FLORENTINA HRISTEA, *Introducere în procesarea limbajului natural cu aplicații în Prolog* [*Introduction to Natural Language Processing With Applications in Prolog*], Bucureşti, Editura Universității din Bucureşti, 2000, 318 p.

The present volume represents an introduction to the highly interdisciplinary field of natural language processing (NLP), viewed, with the eyes of a computer scientist, primarily as a subfield of artificial intelligence. The book adresses both computer scientists, and linguists who are familiar with the fundamentals of logic programming, as well as all those who are interested in the basic computational aspects of the study of natural language. The author performs a thorough analysis of the basic techniques used in computational linguistics and in natural language processing, with special focus on syntax and semantics.

This introduction to the main topics of computational linguistics and natural language processing is conceived as an original synthesis of three fundamental works, those of J. F. Allen[1], M. A. Covington[2], and G. Gazdar, and C. Mellish[3], with a gradual structuring of the corresponding information, performed in the author's own view, and benefitting from her own experience and contributions in the field. We have noticed with increasing satisfaction that the volume successfully integrates Romanian research in the field as well, such as contextual grammars, introduced by Prof. Solomon Marcus, and the use of WG grammars in the case of the Romanian language, a study performed by the author herself. The main theories and techniques of the field, conceived and tested primarily for English, are presented, while comments are made regarding their adaptation to other languages and especially to Romanian. All programs included in the present book have been tested in the case of the Romanian language, and all the theoretical issues discussed are illustrated with examples for both English and Romanian (wherever this is possible). Last but not least, from a computational perspective, we highly appreciate the author's effort to adapt and to develop Prolog programs for performing natural language processing in Romanian.

Going beyond the analysis of the main techniques of computational linguistics and natural language processing respectively, the author manages to clarify basic concepts of the field (see the numerous footnotes inserted in the book) and tries to establish the corresponding Romanian terminology. This is, in fact, to our knowledge, the first Romanian unitary NLP text with broad coverage, which not only establishes the Romanian terminology, but also helps the beginner get the overall picture of a somewhat controversial field, taking into account its highly interdisciplinary nature.

[1] See J. F. Allen, *Natural Language Understanding*. Menlo Park, California, Benjamin - Cummings, 1987.

[2] See M. A. Covington, *Natural Language Processing for Prolog Programmers*. Englewood Cliffs, NJ, Prentice Hall, 1994.

[3] See G. Gazdar, C. Mellish, Natural Language Processing in Prolog: an Introduction to Computational Linguistics. Wokingham, Addison - Wesley, 1989.

RRL, **LI**, *3–4*, p. 517–521, Bucureşti, 2006

The book (published in 2000) is structured in six chapters and one annex, and benefits from a comprehensive bibliography, its reference list being one of the most complete at the time. Of the different levels at which language analysis can be performed, the author concentrates on the syntactic and the semantic level respectively, as well as on the link between the two, having the belief that a study of this type is compulsory when trying to acquaint the reader with the fundamentals of the field within the framework of an introduction. Other aspects, such as those related to the pragmatic level of language analysis or to the statistical processing of natural language, are only touched here and are left for future study.

The first chapter of the book introduces some basic concepts and mentions some of the most important Romanian contributions in the field. Within the second chapter, some fundamental techniques for natural language processing, based on the finite automata model, are described. The third chapter is entirely dedicated to the presentation of several classes of grammars, as well as to their representation in Prolog. Let us note the presence, within this chapter, not only of Chomsky's generative grammars, but also that of contextual grammars, stochastic grammars, and WG grammars. The next chapter covers the main theoretical and practical issues concerning syntactical analysis. The most significant algorithms for performing one of the most important tasks of natural language processing, that of syntactic parsing, are described here and fully implemented in Prolog. The fifth chapter extends the topics and corresponding algorithms of the previous one, dealing with features and augmented grammars and introducing the syntactic parsing based on unification. Finally, the sixth and last chapter discusses issues of computational semantics, while presenting several concepts and elements which are necessary for various natural language understanding applications.

The work under examination deals with classical concepts, but equally touches some of the most modern (at the time) topics of computational linguistics, often placing the discussion at the level of the most recent advances in the field. Some of the applications which it mentions or describes (the international research-development project DBR-MAT, of which the author was the local coordinator, statistical parsing for the Romanian language, a parser based on contextual grammars, WordNet as an interactive lexical database and as a semantic network) represent topics for further thought and reflection, as well as possible research themes and projects, offering great possibilities of extension and generalization, as was proven in recent years.

Marius Popescu
Faculty of Mathematics and Computer Science
University of Bucharest

LIVIU P. DINU, *Formal and Categorization Methods in Mathematical and Computational Linguistics*, Bucureşti, Editura Universităţii din Bucureşti, 2004, 183 p.

