

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 *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.

Erlangen, June 1999

Roland Hausser

Table of Contents

Introduction	1
---------------------------	---

Part I. Theory of Language

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

Part II. Theory of Grammar

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
<hr/>	
Part IV. Semantics and Pragmatics	
<hr/>	
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 (<i>LA-E4</i>)	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.

- a theory of language which explains the natural transfer of information in a way that is functionally coherent, mathematically explicit, and computationally efficient,
- a description of language data which is empirically complete for all components of this theory of language, i.e., the lexicon, the morphology, the syntax, and the semantics, as well as the pragmatics and the representation of the internal context,
- a degree of precision in the description of these components which is sufficient for computation.

Fulfilling these requirements will take hard, systematic, goal-oriented work, but it will be worth the effort.

For theory development, the construction of talking robots is of interest because an electronically implemented model of communication may be tested both externally in terms of the verbal behavior observed, and internally via direct access to its cognitive states. The work towards realizing unrestricted human-computer communication in natural language is facilitated by the fact that the functional model may be developed incrementally, beginning with a simplified, but fully general system to which additional functions as well as additional natural languages are added step by step.

For practical purposes, unrestricted communication with computers and robots in natural languages will make the interaction with these machines maximally user friendly and permit new, powerful ways of information processing. Artificial programming languages may then be limited to specialists developing and servicing the machines.

V. USING PARSERS

Computational linguistics analyzes natural languages automatically in terms of software programs called parsers. The use of parsers influences the theoretical viewpoint of linguistic research, distribution of funds, and everyday research practice as follows:

– *Competition*

Competing theories of grammar are measured with respect to the new standard of how well they are suited for efficient parsing and how well they fit into a theory of language designed to model the mechanism of natural communication.

– *Funding*

Computationally efficient and empirically adequate parsers for different languages are needed for an unlimited range of practical applications, which has a major impact on the inflow of funds for research, development, and teaching in this particular area of the humanities.

– *Verification*

Programming grammars as parsers allows testing their empirical adequacy automatically on arbitrarily large amounts of real data in the areas of word form recognition/synthesis, syntactic analysis/generation and semantic-pragmatic interpretation in both the speaker and the hearer mode.

The verification of theories of language and grammar by means of testing electronic models in real applications is a new approach which clearly differs from the methods of traditional linguistics, psychology, philosophy, and mathematical logic.

VI. THEORETICAL LEVELS OF ABSTRACTION

So far there are no electronic systems which model the functioning of natural communication so successfully that one can talk with them more or less freely. Furthermore, researchers do not agree on how the mechanism of natural communication really works. One may therefore question whether achieving a functional model of natural communication is possible in principle. I would like to answer this question with an analogy² from the recent history of science.

Today's situation in computational linguistics resembles the development of mechanical flight before 1903.³ For hundreds of years humans had observed sparrows and other birds in order to understand how they fly. Their goal was to become airborne in a similar manner. It turned out, however, that flapping wings did not work for humans. This was taken by some as a basis for declaring human flight impossible in principle, in accordance with the pious cliché "If God had intended humans to fly, He would have given them wings."⁴

Today human air travel is commonplace. Furthermore, we now know that a sparrow remains air-borne in accordance with the same aero-dynamic principles as a jumbo jet. Thus, there is a certain level of abstraction at which the flights of sparrows and jumbo jets function in the same way.

Similarly, the modeling of natural communication requires an abstract theory which applies to human and artificial cognitive machines alike. Thereby, one naturally runs the risk of setting the level of abstraction either too low or too high. As in the case of flying, the crucial problem is finding the correct level of abstraction.

A level of abstraction which is too low is exemplified by closed signal systems such as vending machines. Such machines are inappropriate as a theoretical model because they fail to capture the diversity of natural language use, i.e., the characteristic property that one and the same expression can be used meaningfully in different contexts.

A level of abstraction which is too high, on the other hand, is exemplified by naive anthropomorphic expectations. For example, a notion of 'proper understand-

² See also CoL, p. 317.

³ In 1903, the brothers Orville and Wilbur Wright succeeded with the first manned motorized flight.

