as’, 6), (‘call’, 10), (‘known as’, 13), (‘calls’, 10), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 5), ('So(-|)called’, 12), (‘QUOTs refers to’, 4)
IIS	3 (T)	6562	0.76	125 (2.24)	26 (20.80)	0.87	0.62	0.73	(‘called’, 38), (‘defined as’, 4), (‘call’, 17), (‘known as’, 13), (‘calls’, 9), (‘term QUOTs', 19), (‘us(e|es) (a|the) (*) QUOTs', 4), ('So(-|)called’, 10), (‘QUOTs refers to’, 3)
Nai	2 (W)	13226	0.81	135 (2.42)	22 (16.30)	0.83	0.64	0.72	(‘called’, 43), (‘defined as’, 6), (‘call’, 8), (‘known as’, 13), (‘calls’, 14), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 4), (‘QUOTs means’, 2), ('So(-|)called’, 13), (‘QUOTs refers to’, 4)
Nai	3 (W)	17554	0.78	130 (2.33)	27 (20.77)	0.82	0.61	0.70	(‘called’, 44), (‘defined as’, 6), (‘known as’, 13), (‘calls’, 14), (‘term QUOTs', 21), (‘us(e|es) (a|the) (*) QUOTs', 6), (‘QUOTs means’, 2), ('So(-|)called’, 13), (‘QUOTs refers to’, 4)

Table J. Stochastic filtering for all patterns in Sociology corpus (sorted by F-Measure)

TYPE:	POSITIONS (TAGS/ WORDS)	# OF FEATURES	ACC.	EXTRACTED (%)	FILTERED (%)	P	R	F-M	PATTERN DISTRIBUTION (PATTERN,# FOUND)
Nai	3 (W)	17554	0.78	40 (0.78)	2 (5.00)	0.90	0.84	0.87	(' called ', 19), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 6)
GIS	1 (W)	4452	0.86	36 (0.70)	6 (16.67)	0.94	0.79	0.86	(' called ', 19), (' termed ', 10), ('So(-|)called ', 4)
Nai	3 (T)	6562	0.82	39 (0.76)	3 (7.69)	0.90	0.81	0.85	(' called ', 19), (' known as ', 2), (' termed ', 10), ('So(-|)called ', 5)
IIS	3 (W)	17554	0.87	42 (0.82)	0 (0.00)	0.86	0.84	0.85	(' called ', 19), (' call ', 2), (' known as ', 2), (' termed ', 10), ('So(-|)called ', 6)
GIS	3 (W)	17556	0.87	42 (0.82)	0 (0.00)	0.86	0.84	0.85	(' called ', 19), (' call ', 2), (' known as ', 2), (' termed ', 10), ('So(-|)called ', 6)
Nai	2 (W)	13226	0.81	39 (0.76)	3 (7.69)	0.90	0.81	0.85	(' called ', 19), (' known as ', 2), (' termed ', 10), ('So(-|)called ', 5)
GIS	1 (T)	238	0.82	39 (0.76)	3 (7.69)	0.90	0.81	0.85	(' called ', 18), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 5)
Nai	1 (W)	4450	0.78	37 (0.72)	5 (13.51)	0.92	0.79	0.85	(' called ', 19), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 4)
Nai	2 (T)	2012	0.82	38 (0.74)	4 (10.53)	0.89	0.79	0.84	(' called ', 19), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 4)
IIS	2 (W)	13226	0.82	40 (0.78)	2 (5.00)	0.88	0.81	0.84	(' called ', 19), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 5)
GIS	2 (W)	13228	0.81	40 (0.78)	2 (5.00)	0.88	0.81	0.84	(' called ', 19), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 5)
Nai	1 (T)	236	0.77	38 (0.74)	4 (10.53)	0.89	0.79	0.84	(' called ', 19), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 4)
IIS	1 (T)	236	0.76	34 (0.66)	8 (23.53)	0.94	0.74	0.83	(' called ', 18), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 2)
IIS	1 (W)	4450	0.81	34 (0.66)	8 (23.53)	0.94	0.74	0.83	(' called ', 19), (' termed ', 10), ('So(-|)called ', 2)
GIS	2 (T)	2014	0.83	37 (0.72)	5 (13.51)	0.89	0.77	0.82	(' called ', 17), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 5)
GIS	3 (T)	6564	0.76	37 (0.72)	4 (10.81)	0.86	0.74	0.80	(' called ', 17), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 4)
IIS	2 (T)	2012	0.76	33 (0.64)	9 (27.27)	0.91	0.70	0.79	(' called ', 17), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 3)
IIS	3 (T)	6562	0.76	35 (0.68)	6 (17.14)	0.86	0.70	0.77	(' called ', 15), (' known as ', 2), (' termed ', 10), ('So(-|) called ', 4)

Table K. Stochastic filtering for all patterns in Histology corpus (sorted by F-Measure)





Figures 5 and 6. Best results for each filtering algorithm.

Legend: P: Precision; R: Recall; F: F-Measure. NB: naive Bayes; IIS: Maximum Entropy trained with Improved Iterative Scaling; GIS: Maximum Entropy trained with Generalized Iterative Scaling. (Positions/Feature type)

V.1.1.2 Discussion of evaluation metrics

One issue that merits special attention is why some of the algorithms and features work well with one corpus, but not so well with another. This fact is in line with observations in Nigam et al. (1999) that naive Bayes and Maximum Entropy do not show fundamental baseline superiorities, but are dependent on other factors. We plan further research into our stochastic approaches to fine tune them for the task. A hybrid approach that combines hand-crafted collocations with classifiers customized to each pattern's behavior and morpho-syntactic contexts in corpora might offer better results in future experiments. As input for the rest of the processing to the MOP system, we settled on the better-understood collocation-based filtering, until further research help narrow down which factors affect classifier performance.

V.1.2 Evaluation of Predicate Processing task

V.1.2.1 Evaluation parameters

For our evaluation of system performance in this area, we have adopted the standard IE parameters of Recall, Precision and F-Measure described in the previous section. Nevertheless, we have introduced a modified measure of performance that allows qualification of partially correct answers for some of the filled slots in the resulting database structure. We considered that a single erroneous fragment in one of the MID slots filled should not completely invalidate the good overall accuracy for a whole entry. To better reflect record-wide performance in all template slots, we introduced a threshold of similarity of 65% for comparison between a golden standard slot entry and the one provided by the application. Thus, if the autonym or the informational segment is at least 2/3 of the correct response, that slot's entry is counted as a positive, in many cases leveling the field for the expected errors in the prepositional phrase- or acronym-attachment algorithms, but accounting for a (basically) correct selection of superficial sentence segments. Only if all slots are evaluated as correct, after application of the above corrections, is the whole entry in the MID counted as correct

For our test runs, we prepared golden standard answer keys. We performed a manual selection of the autonyms, informational segments and markers-operators in each corpus, and compiled a correct MID for each one, including only patterns that the MOP system was designed to find. The β factor used for calculating F-Measure was again 1. The golden standards and the output files produced by MOP, as well as the evaluation scripts, can be found in the enclosed CD-ROM.

V.1.2.2 Evaluation metrics for the system

Figure 7 shows the main results of the output MIDs processed by the MOP system using the candidate extraction files from our 3 test corpora. It shows Precision and Recall numbers for the

autonyms, the informational segments, and a global performance measure reflecting the whole record, rounded to the nearest decimal. Record numbers and global F-Measures are shown in parentheses next to the corpus name. Table L shows other relevant data for these test runs.

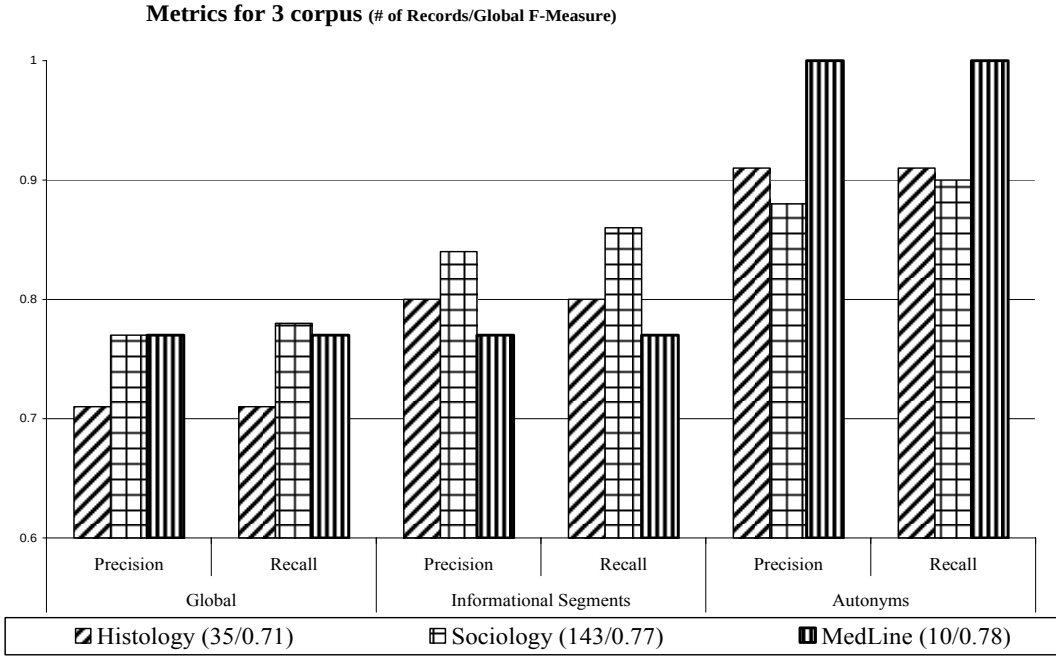


Figure 7. Comparative metrics for test runs

	Corpora:	Histology	Sociology	MedLine
Number of records in golden standard		35	140	9
Number of records in output MID		35	143	9
INFORMATIONAL SEGMENTS	Correct	28	120	7
	Percentage of total	80	85.7	77.7
	Precision	0.8	0.84	0.77
	Recall	0.8	0.85	0.77
	F-measure	0.8	0.84	0.77
AUTONYMS	Correct	32	126	9
	Percentage of total	91.4	90	100
	Precision	0.91	0.88	1.0
	Recall	0.91	0.9	1.0
	F-measure	0.91	0.9	1.0
GLOBAL	Correct	25	110	7
	Percentage of total	71.4	78.5	77.7
	Precision	0.71	0.77	0.77
	Recall	0.71	0.78	0.77
	F-measure	0.71	0.77	0.77

Table L. Global and slot-specific metrics for test runs

The performance of the system showed a wide range of variation, from near 0.7 to a perfect score in the autonym identification task with the small MedLine Sample. The best overall F-Measure was achieved both in that small corpus and in the Sociology one, at 0.77. The two different grammatical contexts where entity identification was performed (informational and autonym roles) gave also a wide range of metrics. In all, autonym identification was the most successful task, aided no doubt by the cognitive saliency we have described as a constitutive property of autonymy, which is reflected, among other things, in the delimitative function of quotation marks. The informational segments presented a more heterogeneous environment in which the system had to decide which items represented valuable information about the autonym. Our self-imposed restriction on the complexity of linguistic processing to be attempted by the MOP system prevented use of high-end parsing machinery capable of identifying complex head nouns or long-distance dependencies, as in the sentence 33:

33) The process whereby sand piles up effectively at random but sooner or later a part of it achieves the needed critical angle is called "self-organised criticality" (Bak et al, 1988).

A deeper grammatical analysis than the one attempted (partial parsing) would be able to obtain, as information about the term *self-organised criticality*, the following definite description: “The process whereby sand piles up effectively at random but sooner or later a part of it achieves the needed critical angle”.

Nevertheless, and within the discussed limitations of the system, results ranged around 0.85 for the best informational segment identification runs. The lowest parameter was obtained in our test run of the Histology corpus, with global precision and recall rates around 0.71, but with high numbers in the autonym identification task (0.91), and midrange ones for the informational segments (0.8). Across domains, we observe that, even though the Health Sciences are supposed to have a more consolidated technical vocabulary than the Social Sciences (a more stable set of naming conventions, to put it another way), results for the MedLine[61] and Histology corpus occupy the extreme positions in the spectrum, with the Sociology corpus in the middle range. The number of sentence candidates analyzed was not a good predictor of system performance, either. Table M displays the unedited MID produced from the MedLine corpus in one of our test runs, as presented by a browser with the referenced xsl file.

[61] The low number of candidate sentences in the MedLine corpus might be due to the fact that even though it consists of running text, it is an abstract database. Generally speaking, abstracts are not ideal places to discuss terminological issues, which are better addressed in the body of the article.