The book has two parts: in the first part (chapter two and three) the author investigates the syllable; the second part (chapter four, five and six) is dedicated to the problems of similarity, classification and aggregation, with applications in the study of the similarity of Romance languages.

Section 2.3 contains an algorithm for syllabification of Romanian words. The algorithm performs well for words that do not contain two or more consecutive vowels (called regular words) and has some difficulties for other kinds of words. The final section of chapter two presents one of the first attempts to formalize the syllable[4].

Chapter three contains a construction of a mathematical model for the Romanian syllable. The main idea is the similarity between the syllabification of a word and the generation of a word by a total Marcus-contextual grammar. In sections 3.1-3.5 a total contextual grammar of syllabification is introduced. In order to optimize the grammar, some restrictions to the derivation rule are imposed:

[4] T. Vennemann, 1978, "Universal syllabic phonology", *Theoretical Linguistics* 5, 2–3, 175–215.

total leftmost derivation and total leftmost derivation with maximum/minimum use of selectors. In subsection 3.3 one analyze the number of regular words of length n (which are the axioms in the upper grammars) and a relation between this number and the n-th term of the Fibonacci sequence. The author generalizes the grammars from 3.5 and introduces a new class of contextual grammars, called syllabic grammars. This new class of grammars is placed between internal contextual grammars with finite choice and total contextual grammars with choice. The relation between the upper grammars and go through automata (GTA) is investigated, proving that GTA accept the language generated by the syllabic grammars. The final section of chapter three is dedicated to the cognitive aspects of contextual grammars and to their relation with speech production.

In chapter four Liviu P. Dinu introduces the rank distance (RD), a metric that measures the similarity between two rankings based on the ranks of the objects. The rankings may differ not only by the position of the elements, but also by the elements themselves: some element in a ranking may be absent in another ranking. In natural languages, in the frame of lexical units, the most important information is carried by the first part of the unit. In genomics, the difference on the first positions between two codons is more important than the difference on the last positions. By analogy to natural language and genomics, the difference on the first positions between two rankings is more important than the difference on the last positions. This was the starting point of the author in the construction of RD. He extends RD to rankings of unequal length and to words. In section 4.3 a measure of similarity of trees based on RD is proposed. This extension measures not only the lexical differences between two words, but also the differences of generation between them.

Using RD, in chapter five an aggregation method of rankings is introduced. The aggregation of n rankings is defined as the ranking for which the sum of RDs from it to each of the n rankings is minimal; this ranking is not unique and the set of all such rankings is denoted by Rank Distance Aggregation (RDA). A polynomial algorithm to compute RDA is given. Also, the 17 rationality conditions of aggregation introduced by Păun[5] are checked for RDA. In section 5.3 the author introduces the Rank Distance Categorization (RDC), a categorization method based on combining of decisions. Each classifier gives a ranking of classes; these rankings are aggregated using RDA and the set $\{A_1,A_2,...,A_k\}$ is obtained. The class of the object predicted by the RDC method is the one that occupies most frequently the first position in the rankings $A_1,A_2,...,A_k$.

Chapter six is dedicated to the experimental results. In section 6.1 the RDC method is tested; it improves the mean of individual performance of classifiers. In section 6.2 one investigates the "syllabic" similarity of the Romance languages. The corpus contains the representative vocabularies of Romance languages. The author syllabicates the vocabularies and for each vocabulary he constructs a ranking of syllables based on frequencies of their occurrences. The RDs between all pairs of such rankings is computed, obtaining a series of results and graphics that express the "syllabic" similarity between Romance languages.

In the final section of each chapter a series of open problems and future investigations are presented.

Solomon Marcus
Romanian Academy

ERICA NISTOR DOMONKOS, *Automated Translation Algorithm from English to Romanian*, Bucureşti, Editura Didactică şi Pedagogică, 1966, 302 p.

The fifties are generally accepted like the beginning of the history of machine translation (MT). Its earliest development was strongly encouraged by the the Georgetown-IBM experiment, which took place on January 7, 1954 and was an influential demonstration of MT. Developed jointly

[5] Gh. Păun, 1987, *Paradoxurile clasamentelor*, Bcureşti, Editura Ştiinţifică şi Enciclopedică.

by the Georgetown University and IBM, the experiment involved fully automatic translation of more than sixty Russian sentences into English. Conceived and performed primarily in order to attract governmental and public interest and funding by showing the possibilities of MT, it was by no means a fully-featured system. It had only six grammar rules and 250 items in its vocabulary. Moreover, the system was merely specialised in the domain of organic chemistry. The translation was done using a IBM 701 mainframe computer. Widely covered by the media and perceived as a success, the experiment did encourage governments (not only the U.S. one, but Western Europe and Russia too) to invest into the field of computational linguistics. The authors and researchers from entire world claimed that high-quality MT of scientific and technical documents would be possible within a very few years (three or five years in the most optimistic way). Hutchins (1986) collects the major MT pioneer works from the United States, Russia, East and West Europe (including first Romanian research in MT field), and Japan, with recollections of personal experiences, colleagues and rivals, the political and institutional background, the successes and disappointments, and above all the challenges and excitement of a new field with great practical importance.

