# Foundations of Computational Linguistics

*man-machine communication in natural language*

ROLAND HAUSSER
Computational Linguistics
Universität Erlangen Nürnberg
Germany


CLUE

## Part I Theory of Language

# Introduction

## 1. Requirements for modeling natural communication

- 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,

## 2. Consequences of using parsers

- *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 arbirarily 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.

## 3. Principles of the SLIM theory of language

1. ***S**urface compositional* (methodological principle)
   Syntactic-semantic composition assembles only concrete word forms, excluding the use of zero-elements, identity mappings, or transformations.
2. ***L**inear* (empirical principle)
   Interpretation and production of utterances is based on a strictly time-linear derivation order.
3. ***I**nternal* (ontological principle)
   Interpretation and production of utterances is analyzed as cognitive procedures located inside the speaker-hearer.
4. ***M**atching* (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.

# Part I
# Theory of Language

# 1. Computational language analysis

## 1.1 Man-machine communication

### 1.1.1 Restricted vs. nonrestricted communication

### 1.1.2 Example of restricted communication: a record-based database

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

### 1.1.3 Database query

Query:

```
select A#
where city = ‘Schwabach’
```

Result:

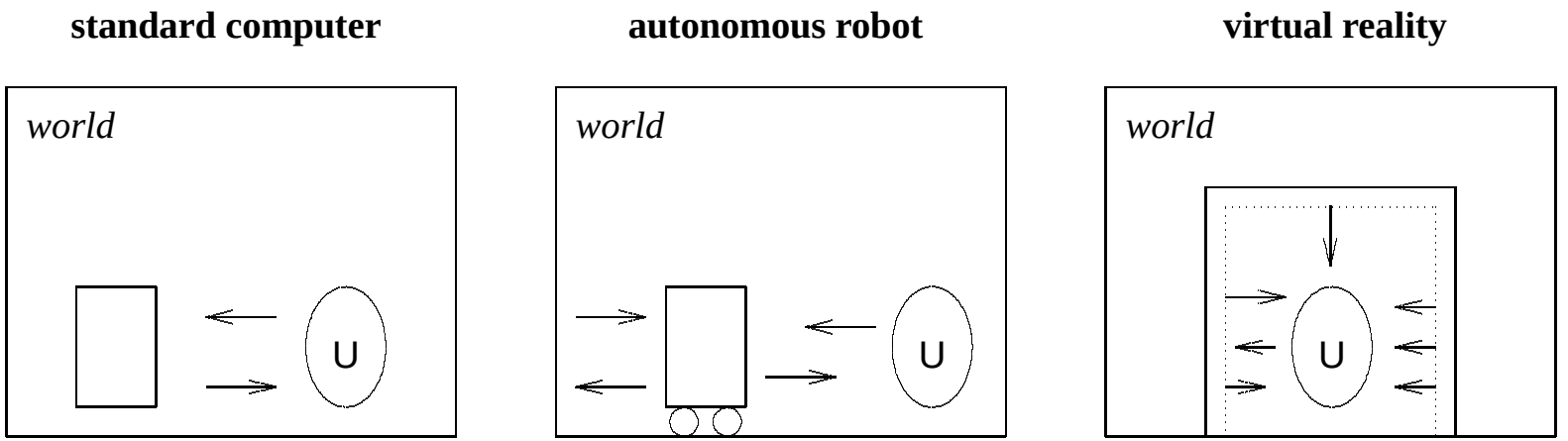
```
A3 Sanders Reinhard
```

### 1.1.4 Classic AI vs Nouvelle AI

Classic AI analyzes intelligent behavior as manipulating symbols. For example, a chess playing program 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 is astronomical. Yet a standard computer is sufficient because the world of chess is closed.

Nouvelle AI aims at the development of autonomous agents. In order interact with their real world environment, they must continually keep track of changes by means of sensors. For this, nouvelle AI uses robots. The strategy of task level decomposition defines inferencing to operate directly on the local perception data.

### 1.1.5 Three types of man-machine communication



# 1.2 Language science and its components

## 1.2.1 Variants of language science

- *Traditional Grammar*
- *Theoretical Linguistics*
- *Computational Linguistics*

## 1.2.2 The components of grammar

- *Phonology*: Science of language sounds
- *Morphology:* Science of word form structure
- *Lexicon:* Listing analyzed words
- *Syntax:* Science of composing word forms
- *Semantics:* Science of literal meanings
- *Pragmatics:* Science of using language expressions

# 1.3 Methods and applications of computational linguistics

## 1.3.1 Methodology of parsing

1. *Decomposition* of a complex sign into its elementary components,
2. *Classification* of the components via lexical lookup, and
3. *Composition* of the classified components via syntactic rules in order to arrive at an overall grammatical analysis of the complex sign.

## 1.3.2 Practical tasks of computational linguistics

- Indexing and retrieval in textual databases
- Machine translation
- Automatic text production
- Automatic text checking
- Automatic content analysis
- Automatic tutoring
- Automatic dialog and information systems

# 1.4 Electronic medium in recognition and synthesis

## 1.4.1 Media of language

Nonelectronic media:

- *sounds* of spoken language
- *letters* of handwritten or printed language
- *gestures* of sign language

Electronic medium:

- Realization-dependent representations:
  tape recording of spoken language
  bitmap of written language
  video recording of signed language
- Realization-independent representations:
  digitally coded electronic sign sequences, e.g. ASCII

### 1.4.2 Transfer between realization-dependent and -independent representations

*recognition*: i⇒d transfer
Realization-dependent representations must be mapped into realization-independent ones.

*synthesis*: d⇒i transfer
Realization-independent representations must be mapped into realization-dependent ones.

### 1.4.3 Methods of d⇒i transfer

Nonautomatic: typing spoken or written language into the computer

Automatic: Acoustic or optical pattern recognition

### 1.4.4 Desiderata of automatic speech recognition

The quality of automatic speech recognition should be at least equal to that of an average human hearer.

- *Speaker independence*
  The system should understand speech of an open range of speakers with varying dialects, pitch, etc. – without the need for an initial learning phase to adapt the system to one particular user.

- *Continuous speech*
  The system should handle continuous speech at different speeds – without the need for unnatural pauses between individual word forms.
- *Domain independence*
  The system should understand spoken language independently of the subject matter – without the need of telling the system in advance which vocabulary is to be expected and which is not.
- *Realistic vocabulary*
  The system should recognize at least as many word forms as an average human.
- *Robustness*
  The system should recover gracefully from interruptions, contractions, and slurring of spoken language, and be able to infer the word forms intended.

### 1.4.5 The crucial question for designing truly adequate speech recognition

*How should the domain and language knowledge best be organized?*

The answer is obvious:

*The domain and language knowledge should be organized within a functional theory of language which is mathematically and computationally efficient.*

# 1.5 Second Gutenberg Revolution

## 1.5.1 The First Gutenberg Revolution

Based on the technological innovation of printing with movable letters, it made a wealth of knowledge available to a broader public.

## 1.5.2 The Second Gutenberg Revolution

Based on the automatic processing of natural language in the electronic medium, it aims at facilitating access to specific pieces of information.

## 1.5.3 SGML: *standard generalized markup language.*

> A family of ISO standards for labeling electronic versions of text, enabling both sender and receiver of the text to identify its structure (e.g. title, author, header, paragraph, etc.)
>
> Dictionary of Computing, p. 416 (ed. Illingworth et al. 1990)

## 1.5.4 Newspaper text with SGML control symbols (excerpt)

```
<HTML>
<HEAD>
<TITLE>9/4/95 COVER: Siberia, the Tortured Land</TITLE>
</HEAD>
<BODY>
<!-- #include "header.html" -->
<P>TIME Magazine</P>
<P>September 4, 1995 Volume 146, No. 10</P>
<HR>
Return to <A href="../../../../../time/magazine/domestic/toc/
950904.toc.html">Contents page</A>
<HR>
<BR>
<!-- end include -->
<H3>COVER STORY</H3>
<H2>THE TORTURED LAND</H2>
<H3>An epic landscape steeped in tragedy, Siberia suffered
grievously under  communism. Now the world’s capitalists covet
its vast riches </H3>
<P><EM>BY <A href="../../../../../time/bios/eugenelinden.html">
EUGENE LINDEN</A>/YAKUTSK</EM>
<P>Siberia has come to mean a land of exile, and the place
easily fulfills its reputation as a metaphor for death and
```

### 1.5.5 Different types of text

- article
- book
- theater play
- movie script
- dictionary

### 1.5.6 TEI

Text encoding initiative: defines a DTD (*document type definition*) for the markup of different types of text in SGML.

### 1.5.7 Different goals of markup

- Function-oriented: SGML and TEI
- Print-oriented: TEXand LATEX
- User-oriented: Winword, WordPerfect, etc.

## 1.5.8 Alphabetical list of word forms (excerpt)

| | | |
|---|---|---|
| 10 | in | STORY |
| 146 | in | suffered |
| 1995 | in | sun |
| 20 | its | than |
| 4 | its | that |
| a | LAND | The |
| a | land | the |
| a | landscape | the |
| a | like | the |
| a | LINDEN | the |
| above | Magazine | the |
| across | markers | the |
| and | mean | the |
| and | metaphor | the |
| and | midnight | the |
| and | midsummer | the |
| Arctic | million | through |
| as | mist | Throughout |
| as | more | to |
| barracks | mossy | to |
| bits | muting | TORTURED |

# 2. Technology and grammar

## 2.1 Indexing and retrieval in textual databases

### 2.1.1 Indexing

The indexing of a textual database is based on a table which specifies for each letter all the positions (addresses) where it occurs in the storage medium of the database.

### 2.1.2 Advantages of an electronic index

- Power of search
- Flexibility
  - General specification of patterns
  - Combination of patterns
- Automatic creation of the index structure
- Ease, speed, and reliability
  - Query
  - Retrieval

### 2.1.3 Definition of recall and precision

*Recall* measures the percentage of relevant texts retrieved as compared to the total of relevant texts contained in the database.

For example: a database of several million pieces of text happens to contain 100 texts which are relevant to a given question. If the query returns 75 texts, 50 of which are relevant to the user and 25 are irrelevant, then the recall is 50 : 100 = 50%.

*Precision* measures the percentage of relevant texts contained in the result of a query.

For example: a query has resulted in 75 texts of which 50 turn out to be relevant to the user. Then the precision is 50 : 75 = 66.6%.

# 2.2 Using grammatical knowledge

## 2.2.1 Linguistic methods of optimization

A. Preprocessing the query

- Automatic query expansion
  (i) The search words in the query are automatically ‘exploded’ into their full inflectional paradigm and the inflectional forms are added to the query.
  (ii) Via a thesaurus the search words are related to all synonyms, hypernyms, and hyponyms. These are included in the query – possibly with all their inflectional variants.
  (iii) The syntactic structure of the query, e.g. A sold x to B, is transformed automatically into equivalent versions, e.g. B was sold x by A, x was sold to B by A, etc., to be used in the query.
- Interactive query improvement
  Prior to the search, the result of a query expansion is presented to the user to allow elimination of useless aspects of the automatic expansion and allow for an improved formulation of the query.

## B. Improving the indexing

- Letter-based indexing
  This is the basic technology of search, allowing to retrieve the positions of each letter and each letter sequence in the database.
- Morphologically-based indexing
  A morphological analyzer is applied during the reading-in of texts, relating each word form to its base form. This information is coded into an index which for any given word (base form) allows to find all corresponding (inflected) forms.
- Syntactically-based indexing
  A syntactic parser is applied during the reading-in of texts, eliminating morphological ambiguities and categorizing phrases. The grammatical information is coded into an index which allows to find all occurrences of a given syntactic construction.
- Concept-based indexing
  The texts are analyzed semantically and pragmatically, eliminating syntactic and semantic ambiguities as well as inferring special uses characteristic of the domain. This information is coded into an index which allows to find all occurrences of a given concept.

## C. Postquery processing

- The low precision resulting from a nonspecific formulation of the query may be countered by an automatic processing of the data retrieved. Because the raw data retrieved are small as compared to the database as a whole they may be parsed after the query and checked for their content. Then only those texts are given out which are relevant according to this post query analysis.

# 2.3 Smart versus solid solutions

## 2.3.1 Smart solutions

avoid difficult, costly, or theoretically unsolved aspects of the task at hand, such as

- Weizenbaum's Eliza program, which appears to understand natural language, but doesn't.
- Statistically-based tagging, which can guess the part of speech.
- Direct and transfer approaches in machine translation, which avoid understanding the source text.

## 2.3.2 Solid solutions

aim at a complete theoretical and practical understanding of the phenomena involved. Applications are based on ready-made off-the-shelf components such as

- Automatic word form analysis based on an online lexicon of the language and a rule-based morphological parser handling inflection, derivation and composition
- Syntactic parsing of free text based on the automatic word form analysis of the language
- Semantic interpretation of the syntactic analysis deriving the literal meaning
- Pragmatic interpretation relative to a context of use deriving the speaker meaning.

