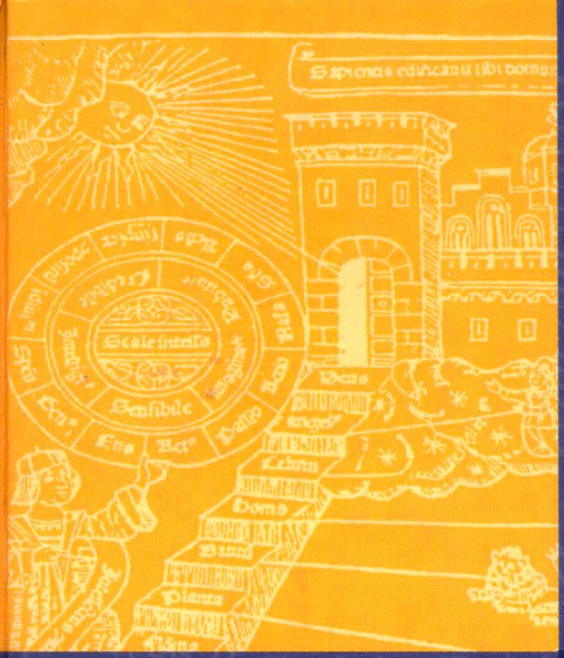

Roland Hausser

Foundations of Computational Linguistics

Human-Computer Communication in Natural Language

2nd Edition, Revised and Extended

Springer

Foundations of Computational Linguistics

Human-Computer Communication in Natural Language

Second Edition, Revised and Extended

Springer

Roland Hausser
Professor of Computational Linguistics
Friedrich Alexander University Erlangen Nürnberg
Bismarckstr. 12, 91054 Erlangen, Germany
rrh@linguistik.uni-erlangen.de

Library of Congress Cataloging-in-Publication Data applied for
Die Deutsche Bibliothek - CIP-Einheitsaufnahme
Hausser, Roland
Foundations of computational linguistics: Human-computer communication in natural language/Roland Hausser. - 2.ed. - Berlin; Heidelberg; New York; Barcelona; Hong Kong; London; Milan; Paris; Tokyo: Springer, 2001
ISBN 3-540-42417-2

ACM Computing Classification (1998): J.5, I.2.7, I.5.4, H.3.1, F.4.2-3, F.1.3

ISBN 3-540-42417-2 Springer-Verlag Berlin Heidelberg New York
ISBN 3-540-66015-1 Springer-Verlag Berlin Heidelberg NewYork (1st ed.)

Springer-Verlag Berlin Heidelberg New York,
a member of BertelsmannSpringer Science+Business Media GmbH
http://www.springer.de

Printed in Germany

Cover design: KünkelLopka, Heidelberg
Cover picture: The ladder of ascent and descent, Ramon Lull, Valencia 1512
Typesetting: Computer to film from author's data
Printing and binding: Stürtz, Würzburg
Printed on acid-free paper SPIN 10844870 45/3142PS 5 4 3 2 1 0

Foundations of Computational Linguistics

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Tokyo

Preface to the second edition

As an interdisciplinary field, computational linguistics has its sources in several areas of science, each with its own goals, methods, and historical background. Thereby, it has remained unclear which components fit together and which do not. This suggests three possible approaches to designing a computational linguistics textbook.

The first approach proceeds from one's own school of thought, usually determined by chance, such as one's initial place of study, rather than by a well-informed, deliberate choice. The goal is to extend the inherited theoretical framework or method to as many aspects of language analysis as possible. As a consequence, the issue of compatibility with other approaches in the field need not be addressed and one's assumptions are questioned at best in connection with 'puzzling problems.'

The second approach takes the viewpoint of an objective observer and aims to survey the field as completely as possible. However, the large number of different schools, methods, and tasks necessitates a subjective selection. Furthermore, the presumed neutrality provides no incentive to investigate the compatibility between the elements selected.

The third approach aims at solving a comprehensive functional task, with the different approaches being ordered relative to it. To arrive at the desired solution, suitability and compatibility of the different elements adopted must be investigated with regard to the task at hand.

In this textbook, the survey Chapters 1 and 2 are based on the second approach, while the remaining Chapters, 3 to 24, are based on the third. The comprehensive task chosen is the design of a robot which can freely communicate in natural language.

The most difficult aspects of this task are treated in Chapters 22–24, which present a declarative outline for programming the semantic and pragmatic interpretation of natural language. Based on a new formulation in a recent article in *Artificial Intelligence* (Hausser 2001c), these chapters have been completely rewritten for the second edition. Sections 22.5, 24.4, and 24.5 go even further than Hausser 2001c, and are followed by a new schematic summary and a new conclusion. Examples and explanations which were contained in the old versions of Chapters 22, 23, and 24 have been moved to the new appendices A, B, and C, respectively.

