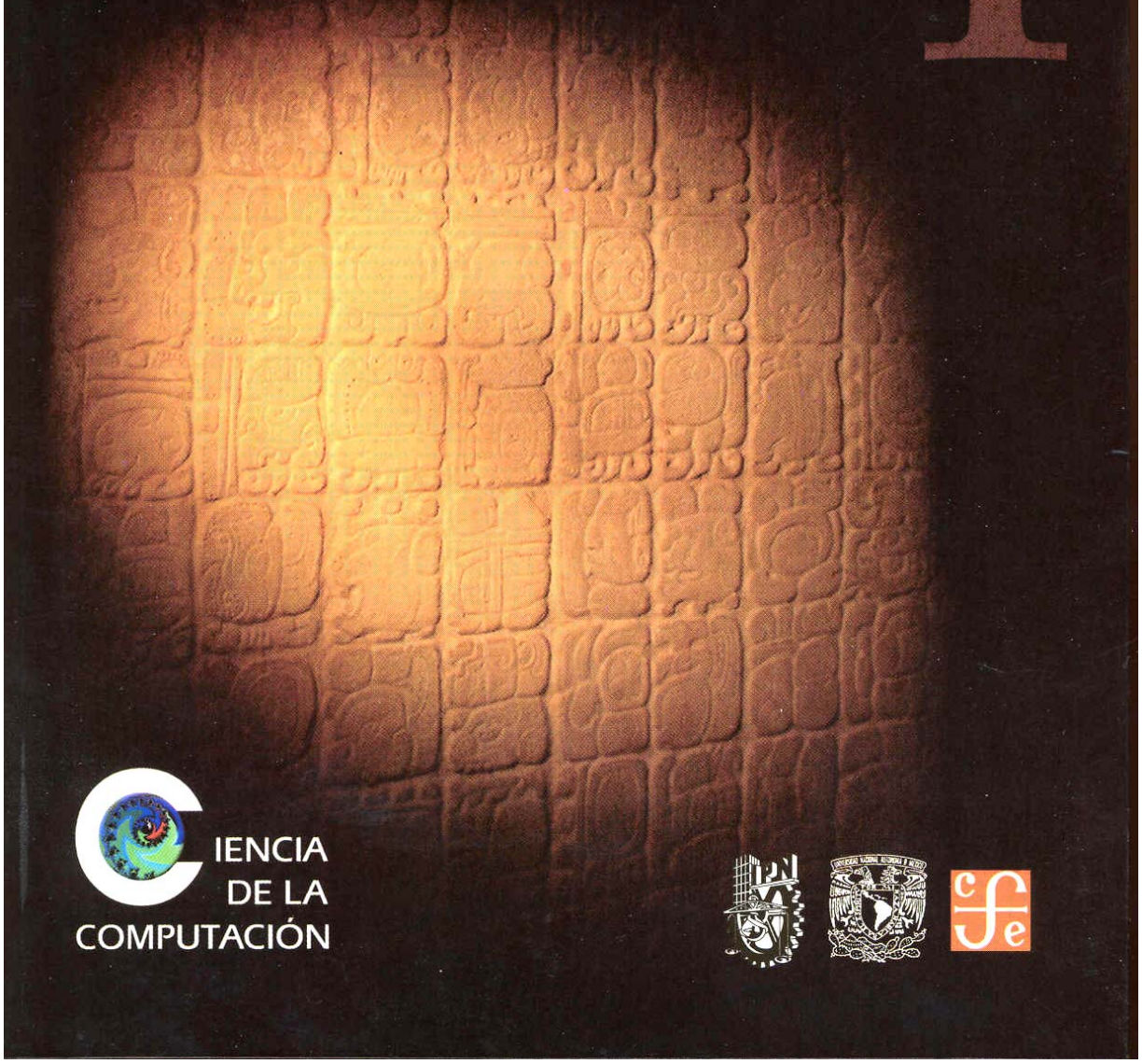
Computational Linguistics

Models, Resources, Applications

Igor Bolshakov
Alexander Gelbukh



CIENCIA DE LA COMPUTACIÓN

CIENCIA DE LA COMPUTACIÓN

COMPUTATIONAL LINGUISTICS
Models, Resources, Applications

COMPUTATIONAL LINGUISTICS
Models, Resources, Applications

Igor A. Bolshakov and Alexander Gelbukh

FIRST EDITION: 2004

ISBN: 970-36-0147- 2

Impreso en México / *Printed in Mexico*

Serie Ciencia de la Computación

The growth of the amount of available written information originated in the Renaissance with the invention of printing press and increased nowadays to unimaginable extent has obliged the man to acquire a new type of literacy related to the new forms of media besides writing. One of such forms is the computer—an object of the modern world that increases the degree of freedom of human action and knowledge, where the fantasy becomes reality, and the new common alphabet penetrates the presence marked by such a phenomenon as computing.