### 2.3.3 Comparison

- As long as the off-the-shelf components are not available, a smart solution seems initially cheaper and quicker. But smart solutions are costly to maintain and their accuracy cannot be substantially improved.
- The components of grammar are a long term investment that can be used again and again in a wide variety of different solid solutions. Improvements in the components of grammar lead directly to better applications.

### 2.3.4 Choice between smart or solid solution depends on application

- A smart solution providing a 70% recall in a large database is more than the user could hope to find by hand. Also, the user is not aware of what is missing in the query result.
- In contrast, the deficiencies of a smart solution providing 70% accuracy in machine translation are painfully obvious to the user. Furthermore, there are human translators available which are able to do a much better job.

# 2.4 Beginnings of machine translation

## 2.4.1 Language pairs

*French* $\rightarrow$ *English* and *French* $\leftarrow$ *English* are two different language pairs.

## 2.4.2 Formula to compute the number of language pairs

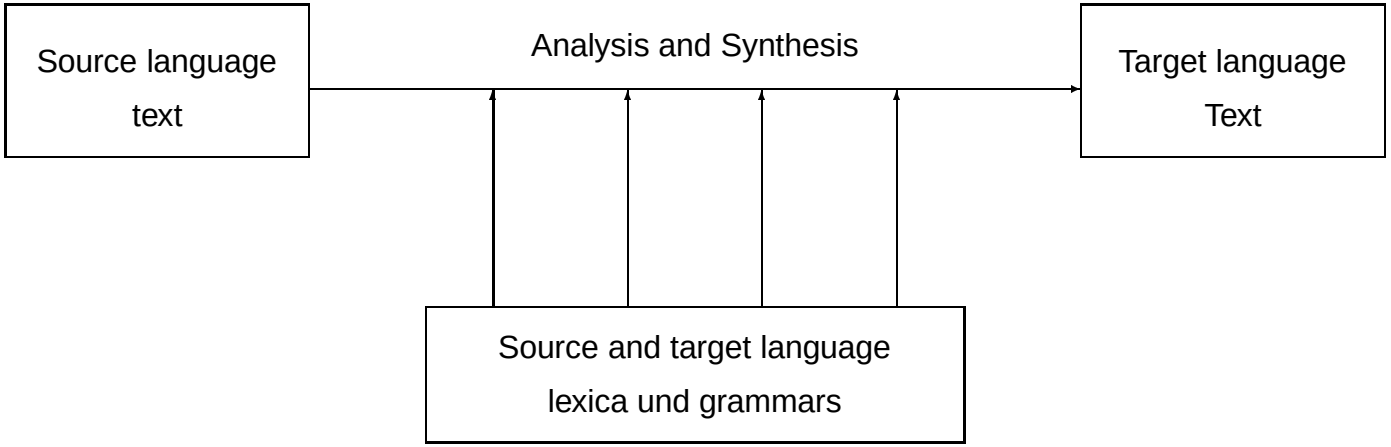
$n \cdot (n - 1)$, *where n = number of different languages*

For example, an EU with 11 different languages has to deal with a total of $11 \cdot 10 = 110$ language pairs.

## 2.4.3 Translating a French document in the EU

| | |
|---|---|
| French $\rightarrow$ English | French $\rightarrow$ Spanish |
| French $\rightarrow$ German | French $\rightarrow$ Portugese |
| French $\rightarrow$ Italian | French $\rightarrow$ Greek |
| French $\rightarrow$ Dutch | French $\rightarrow$ Danish |
| French $\rightarrow$ Swedish | French $\rightarrow$ Finnish |

## 2.4.4 Schema of direct translation




## 2.4.5 What is FAHQT?

FULLY AUTOMATIC HIGH QUALITY TRANSLATION

## 2.4.6 Examples of automatic mis-translations

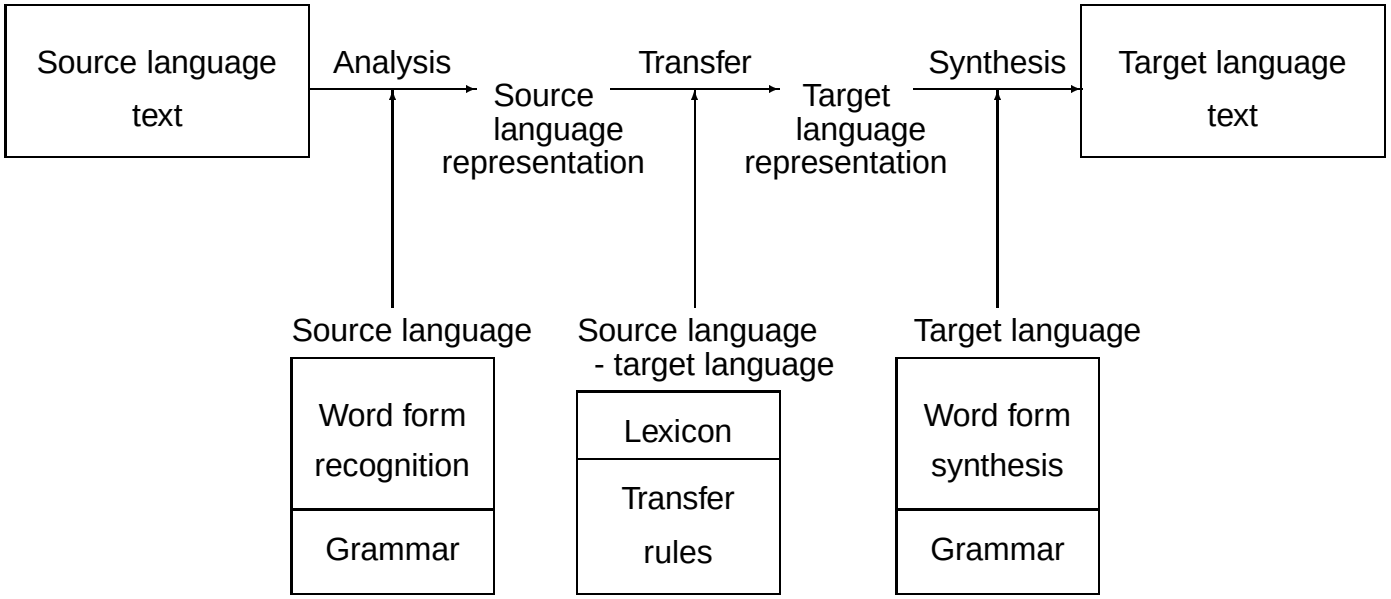
Out of sight, out of mind. $\Rightarrow$ *Invisible idiot.*
The spirit is willing, but the flesh is weak. $\Rightarrow$ *The whiskey is alright, but the meat is rotten.*
La Cour de Justice considère la création d'un sixième poste d'avocat général. $\Rightarrow$ *The Court of Justice is considering the creation of a sixth avocado station.*

## 2.4.7 The transfer approach

aims for a modular separation of the

- source language analysis and target language synthesis, of
- linguistic data and processing procedures, and of the
- lexica for source language analysis, target language transfer, and target language synthesis.

## 2.4.8 Schema of the transfer approach



## 2.4.9 Three phrases of a word form transfer *English-German*

1. Source language analysis:

   Unanalyzed surface: `knew`

   Morphological and lexical analysis: `(knew (N A V) know)`

   The source language analysis produces the syntactic category (N A V) of the inflectional form (categorization) and the base form know (lemmatization).

2. Source-target language transfer:

   Using the base form resulting from the source language analysis, a source-target language dictionary provides the corresponding base forms in the target language.

   `know` ⟹ `wissen`
   `kennen`

3. Target language synthesis

   Using the source language category (resulting from analysis) and the target language base forms (resulting from transfer), the desired target language word forms are generated based on target language morphology.

| | |
|---|---|
| wußte | kannte |
| wußtest | kanntest |
| wußten | kannten |
| wußtet | kanntet |

## 2.4.10 Shortcomings of the direct and the transfer approach

- Each language pair requires a special source-target component.
- Analysis and synthesis are limited to single sentences.
- Semantic and pragmatic analysis are avoided, attempting automatic translation without understanding the source language.

# 2.5 Machine translation today

## 2.5.1 Illustrating the importance of language understanding

- **Syntactic ambiguity in the source language**
  1. Julia flew and crashed the air plane.
     Julia (flew and crashed the air plane)
     (Julia flew) and (crashed the air plane)
  2. Susan observed the yacht with a telescope.
     Susan observed the man with a beard.
  3. The mixture gives off dangerous cyanide and chlorine fumes.
     (dangerous cyanide) and (chlorine fumes)
     dangerous (cyanide and chlorine) fumes

- **Lexical differences between source and target**
  1. The men killed the women. Three days later they were caught.
     The men killed the women. Three days later they were buried.
  2. know: wissen savoir
     kennen connaître
  3. The watch included two new recruits that night.

- **Syntactic differences between source and target**
  - German:
    Auf dem Hof sahen wir einen kleinen Jungen, der einem Ferkel nachlief.
    Dem Jungen folgte ein großer Hund.

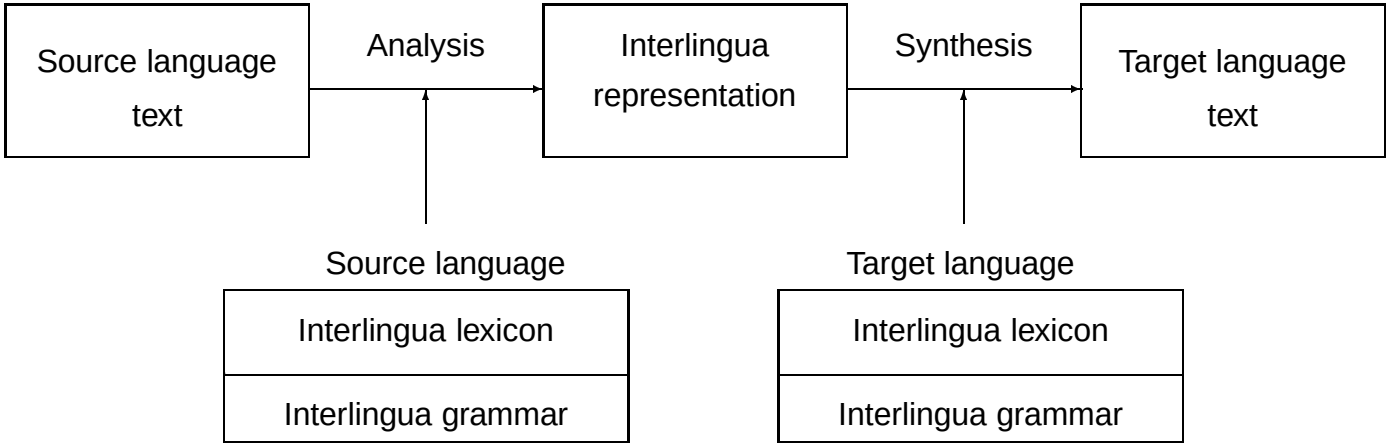
  - English:
    In the yard we saw a small boy running after a piglet.
    A large dog followed the boy.
    The boy was followed by a large dog.
- **Collocation and idiom**

  strong current | high voltage (but: *high current | *strong voltage)
  bite the dust | ins Gras beißen (but: *bite the grass | *in den Staub beißen)

### 2.5.2 Partial solutions for practical machine translation

1. *Machine aided translation* (MAT) supports human translators with comfortable tools such as on-line dictionaries, text processing, morphological analysis, etc.
2. *Rough translation* – as provided by an automatic transfer system – arguably reduces the translators' work to correcting the automatic output.
3. *Restricted language* provides a fully automatic translation, but only for texts which fulfill canonical restrictions on lexical items and syntactic structures.

## 2.5.3 Schema of the interlingua approach




## 2.5.4 Candidates proposed as interlingua

- an artificial logical language,
- a semi-natural language like Esperanto which is man-made, but functions like a natural language,
- a set of semantic primitives common to both, the source and the target language, serving as a kind of universal vocabulary.

# 3. Cognitive foundation of semantics

## 3.1 Prototype of communication

### 3.1.1 Variants of language communication

- two speakers are located face to face and talk about concrete objects in their immediate environment
- two speakers talk on the telephone about events they experienced together in the past
- a merchant writes to a company to order merchandise in a certain number, size, color, etc., and the company responds by filling the order
- a newspaper informs about a planned extension of public transportation
- a translator reconstructs an English short story in German
- a teacher of physics explains the law of gravitation
- a registrar issues a marriage license
- a judge announces a sentence
- a conductor says: Terminal station, everybody please get off.
- a sign reads: Do not step on the grass!
- a professor of literature interprets an expressionistic poem
- an author writes a science fiction story
- an actor speaks a role

### 3.1.2 Prototype of communication

> The basic prototype of natural communication is the direct face to face discourse of two partners talking about concrete objects in their immediate environment.

Possible alternatives: complete texts or signs of nature, such as smoke indicating fire.