Output for EMO MID for file: **MedLine_samples.xml** @ Sun Mar 14 18:16:19 2004

Autonym:	AML1 or PEBP2alphaB
Info:	CBFbeta , and CBFalpha2
Reference: MedLine_samples # 1	**Markers/Operators:** known as
Text:	We used transient-transfection assays , in combination with immunofluorescence and green fluorescent protein-tagged proteins , to monitor subcellular localization of CBFbeta-SMMHC , CBFbeta , and CBFalpha2 (also known as AML1 or PEBP2alphaB) .

Autonym:	leucine zipper-like motifs
Info:	putative amphipathic alpha-helices
Reference: MedLine_samples # 2	**Markers/Operators:** termed
Text:	We determined , via a variety of hydrodynamic measurements as well as protein cross-linking , that native Gro is a tetramer in solution and that tetramerization is mediated by two putative amphipathic alpha-helices (termed leucine zipper-like motifs) found in the N-terminal region .

Autonym:	cAMP-activated protein kinase
Info:	protein kinase
Reference: MedLine_samples # 3	**Markers/Operators:** known as
Text:	Here we report that activation of Rap1 by forskolin and cAMP occurs independently of protein kinase A (also known as cAMP-activated protein kinase) .

Autonym:	the equator
Info:	a line of mirror image symmetry
Reference: MedLine_samples # 4	**Markers/Operators:** called
Text:	A dorsal/ventral boundary established by Notch controls growth and polarity in the Drosophila eye M. Dominguez and J. F. de Celis Nature 396 276-8 1998 In the Drosophila compound eye the dorsal and ventral fields of eye units (ommatidia) meet along the dorsoventral midline , forming a line of mirror image symmetry called the equator .

Autonym:	postsynaptic density-95
Info:	a protein
Reference: MedLine_samples # 5	**Markers/Operators:** known as
Text:	The NMDA receptor can bind a protein known as postsynaptic density-95 (PSD-95) , which may regulate the localization of and/or signalling by the receptor .

Autonym:	decoy receptor
Info:	the discovery of a soluble decoy receptor
Reference: MedLine_samples # 6	**Markers/Operators:** termed
Text:	Here we report the discovery of a soluble decoy receptor , termed decoy receptor 3 (DcR3) , that binds to FasL and inhibits FasL-induced apoptosis .

Autonym:	Zfp106
Info:	The H3a gene
Reference:	**Markers/Operators:** called

Text:	The H3a gene , now called Zfp106 , encodes a 1888-amino acid protein with three zinc fingers and a

Autonym:	public reason
Info:	what
Reference: MedLine_samples # 8	**Markers/Operators:** termed
Text:	We draw certain lines grounded in what Rawls has termed " public reason " beyond which we do not give effect to the autonomous self-regarding decisions of individuals .

Autonym:	neurogenic intermittent claudication
Info:	The characteristic syndrome associated with lumbar stenosis
Reference: MedLine_samples # 9	**Markers/Operators:** termed
Text:	The characteristic syndrome associated with lumbar stenosis is termed neurogenic intermittent claudication .

Table M. MID for MedLine samples

V.1.2.3 Discussion of results

V.1.2.3.1 The MOP system as an Information Extraction system

The MOP system's goals and methods correlate only partially to that of full-blown IE systems of the kind evaluated in the MUC series of conferences. The breadth of entities identified and the kinds of scenarios involved in IE are much more complex than the ones dealt with MOP. For example, in MOP no temporal markup was done, and the semantic tagging involved was very idiosyncratic, with quantities, currencies, organizations, proper names and the like not labeled specifically as such. Only a subset of the tasks could be compared with what the MOP system does. Named Entity and Template Element tasks identify strings that represent entities mentioned in text, and try to establish the relationships among them. What follows is an example of the kind of markup expected from a MUC 7 system:

> The <ENAMEX TYPE="LOCATION">U.K.</ENAMEX> satellite television broadcaster said its subscriber base grew <NUMEX TYPE="PERCENT">17.5 percent</NUMEX> during <TIMEX TYPE="DATE">the past year</TIMEX> to 5.35 million

Some IE evaluations require filling of more than 18 related slots in a database record. In our experiments only two kinds of textual entities were explored (marker-operators were identified at an earlier extraction phase): autonyms and informational segments. Even though these differences are significant, the MOP system can be considered a true (but special and much focused) kind of Information Extraction system dealing exclusively with relationships and entities involved in metalinguistic discourse. Such a system could in theory process text further

to find more entities and parameters, like terminology-modification agents, identify hedging information, and perform coreference resolution, etc. We chose to narrow down the scope of the application to present a practical proof-of-concept system that could be enhanced later on. Table N (modified from Chinchor, 1998) presents best results for the MUC conference evaluations[62] for reference, but our caveats about direct comparison should be taken into consideration when attempting to draw any conclusions about benchmarking the MOP system.

Evaluation/ Tasks	Named Entity	Coreference	Template Element	Template Relation	Scenario Template
MUC-3					R < 50% P < 70%
MUC-4					F < 56%
MUC-5					(Joint Ventures) F < 53% (Microelectronics) F < 50%
MUC-6	F < 97%	R < 63% P < 72%	F < 80%		F < 57%
MUC-7	F < 94%	F < 62%	F < 87%	F < 76%	F < 51%

Table N: Best results reported in MUC-3 through MUC-7, by Task

Legend: R = Recall P = Precision F = F-Measure with 1.0 β

Our results are in line with some of the metrics reported in the IE literature, with some of them clearly improving on what could be perceived as the generics of a particular task. If we consider the global score as a measure of how well this particular "scenario" of metalinguistic predication was filled, these scores stand out. But, as mentioned above, this is not an altogether fair comparison as the complexity of the labeling and of the record structure is much higher in any of the MUC evaluations. This statement notwithstanding, the MOP system performs fairly well in its limited-scope task, especially considering that it does not required labor-intensive tools and resources, like customized lexicons, Gazetteers, training datasets, syntactic parsers, all of them used by the latest versions of the MUC competing systems. The MOP system is indeed highly focused on a specific kind of data, metalinguistic information, and does not use all possible extraction patterns, as it is not striving at this point for exhaustiveness. But on the other hand, as the 0.07 spread in global F-Measure for three very different textual corpora shows, it is very portable across domains, as long as the texts belong to a technical sublanguage.

The DEFINDER system (Klavans et al, 2001) at Columbia University is to my knowledge the only system that is fully comparable with MOP, both in scope and goals, although there exist a few basic differences between them. First, DEFINDER examines user-oriented documents that are bound to contain fully-developed definitions for the layman, as the general goal of the PERSIVAL project is to present medical information to patients in a less technical

[62] Not all IE tasks were evaluated in all conferences

language than the one in the reference literature. MOP, in contrast, focuses on leading-edge research papers that present the less predictable informational templates of highly technical language. Secondly, by the very nature of DEFINDER's goals their qualitative evaluation criteria include readability, usefulness and completeness as judged by lay subjects, criteria which we have not adopted here. Neither have we determined coverage against existing on-line dictionaries, as the Columbia group has done. Taking into account the above-mentioned differences between the two systems' methods and goals, MOP compares well with the 0.8 precision and 0.75 recall of DEFINDER. While the resulting MOP "definitions" generally do not present high readability or completeness, these informational segments are not meant to be read by laymen, but used by domain lexicographers reviewing existing glossaries for neological change, or, for example, in machine-readable form by applications that attempt automatic categorization. Low recall rates in our tests are in part due to the fact that we are dealing with the wider realm of metalinguistic information, as opposed to structured definitional sentences that have been distilled by an expert for consumer-oriented documents.

We have opted for the exploitation of less standardized, non-default metalinguistic information that, as we have argued in Section II.2, is being put forward in text precisely because it cannot be assumed to be part of the ongoing collective expert-domain competence. In doing so, we have exposed our system to the less predictable and highly charged lexical environment of leading-edge research literature, the cauldron where knowledge and terminological systems are forged in real time, and where scientific meaning and interpretation are constantly debated, modified and agreed. We have not performed major customization of the system (like extensively enriching the tagging lexicon with medical terms), in order to preserve the ability to use the system across different domains.[63] Domain customization may improve metrics, but at a cost for portability.

The implementation described here for the task of extracting metalinguistic information from free text undoubtedly shows room for improvement in some areas, including the following: adding other patterns for better overall recall rates, deeper parsing for more accurate semantic typing of sentence arguments, incorporating an anaphora-resolution module, etc. The ability to extend the analysis of the information in sentences, with further processing of the informational segment (to find the head noun, for example, in order to do sortal and categorization operations for domain ontologies), or inclusion of other types of relevant information like agentivity, temporal timelines, hedging, could produce richer and more useful databases for use by humans

[63] "The difference between a viable message understanding technology and a practical message understanding technology lies in the ease with which that technology can be ported across domains." (Lehnert, et al., 1994)

(lexicographers, programmers) or machines. Improving the partial parsing and semantic labeling machinery (for example with third-party, non-Python programs with higher benchmarking than our modules produced almost from scratch), would undoubtedly lead to better precision and recall rates. We will tackle these enhancements in future versions of the MOP system. What we set out to do here was, strictly speaking, to show the feasibility of automating the identification and extraction of the information that technical texts provide about the lexical items and rules of usage of a sublanguage, using both Language Engineering and stochastic methods. We believe that the described system implementation has done so, and fulfils the three conditions that R. Grisham (1997) set for a successful IE system:

> Current methods will be successful if the information to be extracted is expressed directly (so that no complex inference is required), is predominantly expressed in a relatively small number of forms, and is expressed relatively locally within the text.

Most Information Extraction systems are limited in practice to a single topic, to a single domain. However, our specific "domain" here, of language usage and terminological conventions, is present in all areas of knowledge and in all disciplines where an expert community interacts to create a consensus. Therefore, metalinguistic information extraction is potentially applicable to any kind of technical text.

Metalinguistic Information Databases are the ultimate goal of the MOP system, and the final output of this dissertation. They constitute rich knowledge sources about the state and evolution of a field's sublanguage. The next chapter will discuss some of the characteristics, applications and limitations of the Metalinguistic Information Databases that the MOP system creates after processing technical corpora.

VI METALINGUISTIC INFORMATION DATABASES AS NON-STANDARD LEXICAL RESOURCES

Summary

Conventional terminological dictionaries and mental or computational specialized lexicons can be seen as more or less static repositories of the default lexical knowledge of terms used by a linguistic and domain-centered community of expert speakers of a sublanguage. A conventional lexical database fails to represent a sublanguage's dynamicity and open-endness. In contrast, a Metalinguistic Information Database generated with the MOP system contains the multi-textured real-time data produced in the discourse of research papers and technical documents. A MID is not exactly a lexicographic artifact, since it can be viewed as a listing of exceptions, special contexts and specific usages of lexical and terminological items where meaning, value or pragmatic conditions have been spotlighted by discourse for cognitive reasons. Terminological data in MIDs can be more specific and might be better suited for the interpretation of certain texts or utterances than that of lexical knowledge bases and lexicons.

MIDs are semi-structured resources that need to be further processed to become functional taxonomies and lexicons. Nevertheless, they can have many applications for research and technological development; among them: update and fine-tuning of lexicons and ontologies, neology detection, non-default information repositories for inference engines, research and didactics of specialized discourse and scientific activities. MIDs, therefore, are useful theory-neutral data structures, although we have also pointed out that, since some of the information contained in them is not typed and maintains some of its linguistic form, there are some difficulties having to do with their integration and update. Interestingly, for that same reason, they can be considered accumulative records of conceptual and terminological change.

VI.1 Lexical knowledge and computational resources for NLP

VI.1.1 The nature of information in lexicons and dictionaries

The traditional view of language dictionaries is that their definitions contain meaning stereotypes, the semantic content or referents typically associated with a lexical item by a community of users of that language. Lexicographic work on a language would then aim at isolating and discovering those semantic and pragmatic features that uniquely identify the usage of a word within a specific community. A definition will often be created by formulating a descriptive sentence that reflects the common knowledge about what some word means. Lara (1997b) calls this process a "reconstruction" of meaning. Lexicographic entries can be seen as repositories of the default, core lexical information conveyed by words or terms used by a community (that is, the information available to an average, idealized speaker). From this perspective, the difference between a general language dictionary and a terminological dictionary is only in the community whose language is being reflected. In the former case, it would be a whole social group defined by cultural and linguistic parameters, while in the latter case would be a community of users of the sublanguage employed in specialized contexts, in epistemologically-restricted fields and domains.