However, even though this phenomenon has become a part of our everyday life, the printed text has not been substituted by the electronic text; on the contrary, they have become into symbiotic elements that constitute fundamental means for accelerating the transition to the advance society and economy restructured towards the science, technology, and promotion and dissemination of knowledge. Only through such spread of knowledge is it possible to create a scientific culture founded on the permanent quest for the truth, informed criticism, and the systematic, rigorous, and intelligent way of human actions.

In this context, the Computer Science Series published by the Center for Computing Research (CIC) of the National Polytechnic Institute in collaboration with the National Autonomous University of Mexico and the Economic Culture Fund editorial house (Fondo de Cultura Económica) presents the works by outstanding Mexican and foreign specialists—outstanding both in their research and educational achievements—in the areas of tutoring systems, system modeling and simulation, numerical analysis, information systems, software engineering, geoprocessing, digital systems, electronics, automatic control, pattern recognition and image processing, natural language processing and artificial intelligence.

In this way, the publishing effort of the CIC—which includes the journal Computación y Sistemas, the Research on Computing Science series, the technical reports, conference proceedings, catalogs of solutions, and this book series—reaffirms its adherence to the high standards of research, teaching, industrial collaboration, guidance, knowledge dissemination, and development of highly skilled human resources.

This series is oriented to specialists in the field of computer science, with the idea to help them to extend and keep up to date their information in this dynamic area of knowledge. It is also intended to be a source of reference in their everyday research and teaching work. In this way one can develop himself or herself basing on the fundamental works of the scientific community—which promotion and dissemination of science is.

We believe that each and every book of this series is a must-have part of the library of any professional in computer science and allied areas who consider learning and keeping one's knowledge up to date essential for personal progress and the progress of our country. Helpful support for this can be found in this book series characterized first and foremost by its originality and excellent quality.

Dr. Juan Luis Díaz De León Santiago
Center For Computing Research
Director

CONTENTS OVERVIEW

DETAILED CONTENTS

PREFACE

WHY DID WE DECIDE to propose a new book on computational linguistics? What are the main objectives, intended readers, the main features, and the relationships of this book to various branches of computer science? In this Preface, we will try to answer these questions.

A NEW BOOK ON COMPUTATIONAL LINGUISTICS

The success of modern software for natural language processing impresses our imagination. Programs for orthography and grammar correction, information retrieval from document databases, and translation from one natural language into another, among others, are sold worldwide in millions of copies nowadays.

However, we have to admit that such programs still lack real intelligence. The ambitious goal of creating software for deep language understanding and production, which would provide tools powerful enough for fully adequate automatic translation and man-machine communication in unrestricted natural language, has not yet been achieved, though attempts to solve this problem already have a history of nearly 50 years.

This suggests that in order to solve the problem, developers of new software will need to use the methods and results of a fundamental science, in this case linguistics, rather than the tactics of *ad hoc* solutions. Neither increasing the speed of computers, nor refinement of programming tools, nor further development of numerous toy systems for language "understanding" in tiny domains, will suffice to solve one of the most challenging problems of modern science—automatic text understanding.

We believe that this problem, yet unsolved in the past century, will be solved in the beginning of this century by those who are sit-

ting now on student benches. This book on computational linguistics models and their applications is targeted at these students, namely, at those students of the Latin American universities studying computer science and technology who are interested in the development of natural language processing software.

Thus, we expect the students to have already some background in computer science, though no special training in the humanities and in linguistics in particular.

On the modern book market, there are many texts on Natural Language Processing (NLP), e.g. [1, 2, 7, 9]. They are quite appropriate as the further step in education for the students in computer science interested in the selected field. However, for the novices in linguistics (and the students in computer science are among them) the available books still leave some space for additional manuals because of the following shortages:

- Many of them are English-oriented. Meanwhile, English, in spite of all its vocabulary similarities with Spanish, is a language with quite a different grammatical structure. Unlike Spanish, English has a very strict word order and its morphology is very simple, so that the direct transfer of the methods of morphologic and syntactic analysis from English to Spanish is dubious.

- Only few of these manuals have as their objective to give a united and comparative exposition of various coexisting theories of text processing. Moreover, even those few ones are not very successful since the methodologies to be observed and compared are too diverse in their approaches. Sometimes they even contradict each other in their definitions and notations.