### 3.1.3 Three components of the communication prototype

- Specification of the external *task environment*
- Structure of the *cognitive agent* including the internal *problem space*
- Specification of the *language*

### 3.1.4 Objects in the world of CURIOUS

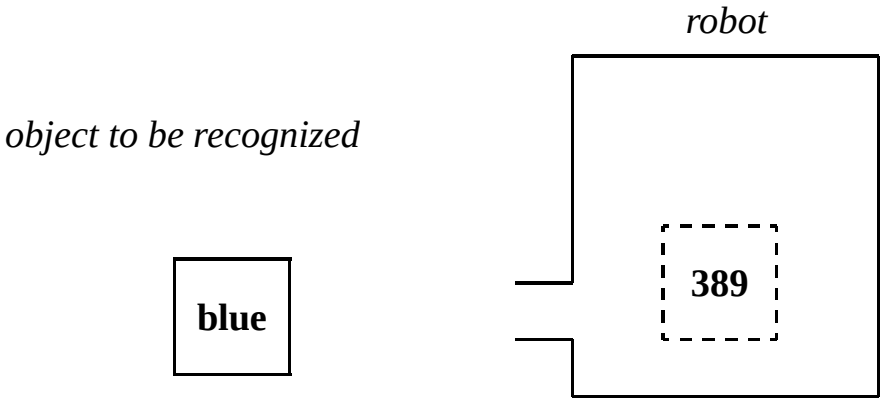
- triangles (scalene, isoceles, etc.)
- quadrangles (square, rectilinear, etc.)
- circles and ellipses

# 3.2 From perception to recognition

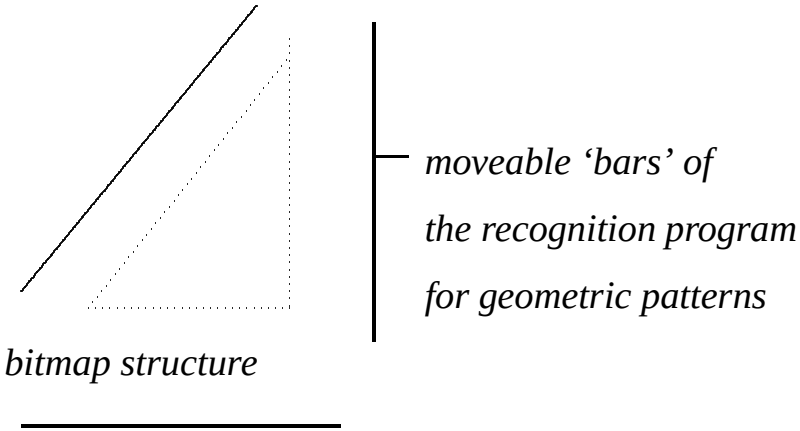
## 3.2.1 Two criteria to evaluate CURIOUS

- Measuring the active and reactive behavior (behavior test)
- Measuring the cognitive processing directly (cognition test)

## 3.2.2 Internal bitmap representation of external object



### 3.2.3 Analysis of an internal bitmap representation




### 3.2.4 Definition of the context

The context of a cognitive agent CA at a given point of time *t* includes

1. the total of all current cognitive parameter values $CA_t$,
2. the logical analyses of the parameter values and their combinations (reconstructed patterns),
3. the conceptual structures used to classify the reconstructed patterns and their combinations.

# 3.3 Iconicity of formal concepts

## 3.3.1 I-concept$_{loc}$ of a square (token)

$$\begin{bmatrix} \text{edge 1: 2cm} \\ \text{angle 1/2: } 90^0 \\ \text{edge 2: 2cm} \\ \text{angle 2/3: } 90^0 \\ \text{edge 3: 2cm} \\ \text{angle 3/4: } 90^0 \\ \text{edge 4: 2cm} \\ \text{angle 4/1: } 90^0 \end{bmatrix}_{loc}$$

## 3.3.2 Definition: I-concept$_{loc}$

An I-concept$_{loc}$ results from successfully matching an M-concept onto a corresponding parameter constellation at a certain space-time location.

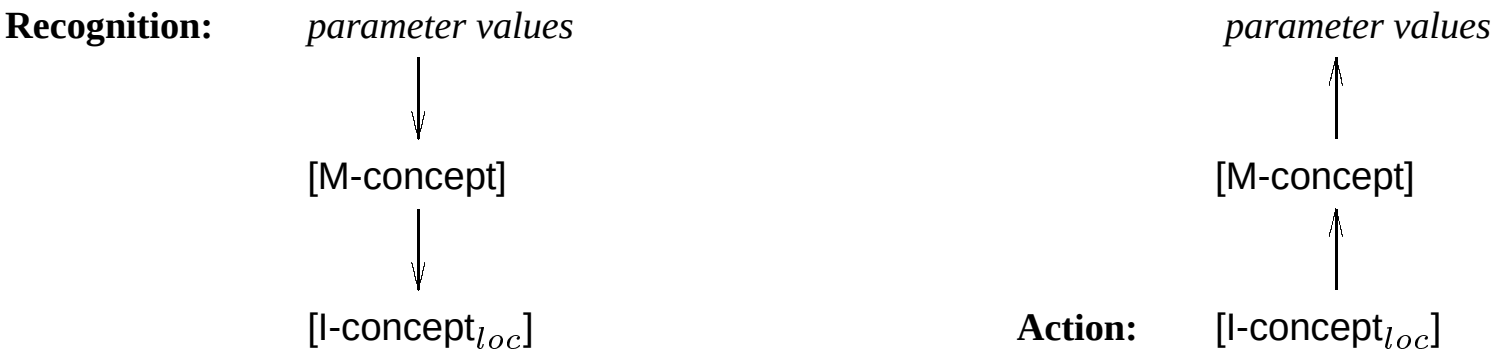
### 3.3.3 M-concept square (type)

$$\begin{bmatrix} \text{edge 1: } \alpha \text{ cm} \\ \text{angle 1/2: } 90^0 \\ \text{edge 2: } \alpha \text{ cm} \\ \text{angle 2/3: } 90^0 \\ \text{edge 3: } \alpha \text{ cm} \\ \text{angle 3/4: } 90^0 \\ \text{edge 4: } \alpha \text{ cm} \\ \text{angle 4/1: } 90^0 \end{bmatrix}$$

### 3.3.4 Definition: M-concept

An M-concept is the structural representation of a characteristic parameter constellation whereby certain parameter values are defined as variables.

## 3.3.5 Contextual recognition and action




## 3.3.6 Aspects of iconicity

- The parameter values of the internal context are images insofar as they reflect the corresponding structures of the real world.
- The reconstructed patterns (I-concepts$_{loc}$) are images of parameter values because they are logical analyses of parameter values.
- The M-concepts of the internal context are images insofar as they (i) originate as abstractions over similar parameter constellations and (ii) characterize associated classes of reconstructed patterns.

### 3.3.7 Fallacious arguments against iconicity

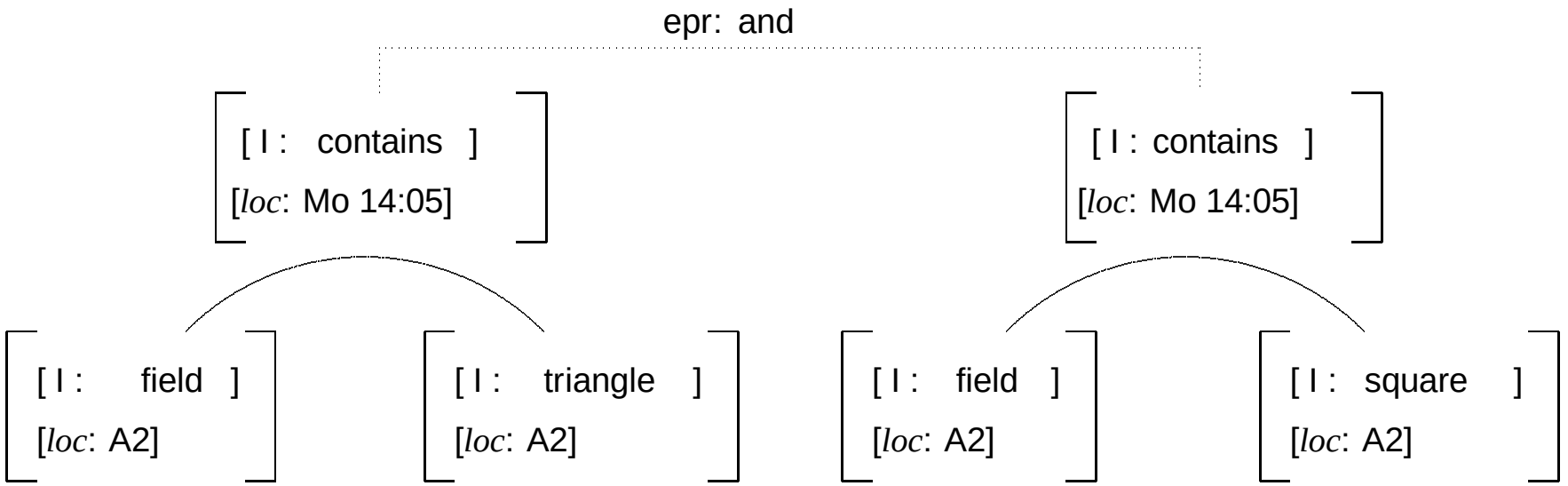
- If one were to surgically search the brain of a person who has seen a tree, one would not find such an image.
- In the case of, e.g., a triangle, what *kind* should the concept should be *exactly* : isoceles, scalene, right-angled? (BERKELEY 1685–1753).
- If there are pictures in the mind, then there must be someone to see them. Yet postulating a little man (homunculus) in the head to see the images would not do because the little man would have images in his little head in turn, requiring another homunculus, and so on. Since postulating a homunculus is of no help to understand the interpretation of images, the images themselves are concluded to be superfluous (HUME 1711–1776).

# 3.4 Contextual I-propositions

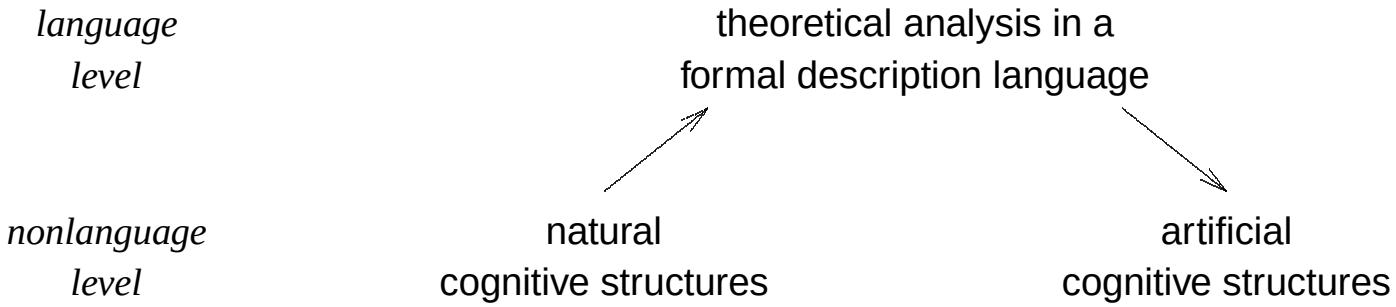
## 3.4.1 The three elements of basic propositions

| | *logic* | *world* | *language* |
|---|---|---|---|
| 1. | functor | relation | verb |
| 2. | argument | object | noun |
| 3. | modifier | property | adjective-adverbial |