Resources like human-readable lexicons and dictionaries create conventional meaning definitions that require fairly sophisticated hermeneutics, and these interpretations can be based on personal linguistic competence, on Corpus Linguistic analysis, or both. Even if they contain a listing of possible distinct senses for a word, lexicographic entries usually cannot reflect accurately all the nuances of meaning and selectional restrictions that a word used in a very specific context can carry. Lexical entries in dictionaries abstract away from any context that is too specific, but aim at preserving general contextual idiosyncrasies as much as possible. Within a modern methodological framework, these definitions would be seen as a condensation of the lexical information gathered from a multi-source analysis of the word's actual usage. As such, lexical entries constitute selections of all the data obtained from actual occurrences of a given word as used in actual sentences. Of course, dictionaries and lexicons often carry more than

definitions, and can provide information on morphology, part of speech, use or domain restrictions, synonyms, argument structure and other kinds of lexical data.[64]

Regardless of its varied nature, the information that dictionaries and lexicons provide can be seen as the common ground that makes communication possible, as it represents a given community's shared knowledge about the linguistic code, either implicitly through their acquired competence or explicitly through the constant negotiation of meaning. As described in section II.1.3, this is high-level, core information that allows a more or less precise characterization of an entry's meaning that is valid for the language-defined community. The information in these conventional resources makes up a set of defaults that a linguistic competence is based on.

Following the lexicalist conception of grammar and meaning adopted by a number of contemporary theoretical frameworks like HPSG, LFG and GL, computational linguists rely more and more on lexicons that represent the richness of the information available "locally" in lexical entries and rules (along with the generative potential concerning phenomena that have been traditionally considered as exclusively syntactic). An accurate representation of lexemes, morphemes and subcategorization frames as they are attested by actual use in extensive and representative linguistic corpora is now a cornerstone of linguistic research. Computational lexicons and dictionaries need to present this information in a format that computer programs can interpret and use correctly. For NLP applications, the lack of an adequate and fairly complete lexicon can produce a veritable bottleneck for processing discourse.

Lexical Knowledge bases in machine-readable form might contain POS tags, valency frames, selectional restrictions or lexical relations such as the hierarchical ones inherent in ontologies and thesauri. Computational lexicons generally contain information useful for syntactic analysis or semantic interpretation. Unlike simple electronic versions of conventional dictionaries, NLP resources are generally not intended for human users, but for computer applications that process, interpret or generate language. Some of these resources are more conceptually-oriented, like the taxonomies in WordNet, or more linguistically-oriented, like tagging lexicons or FrameNet. What sets these apart from traditional dictionaries is the fact that they are compiled semi- or full-automatically from corpora, or using previously human-compiled resources to bootstrap the acquisition process.

As we have stated in IV.2.2, semantic networks and ontologies restrict themselves to very specific aspects of lexical knowledge, while knowledge bases represent fully structured databases with machine-readable, multidimensional linguistic information. These data structures

[64] Other information, like example sentences using a given word, can sometimes be more illustrative of how this word is used in language than a reformulation of meaning through a definition.

differ both in the kind of information they contain and s in the structural organization of that information. As seen in the various example entries at the closing of chapter IV, Metalinguistic Information Databases created with the MOP system contain a wide variety of semi-structured data that retain some of its linguistic organization, and need to be interpreted by humans or further processed. Unlike the lexical knowledge bases that structure their entries using inheritance mechanisms and typed features, Metalinguistic Information Databases as IE products contain only a structural disposition that reflects the predicative relationships among semantic constituents inherent in the sentential form. While such structural organization can be further processed to replicate more sophisticated databases that incorporate what is basically implicit information, they do not presuppose any specific linguistic theory in their organization that would force an interpretation just by formal constraints. The information in MIDs, unlike that in dictionaries, is much more unconstrained and unstructured, with the disadvantages and potentialities that derive from that fact.

VI.1.2 MIDs as repositories of non-default information

Traditional entries of lexical resources like dictionaries and glossaries contain high level, default information on word usage, and by their very nature miss a lot of the more specific, context-dependent, linguistic knowledge accessible through specialized texts in which terms are being proposed, discussed, defined, modified, or evaluated within the complex dynamics of an community of experts. We have already mentioned, following Boguraev and Levin (1993), that a limitation of lexical databases is their failure to properly represent a language's productivity and open-endness. Machine-readable dictionaries as lexical acquisition sources are static objects (Boguraev & Pustejovsky, 1996) that cannot represent the dynamic nature of language, especially of rapidly evolving technical sublanguages. We have already pointed out that MIDs obtained from automatic extracting and processing EMOs would serve two functions at the same time: on the one hand updating terminological resources and on the other compiling a record of terminological change as is evidenced in the technical and academic texts where those changes take place.

A Metalinguistic Information Database (MID) obtained with the MOP system described in previous chapters compiles real-time data provided by metalanguage analysis of leading-edge research papers. In a general sense, a MID could be conceptualized as an anti-dictionary: a listing of exceptions, special contexts and specific usage, of instances where meaning, value or pragmatic conditions have been spotlighted by discourse for cognitive reasons. Kilgarriff (2001) defines a non-standard lexical use in a way that would be circular in our present context, as a word sense that is not attested in any dictionary. Our consideration of such information is

somewhat more functional. Non-standard data, especially for terminological items, is not in traditional dictionaries because they are not part of a core competence, but constitute new meaning vectors needed for epistemological reasons, like an expansion of a shared cognitive space represented by a consensus-accepted theory. Normally, such highly-specific and unconventional information would not be of much use in normalized communicative exchanges, as our mental lexicon is usually enough to process everyday linguistic utterances. Non-standard data becomes important when either the linguistic code or the conceptual system changes, which in the knowledge-producing contexts our research work has targeted happens fairly often. When interpreting text, regular lexical information is applied by default under normal conditions, but more specific lexical, pragmatic or discursive information can override it if necessary, or if context demands so (Lascarides and Copestake, 1995). A MID with computationally tractable data can either override or enrich the default information of a lexical database. Its role would not be to replace, but to complement and enrich terminological Knowledge Bases, computational lexicons or lexical taxonomies.

Exploiting the metalinguistic dimension of specialized discourse, a MID's entries contain information that has been stated precisely because it cannot be assumed to be available under normal circumstances to the recipient of the message; that is, it is relevant because it is not assumed to be part of the shared mental lexicon in the linguistic exchange. It is in this thin slice of data that language is accomplishing the fundamental intersubjective task of building new knowledge and testing the old one.

We are implementing the MID databases using XML standards and resources to ensure transparency, portability and accessibility across platforms and applications. XML is flexible enough to transfer the responsibility for processing data to querying applications, instead of forcing some kind of interpretation by its very nature or structure. In that sense, a database that encourages further processing lies in between the raw possibilities of pure corpus text, and the (sometimes excessively) structured data of traditional lexical resources that are anchored in fixed theoretical frameworks. In this light, MIDs can be seen either as incompletely structured lexical knowledge bases, or as semi-processed lexicographic corpora.

As structured now, our MIDs contain only a subset of all possible information obtainable from EMO processing (see Section III.2.4). The following document type definition shows that structure:

```
<!ENTITY STATEMENT "Prototipo de Base de Informacion Metalinguistica /
Metalinguistic Information Database Prototype" >
<!ELEMENT MID  (OME+)  >
<!ATTLIST MID
 file  CDATA       #REQUIRED
 timestamp  CDATA       #REQUIRED
```

```
 entriesnum  CDATA        #IMPLIED>
<!ELEMENT OME (Autonym+ | Info |  Operator | line )* >
<!ATTLIST OME
  text       CDATA      #REQUIRED
  n          CDATA      #REQUIRED >
<!ELEMENT Autonym   (#PCDATA) >
<!ELEMENT Info  (#PCDATA) >
<!ELEMENT Operator  (#PCDATA) >
<!ELEMENT line  (#PCDATA) >
```

MID document type definition

A small sample from a MID compiled from the Histology corpus illustrates their typical content:

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE MID SYSTEM 'mid.dtd' []>  <?xml-stylesheet type='text/xsl' href='mid.xsl'?>
 <MID file='HistologyNumbered.xml' timestamp=' Sun Mar 14 18:25:03 2004' entriesnum='39'>
<OME text="HistologyNumbered" n="1">
<Autonym>  fine or submicroscopic structure/ultrastructure </Autonym>
<Info>The detailed morphology revealed by EM</Info>
<Operator> called </Operator>
<line> The detailed morphology revealed by EM may be called fine or submicroscopic structure/ultrastructure .</line>
</OME>

<OME text="HistologyNumbered" n="4">
<Autonym>  cristae </Autonym>
<Info>plates or tubules</Info>
<Operator> called </Operator>
<line> the inner membrane projects inwards as plates or tubules called cristae , studded with small 9 nm wide elementary particles - rounded bodies on stalks .</line>
</OME>

<OME text="HistologyNumbered" n="5">
<Autonym>  a basal body/kinetosome </Autonym>
<Info>A similar cylindrical structure</Info>
<Operator> called </Operator>
<line> ( b ) A similar cylindrical structure is seen at the base of each cilium and is called a basal body/kinetosome .</line>
</OME>

<OME text="HistologyNumbered" n="7">
<Autonym>  the actin cortex </Autonym>
<Info>This zone</Info>
<Operator> known as </Operator>
<line> This zone is now known as the actin cortex because of the actin filaments attached to the cell membrane for locomotion and changes in cell  shape .</line>
</OME>

