Linguistic Resources for Natural Language Processing

new

Vol. 1 - 3

Towards Predicate Driven Grammar

JÖRG SCHUSTER
Centrum für Informations- und Sprachverarbeitung
Ludwig-Maximilians-Universität München

This book is about "Predicate Driven Grammar" (PDG), a new type of linguistic grammar. PDG is strongly influenced by the Sense-Text-Model and by the writings of Zellig Harris and Maurice Gross. Unlike most other grammars, PDG presupposes a language to be a relation over the Cartesian product of a set of texts and a set of meanings. A PDG assigns to each text the set of its meanings and to each meaning the set of its texts and, therefore, relates each two texts that are paraphrases, no matter if they are texts of the same or of different languages.

In other words, a PDG is a theory of intralingal and interlingual paraphrasing (also known as translating). A PDG is supposed to achieve this by respecting certain fundamental properties of language: ambiguity (the property of texts to have several meanings), polymorphism (the property of meanings to have several texts), predicate-basedness and non-modularity. The term "predicate-basedness" is supposed to refer to that fact that each predicate of a natural language comes with its very own set of syntax rules. The term "non-modularity" is supposed to refer to the fact that each syntax rule of a natural-language predicate comes with its very own semantics.

ISBN 978 3 58986 567 1. **Linguistic Resources for Natural Language Processing 01**. 240pp. USD 93.40 / EUR 69.20 / GBP 65.70. 2009.

Funktionsverbgefüge und automatische Sprachverarbeitung

(Support verb constructions and Natural Language Processing)

STEFAN LANGER
Centrum für Informations- und Sprachverarbeitung
Ludwig-Maximilians-Universität München

This book discusses support verb constructions and light verbs from the viewpoint of Natural Language Processing and makes suggestions how to formalize their lexical and semantic description. A wide variety of language specific and also cross-language theoretical approaches to the phenomenon are discussed. Based on the analysis of such constructions by M. Gross and I. Melcuk, the main concepts are distilled, and an example formalization is given in FrameNet. The approach to formalization is described with only a few minimal theoretic requirements, namely the distinction of a semantic and a surface layer of description, such that the basic concepts can be potentially utilized for a wide variety of grammar frameworks.

In a separate chapter, a collection of tests is presented which allow delineating different types of verb-noun-constructions that are situated between fully compositional and completely frozen constructions. The result is a test battery for support verb constructions. Following the description of the linguistic tests, automated approaches to detect support verb construction candidates in large corpora are presented and discussed. Although a large number of experiments have been described in various papers and for various languages, none of the presented algorithms achieve satisfying results, which means that manual lexicographic coding of the constructions is still needed to select support verb construction and to capture their properties.

The main focus of this book is on the description and classification of German support verb constructions, but English and French examples are abundant, and the linguistic concepts are described in a language independent way.

ISBN 978 3 929075 64 9. **Linguistic Resources for Natural Language Processing 03**. 130pp. USD 90.20 / EUR 61.40 / GBP 44.20. 2009/III.

Electronic Dictionaries, Tagsets and Tagging

FRANZ GUENTHNER
Centrum für Informations- und Sprachverarbeitung
Ludwig-Maximilians-Universität München

This monograph examines the state of the art in the area of electronic dictionaries and in particulur their application in natural language processing tasks. The book reviews the (brief) history of electronic dictionaries, the approaches and underlying assumptions of existing attempts at exhaustive lexical coverage for languages like English, German, French and others. The author also seriously challenges the adequacy of current practices in the area of tagging; it is shown that not only is the information in tagsets far from sufficient, the fact that tagging methods are typically restricted to single word forms and to syntactic categories (instead of to complex lexical units and their semantic categories) diminishes their usefulness tremendously when it comes to parsing sentences. Given the kind of lexical units used in tagging, it comes as no surprise that very few parsing results can really be considered to be correct. The book also contains interesting new statistics about the distribution of lexical units in large corpora and the problem of how to deal with ambiguity in texts.

ISBN 978 3 89586 445 2. **Linguistic Resources for Natural Language Processing 02**. 130pp. USD 88.10 / EUR 65.30 / GBP 62.00. 2009/III.

Advanced Programming in PROLOG for Computational Linguistics and Artificial Intelligence

RODRIGUE SABIN MOMPELAT
Digital Preservation Dept., Copenhagen University Library

Prolog is a programming language particularly suitable for developing programs in research areas such as corpus linguistics, machine translation, concordance compilation, testing of grammars and many other areas where statistical counts and analyses are required. We will restrict our linguistic and computational techniques to the elaboration of frequency lists and indexes on the basis of the English and Danish languages.

Prolog is also suitable for solving problems concerned with the computerized implementation of intelligent human behaviour such as presentation of knowledge, teaching, learning, association, planning, explaining, acquisition of language etc. We will examine powerful computational tools, the automata, which can analyse, recognize and generate languages, and whose representation by network can apply to many contexts.