## 3.4.2 An example of two contextual propositions

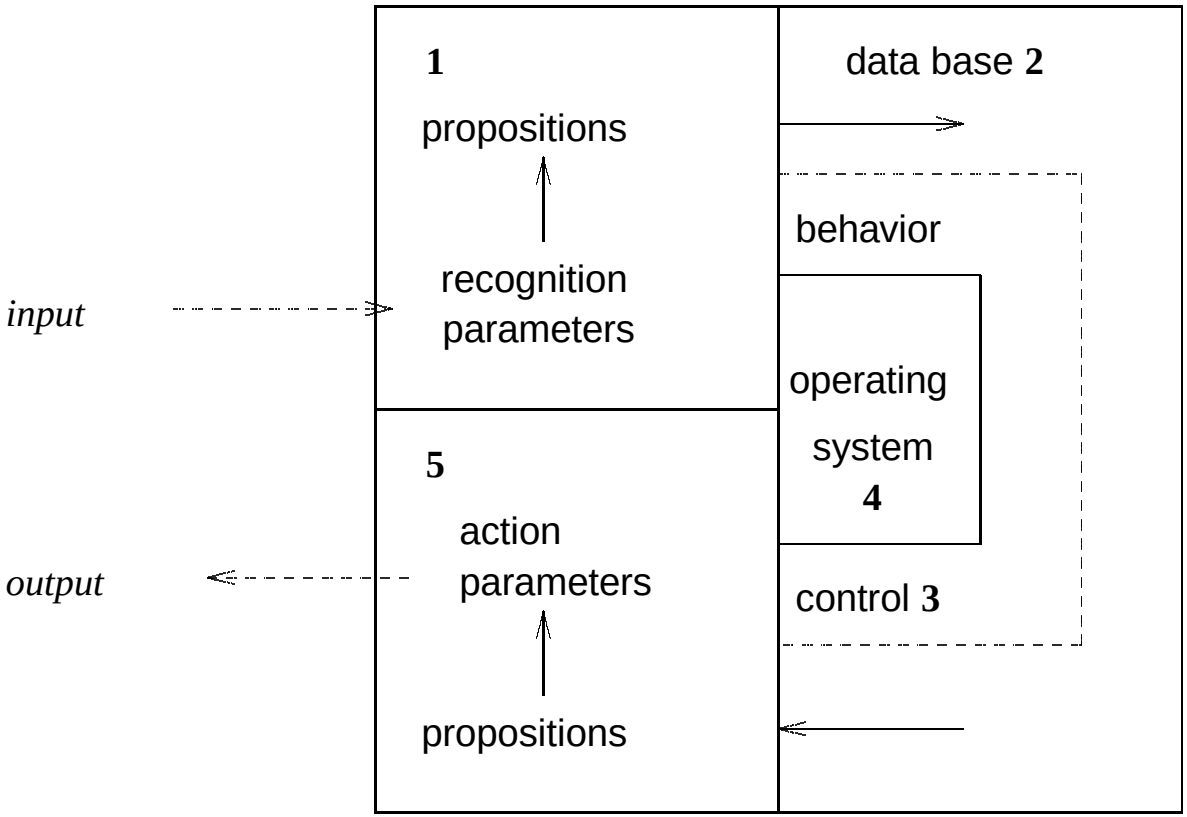


## 3.4.3 Artificial modeling of natural cognition



# 3.5 Recognition and action

## 3.5.1 Schematic structure of context-based cognition



### 3.5.2 Example of a behavior control program

1. Primary analysis of the current task environment:
   (a) Move into the start field A1.
   (b) Analyze the current field:
      i. Approximate bitmap outline with edge program.
      ii. Measure color value inside the bitmap outline.
      iii. Derive I-proposition.
   (c) Write I-proposition at index P-0.1 (present) into the database.
   (d) If current field is not D4, move into the next field and enter state b. Otherwise go to 2.
2. Secondary analysis of current task environment (inferences):
   (a) Count all triangles, rectangles, squares, red triangles, etc., in the primary analysis P-0.1 and write the result at index P-0.2 into the database.
   (b) Compare the current secondary analysis P-0.2 with the previous secondary analysis P-1.2 and write the result (e.g. ‘number of red triangle increased by 2’) at index P-10.3 into the database.
3. Wait for 10 minutes.
4. Return to state 1.

### 3.5.3 CURIOUS as an autonomous agent (nouvelle AI)

Without a carefully built physical grounding any symbolic representation will be mismatched to its sensors and actuators. These groundings provide the constraints on symbols necessary for them to be truly useful.'

R.A. Brooks, 1990, S. 6.

### 3.5.4 CURIOUS as a physical symbol system (classic AI)

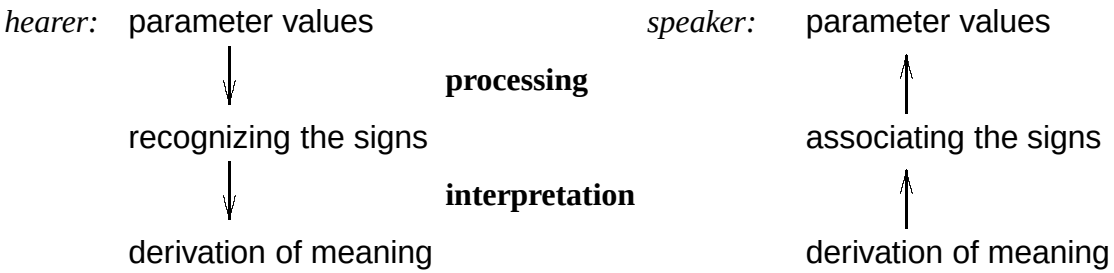
The total concept [of a physical symbol system] is the join of computability, physical realizability (and by multiple technologies), universality, the symbolic representation of processes (i.e. interpretability), and finally, symbolic structure and designation.

A. Newell & H. Simon 1975, p. 46

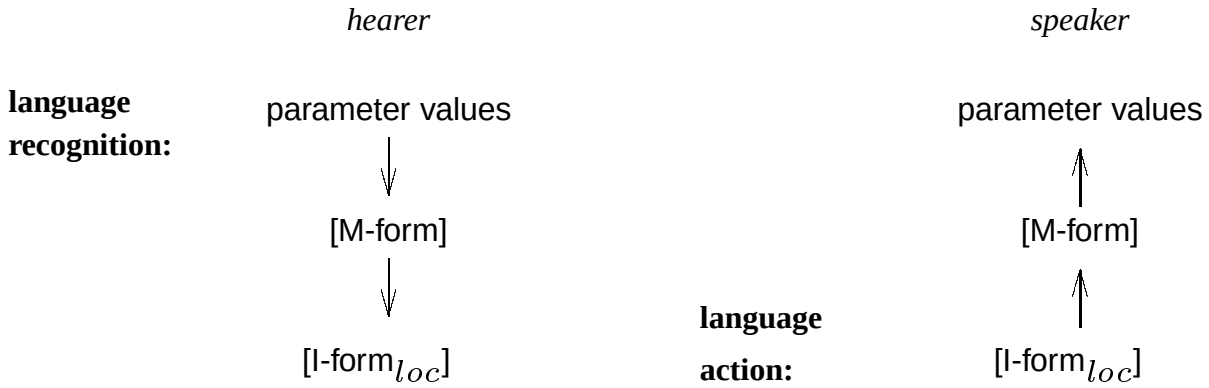
# 4. Language communication

## 4.1 Adding language

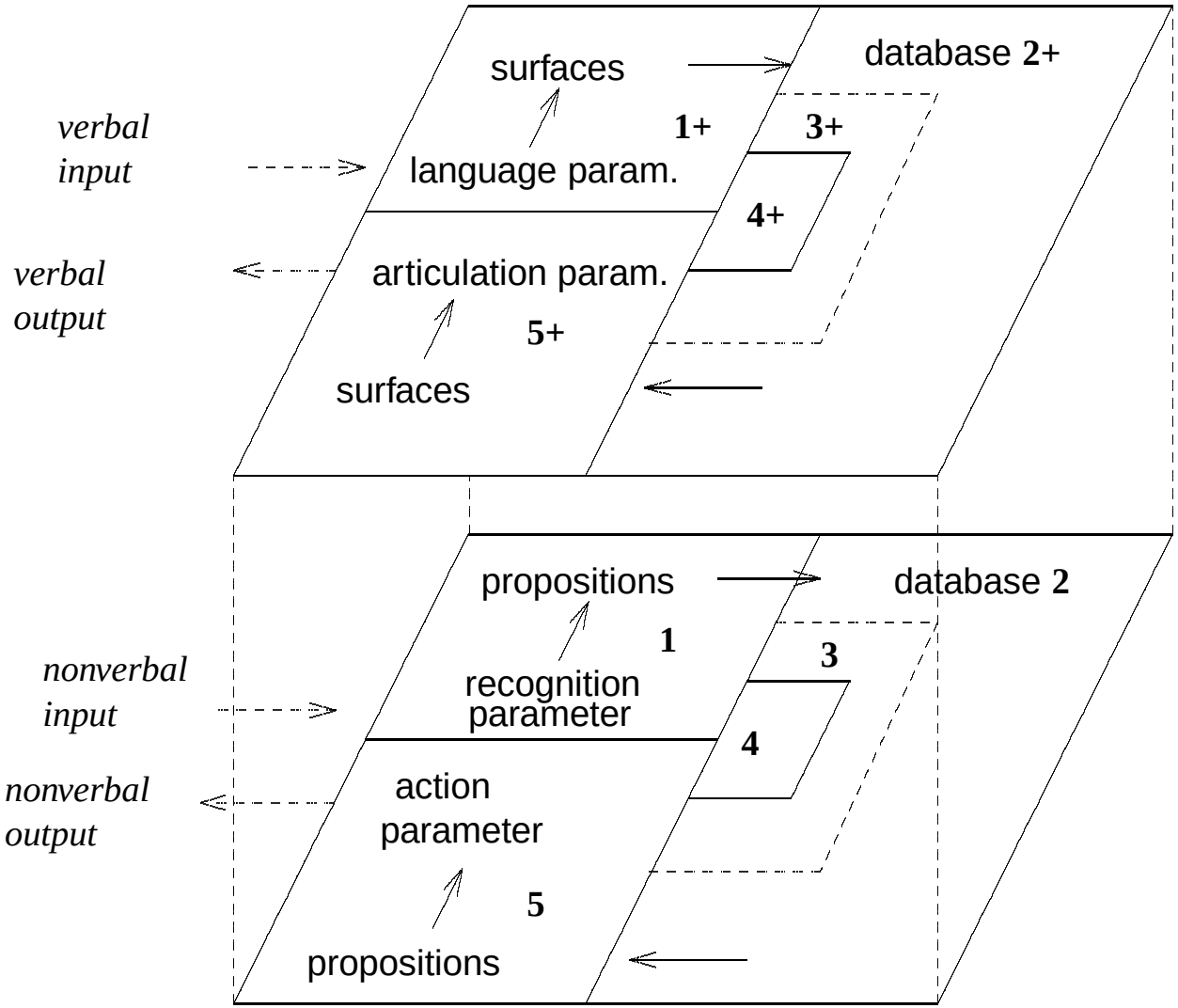
### 4.1.1 Two subprocedures of language use