The first algorithm for English-Romanian automatic translation began in September 1959 at Timişoara. The work was tightly related to the realization of the first electronic computing machine in Romania, MECIPT-1. The specification of the English and Romanian grammars and the programming itself, i.e. the implementation of the algorithm, were finished in 1960, though the MECIPT-1 project only finished in February1962. The first results were reported on 16th of May 1962, in the presence of Gr. Moisil. Here are the translated sentences:

1. Verificând operaţiile, ea a oprit calculatorul.
   Verifying the operations she halted the computer.
2. Dumneavoastră explicaţi dezvoltarea ştiinţei şi noi ajutăm la descrierea exemplelor.
   You explain the development of the science and we help at the description of the examples.

The complete description of this algorithm was published in (Domonkos 1966). The book has two parts which interact with each other during the whole text. The first part describes the algorithm of automatic translation and all the solved or unsolved problems that were encountered during the development of the project. The second part contains the implementation and explicitly refers to the MECIPT-1 machine. While the second part has today merely a historic relevance for the beginning of programming and the development of cybernetics, the first part is still of actuality. Thus consistent parts of it could be reused and continued in the light of the latest development of the linguistic resources (for example, the national project CNCSIS, no.33549/18A "Electronic Morphologic Dictionary", coordinated by Emil Ionescu in the frame of Centre for Computational Linguistics, University of Bucharest).

The method used by Domonkos for the automatic translation can be classified into the ulterior named approach *Dictionary-based machine translation:* the translation process uses dictionary entries, which means that the words will be translated as a dictionary does — word by word, usually without much correlation of the meaning between them.

The translation algorithm is composed of three phases:

1. The input words are searched in the dictionary and if founded they are translated into Romanian words;
2. During the second phase, the machine morphologically analysis and recognizes the flexions and produces their Romanian counterpart;
3. The phrase to be translated is parsed, according to the Romanian syntax.

During the first phase the encountered problems were of technical nature. The aim was to prove that automatic translation for Romanian was possible with the technical means available at the moment. Nevertheless, because of the limited technology available those days, the developers were forced to do a tradeoff between the complexity of the syntactic rules and the dimension of the dictionary. They choose to keep the number of words in the dictionary relatively low: only 80 words.

The most elaborated phase is the second one. Except of the detailed description of the automated noun declension, verb conjugation (where the paper «*Problemele puse de traducerea automată. Conjugarea verbelor în limba română scrisă*» of Gr. Moisil was used without any major

modification), the author was concerned also with problems such as special words or homonymy resolution.

The special words were divided into two classes: a) provisional special words, as personal pronouns and irregular form of English words; and b) permanent special words as *of, the, shall, will, should, would, do, does, did*, which will not appear in the translated text.

Nistor treated the homonymy problem, (although incompletely solved), by using two methods:

1. a morphological method
2. a syntagmatic method

The book lists in the end some incompletely or completely unsolved problems, as: the absolute superlative, invariable adjective, English verbs ended in *e*, English nouns ended in *y* which end in *ies* at plural form, the passive conjugation of verbs, the dative case *to*, etc.

Unfortunately, after the promise of the researches that the MT will be solved had remained unrealized for a decade, the National Academy of Sciences of the United States published the famous report of its Automatic Language Processing Advisory Committee (ALPAC, 1966). The ALPAC Report recommended that the resources that were being expended on MT as a solution to immediate practical problems should be redirected towards more fundamental questions of language processing that would have to be answered before any translation machine could be built. The number of laboratories working in the field was sharply reduced all over the world, and few of them were able to obtain funding for more long-range research programs in what then came to be known as computational linguistics. There was a resurgence of interest in machine translation in the 1980s and, although the approaches adopted differed little from those of the 1960s, many of the efforts were rapidly deemed successful.

As far as our knowledge, Domonkos's work was never continued or improved. We believe that an improvement of this results could be obtained today, in the light of the new linguistic resources developed lately or of the new technics and methods in computational linguistics.

REFERENCES

Domonkos, E. N., 1966, *Automated translation algorithm from Enghlish to Romanian*, Bucureşti, Editura Didactică şi Pedagogică.

Hutchins, W.J., 1986, *Machine Translation: past, present, future.* Ellis Horwood Series in Computers and their Applications, 382; also, available at http://ourworld.compuserve.com/homepages/WJHutchins/PPF-TOC.htm

Pierce, J.R., John B. Carroll *et al.,* 1966, *Language and Machines — Computers in Translation and Linguistics*. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC.

Liviu P. Dinu
University of Bucharest