Many improvements are due to corrections, suggestions, and remarks made in response to the first edition by

Wolfgang Bibel, Darmstadt, Germany
Susan Brennan, Stony Brook, USA
Jaime Carbonell, Pittsburgh, USA
Suk-Jin Chang, Seoul, Korea
Jae-Woong Choe, Seoul, Korea
Key-Sun Choi, KAIST, Korea
Gerald Gazdar, Brighton, UK
Alexander Gelbukh, Mexico City, Mexico
Liu Haitao, Qinghai, China
Yun-Pyo Hong, Cheonan, Korea
Hannu Kangassalo, Tampere, Finland
Ruth Kempson, London, UK
Ferenc Kiefer, Budapest, Hungary
Thomas Künneth, Erlangen, Germany
Kiyong Lee, Seoul, Korea
Minhaeng Lee, Seoul, Korea
Jürgen Lenerz, Köln, Germany
Winfried Lenders, Bonn, Germany
Hans-Heinrich Lieb, Berlin, Germany
Brian MacWhinney, Pittsburgh, USA
Wilfried Meyer-Viol, London, UK
George Miller, Princeton, USA
Mi-Sun Mun, Seoul, Korea
Anne Nicolle, Caen, France
Luis Pineda-Cortez, Mexico City, Mexico
Geoffrey Pullum, Santa Cruz, USA
Teodor Rus, Iowa City, USA
Gérard Sabah, Paris, France
Ivan Sag, Stanford, USA
Geoffrey Sampson, Brighton, UK
Petr Sgall, Prague, Czech Republic
Mark Steedman, Edinburgh, UK
Markus Schulze, Erlangen, Germany
Aesun Yoon, Pusan, Korea

Among the changes is the term "human-computer communication," which is used in the new subtitle and throughout the book.

Last but not least I would like to express my gratitude to Dr. Virginia Swisher, Pittsburgh, for improving the English. As a non-native speaker it never ceases to amaze me how moving the words around a little, adding or removing a comma, etc., can remove subconscious irritation and enhance readability and understanding.

Erlangen, August 2001 Roland Hausser

Preface

The central task of a future-oriented computational linguistics is the development of cognitive machines which humans can freely talk with in their respective natural language. In the long run, this task will ensure the development of a functional theory of language, an objective method of verification, and a wide range of applications.

Natural communication requires not only verbal processing, but also non-verbal perception and action. Therefore the content of this textbook is organized as a theory of language for the construction of talking robots. The main topic is the *mechanism of natural language communication* in both the speaker and the hearer.

The content is divided into the following parts:

I. Theory of Language
II. Theory of Grammar
III. Morphology and Syntax
IV. Semantics and Pragmatics

Each part consists of 6 chapters. Each of the 24 chapters consists of 5 sections. A total of 797 exercises help in reviewing key ideas and important problems.

Part I begins with current applications of computational linguistics. Then it describes a new theory of language, the functioning of which is illustrated by the robot CURIOUS. This theory is referred to with the acronym SLIM, which stands for ***S**urface compositional **L**inear **I**nternal **M**atching*. It includes a cognitive foundation of semantic primitives, a theory of signs, a structural delineation of the components syntax, semantics, and pragmatics, as well as their functional integration in the speaker's utterance and the hearer's interpretation. The presentation refers to other contemporary theories of language, especially those of Chomsky and Grice, as well as to the classic theories of Frege, Peirce, de Saussure, Bühler, and Shannon & Weaver, explaining their formal and methodological foundations as well as their historical background and motivations.

Part II presents the theory of *formal grammar* and its methodological, mathematical, and computational roles in the description of natural languages. A description of categorial grammar and phrase structure grammar is combined with an introduction to the basic notions and linguistic motivation of generative grammar. Further topics are the declarative vs. procedural aspects of parsing and generation, type transparency, as well as the relation between formalisms and complexity classes. It is shown that the

principle of possible *substitutions* causes empirical and mathematical problems for the description of natural language. As an alternative, the principle of possible *continuations* is formalized as LA-grammar. LA stands for the left-associative derivation order which models the time-linear nature of language. Applications of LA-grammar to relevant artificial languages show that its hierarchy of formal languages is orthogonal to that of phrase structure grammar. Within the LA-hierarchy, natural language is in the lowest complexity class, namely the class of C1-languages which parse in linear time.

Part III describes the *morphology* and *syntax* of natural language. A general description of the notions word, word form, morpheme, and allomorph, the morphological processes of inflection, derivation, and composition, as well as the different possible methods of automatic word form recognition is followed by the morphological analysis of English within the framework of LA-grammar. Then the syntactic principles of valency, agreement, and word order are explained within the left-associative approach. LA-grammars for English and German are developed by systematically extending a small initial system to handle more and more constructions such as the fixed vs. free word order of English and German, respectively, the structure of complex noun phrases and complex verbs, interrogatives, subordinate clauses, etc. These analyses are presented in the form of explicit grammars and derivations.