Moreover, Prolog has intrinsic qualities that make its success. It enables to construct databases and their corresponding management system. For that matter, we will develop not only different kinds of procedures of manipulation of information in the databases, but also a module-based database management system which will have to be reliable, robust and user-friendly.

Our approach of Prolog endeavours to make the user capable of using and building systems such as natural language interfaces, terminology and dictionary databases or text-processing programs of great machine-readable text corpora.

ISBN 978 3 89586 359 2 (Hardbound). **LINCOM Studies in Computational Linguistics 01**. 246pp. USD 140.50 / EUR 104.10 / GBP 98.90. 2007.

Mathematical Foundations of Linguistics

H. MARK HUBEY
Montclair State University

Only a few decades ago, only mathematicians, physicists and engineers took calculus courses, and calculus was tailored for them using examples from physics. This made it difficult for students from the life sciences including biology, economics, and psychology to learn mathematics. Recently books using examples from the life sciences and economics have become more popular for such students. Such a math book does not exist for linguists. Even the computational linguistics books (Formal Language Theory) are written for mathematicians and computer scientists.

This book is for linguists. It is intended to teach the required math for a student to be a scientific linguist and to make linguistics a science on par with economics, and computer science.

There are many concepts that are central to the sciences. Most students never see these in one place and if they do, they have to wait until graduate school to obtain them in the often-dreaded "quantitative" courses. As a result sometimes it takes years or even decades before learners are able to integrate what they have learned into a whole, if ever. We have little time and much to do.

In addition to all of these problems we are now awash in data and information. It is now that the general public should be made aware of the

solution to all of these problems. The answer is obviously "knowledge compression". Knowledge is structured information; it is a system not merely a collection of interesting facts.

What this book does, and what all other math books do is teach people the tools with which they can structure and thus compress information and knowledge around them. It has also been said that mathematics is the science of patterns; it is exactly by finding such patterns that we compress knowledge. We can say that mathematics is the science of knowledge compression or information compression.

This book provides the basic tools for mathematics (even including a short and intuitive explanation of differential and integral calculus). The broad areas of linguistics, probability theory, speech synthesis, speech recognition, computational linguistics (formal languages and machines), historical linguistics require mathematics of counting/combinatorics, Bayesian theory, correlation-regression analysis, stochastic processes, differential equations, vectors/tensors. These in turn are based on set theory, logic, measurement theory, graph theory, algebra, Boolean algebra, harmonic analysis etc.

The mathematical fields introduced here are all common ideas from one which one can branch off into more advanced study in any of these fields thus this book brings together ideas from many disparate fields of mathematics which would not normally be put together into a single course. This is what makes this a book especially written for linguists.

ISBN 3 89586 641 5. **LINCOM Handbooks in Linguistics 10**. 260pp. USD 83.30 / EUR 61.70 / GBP 58.60. 1999.

Mathematical and Computational Linguistics

H. MARK HUBEY
Montclair State University

As Lass (1980) has remarked, "system" is something talked about constantly in linguistics but never beyond paying just lip-service to the concept. This book shows how linguistics constitutes a "system". Linguists (except those who study Formal Language Theory) are confronted with a dilemma. What they study is partially based on physics and is in many respects mathematical; yet the mathematics books are divorced from linguistics and linguistics books are divorced from mathematics and physics. There are no books that teach mathematics for linguists or linguistics with mathematics. This book goes a long way toward accomplishing the integration of mathematics, physics and linguistics into a whole, in other words "a system", just like those that are studied by others in the quantitative disciplines such as physics, engineering, computer science or economics.

The methods of mathematics which are used in the books to elucidate system concepts and others that are needed in linguistics includes boolean algebra, differential equations, and fuzzy logic.

Furthermore it also explains in an intuitive manner,those concepts are not only from mathematics but also from the underlying physics and engineering up to and including acoustic theory of speech, speech recognition, and even nonlinearity/catastrophe theory and quantality of phonemic systems.

All the mathematics needed to form the mathematical foundations of linguistics is illustrated with examples from linguistics and thus may be thought of as "theories", those that should replace the standard literary linguistics tradition in the same way that literary economics is no longer the de facto standard. Physical/acoustic theory of speech is blended naturally into the phonological and phonetic standard, and the standard works are used as springboards to the development of vector space concepts that are necessary for comprehension of new works in speech synthesis and speech recognition. It is rather easy then to show how seemingly nonrelated topics such as sonority scales, child language development, and various linguistics processes such as assimilation, metathesis, fortition/lenition can be seen to be a part of the greater whole. Historical processes are also treated in terms of sound change and also in terms of the most basic ideas which are needed for a thorough understanding of the problems such as multiple scale phenomena, distance and similarity, probability theory, and stochastic processes.

