Unification Grammars

Like computer programs, the grammars of natural languages can be expressed as mathematical objects. Such a formal presentation of grammars facilitates mathematical reasoning with grammars (and the languages they denote), on one hand, and the computational implementation of grammar processors, on the other hand. This book presents one of the most commonly used grammatical formalisms, unification grammars, which underlies such contemporary linguistic theories as lexical-functional grammar (LFG) and head-driven phrase structure grammar (HPSG). The book provides a robust and rigorous exposition of the formalism that is both mathematically well founded and linguistically motivated. Although the material is presented formally, and much of the text is mathematically oriented, a core chapter of the book addresses linguistic applications and the implementation of several linguistic insights in unification grammars.

The authors provide dozens of examples and numerous exercises (many with solutions) to illustrate key points. Graduate students and researchers in both computer science and linguistics will find this book a valuable resource.

NISSIM FRANCEZ is professor emeritus of computer science at the Technion-Israel Institute of Technology. His research for the last twenty years has focused on computational linguistics in general, and on the formal semantics of natural language in particular, mainly within the framework of type-logical grammar. His most recent research topic is proof-theoretic semantics for natural language. He is the author of several books and approximately 150 scientific articles. He regularly serves on editorial boards and program committees of several major conferences, including the committee of the FoLLI (Association for Logic, Language, and Information) Beth Dissertation Award.

SHULY WINTNER is associate professor of computer science at the University of Haifa, Israel. His areas of interest include computational linguistics and natural language processing, with a focus on formal grammars, morphology, syntax, and Semitic languages. He is the author or editor of nearly 100 publications. He has served on the program committees of numerous conferences, as a member of the standing committee overseeing the yearly Formal Grammar conference, and as the editor-in-chief of the journal *Research in Language and Computation*.

Unification Grammars

NISSIM FRANCEZ
Technion-Israel Institute of Technology, Haifa, Israel

SHULY WINTNER
University of Haifa, Haifa, Israel

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9781107014176

First published 2012

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data
Francez, Nissim.
Unification grammars / Nissim Francez, Shuly Wintner.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-107-01417-6 (hardback)
1. Grammar, Comparative and general–Mathematical models. 2. Unification grammar.
3. Lexical-functional grammar. 4. Head-driven phrase structure grammar.
I. Wintner, Shuly, 1963- II. Title.
P151.F6775 2011
415.01′51–dc23 2011031353

ISBN 978-1-107-01417-6 Hardback

To Tikva, for a never failing unification

To Galia, Tomer and Inbal, with all my love

וְקָרַב אֹתָם אֶחָד אֶל אֶחָד לְךָ לְעֵץ אֶחָד וְהָיוּ לַאֲחָדִים בְּיָדֶךָ׃

יחזקאל לז יז

Contents

Preface

This book grew out of lecture notes for a course we started teaching at the Department of Computer Science at the Technion, Haifa, in the spring of 1996, and later also taught at the University of Haifa. The students were advanced undergraduates and graduates who had good knowledge of formal languages but limited background in linguistics. We intended it to be an introductory course in computational linguistics, but we wanted to focus on contemporary linguistic theories and the necessary mechanisms for reasoning about and implementing them, rather than on traditional (statistical, corpus-based) natural language processing (NLP) techniques.

We realized that no good textbook existed that covered the material we wanted to teach. Although quite a few good introductions to NLP exist, including Pereira and Shieber (1987), Gazdar and Mellish (1989), Covington (1994), Allen (1995), and more recently, Manning and Schütze (1999), Jurafsky and Martin (2000), and Bird et al. (2009), none of them provides the mathematical and computational infrastructure needed for our purposes.

The focus of this book is two dimensional. On one hand, we focus on a certain formalism, *unification grammars*, for presenting, studying, and reasoning about grammars. Although it is not the sole formalism used in computational linguistics,[1] it has gained much popularity, and it now underlies many ongoing projects. On the other hand, we also focus on fundamental natural language syntactic constructions, and the way they are specified in grammars expressed in this formalism. Other monographs are either too informal (Shieber, 1986) or too advanced for the students we are addressing (Carpenter, 1992; Shieber, 1992). Of course, the linguistic texts (Pollard and Sag, 1994; Dalrymple et al.,

[1] Notably, type-logical (categorial) grammars (Moortgat, 1997; Steedman, 2000) promote a radically different view of syntax.