⁴ Irrational reasons against a modeling of natural communication reside in the subconscious fear of creating artificial beings resembling humans and having superhuman powers. Such *homunculi*, which occur in the earliest of mythologies, are regarded widely as violating a tabu. The tabu of Doppelgänger-similarity is described in Girard 1974.

Besides dark versions of homunculi, such as the cabalistically inspired Golem and the electrically initialized creature of the surgeon Dr. Frankenstein, the literature provides also more lighthearted variants. Examples are the piano-playing doll automata of the 18th century, based on the anatomical and physical knowledge of their time, and the mechanical beauty singing and dancing in *The Tales of Hoffmann*. More recent is the robot C3PO in George Lucas' film *Star Wars*, which represents a positive view of human-like robots.

ing' which requires that the computational system be subtly amused when scanning *Finnegan's Wake* is as far off the mark as a notion of 'proper flying' which requires mating and breeding behavior from a jumbo jet.⁵

VII. ANALYZING HUMAN COGNITION

The history of mechanical flight shows how a natural process (bird flight) poses a conceptually simple and obvious problem to science. Despite great efforts it was unsolvable for a long time. In the end, the solution turned out to be a highly abstract mathematical theory. In addition to being a successful foundation of mechanical flight, this theory is able to explain the functioning of natural flight as well.

This is why the abstract theory of aero-dynamics has led to a new appreciation of nature. Once the development of biplanes, turboprops, and jets resulted in a better theoretical and practical understanding of the principles of flight, interest was refocused again on the natural flight of animals in order to grasp their wonderful efficiency and power. This in turn led to major improvements in artificial flight, resulting in less noisy and more fuel-efficient air planes.

Applied to computational linguistics, this analogy illustrates that our highly abstract and technological approach does not imply a lack of interest in the human language capacity. On the contrary, investigating the specific properties of human language communication is theoretically meaningful only *after* the mechanism of natural language communication has been modeled computationally and proven successful in concrete applications on massive amounts of data.

VIII. INTERNAL AND EXTERNAL TRUTHS

In science we may distinguish between internal and external truths. Internal truths are conceptual models, developed and used by scientists to explain certain phenomena, and held true by relevant parts of society for limited periods of time. Examples are the Ptolemaic (geocentric) view of planetary motion or Bohr's model of the atom.

External truths are the bare facts of external reality which exist irrespective of whether or not there are cognitive agents to appreciate them. These facts may be measured more or less accurately, and explained using conceptual models.

Because conceptual models of science have been known to change radically in the course of history, internal truths must be viewed as *hypotheses*. They are justified mainly by the degree to which they are useful for arriving at a systematic description of external truths, represented by sufficiently large amounts of real data.

Especially in the natural sciences, internal truths have improved dramatically over the last five centuries. This is shown by an increasingly close fit between theoretical predictions and data, as well as a theoretical consolidation exhibited in the form of

⁵ Though this may seem quite reasonable from the viewpoint of sparrows.

greater mathematical precision and greater functional coherence of the conceptual (sub)models.

In contrast, contemporary linguistics is characterized by a lack of theoretical consolidation, as shown by the many disparate theories of language⁶ and the overwhelming variety of competing theories of grammar.⁷ As in the natural sciences, however, there is external truth also in linguistics. It may be approximated by completeness of empirical data coverage and functional modeling.

IX. LINGUISTIC VERIFICATION

The relation between internal and external truth is established by means of a *verification method*. The verification method of the natural sciences consists in the principle that experiments must be repeatable. This means that, given the same initial conditions, the same measurements must result again and again.

On the one hand, this method is not without problems because experimental data may be interpreted in different ways and may thus support different, even conflicting, hypotheses. On the other hand, the requirements of this method are so minimal that by now no self-respecting theory of natural science can afford to reject it. Therefore the repeatability of experiments has managed to channel the competing forces in the natural sciences in a constructive manner.

Another aspect of achieving scientific truth has developed in the tradition of mathematical logic. This is the principle of formal consistency, as realized in the method of axiomatization and the rule-based derivation of theorems.