<OME text="HistologyNumbered" n="9">
<Autonym>  Fibril-Associated Collagens with InterrupTed helices - FACIT</Autonym>
```

<Info>Some of the scaffold-glueing ones , e.g. , types IX , XII , and XIV</Info>

<Operator> termed </Operator>

<line> Some of the scaffold-glueing ones , e.g. , types IX , XII , and XIV , are termed Fibril-Associated Collagens with InterrupTed helices - FACIT .</line>

(…)

</MID>

Sample MID from the Histology corpus

Pustejovsky (1999) states that:

> Computational lexicons are typically evaluated in terms of: (i) coverage: both breadth of the lexicon and depth of lexical information; (ii) extensibility: how easily can information be added to the lexical entry? How readily is new information made consistent with the other lexical structures? (iii) utility: how useful are the lexical entries for specific tasks and applications?

As stated before, that MIDs are not (and do not intend to be) fully functional computational lexicons. However, in the light of the above quote, the following points can be made. With regard to coverage MIDs are not meant to represent a whole lexicon or a field's terminology, but only those terms or words that are being introduced, modified, or otherwise specified in discourse. Also MID entries do not claim to be complete records of all linguistic information for an item, as they represent an specific aspect of that information being mentioned explicitly in the source Explicit Metalinguistic Operation. A MID's claim to relevance, more that exhaustiveness, is its precision and the cognitive relevance of the identified terms. MIDs can help computational lexicons attain adequate coverage of a lexicon by maintaining it updated within the fluidity and dynamics they are design to represent. With regard to extensibility, MIDs are continually being updated by the processing of new text, and previously unencountered terms and their variations can be added to the database instantly. Unfortunately, given that (a) the informational segments being provided as relevant information generally retain their linguistic expression and (b) the kind of information being provided about the sublanguage is an open-ended set, it is very difficult to integrate new with existing information for the same item in the database, or to drive inferences to attain a single data structure for a specific item. MIDs are also records of terminological change, and as such it might be unadvisable to do any more integration that what is strictly needed for a particular task or application (see VI.2.4 below). Although we have not attempted further processing of MIDs to create or update true lexical knowledge bases, we can envision such applications as falling squarely within the present state of the art in Human Language Technology (HLT).

In the following sections, we focus on utility, Pustejovsky's third and definitive criterion. We discuss some specific uses and applications of MIDs, although others might certainly be possible. We are not claiming that MIDs could replace any of the more mainstream data repositories like lexical or terminological knowledge bases, ontologies or marked-up corpora, but we do pretend that our databases could certainly be used concurrently and in certain cases could improve those resources. MIDs provide privileged, knowledge-rich contexts and systemically important terms to drive forward the lexical and knowledge acquisition question we have talked about in IV.2.2 (Lauer, 1994; Boguraev and Pustejovsky, 1996). Unlike more conventional resources, MIDs do not contain homogeneous data of one kind or the other in a strictly enforced data structure, but a variety of information on specific lexical items that is constitutive of high relevance for ongoing discourse.

We have stated that MIDs in their present form are not, in full justice, lexical knowledge bases comparable with the highly-structured and sophisticated resources that use inheritance and typed features, like Lexical Knowgledge Bases. MIDs are semi-structured resources that can be further processed to convert them into auxiliary sources for functional taxonomies and lexicons, using algorithms and techniques along the lines of those developed for the Acquilex project (Briscoe, de Paiva, Copestake, et al., 1993). The task of outputting MIDs is made more difficult by the fact that informative predicates have no single syntactic or semantic realization; although most of the time they are materialized in nominal clauses, sometimes they appear as standardized (pragmatic) values, single-item lexical descriptors, or even adopt a full sentential form. Nevertheless, some practical applications for the MIDs described here and in the previous chapter can be suggested, although different ones can surely be envisioned.

VI.2 Uses and exploitation of MIDs

VI.2.1 Lexicography and neology detection

Although some of the EMOs retrieved by our application perform modification or evaluation of a lexical item that previously existed in the domain's vocabulary, in most cases terms in those sentences are being proposed for the first time. For terminology and specialized lexicography it is very important to recognize these new items accurately in order to update knowledge about the domain's term set, or compile specialized dictionaries and vocabularies. Terminology extraction programs use syntactic patterns and compare against lists of known terms to find new items being used. These use-based contexts can be very useful to discover neology, but mention-based contexts like explicit definitions can provide richer information about the term's behavior, meaning and interrelation to other conceptual and lexical elements in the knowledge

systems. Term extraction systems that only use statistical information might miss low occurrence terms that nonetheless EMOs might show to be vital for a conceptual system or a theory. An information extraction system like MOP could scour scientific journals or even online press to find new terms, maybe even in synchronization with a lookup system that searches new term candidates in a database of previously-known terms.

Two other possible advantages of this method merit mention here. First, the information for discovered terminological items is generally more specific than the information obtainable by lookup in domain glossaries. Most likely it is also better suited for the interpretation of certain texts, or collections of texts that deal with concrete problems or subscribe to a given theoretical framework. The locality of such information is advantageous for specialized lexicography. Secondly, the information that comes to light will better reflect the dynamic nature of specialized text, and will be very much up-to-date, as compared with human-compiled vocabularies that take years to review and create, and might reflect an outmoded conceptual configurations of a given domain. One of the recurring problems of term extraction systems is how to delimit terms in ongoing text, that is, how to identify terminological units. In EMOs, that delimitation is usually done by markers/operators with cognitive and formal devices, and this problem arises in a very tractable form. Resulting terminology, then, can guide term segmentation in conventional term extraction systems.

As mentioned earlier, MIDs cannot be viewed as end-user products, but as semi-processed resources. They are best characterized as auxiliary lexical knowledge resources, rather than core lexical references. Lexicographers and terminologist can use them as tools for their own labor-intensive work of reviewing and compiling dictionaries and domain-specific vocabularies. An interesting term discovered by a MOP-like system can become the focus of a more thorough review or follow-up using familiar Corpus Linguistics or terminology extraction techniques. Computational lexicography can employ MIDs as raw material for further processing that can yield an efficient and reliable lexicon or ontology for NLP.

Automating these processes is not a trivial task. The SPECIALIST and MeSH lexicons developed at the National Library of Medicine are general language and domain-specific lexical resources used to aid NLP processing and indexing of Health Sciences literature. Along with the UMLS metathesaurus, they are used to map the complex and changing conceptual architecture of modern biomedical research, as expressed in the copious literature of the field. As mentioned before, in order to maintain relevance and consistency of these vast resources (the MeSH controlled-vocabulary alone has 22,568 entries, and the MedLine abstract database incorporates around 40,000 records each month) the NLM staff needs to manually review 400,000 highly-technical papers each year (Powell et al, 2002). Lexical knowledge-bases such as MIDs can aid

these efforts by providing reliable information to be used by staff lexicographers and domain specialists for lexical items already known, but could also mine text for unaccounted terms or unknown conceptual relationships.

VI.2.2 Ontology bootstrapping and rerendering, and semantic typing

An example of further processing of an MID is the creation and maintenance of special domain taxonomies or ontologies that are needed for managing the vast scientific knowledge-bases and torrential literature of some fields, like the biobibliome of the Biomedical Sciences described in previous sections. Most of the work in mining Bio-Medicine text has focused on finding specific sets of entity relations (protein interactions, bio-entity bindings). Our research suggests that it is possible to track theoretical and descriptive evolution of the conceptual architecture of the domain by focusing on system-wide data, such as type-hierarchies, terminological change and ontological commitment.

We have also stated that definitional contexts provide sortal information as a natural part of the process of precisely situating a term or concept against the meaning network of interrelated lexical items, establishing a place for that item in the symbolic and mental systems created by theories. For example, a term like “bioconjugation chemistry” could be categorized using an MID entry such as:

Term	bioconjugation chemistry
Marker/Operator	known as "
Type/Definition (is_a)	synthetic procedures, all characterized by high specificity and mild condition of reaction
Context	*The various and often difficult chemical problems encountered in conjugation of so many different products prompted the development of many synthetic procedures , all characterized by high specificity and mild condition of reaction , now known as " bioconjugation chemistry " .*

A fairly accurate definition results for “bioconjugation chemistry”, and its syntactic kernel provides a semantic typing that can help in placing it under the category of *synthetic procedures*. A lookup in existing thesauri like the UMLS can validate its place in the ontology, or, if it is not found there, as in this case, it can be added to it, enriching that branch of the taxonomy.

Pioneer work in this direction is the Medstract project, a joint effort by Brandeis University and Tufts University researchers (Pustejovsky et al., 2002a & 2002b) that mines biomedical abstracts to create specialized resources, like the AcroMed acronym database, and aims at performing semantic “rerendering” of the UMLS ontology based on interesting predicative

patterns. Tools, techniques and resources already developed for the Medstract and MOP systems could be reused and customized, with new modules developed reliably using existing architectures. Lexicographic deliverables obtained with these techniques could help extend and maintain existing UMLS Knowledge Sources (KS), and could become KS themselves after manual validation by the NLM staff or domain experts. System evaluation could be done using standard IE metrics over existing and newly-compiled golden standards, and resulting databases could be checked for accuracy both against manually-validated reference resources, and by domain expert review.

Use of full-text corpora (as opposed to mining of pre-compiled databases or abstracts as is presently done) is vital to ensure a good panoramic coverage of these linguistic phenomena across the domain. These knowledge-rich contexts have not been exploited on the scale possible with the MOP system, and exploiting metalanguage could improve the resources used for indexing and retrieving of relevant information in large-scale textual repositories of Bio-Medicine. Lexical knowledge bases thus obtained can also be useful for other NLP tasks, such as anaphora resolution, topic detection, entity recognition and semantic disambiguation to improve Information Retrieval indexing of relevant information from biobibliome corpora.

Another interesting possible use over full-text biobibliome is to employ a metalinguistic marker such as “defined as” to create a database of how the different clinical trails define their study variables, as it would allow researchers to check how some standard criteria or protocol was applied in each case, and see if a) it is compatible with their own protocols (and thus, comparable with their own results) and b) how well each study’s criteria fits with standardized parameters for the technique, condition or observation, as in the next two instances extracted from the MedLine abstract database (boldface is mine):

- The **value for the recovery of heart rate** was defined as the decrease in the heart rate from peak exercise to one minute after the cessation of exercise.

- A **missed injury** was defined as an injury not identified during assessment in the Emergency department, but identified later in the hospital or rehabilitation centre.

VI.2.3 Backup and update of knowledge sources for inference engines

The non-default and highly relevant information from MIDs could also provide the material for new interpretation rules in reasoning applications, when inferences will not succeed because the state of the lexico-conceptual system, as reflected in their reference lexicons and taxonomies, has changed. A neologism or a word in an unexpected technical sense could stump a NLP system that assumes it will be able to use default information from a machine-readable

