



Roland Hausser

Foundations of Computational Linguistics

Human-Computer Communication
in Natural Language

2nd Edition, Revised and Extended

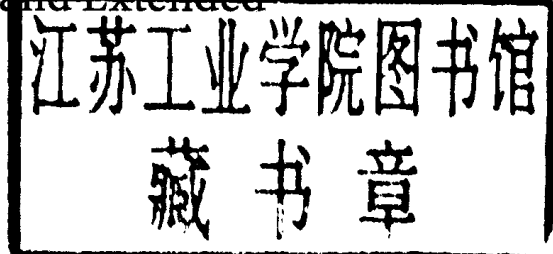


Springer

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 New York (1st ed.)

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

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

© Springer-Verlag Berlin Heidelberg 1999, 2001

Printed in Germany

The use of designations, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

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 Künne, 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.

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 *Surface compositional Linear Internal Matching*. 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.

Table of Contents

Introduction	1
<hr/>	
Part I. Theory of Language	
<hr/>	
1. Computational language analysis	13
1.1 Human-computer communication	13
1.2 Language science and its components	16
1.3 Methods and applications of computational linguistics	21
1.4 Electronic medium in recognition and synthesis	23
1.5 Second Gutenberg revolution	26
<i>Exercises</i>	31
2. Technology and grammar	33
2.1 Indexing and retrieval in textual databases	33
2.2 Using grammatical knowledge	36
2.3 Smart versus solid solutions	39
2.4 Beginnings of machine translation	41
2.5 Machine translation today	45
<i>Exercises</i>	49
3. Cognitive foundations of semantics	51
3.1 Prototype of communication	51
3.2 From perception to recognition	53
3.3 Iconicity of formal concepts	56
3.4 Context propositions	61
3.5 Recognition and action	65
<i>Exercises</i>	67
4. Language communication	69
4.1 Adding language	69
4.2 Modeling reference	72
4.3 Using literal meaning	75

4.4	Frege's principle	77
4.5	Surface compositionality	80
	<i>Exercises</i>	87
5.	Using language signs on suitable contexts	89
5.1	Bühler's organon model	89
5.2	Pragmatics of tools and pragmatics of words	91
5.3	Finding the correct subcontext	93
5.4	Language production and interpretation	96
5.5	Thought as the motor of spontaneous production	99
	<i>Exercises</i>	101
6.	Structure and functioning of signs	103
6.1	Reference mechanisms of different sign-types	103
6.2	Internal structure of symbols and indexicals	107
6.3	Repeating reference	110
6.4	Exceptional properties of icon and name	114
6.5	Pictures, pictograms, and letters	118
	<i>Exercises</i>	121
<hr/>		
Part II. Theory of Grammar		
<hr/>		
7.	Generative grammar	125
7.1	Language as a subset of the free monoid	125
7.2	Methodological reasons for generative grammar	129
7.3	Adequacy of generative grammars	131
7.4	Formalism of C-grammar	132
7.5	C-grammar for natural language	136
	<i>Exercises</i>	139
8.	Language hierarchies and complexity	141
8.1	Formalism of PS-grammar	141
8.2	Language classes and computational complexity	144
8.3	Generative capacity and formal language classes	146
8.4	PS-Grammar for natural language	152
8.5	Constituent structure paradox	157
	<i>Exercises</i>	161
9.	Basic notions of parsing	163
9.1	Declarative and procedural aspects of parsing	163
9.2	Fitting grammar onto language	165
9.3	Type transparency between grammar and parser	170

9.4	Input-output equivalence with the speaker-hearer	176
9.5	Desiderata of grammar for achieving convergence	178
	<i>Exercises</i>	181
10.	Left-associative grammar (LAG)	183
10.1	Rule types and derivation order	183
10.2	Formalism of LA-grammar	186
10.3	Time-linear analysis	190
10.4	Absolute type transparency of LA-grammar	192
10.5	LA-grammar for natural language	195
	<i>Exercises</i>	200
11.	Hierarchy of LA-grammar	203
11.1	Generative capacity of unrestricted LAG	203
11.2	LA-hierarchy of A-, B-, and C-LAGs	206
11.3	Ambiguity in LA-grammar	209
11.4	Complexity of grammars and automata	212
11.5	Subhierarchy of C1-, C2-, and C3-LAGs	215
	<i>Exercises</i>	221
12.	LA- and PS-hierarchies in comparison	223
12.1	Language classes of LA- and PS-grammar	223
12.2	Subset relations in the two hierarchies	225
12.3	Non-equivalence of the LA- and PS-hierarchy	227
12.4	Comparing the lower LA- and PS-classes	229
12.5	Linear complexity of natural language	232
	<i>Exercises</i>	237
<hr/>		
Part III. Morphology and Syntax		
<hr/>		
13.	Words and morphemes	241
13.1	Words and word forms	241
13.2	Segmentation and concatenation	245
13.3	Morphemes and allomorphs	249
13.4	Categorization and lemmatization	250
13.5	Methods of automatic word form recognition	253
	<i>Exercises</i>	257
14.	Word form recognition in LA-Morph	259
14.1	Allo-rules	259
14.2	Phenomena of allomorphy	263
14.3	Left-associative segmentation into allomorphs	269