Taken by itself the quasi-mechanical reconstruction of mathematical intuition in the form of axiom systems is separate from the facts of scientific measurements. As the logical foundation of natural science theories, however, the method of axiomatization has proven to be a helpful complement to the principle of repeatable experiments.

In linguistics, corresponding methods of verification have been sorely missed. To make up for this shortcoming there have been repeated attempts to remodel linguistics into either a natural science or a branch of mathematical logic. Such attempts are bound to fail, however, for the following reasons:

- The principle of repeatable experiments can only be applied under precisely defined conditions suitable for measuring. The method of experiments is not suitable for the objects of linguistic description because they are *conventions* that have developed over the course of centuries and exist as the intuitions (‘Sprachgefühl’) of the native speaker-hearer.

⁶ Examples are nativism, behaviorism, structuralism, speech act theory, model theory, as well Givón’s iconicity, Lieberman’s neostructuralism, and Halliday’s systemic approach.

⁷ Known by acronyms such as TG (with its different manifestations ST, EST, REST, and GB), LFG, GPSG, HPSG, CG, CCG, CUG, FUG, UCG, etc. These theories of grammar concentrate mostly on an initial foundation of internal truths such as ‘psychological reality,’ ‘innate knowledge,’ ‘explanatory adequacy,’ ‘universals,’ ‘principles,’ etc., based on suitably selected examples. Cf. Section 9.5.

- The method of axiomatization can only be applied to theories which have consolidated on a high level of abstraction, such as Newtonian mechanics, thermodynamics, or the theory of relativity. In today's linguistics, there is neither the required consolidation of theory nor completeness of data coverage. Therefore, any attempt at axiomatization in current linguistics is bound to be empirically vacuous.

Happily, there is no necessity to borrow from the neighboring sciences in order to arrive at a methodological foundation of linguistics. Instead, theories of language and grammar are to be implemented as electronic models which are tested automatically on arbitrarily large amounts of real data as well as in real applications of spontaneous human-computer communication. This method of verifying or falsifying linguistic theories objectively is specific to computational linguistics and may be viewed as the counterpart of the repeatability of experiments in the natural sciences.

X. EMPIRICAL DATA AND THEIR THEORETICAL FRAMEWORK

The methodology of computational linguistics presupposes a theory of language which defines the goals of empirical analysis and provides the framework into which components are to be embedded without conflict or redundancy. The development of such a framework can be extraordinarily difficult, as witnessed again and again in the history of science.

For example, in the beginning of astronomy scientists wrestled for centuries with the problem of providing a functional framework to explain the measurements that had been made of planetary motion and to make correct predictions based on such a framework. It was comparatively recently that Kepler (1571–1630) and Newton (1642–1727) first succeeded with a description which was both empirically precise and functionally simple. This, however, required a radical revolution in the theory of astronomy.

The revolution affected the *structural hypothesis* (transition from geo- to heliocentrism), the *functional explanation* (transition from crystal spheres to gravitation in space), and the *mathematical model* (transition from a complicated system of epicycles to the form of ellipses). Furthermore, the new system of astronomy was constructed at a level of abstraction where the dropping of an apple and the trajectory of the moon are explained as instantiations of one and the same set of general principles.

In linguistics, a corresponding scientific revolution has long been overdue. Even though the empirical data and the goals of their theoretical description are no less clear in linguistics than in astronomy, linguistics has not achieved a comparable consolidation in the form of a comprehensive, verifiable, functional theory of language.⁸

⁸ From a history of science point of view, the fragmentation of today's linguistics resembles the state of astrology and astronomy before Kepler and Newton.

XI. PRINCIPLES OF THE SLIM THEORY OF LANGUAGE

The analysis of natural communication should be structured in terms of methodological, empirical, ontological, and functional principles of the most general kind. The SLIM theory of language presented in this book is based on surface compositional, linear, internal matching. These principles are defined as follows.

1. *Surface compositional* (methodological principle)
Syntactic-semantic composition assembles only concrete word forms, excluding the use of zero-elements, identity mappings, or transformations.
2. *Linear* (empirical principle)
Interpretation and production of utterances are based on a strictly time-linear derivation order.
3. *Internal* (ontological principle)
Interpretation and production of utterances are analyzed as cognitive procedures located inside the speaker-hearer.
4. *Matching* (functional principle)
Referring with language to past, current, or future objects and events is modeled in terms of pattern matching between language meaning and context.