lexicon. A domain's MID could efficiently store and make available lexical and pragmatic information to update terminological Knowledge Bases or machine-readable dictionaries, but it could also be used directly by AI systems that need unorthodox information (not readily found in semantic networks, lexicons or traditional ontologies) to drive inferences or disambiguate. When the normalized default information of conventional resources is insufficient or contradictory with other interpretation conditions, use of MIDs as an alternate knowledge repository could provide access to alternative and highly-specific information to allow linguistic generation or understanding of an unorthodox context. As we have pointed out, following Lascarides and Copestake (1995), default knowledge for lexical interpretation is overwritten under certain circumstances by more specific information.

VI.2.4 Research on the evolution and production of expert knowledge

This dissertation's empirical work focuses on the linguistically-based mechanisms with which theoretical knowledge is created, modified and negotiated in expert communities, and how that knowledge is enacted linguistically through a sublanguage and a terminology. It is an epistemological concern, but also a linguistic one, inasmuch linguistics is concerned with the mechanisms and principles whereby language is used to convey meaning, and how precisely it is possible to map from linguistic constructions to conceptual structure (Vallduví, 1992).

Besides creating machine-readable resources of non-standard linguistic information for NLP applications, the MOP system can be useful as a tool to do empirical research into the nature of expert knowledge, as can be evidenced in the interaction of scientists and scholars that struggle to communicate theoretical explanations and descriptions to their colleagues. We believe our findings to be of interest both to theoretical and practical approaches into the related questions of how expert knowledge is created, structured and applied. In particular, MIDs, as containers of dynamically updated information, can be conceived as useful for a visualization tool that displays both the temporal evolution of the meaning of a theory's terms and well as the changing links between those concepts, the source and extent of the changes as they transpire from leading-edge research papers. The following table contains a few key sociological terms from our original exploratory corpus. It illustrates the wide variety of information that can be expected from even a limited number of texts and the fact that most of the time what we are dealing with is fragmentary and ephemeral data that is unlikely to be found in lexicographic definitions. The information is shown in the format of the chunks that were extracted. The table also shows how heterogeneous information is provided for one and the same term because of the evolution of the discipline and the dynamics of theoretical debate. Conceptual reformulation is a key process in all disciplines subjected to consensus. This information can be seen as

differential data that reflects a particular change in a previously-known lexical item. A lexico-conceptual lattice could be created that would allow a privileged perspective of rapidly-changing disciplines through timelines of terminological evolution.

Term/Concept	**Informational Predicates**
FAMILY	[extend the meaning to include same-sex couples, single-parents, nannies, adoptive and step children, and so on]
FAMILY	[two adults of opposite sex, married to each other, and living with their common children]
IDENTITY	[There are two typical contexts]
IDENTITY	[an emotional attachment and a sense of belonging of a semi-sacred kind]
IDENTITY	[consideration to what exactly] [is supposed to mean]
NATIONALISM	[used here, deliberately, to describe both aspects of the phenomenon]
NATIONALISM	[is used for both of these things - world view and activism]

These variations and enrichments in the meaning conditions for key terms in a discipline can be time-stamped and linked to a textual instance (a document where they are presented), an author,[65] or a specific theoretical framework. From a sophisticated MID that updates its records but preserves an historical archive indexed to each term, a visualization of this evolution might be possible, along the lines suggested in Figure 8, representing in this case the field of Sociology. In that figure, theories and authors share some concepts and terms, but follow different evolutionary routes and present alternate orderings of their conceptual structures, while influencing each other's theoretical and semantic choices:

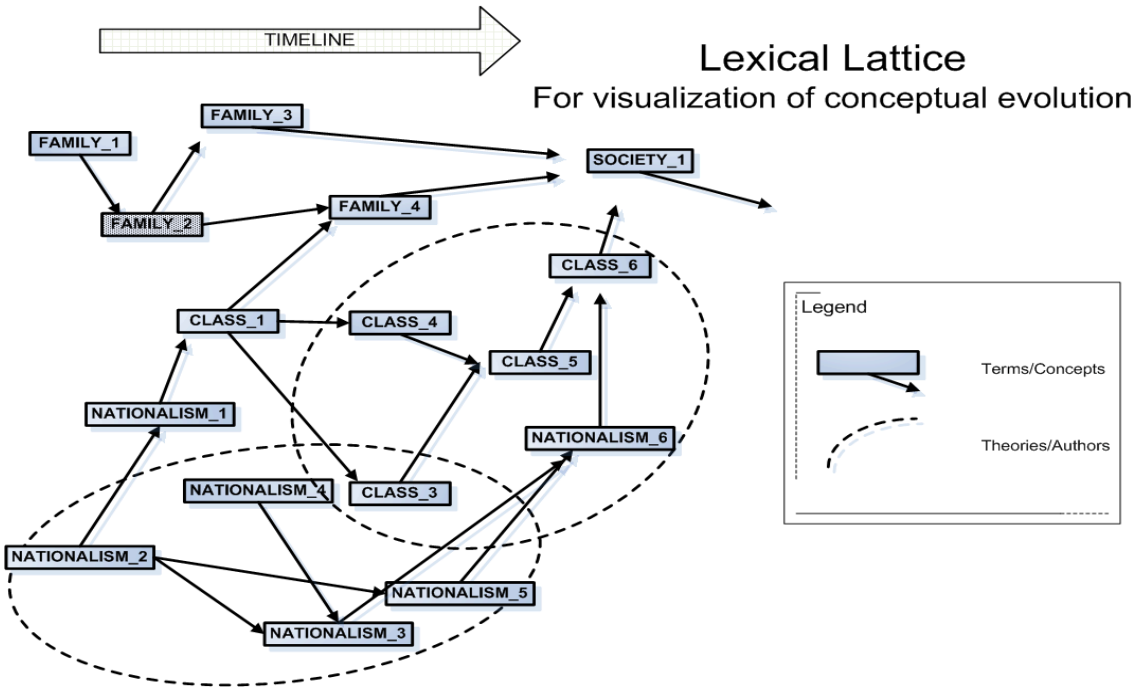


Figure 8. Visualization of conceptual evolution

[65] Or cluster of authors representing a school of thought

We have focused as much as possible on the realm of theory-building, since it is through explanatory theories that modern scientific and disciplinary knowledge is enacted and circulated within the community of experts that define and modify those theories. We can follow in textual corpora the evolution and development of specialized knowledge, through the public exchanges that aim at term introduction or modification, that is, through the debates about terminology and language use included in almost all important theoretical scientific papers. This data-driven, empirical methodology should complement other more traditional philosophical approaches to the study of the nature of scientific activity.

VI.2.5 Didactics of domain-specific sublanguages

Of course, an empirical-based representation of a discipline's state of the art and history would be very useful to teach its intricacies to students. Acquiring a terminology is one of the paths to the mastery of an academic field. Following the establishment of terms and concepts in the literature should enable students to better understand theoretical issues and the discursive processes that model them. Use of these techniques would also endow students with more accurate and fine-grained domain vocabularies to support a more thorough study of their disciplines, and the different contexts where that discipline advances through a process of rational argumentation and descriptive efficacy. Incidentally, the benefits of all of this for translation studies are also obvious.

The MOP system and resulting MIDs, as well as Corpus Linguistic analysis in general, can also provide students with an overview of a field's stylistic conventions, and would teach them to be good producers of scientific literature. This aspect of science training is seldom addressed directly, even though producing specialized texts is an important part of the activities of scientists as creators and validators of new knowledge. A mastery of the sublanguage employed to enact new knowledge is a vital part of becoming a member of an expert community, of truly becoming a peer among peers.

VI.3 Enriching MIDs: Integration of metalinguistic information into a useful data structure

The real challenge facing the work presented here lies not in retrieving EMOs from text to populate a MID but in the successful formalization of heterogeneous linguistic information into a robust and manageable data structure. We have already mentioned this (VI.1.2) when we presented Pustejovsky's (1999) extensibility criterion for the evaluation of lexical resources in general, and pointed out that although information can be easily added, it is very difficult to integrate it fully with previously-culled one. This objective might require redefinition, within

this context, of traditional semantic and pragmatic notions of meaning and conditions of use of terminological items. We have not even attempted to retrieve from EMOs rich information about scope of term modification or proposal tentativity, for reasons explained elsewhere, but such information is available there contingent on further processing. An effective and efficient computational mapping of such heterogeneous and diverse linguistic information into a robust and manageable data structure is not a trivial task, and we cannot claim to have achieved it here. In MIDs the information is very heterogeneous and, unlike in true knowledge bases, it is not typed and maintains some of its linguistic expression. Notions that present more or less clear-cut choices within their native theories, such as lexical meaning, semantic content, sense restrictions, contextual conditions, community consensus, etc., have to be dealt with in a consistent and reasonable manner, and have to be interpreted from a motley set of predicates and informational segments obtained from an empirical study with Corpus Linguistics methods. As we stated before, MIDs are, for the time being, semi-processed resources, midway between raw corpora and structured lexical bases.

A collateral effect of compiling large enough MIDs is to make available to theory-driven linguistic research a comprehensible sampling of the real-world diversity of lexical data. As Kilgarriff (2001) shows, even a sophisticated model like the Generative Lexicon cannot account for some of the non-standard lexical uses encountered in corpora. A Metalinguistic Information Database, as we envision it, must be able to integrate some features from diverse (and perhaps conflicting) lexical representation systems. Again, our prototype system does not claim to do that yet. We have only attempted to sketch a proof-of-concept implementation that could eventually lead to real-world, robust systems. Nevertheless, our corpus-based survey of the metalinguistic aspect of scientific discourse has pointed to the need of exploring the common ground that all semantic and pragmatic theories share. In one sense creating an adequate data structure from MIDs tests the representational limits of normalized lexical knowledge bases, which contain word properties and values as well as "constraints on word behavior, dependence of word interpretation on context and distribution of linguistic generalizations" (Pustejovsky, 1999). Attempting to combine the strong points of different computational implementations of diverse (and sometimes antagonistic) theoretical principles can quickly become a walk through an intellectual minefield, although it could be, it has to be said, a very interesting stroll indeed. Due to self-imposed constraints for this dissertation, we will not attempt that theoretical exercise at this time.

VII RECAPITULATION AND CONCLUSIONS

A very brief summary of the work done in this Ph.D. dissertation is that we have studied a very specific aspect of specialized discourse in order to apply Corpus and Computational Linguistic techniques to the automatic extraction of sublanguage information from unstructured text. The goal of the Metalinguistic Operator Processor (MOP) system is the automatic compilation of Metalinguistic Information Databases that are useful for a variety of academic and technological tasks as described in the previous chapter that run the gamut from updating computational lexicons to driving graphical representation of conceptual change in Science. The MOP system operates over certain textual and discursive instances that we have called Explicit Metalinguistic Operations (EMOs), that we have described in detail. Besides the actual coding of our proof-of-concept application, the core claims that I have made in this dissertation are:

(1) that EMOs as discourse processes have a foundational nature in the intersubjective construction of scientific and technical knowledge, as materialized in the sublanguages that constitute their material support;

(2) that metalinguistic predication, by virtue of its formal, cognitive and pragmatic properties, can be processed in a robust and theoretically-motivated manner by computer applications that extract and structure domain-specific terminological and linguistic information;

(3) that the computational resources obtained with these methods (termed MIDs) can be useful for empirical studies of scientific knowledge and expert communities, as well as for Natural Language Processing applications and specialized lexicography.

There have been many efforts at automatic extraction of terminology and definitional information from semi- or non-structured text; however, our work includes the processing of metalinguistic information of a wider spectrum than conventional definitional structures. Our central claims to novelty lie, then, in presenting an adequate theoretical object (EMOs) on which to build an application that can retrieve metalinguistic information in general (including, but not limited to, definitions). Some of the methods employed here have been borrowed from Information Extraction, but our use of some of those computational techniques on the metalinguistic dimension of free-form text is certainly novel. Even without the theoretical apparatus we have presented (in chapters II and III) for the phenomena over which our MOP

system operates, we believe that its implementation and the Metalinguistic Information Databases produced constitute a clear advance of the state of the art of Human Language technology, especially with regard to the study and processing of domain-specific sublanguages. The MOP system is a proof-of-concept implementation aimed at showing the practical applications and basic soundness of the theoretical description of EMOs; it does not claim to be the ultimate or most efficient way to carry out these tasks from a computational point of view. An enclosed CDROM contains the corpora, datasets and application codes for this dissertation, along with an electronic version of this document.

In what follows, we present a chapter by chapter summary and a set of conclusions.

Chapter I motivates and introduces the subject and the methodology of the research described in this dissertation, including the document sets used for analysis. It argues for the need to engage in a data-driven and empirical study of specialized discourse that goes beyond the conceptualist and mentalist analysis preferred by traditional onomasiological Wüsterian Terminology. It also argues for the need for extensive NLP resources such as can be seeded and updated with the Metalinguistic Information Databases advocated here. The initial chapter also attempts to formulate the main issues tackled in this work within four main perspectives: epistemological, linguistic, computational and cognitive.