- The majority of these manuals are oriented only to the formalisms of syntax, so that some of them seemingly reduce computational linguistics to a science about English syntax. Nevertheless, linguistics in general investigates various linguistic levels, namely, phonology, morphology, syntax, and semantics. For each of these levels, the amount of purely linguistic knowledge rele-

vant for computational linguistics seems now much greater than that represented in well-known manuals on the subject.

Reckoning with all complications and controversies of the quickly developing discipline, the main methodological features of this book on computational linguistics are the following:

- Nearly all included examples are taken from Spanish. The other languages are considered mainly for comparison or in the cases when the features to be illustrated are not characteristic to Spanish.
- A wide variety of the facts from the fundamental theory—general linguistics—that can be relevant for the processing of natural languages, right now or in the future, are touched upon in one way or another, rather than only the popular elements of English-centered manuals.
- Our educational line combines various approaches coexisting in computational linguistics and coordinates them wherever possible.
- Our exposition of the matter is rather slow and measured, in order to be understandable for the readers who do not have any background in linguistics.

In fact, we feel inappropriate to simply gather disjoint approaches under a single cover. We also have rejected the idea to make our manual a reference book, and we do not have the intention to give always well-weighted reviews and numerous references through our texts. Instead, we consider the coherence, consistency, and self-containment of exposition to be much more important.

The two approaches that most influenced the contents of this book are the following:

- The Meaning ⇔ Text Theory (MTT), developed by Igor Mel'čuk, Alexander Žolkovsky, and Yuri Apresian since the mid-sixties, facilitates describing the elements, levels, and structures of natural languages. This theory is quite appropriate for any language,

but especially suits for languages with free word order, including Spanish. Additionally, the MTT gives an opportunity to validate and extend the traditional terminology and methodology of linguistics.

- The *Head-driven Phrase Structure Grammar* (HPSG), developed by Carl Pollard and Ivan Sag in the last decade, is probably the most advanced practical formalism in natural language description and processing within the modern tradition of generative grammars originated by Noam Chomsky. Like the MTT, HPSG takes all known facts for description of natural languages and tries to involve new ones. As most of the existing formalisms, this theory was mainly tested on English. In recent years, however, HPSG has acquired numerous followers among researchers of various languages, including Spanish. Since the main part of the research in NLP has been fulfilled till now in the Chomskian paradigm, it is very important for a specialist in computational linguistics to have a deeper knowledge of the generative grammar approach.

The choice of our material is based on our practical experience and on our observations on the sources of the problems which we ourselves and our colleagues encountered while starting our careers in computational linguistics and which still trouble many programmers working in this field due to the lack of fundamental linguistic knowledge.

After coping with this book, our reader would be more confident to begin studying such branches of computational linguistics as

- Mathematical Tools and Structures of Computational Linguistics,
- Phonology,
- Morphology,
- Syntax of both surface and deep levels, and
- Semantics.

The contents of the book are based on the course on computational linguistics that has been delivered by the authors since 1997

at the Center for Computing Research, National Polytechnic Institute, Mexico City. This course was focused on the basic set of ideas and facts from the fundamental science necessary for the creation of intelligent language processing tools, without going deeply into the details of specific algorithms or toy systems. The practical study of algorithms, architectures, and maintenance of real-world applied linguistic systems may be the topics of other courses.

Since most of the literature on this matter is published in English regardless of the country where the research was performed, it will be useful for the students to read an introduction to the field in English. However, Spanish terminological equivalents are also given in the Appendix (see page 173).

The book is also supplied with 54 review questions, 58 test questions recommended for the exam, with 4 variants of answer for each one, 30 illustrations, 58 bibliographic references, and 37 references to the most relevant Internet sites.

The authors can be contacted at the following e-mail addresses: igor@cic.ipn.mx, gelbukh@gelbukh.com (gelbukh@cic.ipn.mx); see also www.Gelbukh.com (www.cic.ipn.mx/~gelbukh). The webpage for this book is www.Gelbukh.com/clbook.

OBJECTIVES AND INTENDED READERS OF THE BOOK

The main objectives of this book are to provide the students with few fundamentals of general linguistics, to describe the modern models of how natural languages function, and to explain how to compile the data—linguistic tables and machine dictionaries—necessary for the natural language processing systems, out of informally described facts of a natural language. Therefore, we want to teach the reader how to prepare all the necessary tools for the development of programs and systems oriented to automatic natural language processing. In order to repeat, we assume that our readers are mainly students in computer sciences, i.e., in software development, database management, information retrieval, artificial intelligence or computer science in general.