Part IV describes the *semantics* and *pragmatics* of natural language. The general description of language interpretation begins by comparing three different types of semantics, namely those of logical languages, programming languages, and natural languages. Based on Tarski's foundation of logical semantics and his reconstruction of the Epimenides paradox, the possibility of applying logical semantics to natural language is investigated. Alternative analyses of intensional contexts, propositional attitudes, and the phenomenon of vagueness illustrate that different types of semantics are based on different ontologies which greatly influence the empirical results. It is shown how a semantic interpretation may cause an increase in complexity and how this is to be avoided within the SLIM theory of language. The last two chapters, 23 and 24, analyze the interpretation by the hearer and the conceptualization by the speaker as a time-linear navigation through a database called *word bank*. A word bank allows the storage of arbitrary propositions and is implemented as an extension of a classic (i.e., record-based) network database. The autonomous navigation through a word bank is controlled by the explicit rules of suitable LA-grammars.

As supplementary reading the *Survey of the State of the Art in Human Language Technology*, Ron Cole (ed.) 1998 is recommended. This book contains about 90 contributions by different specialists giving detailed snapshots of their research in language theory and technology.

Erlangen, June 1999 Roland Hausser

Table of Contents

Introduction

I. Basic goal of computational linguistics

Transmitting information by means of a natural language like Chinese, English, or German is a real and well-structured procedure. This becomes evident when we attempt to communicate with people who speak a foreign language. Even if the information we want to convey is completely clear to us, we will not be understood by our hearers if we fail to use their language adequately.

The goal of computational linguistics is to reproduce the natural transmission of information by modeling the speaker's production and the hearer's interpretation on a suitable type of computer. This amounts to the construction of autonomous cognitive machines (robots) which can communicate freely in natural language.

The development of speaking robots is not a matter of fiction, but a real scientific task. Remarkably, however, theories of language have so far avoided a functional modeling of the natural communication mechanism, concentrating instead on peripheral aspects such as methodology (behaviorism), innate ideas (nativism), and scientific truth (model theory).

II. Turing test

The task of modeling the mechanism of natural communication on the computer was described in 1950 by Alan Turing (1912–1954) in the form of an 'imitation game' known today as the Turing test. In this game, a human interrogator is asked to question a male and a female partner in another room via a teleprinter in order to determine which answer was given by the man and which by the woman. The people running the test count how often the interrogator classifies his communication partners correctly and how often (s)he is fooled by them.

Subsequently one of the two humans is replaced by a computer. The computer passes the Turing test if it simulates the man or the woman which it replaced so well that the guesses of the interrogator are just as often right and wrong as with the previous set of partners. In this way Turing wanted to replace the question "Can machines think?" by the question "Are there imaginable digital computers which would do well in the imitation game?"

III. ELIZA PROGRAM

In its original intention, the Turing test requires the construction of an artificial cognitive agent with a verbal behavior so natural that it cannot be distinguished from that of a human native speaker. This presupposes complete coverage of the language data and of the communicative functions in real time. At the same time, the test tries to avoid all aspects not directly involved in verbal behavior.[1]

However, the Turing test does not specify what cognitive structure the artificial agent should have in order to succeed in the imitation game. For this reason, it is possible to misinterpret the aim of the Turing test as fooling the interrogator rather than providing a functional model of communication on the computer. This was shown by the Eliza program of Weizenbaum 1965.

The Eliza program simulates a psychiatrist encouraging the human interrogator to talk more and more about him- or herself. The structure of Eliza is based on sentence templates into which certain words used by the interrogator, now in the role of a patient, are inserted. For example, if the interrogator mentions the word mother, Eliza uses the template Tell me more about your ___ to generate the sentence Tell me more about your mother.

Because of the way in which Eliza works, we know that Eliza has no understanding of the dialog with the interrogator/patient. Thus, the construction of Eliza is not a model of communication. If we regard the dialog between Eliza and the interrogator/patient as a modified Turing test, however, the Eliza program is successful insofar as the interrogator/patient *feels* him- or herself understood and therefore does not distinguish between a human and an artificial communication partner in the role of the psychiatrist.

The purpose of computational linguistics is the real modeling of natural language communication, and not a mimicry based on exploiting particular restrictions of a specific dialog situation, as in the Eliza program. Thus, computational linguistics must (i) explain the mechanism of natural communication theoretically and (ii) verify this explanation in practice. The latter is done in terms of a complete and general implementation which must prove its functioning in everyday communication rather than in the Turing test.

IV. MODELING NATURAL COMMUNICATION

Designing a talking robot provides an excellent occasion for systematically developing the basic notions as well as the philosophical, mathematical, grammatical, methodological, and programming aspects of computational linguistics. This is because modeling the mechanism of natural communication requires

[1] As an example of such an aspect, A. Turing 1950, p. 434, mentions the artificial recreation of human skin.