The set of corpora used in our work both for empirical analysis and for application development and evaluation included a corpus of Sociology research papers and another corpus created from systematic searches with selected lexical patterns on the British National Corpus, as well as other corpora from technical documentations, manuals and abstract databases.

In chapter II, an in-depth analysis of metalanguage grounds the descriptive model for EMOs that will be presented in chapter III. That section highlights some of the formal, cognitive and pragmatic features of metalinguistic sentences that facilitate their automatic processing. From a formal point of view all symbolic systems like language are founded and enacted through the use of a metalanguage that describes its elements and rules. Autonyms are self-referential words functioning as signs for themselves, as when they are being mentioned instead of being used normally. Metalanguage, from a cognitive point of view, needs to be highly marked to establish special processing conditions for interpretation. Its saliency is one of its defining features. This constitutive markedness is enacted through various linguistic and paralinguistic devices in language, like quotation marks and metalinguistic verbs and descriptors. Metalinguistic sentences create and materialize specialized knowledge through linguistic means. Metalinguistic information, in order to be relevant and informative, cannot be inferable from previous knowledge or from regular language competence. Regular words and technical terms differ, among other things, in that terms require volitional creation and introduction into generalized

use for specific purposes, unlike the everyday lexicon, and also in that terms evolve rapidly by means of conscious and volitional consensus attained through metalinguistic devices.

Conventional definitions are by no means the only possible kind of metalinguistic sentences and they are not even the most common ones. Conventional dictionary definitions represent highly processed compilations of meaning that cannot fully realize the spectrum of possible information about a sublanguage's denotation, connotation, usage conditions and meaning that metalanguage can convey. The information supplied by Explicit Metalinguistic Operations can be non-complete and heterogeneous but still highly relevant for establishing a coherent and complete conceptual or terminological system.

In chapter III, the notion of EMO is defined in contrast to definition-like structures as a more adequate and formalized descriptive object for metalanguage in specialized contexts. EMOs perform two basic functions in the code specification inherent in all metalinguistic acts; informing on the standing conventions for the system and directing the interpretation or elaboration of linguistic messages done through that code. As specified here, EMOs consist of three basic elements:

(A) a single or multiple lexical unit that stands in autonymical condition as the subject of the metalinguistic predication;

(B) an informational segment in the sentence that provides information or instructions on the self-referential unit in A; and

(C) lexical, punctuation or paralinguistical elements acting as a predicative articulation between A and B, and functioning as flagging devices for the metalinguistic nature of the utterance.

The inventory of elements acting as markers/operators is in principle finite and our corpus analysis has brought forward a number of them (with statistics of use) that have been used to construct our EMO Corpus of metalinguistic speech acts and to implement the first extraction phase of our MOP system. Realizations and general kinds of possible informational segments are varied, but some common categories are listed here, among them: unexpressed existential variables, full clauses with intensional or denotational references, etc. Other types of information that EMOs can carry but that the MOP system does not attempt to extract, are: agentivity, hedging, illocutionary force, participants in the exchange, etc.

The MOP system is described in detail in chapter IV, and its evaluation makes up the bulk of chapter V. MOP is coded in Python, using the NLTK development platform, and employs XML standards for its output and operational data files. MOP applies standard pre-processing techniques (tokenization, POS tagging and partial parsing) on specialized, non-structured texts, and some of its tasks resemble Information Extraction tasks, but some significant differences exist between them, both with regard to goals, methods and delivered information.

MOP has two stages: Candidate Extraction and a Predicate Processing. It first locates and extracts candidate metalinguistic fragments employing as lexical triggers patterns of lexis and punctuation compiled from extensive analyses of EMO corpora; The MOP system's disambiguates metalinguistic sentence candidates using two different approaches, one involving machine-learning algorithms and another one using pattern-matching over collocations obtained from corpora. After obtaining EMO sentences, the MOP system performs semantic labeling of the chunks, and parses the ones that might fulfill desirable semantic roles into a database structure by using heuristic rules derive from manual analysis of such sentences and of their lexical markers, as well as from relevant semantic frames from the FrameNet project. The MOP system carries out this processing avoiding overly complex and sophisticated NLP machinery, and consequently presents limited parsing and co-reference resolution capabilities. Final output for this processing is a Metalinguistic Information Database (MID) containing a three-entry record of informational segments, markers/operators and autonyms.

In chapter V, the overall evaluation of the MOP system follows standard IE and IR metrics like Precision, Recall and F-measure. It uses manually-compiled Golden standards of three corpora from different domains and with different characteristics: the Sociology corpus described before, an online Histology textbook and a sample of the MedLine abstract database of Bio-Medical papers.

On the Candidate Extraction phase using all EMOs identified in our test corpora (whether their distinctive marker/operators pattern were included or not in the extraction list used by MOP), gave excellent precision but low recall rates in our evaluation runs. We believe a non-exhaustive list of extraction patterns is the cause of such difference in metrics. Using only tagged examples that contained patterns that could actually be recognized, precision was maintained at high levels (0.94 and more), but recall increased substantially (0.79 and more), with F-measures at 0.87 for β of 1 to balance out P & R.

Tests with learning algorithms trained on a subset of the EMO Corpus yielded good metrics also, but were otherwise inconclusive with regard to algorithm and feature set baseline superiorities. With the Sociology corpus, Maximum Entropy algorithms using as features a single word form to left and right of markers/operators presented the following numbers: P = 0.9, R = 0.7, while best results for the Histology corpus were attained with a Bayesian network using three word forms at each side of the markers (P=0.9, R=0.84). In short, although results were very good for this classification task, they were inconsistent and inconclusive, and merit further research. The expectation of improved performance with POS contexts and a wider context did not bore out in our tests.

Evaluations of the Predicate Processing task were also very good if compared with similar IE tasks. The global performance of the MOP system ranged from near 0.7 to a perfect score in the autonym identification task in the small MedLine Sample. The best overall F-measure was achieved both in that small corpus and in the Sociology one, at 0.77. The two different grammatical contexts where entity identification was performed gave also a wide range of metrics. Autonym identification was the most successful (averaging 0.9 P and 0.91 R), while Informational segments averaged 0.8.5 P and 0.8 R. Overall, these are very good numbers (regardless of domain involved), albeit for a very simple extraction with few database slots and a very basic structure (with almost no inference mechanisms). Improvements on coverage and precision are possible, but the system will need to increase its processing complexity significantly, adding deeper parsing and a coreference and anaphora resolution module.

In Chapter VI we argued that conventional terminological dictionaries and mental or computational specialized lexicons can be seen as static repositories of the condensed default lexical knowledge of the terms used by a domain-centered community of expert speakers of a sublanguage. The Metalinguistic Information Database generated with the MOP system contains the multi-textured real-time data produced in the discourse of research papers and technical documents. This sets MIDs apart from conventional lexical databases, which fail to represent a sublanguage's dynamicity and open-endness. A MID is not really a lexicographic artifact, since it can be viewed as an *anti-dictionary*, a listing of exceptions, special contexts and specific usages of lexical and terminological items where meaning, value or pragmatic conditions have been spotlighted by discourse for cognitive reasons. Terminological data in MIDs can be more specific and might be better suited for the interpretation of certain texts or utterances than that of lexical knowledge bases and lexicons.

MIDs are semi-structured resources that need to be further processed to turn into functional taxonomies and lexicons. Nevertheless MIDs can have many applications for research and technological development; among them: update and fine-tuning of lexicons and ontologies, neology detection, non-default information repositories for inference engines, research and didactics of specialized discourse and scientific activities. Although MIDs are useful theory-neutral data structures, some of the information contained in them is not typed and maintains some of its linguistic form, which causes MIDs to present difficulties for their integration and update. Nevertheless, for that same reason they can be considered accumulative records of conceptual and terminological change. Although less sophisticated and structured than LKB and ontologies, with a non-exhaustive but cognitively important coverage of a sublanguages lexicon, we believe that MIDs —and the MOP system designed to generate them— constitute an advance in the state of the art of Human Language Technology, as well as a novel and

invaluable resource for empirical research in the production and evolution of consensus-based technical and scientific knowledge. There are still a few possible enhancements to implement and some avenues to explore, but we believe our main goals have been attained. By empirically approaching the ways in which linguistic expression interacts with knowledge-buildup through the powerful mechanisms of metalanguage and the group dynamics of expert peers we have grounded a proof-of-concept computational application capable of automatically exploiting EMOs for research and development. In doing so, we believe we have accomplished in a significant way the interdisciplinary goals we set out for this dissertation.

VIII BIBLIOGRAPHY

Achinstein, P. (1964) *On the Meaning of Scientific Terms.* Journal of Philosophy, 61

Achinstein, P. (1968) *Terminos Teóricos.* Filosofía de la ciencia: Teoría y Práctica, Siglo XXI editores, México D. F., 1989

ACQUILEX final report available at: http://www.cl.cam.ac.uk/Research/NL/acquilex/

Ahmad, K. (1996) *A Terminology Dynamic and the growth of knowledge*. TKE 96 Terminology and Knowledge Engineering. Frankfurt INDEKS-Verlag 1996 p. 1-11

Alarcón R., Sierra G. (2002). *Hacia la extracción automática de conceptos.* In Proc. VIII Simposio Iberoamericano de Terminología RITerm. Cartagena, Colombia, 2002.

Amsler, R. & White, J. (1979) *Development of a computational methodology for deriving natural semantic structures via analysis of machine-readable dictionaries*. Technical Report TR MCS77-01315, Linguistics Research Center, University of Texas.

Anscrombre, J.C. & Ducrot, O. (1983) *La Argumentación en la Lengua.* Gredos, Madrid.

Asher, N. & Lascarides, A. (1996) *Lexical desambiguation in a discourse context*. Lexical Semantics (The problem of polysemy). eds. J. Pustejovsky & B. Boguraev, Oxford University Press.

Auger, A. (1997) *Repérage de énoncés d'intérêt définitoire dans les bases de donées textuelles.* Thése de doctorat, Université de Neuchatêl.

Austin, J.L. (1962) *How to do things with words.* Oxford University Press [versión castellana: Cómo hacer cosas con las palabras. Barcelona, Paidós. 1990]

Bach, K. and Harnish, R. (1979) *Linguistic Communication and Speech Acts.* MIT Press

Baker, Collin F., Fillmore, Charles J., and Lowe, John B. (1998): The Berkeley FrameNet project. In Proceedings of the COLING-ACL, Montreal, Canada.

Basili R., Pazienza M-T , Velardi P. (1996). *A context driven conceptual clustering method for verb classification*, In: Corpus processing for lexical acquisition, MIT Press, Cambridge, MA, 1996

Baudet (1991) *Editologie.* Cahiers de Linguistique Sociale, Num. 18, 1991

de Beaugrande R. & Dressler W. U. (1981) *Introducción a la lingüística del texto.* Ariel, Barcelona, 1997

Benveniste E. (1969) *Génesis del término "scientifique"*. Problemas de lingüística general, Vol. II, Siglo XXI. México 1977

Berger, A., S. Della Pietra et al., (1996). *A Maximum Entropy Approach to Natural Language Processing.* Computational Linguistics, vol. 22, no. 1.

Bierwisch M. & Kiefer F. (1969) *Remarks on definitions in general language*, Studies in Syntax and Semantics. D. Riedel Humanities Press, NY

Briscoe T. , de Paiva V., Copestake A., et al., (1993). *Inheritance, Defaults and the Lexicon.* Studies in Natural Language Processing. Cambridge University Press.

Bronckart, J.-P. (1985) *Le Fonctionnement des Discours,* Neuchatel-París, Delachaux-Niestlé Éditeurs.

Boguraev, B. and Levin, B. (1993) *Models for Lexical Knowledge Bases*, in J. Pustejovsky, ed., Semantics and the Lexicon, Kluwer, Dordrecht.