### 4.1.2 Processing signs of language

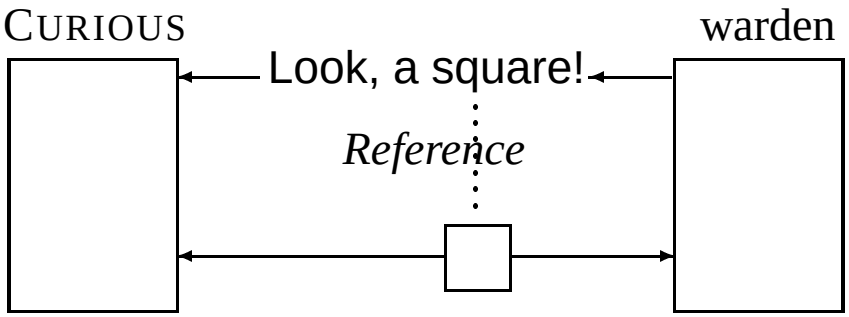


## 4.1.3 Expanded structure of CURIOUS

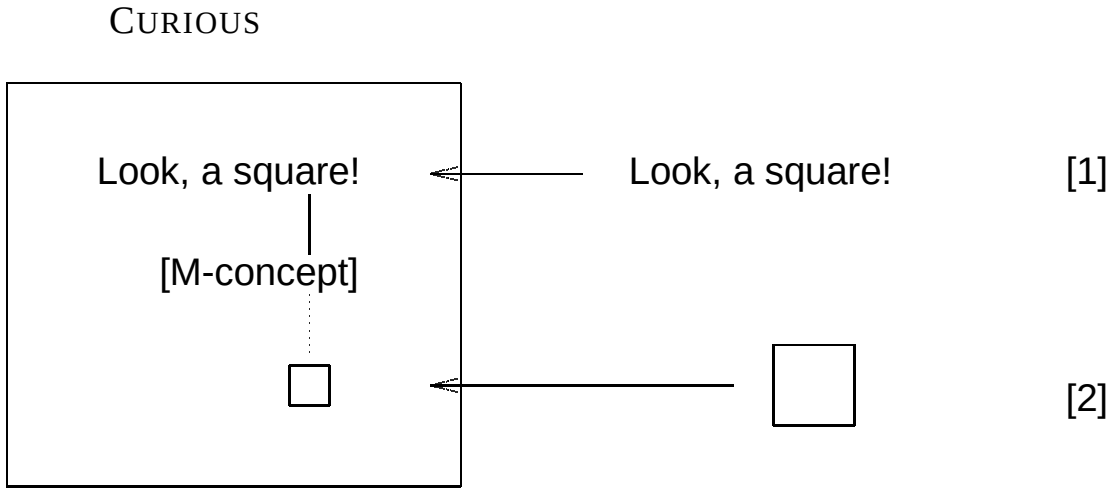


# 4.2 Modeling reference

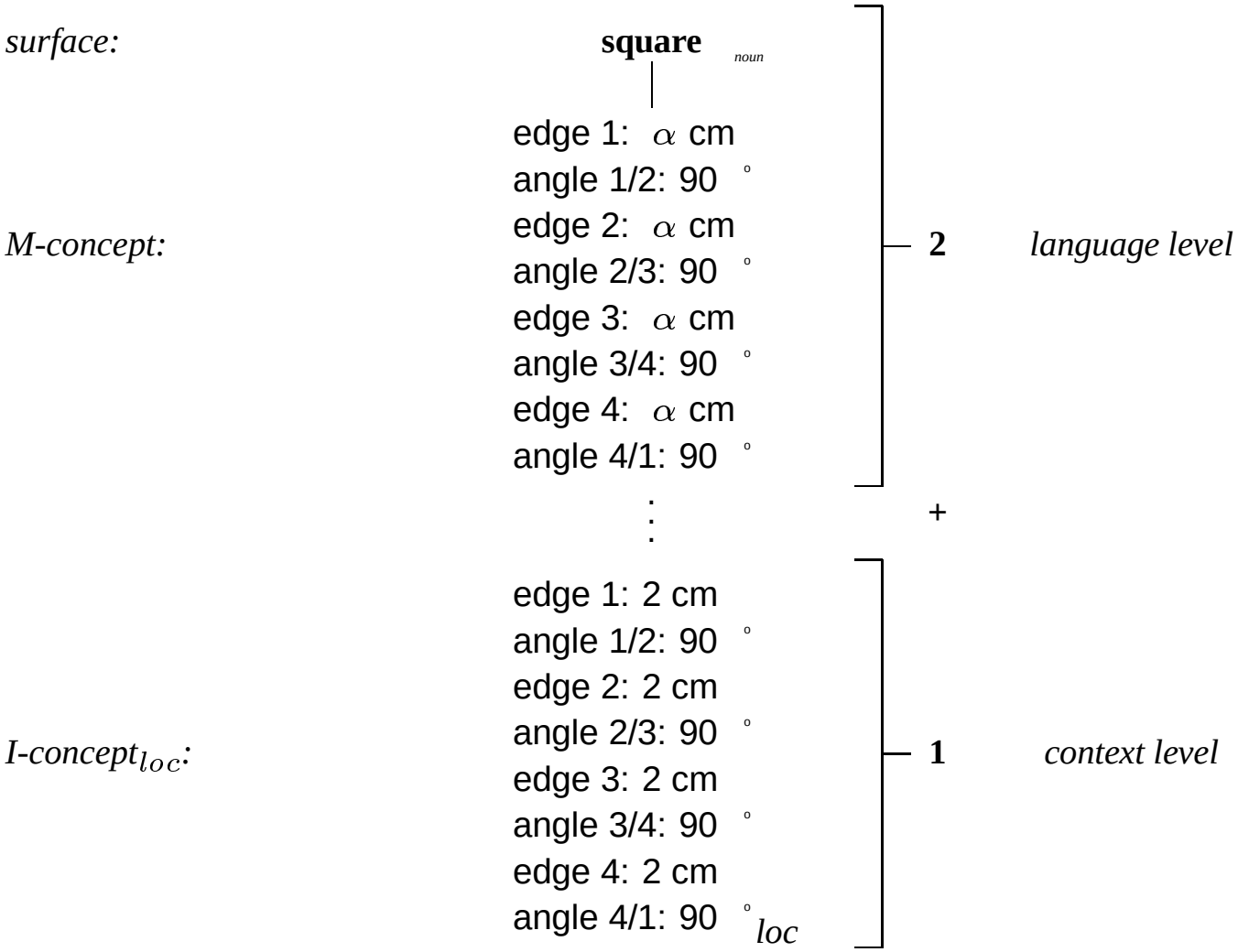
## 4.2.1 An external view of reference




## 4.2.2 Internal and external aspects of reference



## 4.2.3 Cognitive 2+1 level analysis of reference



# 4.3 Using literal meaning

## 4.3.1 Immediate and mediated reference

- *Immediate reference* is the speaker's or the hearer's reference to objects in the current task environment.
- *Mediated reference* is the speaker's or hearer's reference to objects which are not in the current task environment.

## 4.3.2 Two notions of meaning

- meaning$_1$ = property of signs, also called literal meaning
- meaning$_2$ = property of utterances, also called speaker meaning

## 4.3.3 First principle of pragmatics (PoP-1)

The speaker's utterance meaning$_2$ is the use of the sign's literal meaning$_1$ relative to an internal context.

# 4.4 Frege's principle

## 4.4.1 Frege's principle

The meaning of a complex expression is a function of the meaning of the parts and their mode of composition.

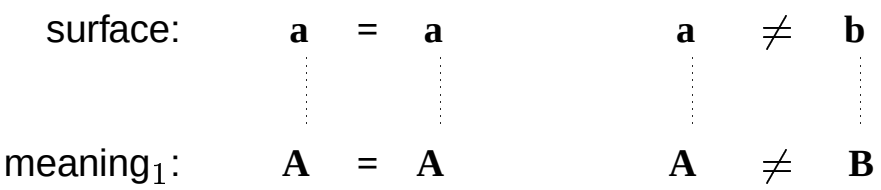
## 4.4.2 Different parts

$a$. The dog bites the *man*
$b$. The dog bites the *bone*

## 4.4.3 Different composition

$a$. The *dog* bites the *man*
$a'$. The *man* bites the *dog*

### 4.4.4 Standard interpretation of Frege's principle



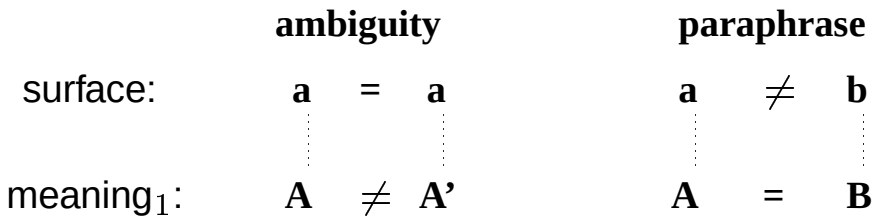

### 4.4.5 Syntactic ambiguity

They don't know how good meat tastes

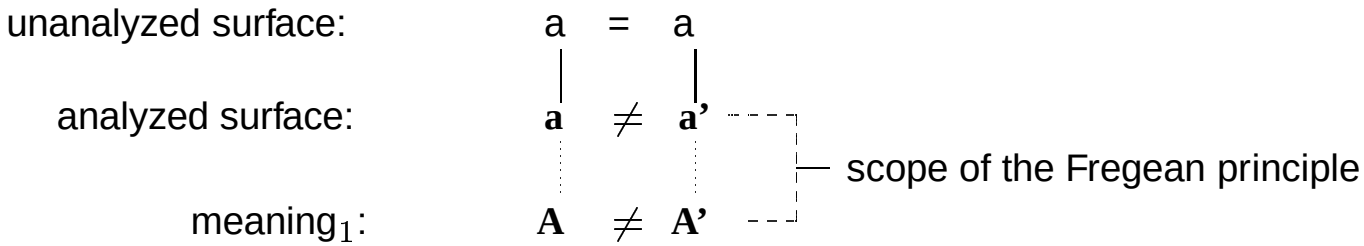
### 4.4.6 Paraphrase

The dog bit the man (active)
The man was bitten by the dog (passive)

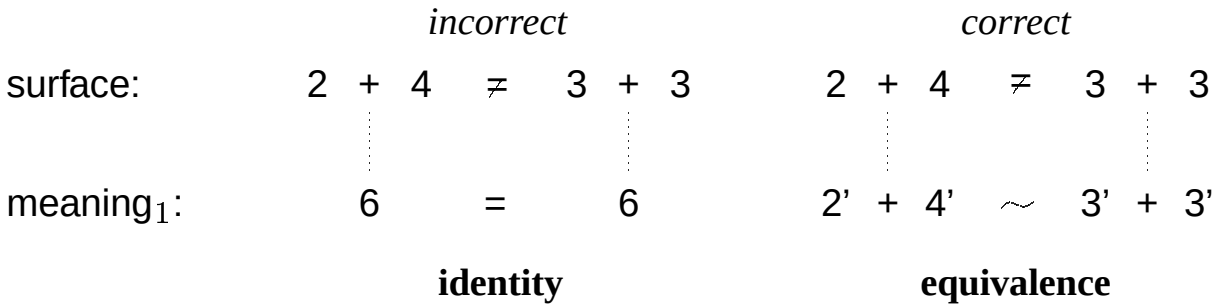
## 4.4.7 Apparent exceptions (incorrect analysis)




## 4.4.8 Syntactic ambiguity (correct analysis)




## 4.4.9 Syntactic paraphrase



# 4.5 Surface compositionality

In its standard interpretation, Frege's principle corresponds to the principle of surface compositionality.

### 4.5.1 Surface compositionality I (SC-I principle)

> An analysis of natural language is surface compositional if it uses only concrete word forms as the building blocks such that all syntactic and semantic properties of complex expression derive systematically from the syntactic category and the meaning$_1$ of their building blocks.

### 4.5.2 Consequences of surface compositionality

- Methodologically:
  Syntactic analyses are *concrete* because no kind of zero surface or underlying form may be used,
- Mathematically:
  Syntactic and semantic analyses may be of *low complexity*
- Functionally:
  The internal matching between meaning$_1$ and context may be extended from single words to the systematic syntactic-semantic *combination* of expressions.

# *Violating surface compositionality:* EXAMPLE I

## 4.5.3 Linguistic generalizations with transformational grammar

Transformations are supposed to be innate, yet have no function in communication.

## 4.5.4 Examples of 'classical' transformations

| DEEP STRUCTURE: | SURFACE STRUCTURE: |
|---|---|
| *Passive*: | |
| Peter closed the door | $\Rightarrow$ The door was closed by Peter |
| *Do-support*: | |
| Peter not open the door | $\Rightarrow$ Peter didn't open the door |
| *Reflexivization* | |
| Peter$_i$ shaves Peter$_i$ | $\Rightarrow$ Peter shaves himself |
| *There-insertion* | |
| A hedgehog is in the garden | $\Rightarrow$ There is a hedgehog in the garden |

*Pronominalization*
Peter$_i$ said that Peter$_i$ was tired $\Rightarrow$ Peter said that he was tired

*Relative clause formation*
Peter [Peter was tired] $\Rightarrow$ Peter, who was tired

*Main clause order in German*
Peter die Tür geschlossen hat $\Rightarrow$ Peter hat die Tür geschlossen

*Object raising*
Peter persuaded Jim [Jim sleeps] $\Rightarrow$ Peter persuaded Jim to sleep

*Subject-raising*
Peter promised Jim [Peter sleeps] $\Rightarrow$ Peter promised Jim to sleep

### 4.5.5 Transformations and the standard interpretation of Frege's Principle

For a while, transformational grammar assumed the equivalence of

*active* Everyone in this room speaks at least two languages
*passive* At least two languages are spoken by everyone in this room

### 4.5.6 Transformations and Darwin's law: Form follows function

The structure of, e.g., a duck foot is innate. Good science should explain its form in terms of its function. The same holds for innate cognitive structures, e.g., the language ability.

### 4.5.7 Cognitive variant of Occam's razor

Entities or components of grammar should not be postulated as innate if they have no clearly defined function within natural communication.

### 4.5.8 Applications of the cognitive razor

The cognitive razor applies to transformational grammar as well as all later variants of nativism including LFG, GPSG, and HPSG. Like transformational grammar, their linguistic generalizations are nonfunctional with respect to communication and inherently in violation of surface compositionality.

# *Violating surface compositionality:* EXAMPLE II

## 4.5.9 Definition of meaning by Grice

Definiendum: U meant something by uttering x.
Definiens: For some audience A, U intends his utterance of x to produce in A some effect (response) E, by means of A's recognition of the intention.

## 4.5.10 Explaining the evolution of language

Grice defines sentence meaning as an utterance type
and utterance meaning as a token of this utterance type.

Problem:
How can a type evolve if it is already presupposed by the first utterance meaning?

## 4.5.11 Conflicting uses of convention

Literal use: Conveying intentions by obeying conventions
Metaphoric use: Conveying intentions by violating conventions

### 4.5.12 Elementary notions suitable for computational implemenation?

Recognition of an intention, producing some effect, intending for some audience...

### 4.5.13 Successful man-machine communication

L = a natural language, SH = a human speaker-hearer of L, CA = a cognitive agent.

- *Successful natural language interpretation*
  CA communicates successfully in the hearer mode, if CA understands the L-utterance in the way intended by SH. In technical terms this means that CA correctly recreates the speaker meaning of the L-utterance in its database. The developers of CA can verify the procedure because (i) they themselves can understand the utterance in L and (ii) they can view the interpretation directly in the database of CA.
- *Successful natural language production*
  CA communicates successfully in the speaker mode, if CA formulates its intentions in L in a way that SH can understand. This requires technically that CA maps a certain structure in its database into an L-utterance which SH can correctly reconstruct. The developers of CA can verify the procedure because (i) they have direct access to the database structure to be communicated and (ii) they themselves can understand the utterance in L.

# 5. Using language signs on suitable contexts

## 5.1 Bühler's organon model

### 5.1.1 Theory of pragmatics

Analyzes the general principles of purposeful action.
Describes how a cognitive agent can achieve certain goals.

### 5.1.2 Examples of pragmatic problems

- The use of a screw driver to fasten a screw
- The use of one's legs to go from $a$ to $b$
- The scavenging of the refrigerator in the middle of the night to fix a BLT sandwich and satisfy one's hunger
- The request that someone fix and serve the sandwich

### 5.1.3 Nonlinguistic and linguistic pragmatics

Depending on whether or not the means employed are signs of language we speak of linguistic and nonlinguistic pragmatics.

### 5.1.4 Embedding linguistic in nonlinguistic pragmatics

Just as language recognition and articulation may be analyzed as a phylo- and ontogenetic specialization of contextual (nonverbal) recognition and action (cf. 4.1.3), respectively, linguistic pragmatics may be analyzed as a phylo- and ontogenetic specialization of nonlinguistic pragmatics.