Throughout this book, special emphasis is made on applications to the Spanish language. However, this course is not a mere manual of Spanish. A broader basis for understanding the main principles is to be elucidated through some examples from English, French, Portuguese, and Russian. Many literature sources provide the reader with interesting examples for these languages. In our books, we provide analogous examples for Spanish wherever possible.

Significant difficulties were connected with the fact that Latin American students of technical institutes have almost no knowledge in linguistics beyond some basics of Spanish grammar they learned in their primary schooling, at least seven years earlier. Therefore, we have tried to make these books understandable for students without any background in even rather elementary grammar.

Neither it should be forgotten that the native language is studied in the school prescriptively, i.e., how it is preferable or not recommendable to speak and write, rather than descriptively, i.e., how the language is really structured and used.

However, only complete scientific description can separate correct or admissible language constructions from those not correct and not belonging to the language under investigation. Meantime, without a complete and correct description, computer makes errors quite unusual and illogical from a human point of view, so that the problem of text processing cannot be successfully solved.

COORDINATION WITH COMPUTER SCIENCE

The emphasis on theoretical issues of language in this book should not be understood as a lack of coordination between computational linguistics and computer science in general. Computer science and practical programming is a powerful tool in all fields of information processing. Basic knowledge of computer science and programming is expected from the reader.

The objective of the book is to help the students in developing applied software systems and in choosing the proper models and data structures for these systems. We only reject the idea that the

computer science's tools of recent decades are sufficient for computational linguistics in theoretical aspects. Neither proper structuring of linguistic programs, nor object-oriented technology, nor specialized languages of artificial intelligence like Lisp or Prolog solve by themselves the problems of computational linguistics. All these techniques are just tools.

As it is argued in this book, the ultimate task of many applied linguistic systems is the transformation of an unprepared, unformatted natural language text into some kind of representation of its meaning, and, vice versa, the transformation of the representation of meaning to a text. It is the main task of any applied system.

However, the significant part of the effort in the practical developing of an NPL system is *not* connected directly with creating the software for this ultimate task. Instead, more numerous, tedious, and inevitable programming tasks are connected with *extraction of data* for the grammar tables and machine dictionaries from various texts or human-oriented dictionaries. Such texts can be originally completely unformatted, partially formatted, or formalized for some other purposes. For example, if we have a typical human-oriented dictionary, in the form of a text file or database, our task can be to parse each dictionary entry and to extract all of the data necessary for the ultimate task formulated above.

This book contains the material to learn how to routinely solve such tasks. Thus, again, we consider programming to be the everyday practical tool of the reader and the ultimate goal of our studies.

COORDINATION WITH ARTIFICIAL INTELLIGENCE

The links between computational linguistics and artificial intelligence (AI) are rather complicated. Those AI systems that contain subsystems for natural language processing rely directly on the ideas and methods of computational linguistics. At the same time, some methods usually considered belonging only to AI, such as, for example, algorithms of search for decision in trees, matrices, and

other complex structures, sometimes with backtracking, are applicable also to linguistic software systems.

Because of this, many specialists in AI consider computational linguistics a part of AI [2, 40, 54]. Though such an expansion hampers nothing, it is not well grounded, in our opinion, since computational linguistics has its own theoretical background, scientific neighbors, methods of knowledge representation, and decision search. The sample systems of natural language processing, which wander from one manual on AI to another, seem rather toy and obsolete.

Though the two fields of research are different, those familiar with both fields can be more productive. Indeed, they do not need to invent things already well known in the adjacent field, just taking from the neighborhood what they prefer for their own purposes. Thus, we encourage our readers to deeply familiarize themselves with the area of AI. We also believe that our books could be useful for students specializing in AI.

SELECTION OF TOPICS

Since the MTT described below in detail improves and enriches rather than rejects the previous tradition in general linguistics, we will mainly follow the MTT in description and explanation of facts and features of natural languages, giving specific emphasis on Spanish.

To avoid a scientific jumble, we usually do not mark what issues are characteristic to the MTT, but are absent or are described in a different way in other theories. We give in parallel the point of view of the HPSG-like formalisms on the same issue only in the places where it is necessary for the reader to be able to understand the corresponding books and articles on computational linguistics written in the tradition originated by Chomsky.

We should assert here that the MTT is only a tool for language description. It does not bind researchers to specific algorithms or to specific formats of representation of linguistic data. We can find the