These principles originate in widely different areas (methodology, ontology, etc.), but within the SLIM theory of language they interact very closely. For example, the functional principle of (4) matching can only be implemented on a computer if the overall system is handled ontologically as (3) an internal procedure of the cognitive agent. Furthermore, the methodological principle of (1) surface compositionality and the empirical principle of (2) time-linearity can be realized within a functional mechanism of communication only if the overall theory is based on internal matching (3,4).

In addition to what its letters stand for, the acronym SLIM is motivated as a word with a meaning like *slender*. This is so because detailed mathematical and computational investigations have proven SLIM to be efficient in the areas of syntax, semantics, and pragmatics – both relatively in comparison to existing alternatives, and absolutely in accordance with the formal principles of mathematical complexity theory.

XII. CHALLENGES AND SOLUTIONS

The SLIM theory of language is defined on a level of abstraction where the mechanism of natural language communication in humans and in suitably constructed cognitive machines is explained in terms of the same principles of surface compositional, linear, internal matching.⁹ This is an important precondition for unrestricted human-

⁹ Moreover, the *structural hypothesis* of the SLIM theory of language is a regular, strictly time-linear derivation order – in contrast to grammar systems based on constituent structure. The *functional explanation* of SLIM is designed to model the mechanism of natural communication as a speaking robot – and not some tacit language knowledge innate in the speaker-hearer which excludes language use (performance). The *mathematical model* of SLIM is the continuation-based algorithm of LA-grammar, and not the substitution-based algorithms of the last 50 years.

computer communication in natural language. Its realization requires general and efficient solutions in the following areas.

First, the hearer's *understanding* of natural language must be modeled. This process is realized as the automatic reading-in of propositions into a database and – most importantly – determining the correct place for their storage and retrieval. The foundation of the semantic primitives is handled in terms of natural or artificial recognition and action.

Second, how the speaker determines the contents to be expressed in language must be modeled. This process, traditionally called *conceptualization*, is realized as an autonomous navigation through the propositions of the internal database. Thereby speech production is handled as a direction reflection (internal matching) of the navigation path in line with the motto: *Speech is verbalized thought*.

Third, the speaker and the hearer must be able to draw *inferences* on the basis of the contents of their respective databases. Inferences are realized as a special form of the autonomous time-linear navigation resulting in the derivation of new propositions. Inferences play an important role in the pragmatic interpretation of natural language, both in the hearer and the speaker.

The formal basis of time-linear navigation consists in concatenated propositions stored in a network database as a set of word tokens. A word token is a feature structure with the special property that it explicitly specifies the possible continuations to other word tokens, both within its proposition and from its proposition to others. This novel structure is called a *word bank* and provides the 'railroad tracks' for the navigation of a mental focus point. The navigation is powered and controlled by suitable LA-grammars (motor algorithms) which compute the possible continuations from one word token to the next.

The word bank and its motor algorithms constitute the central processing unit of an artificial cognitive agent called SLIM machine. The word bank is connected to external reality via the SLIM machine's recognition and action. The interpretation of perception, both verbal and nonverbal, results in concatenated propositions which are read into the word bank. The production of action, both verbal and nonverbal, is based on realizing some of the propositions traversed during the autonomous navigation.

CONCLUDING REMARK

In summary, the vision of unrestricted natural language communication between humans and machines is like the vision of motorized flight a hundred years ago: largely solved theoretically, but not yet realized in practical systems. At this point, all it will take to really succeed in computational linguistics is a well-directed, concentrated, sustained effort in cooperation with robotics, artificial intelligence, and psychology.

Part I

Theory of Language

1. Computational language analysis

The practical development of computers began around 1940. From then on there evolved a basic distinction between numerical and nonnumerical computer science.