### 5.1.5 Language as an organon

Embedding of linguistic pragmatics into nonlinguistic pragmatics:

PLATO (427(?)–347 BC)

KARL BÜHLER (1879–1963 AD)

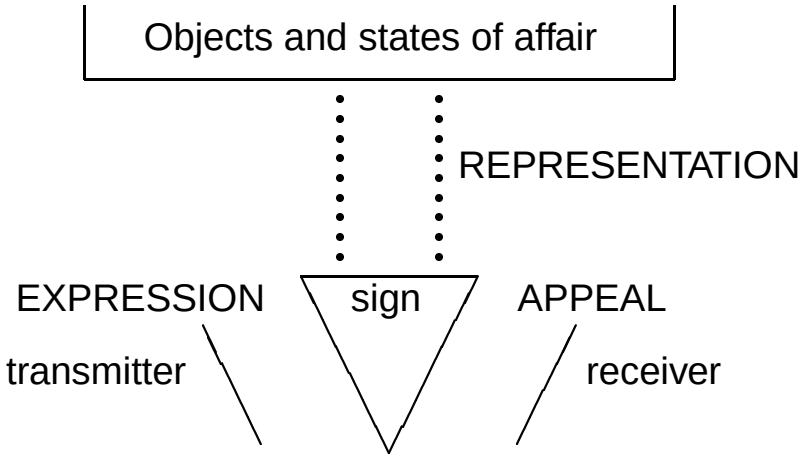
### 5.1.6 The tool character of language

Die Sprache ist dem Werkzeug verwandt; auch sie gehört zu den Geräten des Lebens, ist ein Organon wie das dingliche Gerät, das leibesfremde Zwischending; die Sprache ist wie das Werkzeug ein *geformter Mittler*. Nur sind es nicht die materiellen Dinge, die auf den sprachlichen Mittler reagieren, sondern es sind die lebenden Wesen, mit denen wir verkehren.

[Language is akin to the tool: language belongs to the instruments of life, it is an organon like the material instrument, a body-extraneous hybrid; language is – like the tool – a *purposefully designed mediator*. The only difference is that it is not material things which react to the linguistic mediator, but living beings with whom we communicate.]

K. Bühler 1934, p. XXI

### 5.1.7 Bühler's organon model




Representation refers to the language-based transfer of information. Expression refers to the way the transmitter produces the sign. Appeal refers to the way the sign affects the receiver beyond the bare content of the sign.

### 5.1.8 Shannon & Weaver's information theory 1949

Central notions besides transmitter and receiver are the band width of the channel, the redundancy and relative entropy of the codes, and the noise in the transmission. Its laws hold also in everyday conversation, but background noises, slurring of speech, hardness of hearing, etc., are not components of the natural communication mechanism.

### 5.1.9 Comparing organon model and CURIOUS (4.1.3)

The organon model describes the relation between the ‘transmitter’ and the ‘receiver’ from an external viewpoint and is therefore limited to immediate reference.

The SLIM model of CURIOUS describes the internal structure of the speaker-hearer and can therefore handle mediated reference in addition to immediate reference.

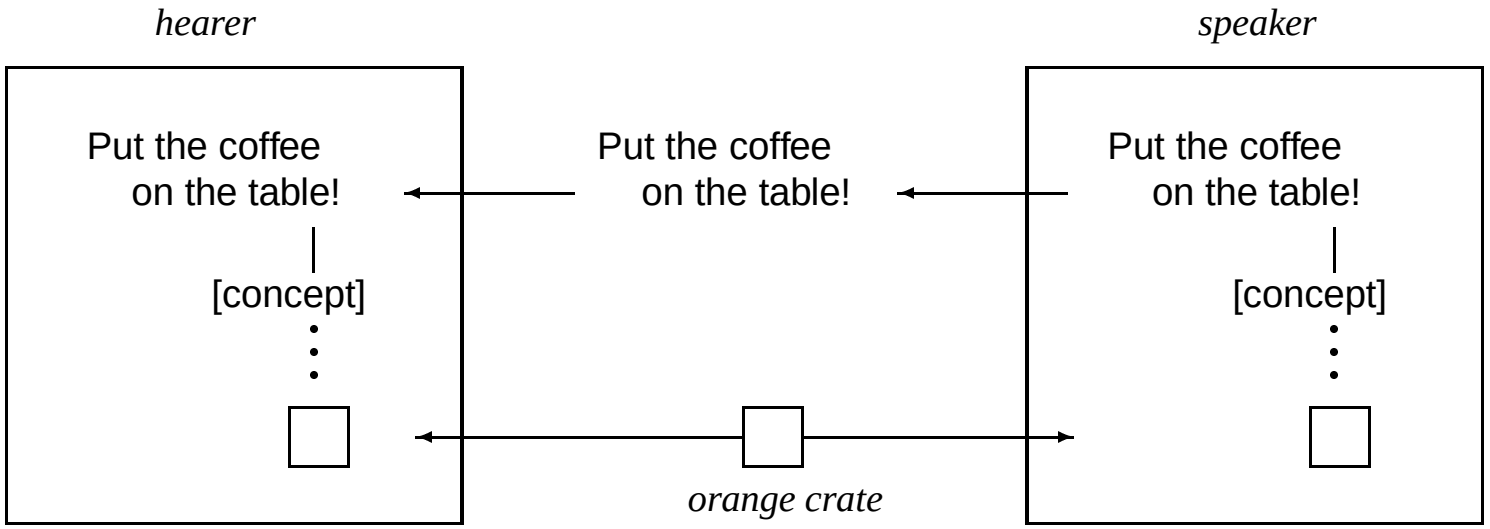
The organon function of ‘expression’ is to be located in component 5+ (language synthesis) of CURIOUS.

The organon function of ‘appeal’ is to be located in component 1+ (language recognition) of CURIOUS.

The organon function of ‘representation’ is performed by CURIOUS in the lexical, syntactic, and semantic components of the language-based database structure 2+ and interpreted in relation to the contextual database structure 2.

# 5.2 Pragmatics of tools and pragmatics of words

## 5.2.1 Nonliteral use of the word table: Principle of best match




## 5.2.2 Central question of linguistic pragmatics

How does the speaker code the selection and delimitation of the used subcontext into the sign and how can these be correctly inferred by the hearer?

# 5.3 Finding the correct subcontext

## 5.3.1 Postcard example

New York, December 1, 1998

Dear Heather,
Your dog is doing fine. The weather is very cold. In the morning he played in the snow. Then he ate a bone. Right now I am sitting in the kitchen. Fido is here, too. The fuzzball hissed at him again. We miss you.

Love,
Spencer

## 5.3.2 Parameters of origin of signs (STAR-point)

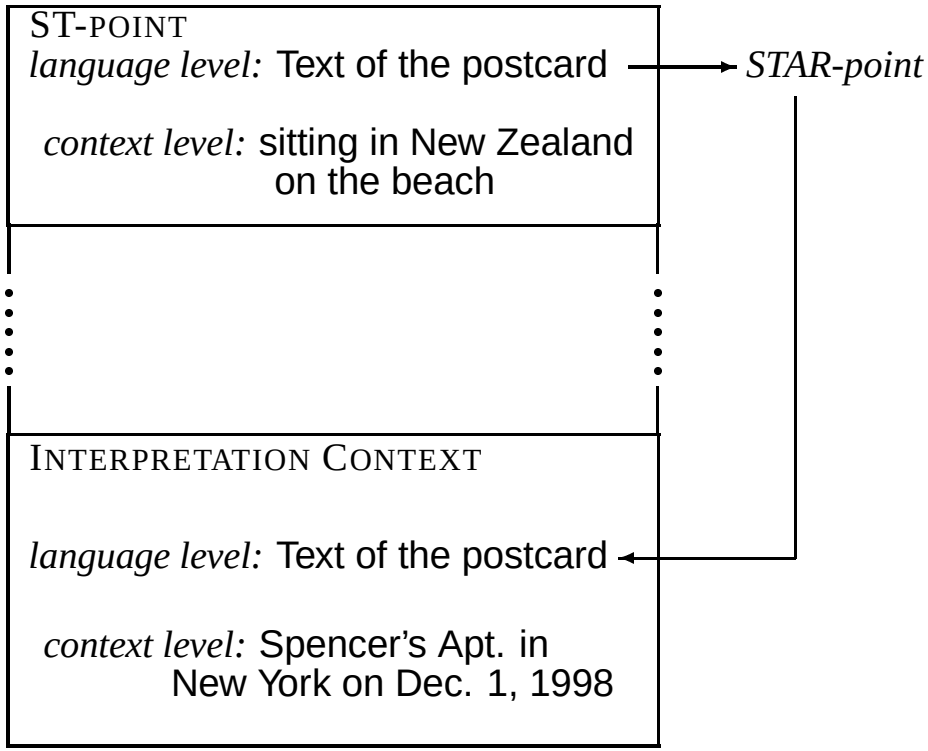
1. S = the **S**patial place of origin
2. T = the **T**emporal moment of origin
3. A = the **A**uthor
4. R = the intended **R**ecipient.

### 5.3.3 Second principle of pragmatics (PoP-2)

The STAR-point of the sign determines its primary positioning in the database by specifying the *entry context* of interpretation.

### 5.3.4 Primary positioning in terms of the STAR-point

Heather's cognitive representation:



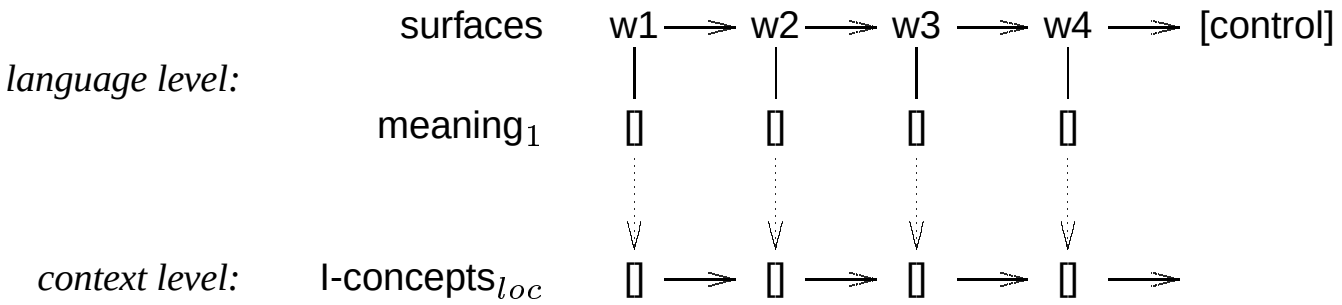
### 5.3.5 Fictitious STAR-point: Beginning of ‘Felix Krull’

Indem ich die Feder ergreife, um in völliger Muße und Zurückgezogenheit – gesund übrigens, wenn auch müde, sehr müde . . .

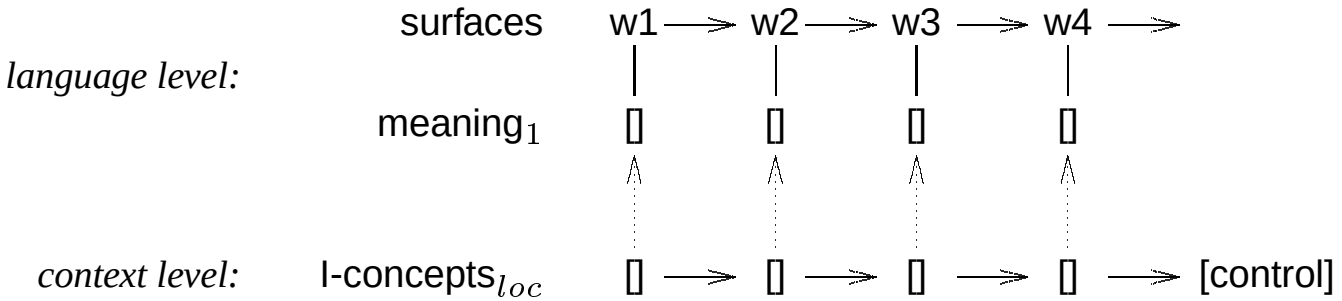
[While I seize the pen in complete leisure and seclusion – healthy, by the way – though tired, very tired . . . ]

# 5.4 Language production and interpretation

## 5.4.1 Schema of language interpretation (analysis)




## 5.4.2 Schema of language production (generation)



### 5.4.3 The time-linear structure of natural language signs

The basic structure of natural language signs is their *time-linear order*. This holds for the sentences in a text, the word forms in a sentence, and the allomorphs in a word form.

*Time-linear* means:

LINEAR LIKE TIME AND IN THE DIRECTION OF TIME.

### 5.4.4 De Saussure's second law: *linear character of signs*

SECOND PRINCIPE; CARACTÈRE LINÉAIRE DU SIGNIFIANT.
Le signifiant, étant de nature auditive, se déroule dans le temps seul et a les caractères qu'il emprunte au temps: a) *représente une étendue*, et b) *cette étendue est mesurable dans une seule dimension*: c'est une ligne.