1995; Sag and Wasow, 1999; Butt et al., 1999) are inadequate for our purposes because their emphasis is on linguistic issues, rather than the underlying mathematical and computational formulations.

Since 1996 we have taught courses and tutorials based on drafts of this book at universities and research institutes worldwide. Unification grammars had become part of the curriculum in several introductory computational linguistics programs, and there was still no adequate textbook on the subject. With this book, we hope to fill a gap on the introductory computational linguistics bookshelf.

Whom is this book for?

Computational linguistics is a relatively young field of research that lies at the interface of linguistics and computer science. Students of computational linguistics usually come from one of these two disciplines and frequently lack background in the other.

This book is an introductory textbook, and as such is oriented toward students with little background in either. However, we do assume at least some basic knowledge of both paradigms, linguistics and mathematics. More specifically, we assume that the reader has a fair knowledge (equivalent to what is usually acquired in one introductory course) of syntax, as well as some acquaintance with its elementary terminology. In the same way, we assume a background of at least one year of undergraduate study in mathematics. In particular, we assume acquaintance with elementary set theory, formal language theory, basic algorithms and graph theory, some basic universal algebra, and some logic. As far as programming is concerned, acquaintance with some logic programming language (such as Prolog) can be helpful, but it is definitely not necessary to understand this book.

As an aid to readers, and to establish a common mathematical terminology, Appendix B contains a brief overview of some of the basic mathematical concepts we use in the book.

The organization of the book

Since this book is intended to be accessible to readers with some background in computer science or a related discipline, the text consists of mathematical presentations of concepts in the study of natural language syntax. After an introductory chapter in which we outline the problems and recapitulate basic concepts (in particular, context-free grammars), Chapter 2 introduces feature structures, the main building blocks of unification grammars. The unification operation is discussed in Chapter 3, and Chapter 4 defines grammars and their

languages. The linguistic applications of the formalism we develop are discussed in Chapter 5, where a series of grammar fragments are presented for various natural language phenomena. The book does not deal with theories of *meaning* (semantics). Chapter 6 discusses computational issues, and in particular, parsing with unification grammars. The discussion is limited to *untyped* unification formalisms; extension to the *typed* case is planned for the future.

As we intend the book to serve as a textbook, we have scattered various exercises throughout the text. Some of them are simple, intended to apply certain ideas discussed in the text; some complete the text (for example, ask the reader to complete some proofs); and some call for a deeper understanding and creative thought. Solving the exercises is, in our mind, a good way of internalizing the concepts discussed in the text. Nevertheless, it is not required in order to follow the text. We provide sketches of solutions to selected exercises (those marked by '(*)') in Appendix C.

At the end of each chapter we list references to the original publications on which our presentation is based, as well as suggestions for further reading.

General conventions Throughout the book, Sans Serif font is used for depicting phrases in natural languages. When the examples are not drawn from English, glosses are provided. Ungrammatical strings are preceded by '$*$', as is common in the linguistic literature. We use *italics* for emphasis and **bold face** for introducing new concepts.

When a symbol is referenced, rather than used, we usually enclose it within single quotes. For example, the symbol '$\sqcup$' denotes unification. When the context in which a symbol appears eliminates possible confusion, we sometimes omit the quotes.

Acknowledgments

Many students and colleagues have read drafts of this book and provided valuable comments. We thank Gilad Ben-Avi, Daniel Feinstein, and Israel Gutter for spotting several errors in earlier versions of this book and for suggesting useful improvements. Special thanks are due to Yael Sygal for not only suggesting numerous improvements, but also actively contributing two proofs to the text.

The work of the first author was partially supported, over the years, by grants from the Israel Academy of Sciences and Arts (Israel Science Foundation), and by the Technion Vice-president's Fund for the Promotion of Research at the Technion. The second author was partially supported by the Israel Science Foundation.