Numerical computer science specializes in the calculation of numbers. In the fields of physics, chemistry, economics, sociology, etc., it has led to a tremendous expansion of scientific knowledge. Also many applications like banking, air travel, stock inventory, manufacturing, etc., depend heavily on numerical computation. Without computers and their software, operations could not be maintained in these areas.

Nonnumerical computer science deals with the phenomena of perception and cognition. Despite hopeful beginnings, nonnumerical computer science soon lagged behind the numerical branch. In recent years, however, nonnumerical computer science has made a comeback as artificial intelligence and cognitive science. These new, interdisciplinary fields investigate and electronically model natural information processing.

The term computational linguistics refers to that subarea of nonnumerical computer science which deals with language production and language understanding. Like artificial intelligence and cognitive science in general, computational linguistics is a highly interdisciplinary field which comprises large sections of traditional and theoretical linguistics, lexicography, psychology of language, analytic philosophy and logic, text processing, and the interaction with databases, as well as the processing of spoken and written language.

1.1 Human-computer communication

The goal of nonrestricted human-computer communication presupposes solutions to the most basic tasks of natural language analysis. Realizing this goal is therefore the ultimate standard for a successful computational linguistics.

Today, human-computer communication is still limited to highly *restricted* forms. Consider, for example, the interaction between the user and a standard computer, such as a PC or a work station. These machines provide a keyboard for the input of letters and a screen for the output of letters and pictures.¹

¹ For simplicity, we are disregarding additional input and output devices, such as mouse and sound, respectively.

It is conceivable that one could expand the notion of human-machine communication to machines

Computers are comfortable for entering, editing, and retrieving natural language, at least in the medium of writing, for which reason they have replaced electric typewriters. For utilizing the computers' abilities beyond word processing, however, commands using artificial languages must be applied. These are called programming languages, and are especially designed for controlling the computer's electronic operations.

In contrast to natural languages, which are flexible and rely on the seemingly obvious circumstances of the utterance situation, common background knowledge, the content of earlier conversations, etc., programming languages are inflexible and refer directly, explicitly, and exclusively to operations of the machine. For most potential users, a programming language is difficult to handle because (a) they are not familiar with the operations of the computer, (b) the expressions of the programming language differ from those of everyday language, and (c) the use of the programming language requires great precision.

Consider, for example, a standard database² which stores information about the employees of a company in the form of records:

1.1.1 EXAMPLE OF A RECORD-BASED DATABASE

	last name	first name	place	...
A1	Schmidt	Peter	Bamberg	...
A2	Meyer	Susanne	Nürnberg	...
A3	Sanders	Reinhard	Schwabach	...
	⋮	⋮	⋮	

The rows, named by different attributes like *first name*, *last name*, etc., are called the fields of the record type. The lines A1, A2, etc., each constitute a record. Based on this fixed record structure, the standard operations for the retrieval and update of information in the database are defined.

To retrieve the name of the representative in, for example, Schwabach, the user must type in the following commands of the programming language (here, a query language for databases) without mistake:

which do not provide general input/output components for language signs. Consider, for example, the operation of a contemporary washing machine. Leaving aside the loading of laundry and the measuring of detergent, the 'communication' consists in choosing a program and a temperature and pushing the start button. The machine 'answers' by providing freshly laundered laundry once the program has run its course.

Such an expanded notion of human-machine communication should be avoided, however, because it fosters misunderstandings. Machines without general input/output facilities for language constitute the special case of *nonverbal* human-machine communication, which may be neglected for the purposes of computational linguistics.

² As introductions to databases see C. Date 1990⁴ and R. Elmasri & S. Navathe 1989. We will return to this topic in Chapter 22 in connection with the interpretation of natural language.

1.1.2 DATABASE QUERY

Query:

```
select A#
where city = 'Schwabach'
```

Result:

```
result: A3 Sanders Reinhard
```

The correct use of commands such as 'select' initiates quasi-mechanical procedures which correspond to filing and retrieving cards in a filing cabinet with many compartments. Compared to the nonelectronic method, the computational system has many practical advantages. The electronic version is faster, the adding and removing of information is simpler, and the possibilities of search are much more powerful because various different keywords may be logically combined into a complex query.³ Is it possible to gradually extend such an interaction with a computer to natural language?