Ce principe est évident, mais il semble qu'on ait toujours négligé de l'énoncer, sans doute parce qu'on l'a trouvé trop simple; cependent il est fondamental et les conséquences en sont incalculables; son importance est égale à celle de la première loi. Tout le méchanisme de la langue en dépend.

[The designator, being auditory in nature, unfolds solely in time and is characterized by temporal properties: (a) *it occupies an expansion*, and (b) *this expansion is measured in just one dimension*: it is a line.

This principle is obvious, but it seems that stating it explicitly has always been neglected, doubtlessly because it is considered too elementary. It is, however, a fundamental principle and its consequences are incalculable. Its importance equals that of the first law. All the mechanisms of the language depend on it.]

F. de Saussure 1913/1972, p. 103

### 5.4.5 Third principle of pragmatics (PoP-3)

The matching of word forms with their respective subcontexts is incremental whereby in production the elementary signs follow the time-linear order of the underlying thought path while in interpretation the thought path follows the time-linear order of the incoming elementary signs.

# 5.5 Thought as the motor of spontaneous production

## 5.5.1 The once famous motto of behaviorism

THOUGHT IS NONVERBAL SPEECH

## 5.5.2 The motto of the SLIM theory of language

SPEECH IS VERBALIZED THOUGHT.

Thought is defined as the time-linear navigation of a focus point through the concatenated propositions of the internal database.

## 5.5.3 The role of time-linear order for the semantic interpretation

Original order:
In the morning he played in the snow. Then he ate a bone.

Inverted order (incoherent):
Then he ate a bone. In the morning he played in the snow.

### 5.5.4 Alternative navigation through propositional content ( anti-temporal sequencing)

In the morning Fido ate a bone. Before that he played in the snow.

### 5.5.5 Modification of interpretation by changing sequencing

$a$. 1. In February, I visited the Equator. 2. There it was very hot. 3. In March, I was in Alaska. 4. There it was very cold.

$b$. 3. In March, I was in Alaska. 2. There it was very hot. 1. In February, I visited the Equator. 4. There it was very cold.

### 5.5.6 The time-linearity of speech

> Speech is irreversible. That is its fatality. What has been said cannot be unsaid, except by adding to it: to correct here is, oddly enough, to continue.
>
> R. Barthes, 1986, p. 76

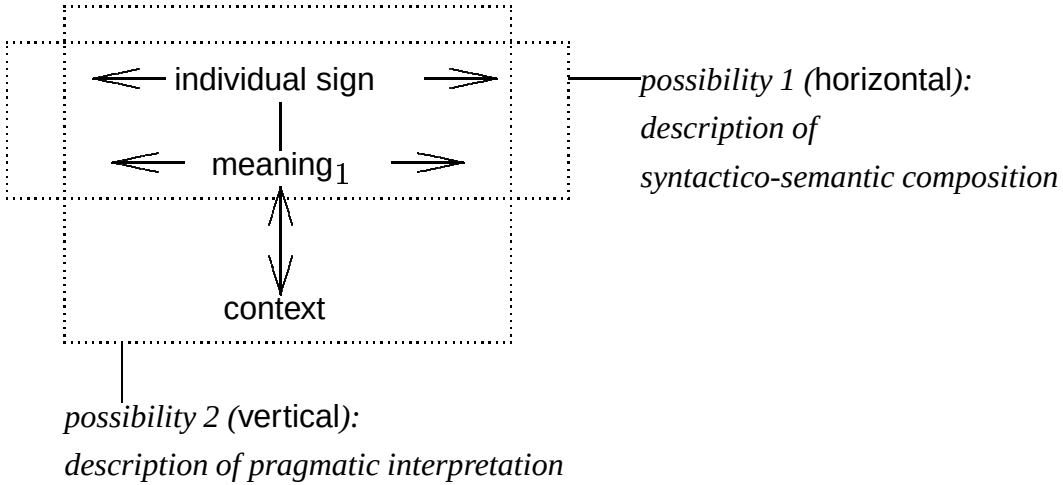
# 6. Structure and functioning of signs

## 6.1 Reference mechanisms of different sign types

### 6.1.1 The mechanism of natural language communication as described so far

1. PoP-1: Language use as an internal matching between meaning$_1$ and a subcontext
2. PoP-2: Entrance subcontext determined by STAR-point of the sign
3. PoP-3: Derivation order is the time-linear sequence of words

### 6.1.2 Alternatives of description



### 6.1.3 Example with minimal syntax

Me up. Weather nice. Go beach. Catch fish. Fish big. Me hungry. Fix breakfast. Eat fish. Feel good.

### 6.1.4 Fourth principle of pragmatics (PoP-4)

The meaning$_1$ of the sign type **symbol** is defined as an M-concept. Symbols refer from their place in a positioned sentence by matching their M-concept with suitable contextual referents (I-concepts$_{loc}$).

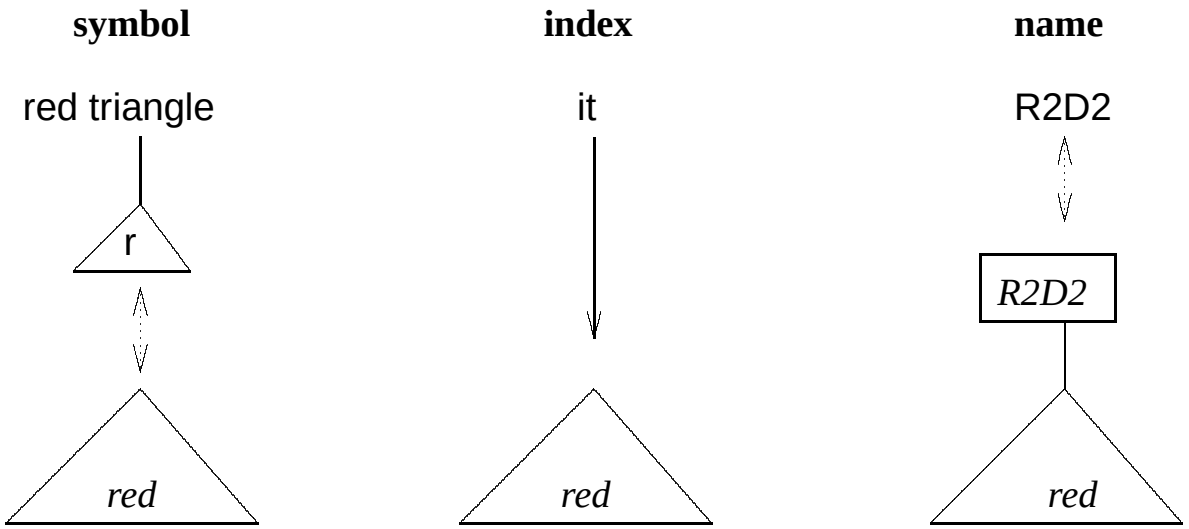
### 6.1.5 Fifth principle of pragmatics (PoP-5)

The meaning$_1$ of the sign type **index** is defined as a pointer. An index refers by pointing from its place in the positioned sentence into appropriate parameter values.

### 6.1.6 Sixth principle of pragmatics (PoP-6)

The reference mechanism of the sign type **name** is based on an act of naming which consists in adding a name-marker to the internal representation of the individual or object in question. A name refers by matching its surface with a corresponding marker.

## 6.1.7 Comparing iconic, indexical, and name-based reference



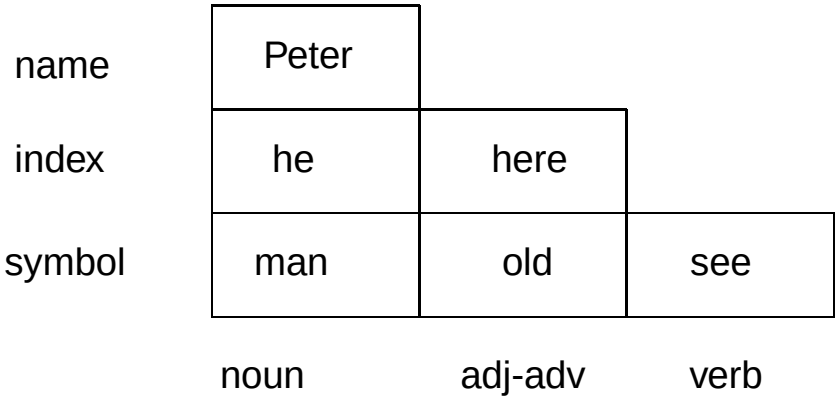

## 6.1.8 Reference in nonverbal and preverbal communication

1. nonverbal iconic reference consists in spontaneously imitating the referent by means of gestures or sounds,
2. nonverbal indexical reference consists in pointing at the referent, and
3. nonverbal name-based reference consists in pointing at the referent while simultaneously pronouncing a name.

### 6.1.9 Seventh principle of pragmatics (PoP-7)

The sign type *symbol* occurs as noun, verb, and adjective-adverbial. The sign type *index* occurs as noun and adjective-adverbial. The sign type *name* occurs only as noun.

### 6.1.10 Relation between sign types and parts of speech



# 6.2 Internal structure of symbols and indices

## 6.2.1 Internal components of the sign types symbol and index

- The *surface* is constrained by the laws of acoustic articulation (in the original medium of spoken language).
- The *category* reflects the combinatorial properties of the part of speech and the inflectional class to which the sign belongs.
- The *meaning*$_1$ reflects the conceptual structures of the internal context and/or contains characteristic pointers to certain contextual parameters.
- The *glue* connecting surface, category, and meaning$_1$ consists in conventions which must be learned by each member of the language community.

### 6.2.2 *De Saussure*'s first law

> PREMIER PRINCIPE; L'ARBITRAIRE DU SIGNE.
> Le lien unissant signifiant au signifiè est arbitraire, ou encore, puisque nous entendons par signe le total résultant de l'association d'un signifiant à un signifié, nous pouvons dire plus simplement: *le signe linguistique est arbitraire*.
> [THE FIRST LAW: ARBITRARINESS OF SIGNS
> The link connecting the designator and the designated is arbitrary; and since we are treating a sign as the combination which results from connecting the designator with the designated, we can express this more simply as: *the linguistic sign is arbitrary.*]
>
> F. de Saussure 1913/1972, p. 100

### 6.2.3 Possible functions of the sign type symbol

1. *Initial* reference to objects which have not been mentioned so far.
2. *Repeating* reference to referents which have already been introduced linguistically.
3. *Metaphorical* reference to partially compatible referents, both in initial and repeating reference.

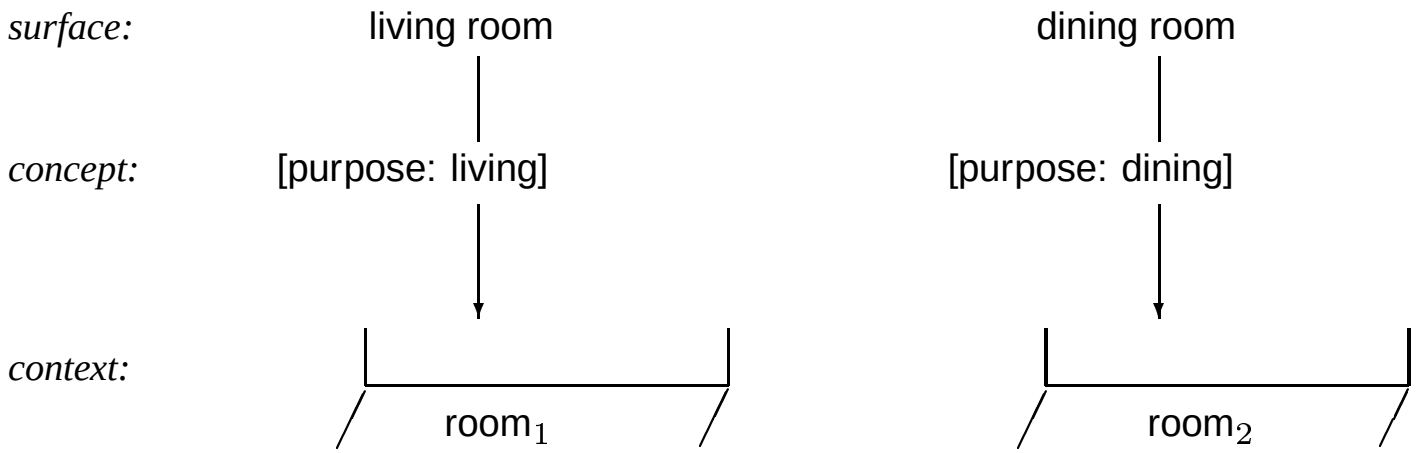
### 6.2.4 Structural basis of symbolic reference

- the minimal meaning$_1$ structure of symbols (M-concepts), and
- the limited selection of compatible referents available in the subcontext.

### 6.2.5 Integrating additional symbolic content to ensure reference

the table, the garden table, the green garden table, the small green garden table, the round small green garden table, the round small green garden table near the pool, etc.

### 6.2.6 Characterizing objects symbolically



# 6.3 Indices for repeating reference

### 6.3.1 Pure indices

Index words which contain no additional grammatical or symbolic meaning components, e.g. here, now, and you.