Boguraev, B. and Pustejovsky, J. (1996). *Issues in Text-based Lexicon Acquisition,* In: Corpus processing for lexical acquisition, MIT Press, Cambridge, MA, 1996.

Budin, G. (1990) *Scientific knowledge structures.* TKE'90 Terminology and Knowledge Engineering. Proceedings from TKE'90. Indeks Verlag, Frankfurt.

Burks, A. W., Goldstine, H. H., and von Neumann, J. (1963) *Preliminary discussion of the logical design of an electronic computing instrument.* In Taub, A. H., editor, John von Neumann Collected Works, The Macmillan Co., New York, Volume V, 34-79.

Cabré, T., Estopà, R., Vivaldi, J. (2001). *Automatic term detection: a review of current systems*. In "Recent Advances in Computational Terminology". Ámsterdam, Philadelphia: John Benjamins. pg. 53-87.

Cabré, M. T. (1998) *Elementos para una teoría de la terminología: hacia un paradigma alternativo*. El Lenguaraz, 1 (1) 59-78. B.B.A.A.

Cabré, M. T. (1999) *La Terminología: Representación y comunicación. Elementos para una teoria de base comunicativa y otros artículos.* Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Barcelona

Cabré, M. T. (1993) *La terminología: teoría, metodología y aplicaciones,* Ed. Antártida/Empúries, Barcelona.

Califf, M. E. and R. J. Mooney. (1999). Relational learning of pattern-match rules for information extraction. In Proceedings of AAAI99, pages 328–334.

Candel, D. (1993) *Le Discours Définitoire: variations discursives chez les scientifiques.* Parcours Linguistiques de Discours Spécialisés, Peter Lang, Berna.

Carnap, R. (1934). *The Logical Syntax of Language*. Routledge and Kegan, Londres 1964.

Cartier, E. 1998. *Analyse Automatique des textes: l'example des informations définitoires.* RIFRA'98. Sfax, Tunisia.

Chieu, H., Ng, H., & Lee, Y. (2003) *Closing the Gap: Learning-Based Information Extraction Rivaling Knowledge-Engineering Methods.* Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03). Sapporo, Japan.

Chinchor, N. (1998) *OVERVIEW OF MUC-7/MET-2.* Proceedings of the Seventh Message Understanding Conference.

Chodorow, M. S.; Byrd, R.; and Heidorn, G. (1985). *Extracting semantic hierarchies from a large on-line dictionary.* Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics, Chicago, 299--304.

Chomski, N. (1993) *Language and Thought.* Moyer Bell , Londres

Clark, H. H. (1998) *Communal Lexicons.* In: Context in language learning and language understanding, eds. K. Malmkjoer & J. Williams. Cambridge University Press.

Clark, H. H. (1996) *Using language.* Cambridge University Press.

Clark, H.H. and Wilkes-Gibbs, D. (1986). *Referring as a collaborative process.* Cognition, 22, 1-39

Condamines, A and Rebeyrolle, J. (2001). *Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB): Method and Results.* In Bourigault, Didier, Christian Jacquemin and Marie-Claude L'Homme (eds.), Recent Advances in Computational Terminology. John Benjamins, 2001.

Copestake, A.. Sanfilippo, A., Briscoe, T. and de Pavia, V. (1993) *The ACQUILEX LKB: An introduction.* In: Inheritance, Defaults and the Lexicon. Cambridge University Press.

Coseriu, E. (1986) *Principios de semántica estructural*. Madrid. Gredos,

Coseriu, E. (1967) *Teoría del Lenguaje y Lingüística General.* Ed. Gredos, Madrid ("Sistema, norma y habla, 1952").

Cuouto, J., Crispino et al. (1999) *Estructuración de Índices Gramaticales y Léxicos para la Extracción y Recuperación de Información.* Procesamiento del Lenguaje Natural. SEPLN. No. 25

Darien, S. (1981) *The role of definitions in scientific and technical writing: forms, functions and properties*. English Language Research Journal 2: 41-56

Davidson L., Kavanagh K., Mackintosh I., Meyer I. & Skuce D. (1998) *Semi-automatic Extraction of Knowledge-rich Contexts from Corpora*. In Computerm'98: First Workshop on computational terminology, COLING-ACL'98 Montreal.

Desclés, J-P. Et al. (1997) *Textual Processing and Contextual Exploration Method.* In CONTEXT'97, Río de Janeiro, Brasil

Dolan, William B., L. Vanderwende, et al. (1993). *Automatically Deriving Structured Knowledge Base from On-line Dictionaries*. Proceedings of the Pacific Association for Computational Linguistics, April 21-24, 1993, Vancouver, British Columbia.

Dubois, D. (1999) *Le lexique, fixateur de représentations et producteur d'ontologie.* (forthcoming). Seminario de Terminología Teórica. January 1999, IULA/UPF, Barcelona

Eco, U. (1997) *Kant y el Ornitorrinco.* Lumen, Barcelona.

Estopà, R., Vivaldi, J. and Cabré, T. (1998) *Sistemes d'extracció automàtica de (candidats a) termes: Estat de la qüestió* Serie Informes 22, IULA, Universitat Pomeu Fabra. Barcelona.

Fetzer, A. 2003. *Reformulations and Common Ground.* Eight International Pragmatics Conference of the International Pragmatics Association (IPrA). Panel on Lexical Markers of Common Ground. Toronto, 2003.

Feyerabend, P. (1965) *Problemas del empirísmo,* en Filosofía de la ciencia.... (1989)

Fillmore, Ch. (1969) *Types of lexical information.* Studies in Syntax and Semantics. D. Riedel Humanities Press, NY

Fisher, D., S. Soderland, J. McCarthy, F. Feng, and W. Lehnert. (1995). *Description of the UMass system as used for MUC-6.* In Proceedings of MUC-6, pages 127–140.

Flowerdew, J. (1992) *Definitions in science lectures.* Applied Linguistics (13)2

Flowerdew, J. (1991) *Pragmatic modifications on the 'representative' speech act of defining.* Journal of Pragmatics 15. [pags. 253-264]

Fowler R., Hodge B., et.al. (1979) *Language and Control*. Routledge & Kegan Paul, London.*[Lenguaje y Control*, Fondo de Cultura Económica, México, 1983]

Fox, N. (1995) *Intertextuality and the Writing of Social Research.* Electronic Journal of Sociology 1, 2. <http://www.sociology.org/content/vol001.002/fox_d.html>

Fujii, A. and Ishikawa, T. (2004) Summarizing Encyclopedic Term Descriptions on the Web. Proceedings of the 20th International Conference on Computational Linguistics (to appear)

Fujii, A. and Ishikawa, T. (2000) *Utilizing theWorldWideWeb as an encyclopedia: Extracting term descriptions from semi-structured texts.* In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pages 488–495.

Gambier, I. (1991) *Pressuposes de la Terminologia.* Cahiers de Linguistique Sociale, Num. 18.

Garrod, S. & Anderson, A. (1987). *Saying what you mean in dialogue: A study in conceptual and semantic co-ordination*. Cognition, 27, 181-218.

Geeraerts, D. (1985) *Les donnés stéréotypiques, prototypiques et encyclopédiques dans le dictionnaire.* Cahiers de Lexicologie. Vol. XLVII

Gentilhomme, Y. (1994) *L'éclatement du signifié dans les discours techno-scientifiques*. Cahiers de Lexicologie, Vol. LXIV-1, p.5-53.

González, M-I & Soláns, M-A (1997) *Do scientific writers criticize?* Pragmalingüística 3-4 (1995-1996) Universidad de Cádiz

Green, G. (1988) *Pragmatics and natural language understanding.* Lawrence Erlbaum Associates, Hillsdale (N.J.)

Grice, H. P. (1991) *Studies in the way of words.* Harvard University Press. Cambridge, Mass.

Grisham, R. (1997) *Information Extraction: Techniques and Challenges*. In Maria Teresa Pazienza, editor, Information Extraction. Springer-Verlag, Lecture Notes in Artificial Intelligence, Rome

Gross, A. (1991) *The Rhetoric of science,* Harvard University Press, London

Gildea, D. and Jurafsky, D. (2002). *Automatic labeling of semantic roles*. Computational Linguistics, 28(3):245--288.

Gülich, E. & Kotschi, T. (1995) *Discourse Production in Oral Communication,* en Aspects of Oral Communication, W. De Gruyter: Berlin-New York p. 30-66

Guterl, F. (1996) *Riddles in the sand*. Discovery, Noviembre, 1996.

Gutierrez , S. (1989) *Variaciones sobre la atribución.* Colección Contextos #5. Centro de Estudios Metodológicos e Interdisciplinares, Universidad de León.

Gutierrez , S. (1988) *Del uso metalingüístico*, Archivum, T. XXXVII-XXXVIII, Universidad de Oviedo. p. 5-19

Haensch, Wolf, Ettinger *et alia.* (1982) *La Lexicografía: de la lingüística teórica a la lexicografía práctica*. Gredos, Madrid.

Halliday, M. A. K. & Martin, J. R. (1993) *Writing science: literacy and discourse power*, London, The Falmer Press.

Halliday, M.A.K. (1985) *An introduction to Functional Grammar,* Edward Arnold, Londres, 1994.

Harris, Z. (1991) *A Theory of language and information: a mathematical approach.* Clarendon Press, Oxford.

Harris, Z. (1968) *Mathematical Structures of Language*. R. E. Krieger, New York, 1979

Hearst, M. (1998) *Automated discovery of wordnet relations.* In Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA

Hearst M., Schütze H. (1993) *Customizing a Lexicon to Better Suit a Computational Task.* In: Proc. ACL SIGLEX Workshop "Acquisition of Lexical Knowledge from Text".

Hempel, C. (1973) *El significado de los términos teóricos: Una crítica a la concepción empirísta estándar,* In: L. Olivié & A.R. Pérez Ransanz (comp.) Filosofía de la ciencia: teoría y observación. México: Siglo XXI, 1989.

Hempel, C. (1958) *El dilema teórico: un estudio sobre la lógica de la construcción de teorías.* In En: L. Olivié & A.R. Pérez Ransanz (comp.) Filosofía de la ciencia: teoría y observación. México: Siglo XXI, 1989

Hempel, C. (1952) *Fundamentos de la formación de conceptos en ciencia empírica,* Alianza Editorial, Barcelona, 1988.

Hermas, A (1991) *Sociologuie des vocabulaires scientifiques et techniques,* en Cahiers de Linguistique Sociale, N. 18, 1991

Hjelmslev, L. (1943) *Prolegómenos a una teoría del lenguaje.* Gredos, Madrid 1974

Hunston, S. (1996) *Evaluation and ideology in scientific writing,* en Register analysis, Theory and Practice, Printer Publishers, London.

Hyland, K. (1998) *Persuasion and context: The Pragmatics of Academic Metadiscourse*. Journal of Pragmatics 30 (1998) 437-455

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford University Press.

Jacquemin, Ch. (2001). *Spotting and discovering terms through NLP.* The MIT Press

Jakobson, R. (1980) *El Metalenguaje como problema lingüístico.* El Marco del Lenguaje. FCE, México 1988

Jakobson, R. (1963) *Ensayos de Lingüística General*. Ariel, Barcelona. 1984.

Kageura, K. (2002). *The dynamics of terminology. A descriptive theory of term formation and terminological growth.* John Benjamins.

Kenyon, R. (1994) *On the use of Quotation Marks*. A Review of General Semantics, General Semantics Bulletin Vol. 51 No 1, Spring

Kigarriff, A. (2001) *Generative Lexicon Meets Corpus Data: The Case of Nonstandard Word Uses.* The Language of Word Meaning.Cambridge University Press, eds. Bouillon, P. & Busa, F.

Kittredge, R. (1982) *Variation and Homogeneity of sublanguages.* Sublanguage: studies of language in restricted semantic domains. Kittredge & Lehrberger, eds. Berlin De Gruyter

Klavans, J. and S. Muresan. (2001). *Evaluation of the DEFINDER System for Fully Automatic Glossary Construction*, proceedings of the American Medical Informatics Association Symposium 2001

Kleiber, G. (1990) *La semántique du prototype.* Presses Universitaires de France.

Kotschi, T. (1986) *Procedes d'evaluation et de commentaire metadiscursif comme estrategies interactives.* CLF Cahiers de Lange Francaise, 7. (207-230)

Kruijff, G. & Schaake, J. (1995) *Discerning relevant information in discourses using TFA.* CLIN V: Papers from the Fifth CLIN Meeting

Kuhn, T. (1962) *La Estructura de las revoluciones científicas.* Fondo de Cultura Económica. Madrid, 1971

Lachenmayer, Ch. (1971) *El lenguaje de la sociología,* Editorial Labor, Barcelona, 1976.

Lakoff, G. (1987) *Women, fire, and dangerous things.* University of Chicago Press. Chicago, 1996

Landheer, (1989) *L'importance des relationes hyponymiques dans la descripción lexicographique,* Actes du XVIIIe Congrès International de Linguistique et de Philologie Romanes, Université de Trèves. Tübingen: Max Niemeyer Verlag, 139-151; En Pearson (1998).

Lara, L. F. (1997a) *Por una nueva Teoría del Signo.* Varia Lingüística y Literaria-50 años del CELL, El Colegio de México, México DF

Lara, L. F. (1997b) *Teoría del Diccionario Monolingüe* Estudios de Lingüística y Literatura XXXIII, El Colegio de México, México DF

Lara, L. F. (1989) *Une critique du concept de metalangage.* FOLIA LINGUISTICA (Acta societatis Linguistica Europaeae. Tomus XXIII /3-4, Mouton/De Gruyter

Lascarides, A. and Copestake A. (1995) *The Pragmatics of Word Meaning*, Proceedings of the AAAI Spring Symposium Series: Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity, Stanford CA.

Lauer, M. (1994) *Conceptual Association for Compound Noun Analysis.* In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Student Session, Las Cruces, NM.

Lazerowitz, M. & Ambrose, A. (1985) *Necesidad y Filosofía.* Cuadernos del Instituto de Investigaciones Filosóficas #42. Universidad Nacional Autónoma de México, México D.F.

Leech, G. (1980) *Explorations in semantics and pragmatics*. Amsterdam, John Benjamins

Lehnert, W. (1991) *Symbolic/Subsymbolic Sentence Analysis: Exploiting the Best of Two Worlds* In Advances in Connectionist and Neural Computation Theory, Vol. I. John A. Barnden and Jordan Pollack (eds.), Ablex Publishing, Norwood, New Jersey, 135-164.

Lehnert, W. C. Cardie, D. Fisher, J. McCarthy, E. Riloff, & S. Soderland. (1994) *Evaluating an Information Extraction System.* Journal of Integrated Computer-Aided Engineering. 1(6)

Lemke, J. (1998) *Multiplying meaning: visual and verbal semiotics in scientific text*. en "Reading Science: Critical and functional perspective on discourses of science". ed. Martin, J.R. & Robert Veel, Routledge

Liddicoat, A. (1997) *The function of the conditional in French scientific writing*. Linguistics 35-4, 767-780. Walter de Gruyter

Lin, D. and Nalante, Inc. (1998). *Using Collocation Statistics in Information Extraction.* Proceedings of the Seventh Message Understanding Conference.

Loper, E. and Byrd, S., (2002) *NLTK: The Natural Language Toolkit.* ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics

Loper, E., (2003) *NLTK: Classification.* Draft technical report at http://nltk.sourceforge.net.

Losee, R. & Haas, S. (1998) *Sublanguage Terms: Dictionaries, Usage and Automatic Classification.* Journal of the American Society for Information Science, 46(7).

Lucy, J. (1993) *Reflexive language and the human disciplines.* Reflexive Language (Reported speech and metapragmatics). Cambridge University Press, Cambridge.