14.4 Combi-rules	272
14.5 Concatenation patterns	275
<i>Exercises</i>	279
15. Corpus analysis	281
15.1 Implementation and application of grammar systems	281
15.2 Subtheoretical variants	284
15.3 Building corpora	288
15.4 Distribution of word forms	291
15.5 Statistical tagging	295
<i>Exercises</i>	299
16. Basic concepts of syntax	301
16.1 Delimitation of morphology and syntax	301
16.2 Valency	304
16.3 Agreement	307
16.4 Free word order in German (<i>LA-D1</i>)	310
16.5 Fixed word order in English (<i>LA-E1</i>)	316
<i>Exercises</i>	318
17. LA-syntax for English	321
17.1 Complex fillers in pre- and postverbal position	321
17.2 English field of referents	326
17.3 Complex verb forms	328
17.4 Finite state backbone of LA-syntax (<i>LA-E2</i>)	331
17.5 Yes/no-interrogatives (<i>LA-E3</i>) and grammatical perplexity	335
<i>Exercises</i>	340
18. LA-syntax for German	343
18.1 Standard procedure of syntactic analysis	343
18.2 German field of referents (<i>LA-D2</i>)	346
18.3 Verbal positions in English and German	351
18.4 Complex verbs and elementary adverbs (<i>LA-D3</i>)	354
18.5 Interrogatives and subordinate clauses (<i>LA-D4</i>)	360
<i>Exercises</i>	366

Part IV. Semantics and Pragmatics

19. Three system types of semantics	371
19.1 Basic structure of semantic interpretation	371
19.2 Logical, programming, and natural languages	373
19.3 Functioning of logical semantics	375

19.4 Metalanguage-based or procedural semantics?	380
19.5 Tarski's problem for natural language semantics	383
<i>Exercises</i>	387
20. Truth, meaning, and ontology	389
20.1 Analysis of meaning in logical semantics	389
20.2 Intension and extension	392
20.3 Propositional attitudes	395
20.4 Four basic ontologies	399
20.5 Sorites paradox and the treatment of vagueness	402
<i>Exercises</i>	406
21. Absolute and contingent propositions	409
21.1 Absolute and contingent truth	409
21.2 Epimenides in a [+sense,+constructive] system	413
21.3 Frege's principle as homomorphism	416
21.4 Time-linear syntax with homomorphic semantics	420
21.5 Complexity of natural language semantics	423
<i>Exercises</i>	426
22. Database semantics	429
22.1 Database metaphor of natural communication	429
22.2 Descriptive aporia and embarrassment of riches	433
22.3 Communication between the speaker and the hearer	436
22.4 Three kinds of propositions	442
22.5 Spatio-temporal indexing	446
<i>Exercises</i>	453
23. Structure and functions of a SLIM machine	455
23.1 Representing vs. activating propositional content	455
23.2 Motor algorithm for powering the navigation	458
23.3 Autonomous control structure	461
23.4 Contextual cognition as the basis of coherence	464
23.5 The ten SLIM states of cognition	466
<i>Exercises</i>	474
24. A formal fragment of natural language	477
24.1 DBL-LEX and LA-INPUT	477
24.2 LA-MOTOR and LA-OUTPUT	485
24.3 LA-QUERY and LA-INFERENCE	489
24.4 Relating stored content to the current situation	496
24.5 Mapping between meaning ₁ and meaning ₂	500
<i>Exercises</i>	505

Schematic summary	507
Conclusion	511
<hr/>	
Appendix	
<hr/>	
A. Another example of a word bank	515
A.1 Embedding and extracting information	515
A.2 Translating the content of a knowledge base into propositions	516
A.3 An equivalent graphical representation	516
A.4 Word bank representation	517
A.5 Embedding and extracting propositional content	518
B. Interpretation of a complex sentence (LA-E4)	521
B.1 The sample sentence	521
B.2 Definition of <i>LA-E4</i>	521
B.3 Pre-verbal application of DET+N	523
B.4 Application of NOM+FV	524
B.5 Application of FV+MAIN	525
B.6 Reapplication of FV+MAIN	525
B.7 Post-verbal application of DET+N	526
B.8 Transition to the subordinate clause based on ADD-ADP	527
B.9 Beginning of the subordinate clause based on START-SUBCL	528
B.10 Reapplication of NOM+FV	529
B.11 Completing the subordinate clause with FV+MAIN	530
B.12 Result of the derivation	531
C. Subordinating navigation in the speaker mode	533
C.1 Different navigation types	533
C.2 Embedding constructions	534
C.3 Realization of clauses with the verb in final position	535
C.4 Lexical realization of conjunctions	536
C.5 Multiple center embeddings	537
Bibliography	539
Name Index	559
Subject Index	563

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.