### 6.3.2 Nonpure indices

Index words which incorporate symbolic-grammatical distinctions, e.g. between singular (I) and plural (we), between nominative (I, we) and oblique (me, us) case, etc.

### 6.3.3 Pointing area of third person pronouns

It is outside of the STAR-point parameters and comprises all objects and persons that have been activated so far and are neither the speaker nor the hearer.

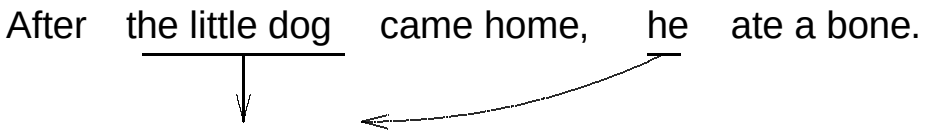
### 6.3.4 Repeating reference

This special task reference arises in longer sentences and texts when a certain referent has already been introduced and needs to be referred to again.

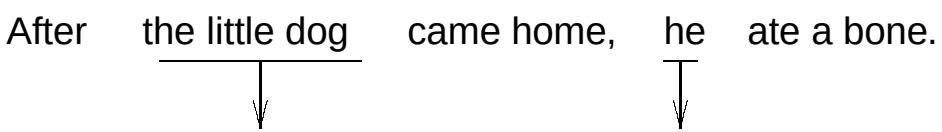
### 6.3.5 Using third person pronouns for repeating reference

Because of their grammatical differentiation and the general nature of their pointing area, third person pronouns are ideally suited for a brief, precise, and versatile handling of repeating reference .

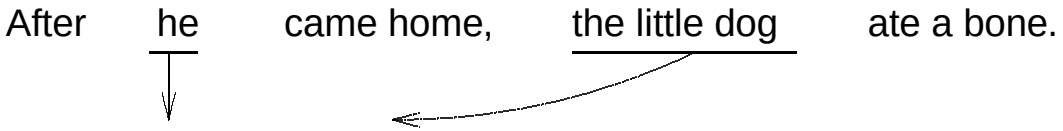
### 6.3.6 Indexically repeating reference



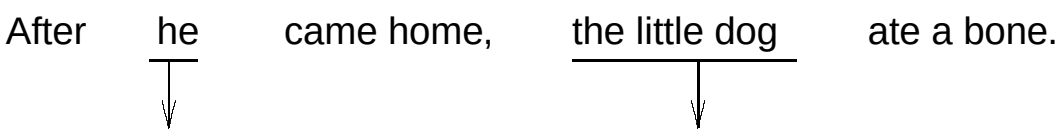

### 6.3.7 Indexical reference without coreference



## 6.3.8 Symbolically repeating reference I




## 6.3.9 Symbolic reference without coreference



### 6.3.10 Sentence structure blocking repeating reference

Anaphorical positioning:

After Peter$_i$ came home he$_i$ took a bath.
Peter$_i$ took a bath after he$_i$ came home.
%! Near Peter$_i$ he$_i$ sees a snake.

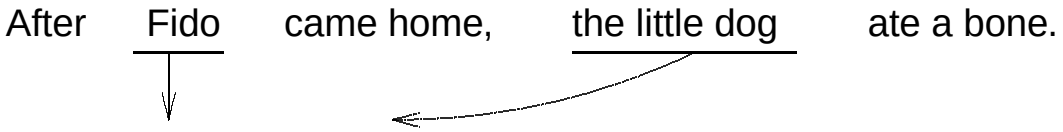
Cataphorical positioning:

After he$_i$ came home Peter$_i$ took a bath.
%! He$_i$ took a bath after Peter$_i$ came home.
Near him$_i$ Peter$_i$ saw a snake.

### 6.3.11 Cross-sentential coreference

Peter$_k$ wanted to drive into the country. He$_k$ waited for Fido$_i$. When the little dog$_i$ came home he$_k$ was glad.

## 6.3.12 Symbolically repeating reference II




## 6.3.13 Initial reference established by symbol, index, and name

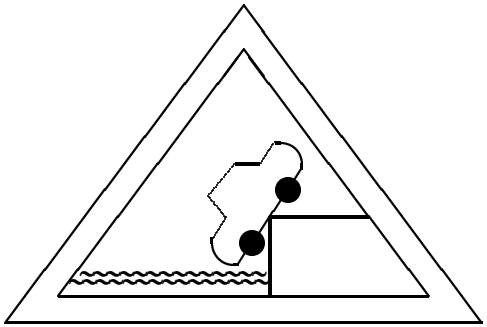
| | | |
|---|---|---|
| | the little dog | |
| After | he | came home, he slept. |
| | Fido | |

## 6.3.14 Repeating reference using symbol, index, and name

| | | |
|---|---|---|
| | the little dog | |
| After he came home, | he | slept. |
| | Fido | |

# 6.4 Exceptional properties of icon and name

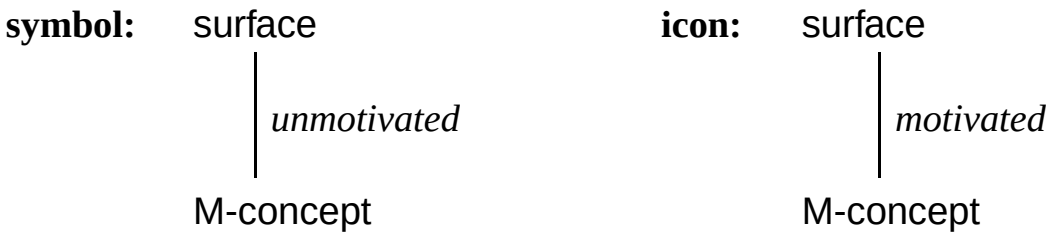
## 6.4.1 Example of an icon (street sign)



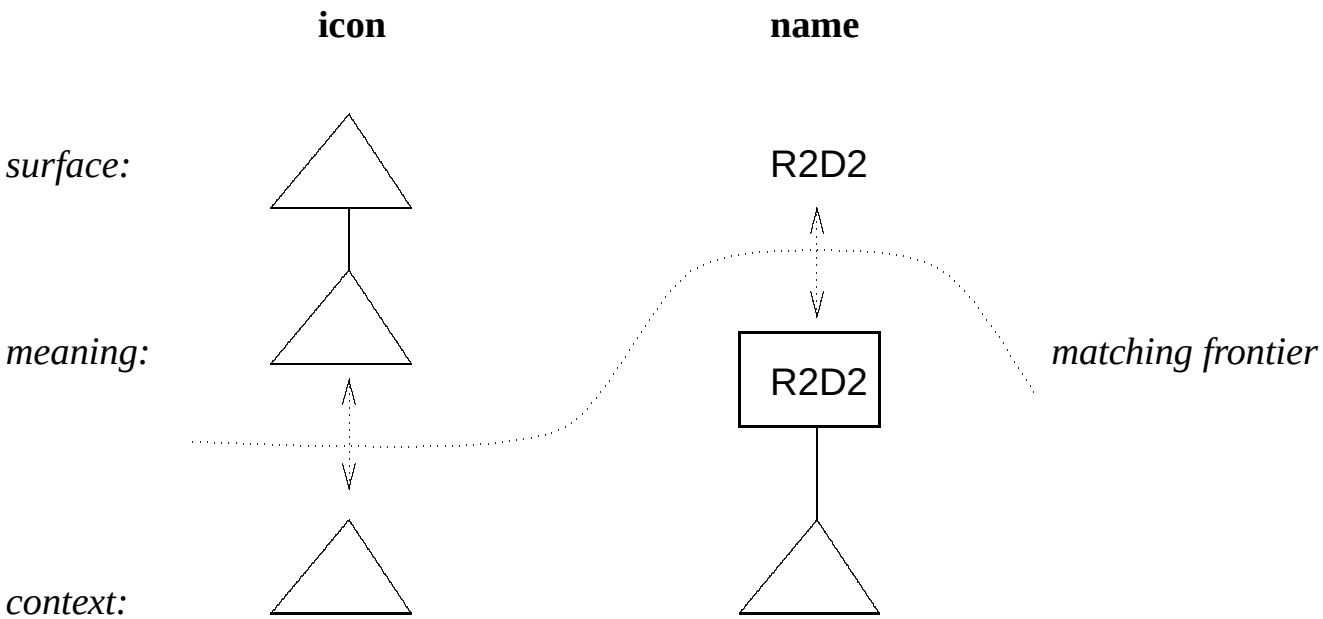
## 6.4.2 The controversy between *naturalists* and *conventionalists*

> [The naturalists] maintained that all words were indeed ‘naturally’ appropriate to the things they signified. Although this might not always be evident to the layman, they would say, it could be demonstrated by the philosopher able to discern the ‘reality’ that lay behind the appearance of things. Thus was born the practice of conscious and deliberate etymology. The term itself (being formed from the Greek stem *etymo-* signifying ‘true’ or ‘real’) betrays its philosophical origin. To lay bare the origin of a word and thereby its ‘true’ meaning was to reveal one of the truths of ‘nature’.
>
> J. Lyons 1968, p. 4 f.

## 6.4.3 Comparing the structure of symbol and icon

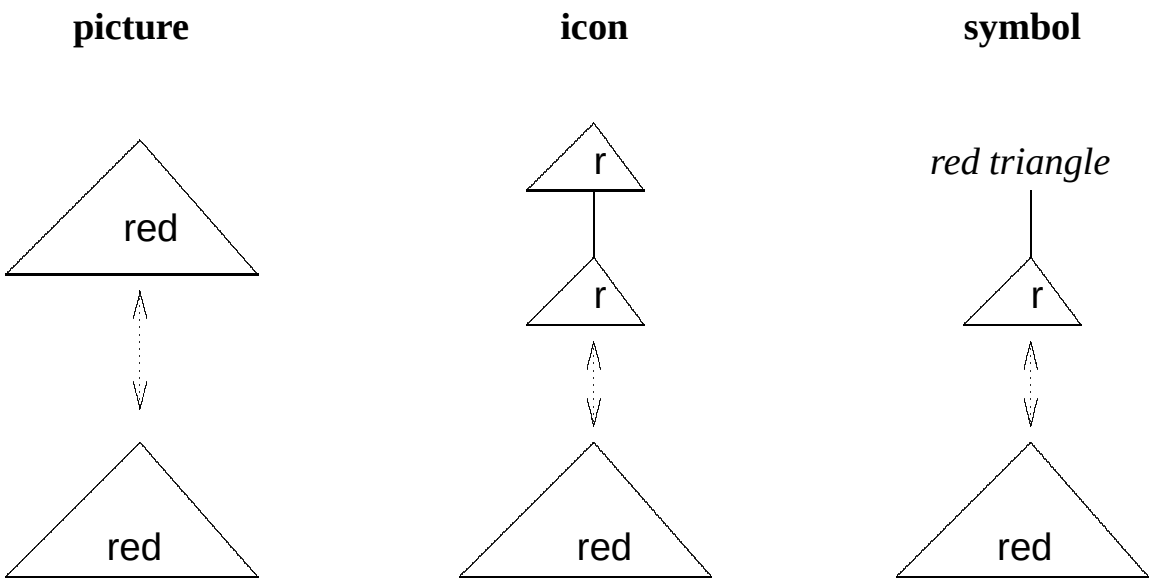



## 6.4.4 Comparing the structure of icon and name

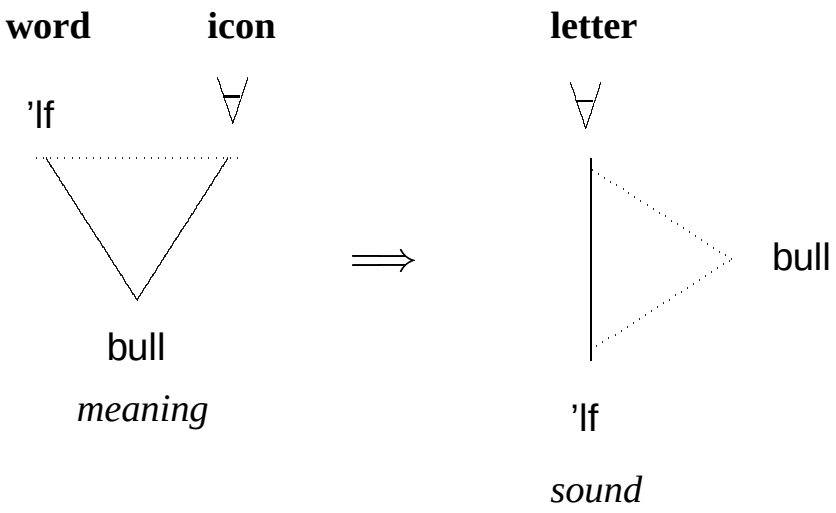


# 6.5 Pictures, pictograms, and letters

## 6.5.1 Transition from picture to icon to symbol



## 6.5.2 Rotation principle underlying transition from icon to letter




## 6.5.3 Aristotle on writing

Spoken words are the signs of mental experience and written words are the signs of spoken words.

DE INTERPRETATIONE, 1