Luscher, J-M (1994) *Les marques de connexion: des guides pour l'interprétaction.* Langage et Pertinence: référence temporelle, anaphore, connecteurs et métaphore. Presses Universitaires de Nancy

Lyons, J. (1981) *Lenguaje, significado, contexto.* Paidós comunicación, Barcelona 1991.

Lyons, J. (1977) *Semantics*. Cambridge University Press

Malaisé, V., Zweigenbaum, P. and Bachimont, B. (2004). *Detecting Semantic Relations between Terms in Definitions*. COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology. Geneva, Switzerland.

Malinowski, B. (1944) *Una teoría científica de la cultura*. Barcelona, Serpe, 1984.

Manning, Ch. (1993) *Automatic acquisition of a large subcategorization dictionary from corpora*, In Proceedings of the 31st ACL, Columbus, OH.

Marcus M., Santorini B., Marcinkiewicz M-A (1993) *Building a large annotated corpus of English: the Penn Treebank*. Computational Linguistics, vol. 19, 1993

Martin, J.R. (1992) *English text: system and structure,* J. Benjamins, Amsterdam.

McEnery, T. and Wilson, A. (1996) *Corpus Linguistics.* Edinburgh University Press

Meyer, I. (2001). *Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework.* In Bourigault, Didier, Christian Jacquemin and Marie-Claude L'Homme (eds.), Recent Advances in Computational Terminology. John Benjamins, 2001.

Mignolo, W. (1986) *Elementos para una teoría del texto literario.* Teoría del texto e interpretación de textos, UNAM, México

Miller G. A. et al., (1990) *WordNet: an online lexical database.* International Journal of Lexicography, 3:235-244.

Muslea, I.. (1999) *Extraction patterns for information extraction tasks: A survey.* In AAAI-99 Workshop on Machine Learning for Information Extraction, American Association for Artificial Intelligence (AAAI).

Myers, G. (1992) *"In this paper we report...": speechs acts and scientific facts.* Journal of Pragmatics 17 North Holland/ Elsevier. Amsterdam

Nigam, K., Lafferty, J., and McCallum, A. (1999) *Using Maximum Entropy for Text Classification*, IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61-67

Nirenburg, S. & Levin, L. (1991) *Syntax-driven and Ontology-driven Lexical Semantics.* Lexical Semantics and Knowledge Representation. First Siglex Workshop; Pustejovsky-Bergler, eds. Springer-Verlag.

Nyan, T. (1996) *Metalinguistic Operators: Why they matter to linguistic theor*y, Le discourse: cohérence et connexion (Actes du colloque international). Etudes Romanes 35, Copenhagen, Museum Tusculanum Press

Ogden, C. K. & Richards, I. A. (1929), *The Meaning of Meaning*, Routledge & Kegan Paul LTD, Londres, 1969.

Pascual, E. & Péry-Woodley, M-P (1997a) *Modèles de texte pour la définition.* Actas de Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue, 15-16 abril, Avignon

Pascual, E. & Péry-Woodley, M-P (1997b) *Définition et Action dans les Textes Procéduraux.* Atelier "Texte et Communication". Programme Recherches Sciencies Cognitives Toulouse. Toulouse.

Pearson, J. (1996) *The Expression of definitions in Specialised Texts: A Corpus-based Analysis.* Euralex'96 proceedings II, VII Euralex International Congress on Lexicography. Göteborg, Suecia

Pearson, J. (1998) *Terms in Context,* John Benjamins (Studies in Corpus Linguistics) Vol 1., Amsterdam

Péry-Woodley, M-P. (1998) *Signalling in written text: a corpus-based approach.* Workshop "Discourse Markers and Discourse Relations", COLING'98, 79-85, Montreal

Phillips, M. (1985). *Aspects of Text Structure: an investigation of lexical organization of text.* North-Holland Linguistic Series 52. Elsevier.

Pickering,M. & Garrod, S. (forthcoming): *Toward a mechanistic Psychology of dialogue.* Behavioral Brain Sciences.

Pollard, C. & Sag, I. (1995) *Head-Driven Phrase Structure Grammar,* University of Chicago Press.

Popper, K. (1959) *La Lógica de la Investigación Científica.* Técnos, Madrid 1977

Powell, T.; Srinivasan, S.; Nelson, S. J.; Hole, W..; Roth, L.; Olenichev, V. (2002) *Tracking Meaning Over Time in the UMLS Metathesaurus.* In: Kohane, Issac S., editor. Biomedical Informatics: One Discipline. Proceedings of the Annual Symposium of the American Medical Informatics Association; 2002 Nov 9-13; San Antonio, TX. Philadelphia: Hanley & Belfus, Inc.; 2002. p. 622-626.

Pustejovsky, J., J. Castaño, R. Saurí, A. Rumshisky, J. Zhang, W. Luo. (2002a), *Medstract: Creating Large-scale Information Servers for Biomedical Libraries.* ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain. Philadelphia, PA.

Pustejovsky J., A. Rumshisky and J. Castaño. (2002b) *Rerendering Semantic Ontologies: Automatic Extensions to UMLS through Corpus Analytics*. LREC 2002 Workshop on Ontologies and Lexical Knowledge Bases. Las Palmas, Canary Islands, Spain.

Pustejovsky, J. (1999) Computational Lexicons, in *The MIT encyclopedia of the cognitive sciences* edited by Robert A. Wilson, Frank C. Keil.

Pustejovsky, James (1995). *The Generative Lexicon.* MIT Press.

Putman H. (1975) *The Meaning of "Meaning"*. Language, Mind and Knowledge. University of Minessota Press.

Putman H. (1960) *Lo que las teorías **no** son*, In En: L. Olivié & A.R. Pérez Ransanz (comp.) Filosofía de la ciencia: teoría y observación. México: Siglo XXI, 1989

Quine, W. V. O. (1962) *Dos dogmas del empirismo.* Desde un punto de vista lógico. Ariel, Barcelona

Quine, W. V. O. (1960) *Word and Object.* The MIT press. Massachsetts Institute of Technology, Mass. 1992.

Quirk, R. & Greenbaum, S. (1973) *A University Grammar of English.* Longman, 1997

Ratnaparkhi A. (1997). *A Simple Introduction to Maximum Entropy Models for Natural Language Processing*, TR 97-08, Institute for Research in Cognitive Science, University of Pennsylvania

Rebeyrolle, J.& Péry-Woodley, M-P (1998) *Repérage d'objets textuels fonctionnels pour le filtrage d'information: le cas de la définition*. Tunisia, Rifra '98

Reichenbach, C. (1947) *Elements of symbolic logic,* Free Press, NY.

Rey, A. (1995) *Essays on terminology,* New York, John Benjamins Publishing

Rey-Debove, J. (1978) *Le Métalangage.* Le Robert, Paris.

Riegel, M. (1987) *Definition directe et indirecte dans le langage ordinaire: les énoncés définitoires copulatifs.* Langue Francaise 73.

Riloff, E. (1993) *Automatically Constructing a Dictionary for Information Extraction Tasks* Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93) , AAAI Press/The MIT Press, pp. 811-816

Riloff, E. and Lehnert, W.G. (1994). *Information Extraction as a Basis for High-Precision Text Classification*. ACM Transactions on Information Systems TC-24: 296-333.

Riloff, E. and Jones, R. (1999). *Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping* Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99).

Rish, I., (2001). *An empirical study of the naive Bayes classifier*, in International Joint Conference on Artificial Intelligence workshop on "Empirical Methods in AI".

Robinson, R. (1950) *The Definition.* Oxford. Clarendon Press.

Rodríguez, C. (forthcoming) *A common ground for knowledge through knowledge of language: A computational research of consensus-based meaning in scientific papers*. In "Lexical Markers of Common Grounds", Studies in Pragmatics. Elsevier.

Rodríguez, C. (2004b) *Metalinguistic Information Extraction for Terminology.* 3rd International Workshop on Computational Terminology (CompuTerm 2004) Coling 2004. Geneve

Rodríguez, C. (2004a) *Mining metalinguistic activity in corpora to create lexical resources using Information Extraction techniques: the MOP system*. ACL 2004, Barcelona.

Rodríguez, C. (2003) *A common ground for knowledge through knowledge of language (Researching research using metalinguistic Information Extraction)*; Panel on Lexical Markers of Common Grounds. 8th International Pragmatics Conference Toronto; July, 2003

Rodríguez, C. (2002) *Automatic Extraction of Non-standard Lexical Data for a Metalinguistic Information Database,* in: Computational Linguistics and Intelligent Text Processing Third International Conference, CICLing 2002, Mexico City, Mexico. Series: Lecture Notes in Computer Science. VOL. 2276 Springer-Verlag.

Rodríguez, C. (2001b) *Las características del conocimiento especializado y la relación con el conocimiento en general*" *& Principios metodológicos de la propuesta teórica (II)*, in La Terminología Científico-Técnica: Reconocimiento, análisis y extracción de información formal y semántica. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Barcelona.

Rodríguez, C. (2001a). *Parsing Metalinguistic Knowledge from Texts*, Selected papers from CICLING-2000 Collection in Computer Science (CCC); National Polytechnic Institute (IPN), Mexico

Rodríguez, C. (2000) *Extraction of knowledge about terms from metalinguistic activity in texts.* In: A. Gelbukh (ed.): Proceedings of the Conference on Intelligent text processing and Computational Linguistics, CICLING-2000. Instituto Politécnico Nacional, Mexico City, Mexico.

Rodríguez, C. (2000) *Extraction of knowledge about terms from metalinguistic activity in texts.* In: A. Gelbukh (ed.): Proceedings of the Conference on Intelligent text processing and Computational Linguistics, CICLING-2000. Instituto Politécnico Nacional, Mexico City, Mexico.

Rodríguez, C. (1999b) *Explicit Metalinguistic Operations in specialized discourse: The construction of lexical meaning in theoretic science.* In P. Sandrini (ed.): Terminology and Knowledge Engineering TKE'99, Innsbruck, Austria.

Rodríguez, C. (1999a) *Operaciones Metalingüísticas Explícitas en textos de especialidad.* (Explicit Metalinguistic Operations in Special Domain Texts) Master's Dissertation. Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. Barcelona.

Rondeau, G. (1984) *Introduction à la terminologie.* Québec: Gaëtan Morin

Rorty, R. (1997) *Thomas Kuhn, Rocks and the Laws of Physics*. Common Knowledge, Spring.

Rorty, R. (1989) *Contingency, Irony, and Solidarity*. Cambridge University Press. [*Contingencia, Ironía y Solidaridad.* Paidós Básica, Barcelona (1991)]

Sager N.& Hirschman, H. (1982) *Automatic information formatting of a medical sublanguage.* In Kittredge & Lehrberger, eds. Berlin De Gruyter.

Sager N, Friedman C, Lyman MS. (1987). *Medical Language Processing: Computer Management of Narrative Data.* Menlo Park, CA: Addison-Wesley.

Sager, J-C & Kageura, K. (1994) *Concept classes and conceptual structures.* Terminology and LSP linguistics. Studies in Specialized Vocabularies and Texts, ALFA 7/8

Sager, J-C (1990) *A Practical course in terminology processing.* John Benjamins, Amsterdam

Sager, J-C (2001) *Essays on Definition.* John Benjamins, Amsterdam.

Sager, J-C., Dungworth, D. & McDonald, P. (1980) *English Special Languages (Principles and practice in Science and Technology)*. Oscar Brandsteter Verlang, Wiesbaden,

Saussure, F. de (1916) *Curso de lingüística general.* Planeta-Agostini, Barcelona 1985

Searle, J. (1969) *Speech Acts. An essay in the philosophy of language.* Cambridge University Press, Cambridge

Sierra G., Alarcón R. (2002). *Identification of recurrent patterns to extract definitory contexts*. In Lecture Notes in Computer Science. Num. 2276, Springer Verlag. pp. 436-438

Silverstein, M. (1993) *Metapragmatic discourse and metapragmatic function*, Reflexive Language (Reported speech and metapragmatics) Ed. J. Lucy. Cambridge University Press

Soderland, S. D. Fisher, and W. Lehnert, (1997) *Automatically Learned vs. Hand-crafted Text Analysis Rules*, CIIR Technical Report

Sosteric, M. (1996a) *Interactive Peer Review: A Research Note,* Electronic Journal of Sociology. 2, 2

Sosteric, M. (1996b) *Electronic Journals: The Grand Information Future?*. Electronic Journal of Sociology: 2, 2. http://www.sociology.org/

Sperber, D. & Wilson, D. (1986) *La Relevancia.* Lingüística y conocimiento, Madrid: Visor. 1994.

Tarski, A. (1944) *La concepción semántica de la verdad y los fundamentos de la semántica.* Nueva Visión, Buenos Aires. 1972

Temmerman, R. (1997) *Questioning the univocity ideal. The difference between socio-cognitive terminology and traditional terminology.* Hermes, Journal of Linguistics, 18: 51-90.

Teufel, S. and Moens, M. (2002). *Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status.* In Computational Linguistics, 28 (4), Dec. 2002.

Thoiron, Ph. & Béjoint, H. (1991) *La place des reformulations dans les textes scientifiques*. META XXXVI, 1.

Trimble, L. (1985) *English for science and technology: a discourse approach*. Cambridge University Press, Cambridge.

Urresti, M.V. (1994) *La definición en un contexto científico-técnico.* en Lenguas para fines específicos ; eds. Sebastián Barrueco, et. al. Universidad de Alcalá de Henares,

Vallduví, E. (1992) *The Informational Component*. Garland Publishing, New York

Virbel, J. (1989) *The contribution of linguistic knowledge to the interpretation of text structures.* Structured Documents. Cambridge University Press, Cambridge

Vossen, P. and Copestake, A. (1993) *Untangling Definition Structure into Knowledge Representation*. In: Inheritance, Defaults and the Lexicon. Cambridge University Press.

Weissenhofer, P. (1995) *Conceptology in Terminology Theory.* Vienna: IITF-Series 6, Termnet.

Wignell, P. (1998) *Technicality and abstraction in social science*, en "Reading Science: Critical and functional perspective on discourses of science." ed. Martin, J.R. & Robert Veel, Routledge

Wilkes-Gibbs, D., & Clark, H. H. (1992) *Coordinating beliefs in conversation.* Journal of Memory and Cognition, 31, 183-194.

Wittgenstein, L. (1981) *Observaciones.* Siglo XXI ediciones. México, D.F. (English title: *Culture and value*)

Wittgenstein, L. (1953) *Investigaciones Filosóficas. Instituto de Investigaciones Filosóficas. UNAM. México, D.F., 1988*

Wüster, E. (1959) *Exposé illustré et terminologique de la nomination du monde (Das Worten der Welt, schaubildich und terminologisch dargestellt)* dans Sprachforum, vol. 3, nos 3-4, p. 183-204, traduit par INFOTERM, Bibliothèque d'INFOTERM.

Xu, F. (2002) *The role of language in acquiring object kind concepts in infancy*. Cognition, 85, 223-250

Yang, Y, (1999). *An evaluation of statistical approaches to text categorization.* Journal or Information Retrieval. 1(1): 69-90;

Zawada, B. & P. Swanepoel (1997) *On the empirical adequacy of terminological concept theories. The case for prototype theory.* Terminology 1 (2) 253-275.

IX INDEX

Figures, Tables and Illustrations