ISBN 3 89586 639 3. **LINCOM Handbooks in Linguistics 09**. 450pp. USD 131.00 / EUR 97.10 / GBP 92.20. 1999.

Forensic Linguistics

Forensic Linguistics
An Introduction to the Study of Language and the Law

FARINDE RAIFU OLANREWAJU
Department of English, Adeyemi College of Education

It has already been affirmed that there are power and asymmetries in courtroom discourse. The courtroom professional such as judges, magistrates, lawyers and prosecutors have power over the defendants and witnesses. (Danet, 1984, Luchjenbroers, 1997). This book attempts to provide an explanatory account of linguistic communication between legal professional such as lawyers and prosecutors and the witnesses with a view to show the power prevalent in the courtroom discourse. To this end, various forms of questions such as WH-questions, alternative questions, yes/no questions and declarative questions were analysed to account for the discoursive practices between the lawyers, prosecutors and the witnesses. The framework of this study is supplied by Luchjenbroers (1993). Additionally, WH-questions and declarative in their various forms are further analysed which reveal further manipulation by lawyers to maintain control over courtroom discourse.
The data are 20 hours of audio-taped cases recorded at the High Courts of Justice and Magistrate Courts in Nigeria. The cases collected include assault, theft, house breaking, land, mutiny and rental.
The key suggestions in this book are that narrative mode is indispensable in the fact-finding process which explains why it is favoured during examination. Also questions that contain propositions and presuppositions are strong weapon for the lawyers in controlling, convincing and persuading the witnesses to endorse their ideas. The four analysis carried out in the thesis suggest the fact that lawyers maintain tight control of courtroom discourse.

ISBN 978 3 89586 192 5**. Linguistics Edition 71**. 326pp. USD 101.20 / EUR 75.00 / GBP 71.20. 2009.

Forensic Speaker Identification
A Likelihood Ratio-based Approach Using Vowel Formants

TONY ALDERMAN
Australian National University

This monograph describes an experiment in Forensic Speaker Identification, showing how speech samples from the same speaker can be discriminated from speech from different speakers with acoustic features commonly used in forensics. It also explains what is now considered the legally and logically correct approach to Forensic Speaker Identification, and presents data that can be used both in real casework and in further testing.

Forensic Speaker Identification is typically concerned with addressing the question of whether two or more speech samples have been produced by the same, or different, speakers. Research over the last decade has shown that the legally and logically correct way of doing this is by using a Bayesian Likelihood Ratio.

The monograph explains what a Likelihood Ratio is; why its use is now considered correct; and how it can be used to successfully discriminate same-speaker pairs from different-speaker pairs. The tests are performed on data from eleven male speakers of Australian English, with non-contemporaneous samples, using formant values at target for their long monophthongal vowels. Likelihood Ratios are estimated using two formulae with different assumptions regarding the distribution of the reference data.

Reference data is also presented which is potentially useful for various permutations of the different-speaker hypothesis for male speakers of Australian English.

ISBN 3 89586 715 2. **LINCOM Studies in Phonetics 01**. 160pp. USD 81.90 / EUR 60.70 / GBP 57.60. 2005.

Neurolinguistics

The Neural Basis of Language

ÁNGEL LÓPEZ-GARCÍA
University of Valencia

The continuity (distributed structure) and the discontinuity (modular structure) are not the exclusive property of neural webs that affect wide areas of the brain, but their possibility is implied in the microscopic base within themselves. This situation is found in all types of psychic activity: emotional, rational and linguistic, although in a different way in each of them. Language maximizes the coexistence of both types of nervous processing maintaining them in equilibrium from the first moment. A theory of language which is capable of assuming the two perspectives is needed, since it is the only one that is neurologically justified.

In this book it is proposed that said theory be constructed on Gestaltic principles; this is because Gestaltic principles formally adjust to Topological rules, which allow us to, at the same time, be informed about the spatial reception of the world, specific to the dominated hemisphere, and its correlative verbalization through some form of natural language, which is specific to the dominant hemisphere.

Angel Lopez-Garcia is Professor of Linguistics at the University of Valencia (Spain), and he has been a visiting Professor at the universities of Virginia, Mainz, Minnesota, Tucumán and Aarhus. He helped to develop the Liminar Grammar model of Cognitive Linguistics as soon as 1980, and elaborated a series of applications to natural languages. Lately he is working on the relation between Linguistics and Biology. He is the author of The grammar of genes. How the Genetic Code Resembles the Linguistic Code (2005).

ISBN 978 3 89586 405 6. **LINCOM Studies in Neurolinguistics 03**. 128pp. 39 illustrations. USD 88.10 / EUR 65.30 / GBP 62.00. 2007.