Standard computers have been regarded as general purpose machines for information processing because any kind of standard program can be developed and installed on them. From this point of view, their capabilities are restricted only by hardware factors like available speed and memory. In another sense, the information processing of standard computers is not general purpose, however, because their input and output facilities are restricted to the language channel.

A second type of computer not subject to this limitation is autonomous robots. In contradistinction to standard computers, robots are not restricted to the language channel, but designed to recognize their environment and to act in it.⁴

Corresponding to the different technologies of standard computers and robots, there have evolved two different branches of artificial intelligence. One branch, dubbed classic AI by its opponents, is based on standard computers. The other branch, which calls itself nouvelle AI,⁵ requires the technology of robots.

Classic AI analyzes intelligent behavior in terms of manipulating abstract symbols. A typical example is a chess-playing program.⁶ It operates in isolation from the rest of the world, using a fixed set of predefined pieces and a predefined board. The search space for a dynamic strategy of winning in chess is astronomical. Yet the technology of a standard computer is sufficient because the world of chess is closed.

Nouvelle AI aims at the development of autonomous agents. In contrast to systems which respond solely to a predefined set of user commands and behave otherwise in isolation, autonomous agents are designed to interact with their real world environment. Because the environment is constantly changing in unpredictable ways they must continually keep track of it by means of sensors.

³ See also Section 2.1.

⁴ Today three different generations of robots are distinguished. Most relevant for computational linguistics are robots of third generation, which are designed as autonomous agents. See D.W. Wloka 1992.

⁵ See for example P. Maes (ed.) 1990.

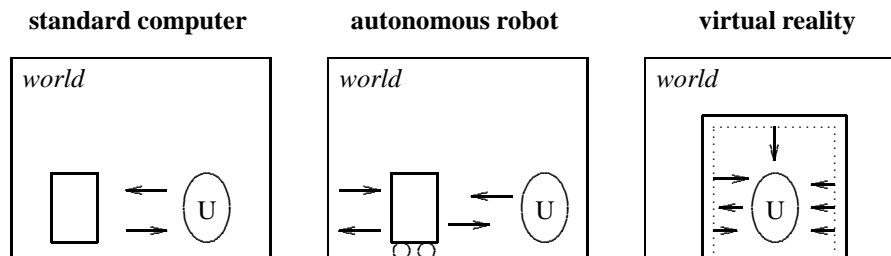
⁶ A. Newell & H. Simon 1972, R. Reddy et al. 1973.

For this, nouvelle AI uses the strategy of task level decomposition. Rather than building and updating one giant global representation to serve as the basis of automatic reasoning, nouvelle AI systems aim at handling their tasks in terms of many interacting local procedures controlled by perception. Thereby low-level inferencing operates directly on the local perception data.

A third type of machine processing information – besides standard computers and robots – is systems of virtual reality (VR).⁷ While a robot analyzes its environment in order to influence it in certain ways (such as moving in it), a VR system aims at creating an artificial environment for the user. Thereby the VR system reacts to the movements of the user's hand, the direction of his/her gaze, etc., and utilizes them in order to create as realistic an environment as possible.

The different types of human-computer communication exemplified by standard computers, robots, and VR systems may be compared schematically as follows:

1.1.3 THREE TYPES OF HUMAN-COMPUTER INTERACTION



The ovals represent the users who face the respective systems in the 'world.' The arrows represent the interaction of the systems with their environment and the user.

A standard computer communicates with users who initiate the interaction. A robot interacts independently with its environment and its users. A VR system does not interact with its environment, but rather creates an artificial environment for the user. In robots and VR systems, communication with the user in terms of language is optional and may be found only in advanced systems. These systems must always have a language-based 'service channel,' however, for the installation and upgrading of the system software.

1.2 Language science and its components

A speaker of English knows the meaning of a word like **red**. When asked to pick the red object among a set of non-red objects, for example, a competent speaker-hearer will be able to do it. A standard computer, on the other hand, does not 'understand' what **red** means, just as a piece of paper does not understand what is written on it.⁸

⁷ For an introduction see A. Wexelblat (ed.) 1993.