Erasmus Lectures: Introduction to Computational Linguistics

ROLAND HAUSSER
Computational Linguistics
Universität Erlangen Nürnberg
Germany

University of Jyväskylä, Finland
March 30–April 3, 2009, 12:15–13:45

CLUE

Table of Contents

1. Day One: Words and Word Forms 3
1.1 Nature of the External Language Signs 3
1.2 Modality Transfer in the Speaker and the Hearer Mode 10
1.3 The Context Component 12

2. Day Two: Reference 21
2.1 Four Different Approaches to Reference 21
2.2 Problems with Approaches to Reference which are not [+sense] 25
2.3 Frege's [+sense] approach to Reference 29
2.4 Problem of [–constructive] approaches in general 30
2.5 Metalanguage-based versus procedural semantics 33
2.6 Basic structure of semantic interpretation 36
2.7 Logical, programming, and natural languages 38

3. Day Three: Semantic Relations 40
3.1 Traditional Notions of Grammar 40
3.2 Compositionality 48
3.3 Parts of Speech at the Elementary, Phrasal, and Clausal level 51
3.4 Semantic Relations 55
3.5 Three Sign-oriented Grammar Formalisms 56

3.6 From a Sign-Oriented to an Agent-Oriented Approach 58
3.7 Verification 61

4. Day Four: The Cycle of Natural Language Communication 65
4.1 The Data Structure of Database Semantics 65
4.2 The Cycle of NL Communication 67
4.3 Coding semantic relations 72
4.4 Matching proplets 75
4.5 Storage in a Database 77
4.6 Database Semantics: Think Mode 78
4.7 Database Semantics: Speaker Mode 82

5. Day Five: Treating the Phrasal and Clausal Levels 85
5.1 Establishing Semantic Relations at the Phrasal Level 85
5.2 Phrasal Level Production in the Speaker Mode 88
5.3 Interpretation at the Clausal Level 94
5.4 Production at the Clausal Level 96
5.5 Conclusion and References 97

1. Day One: Words and Word Forms

1.1 Nature of the External Language Signs

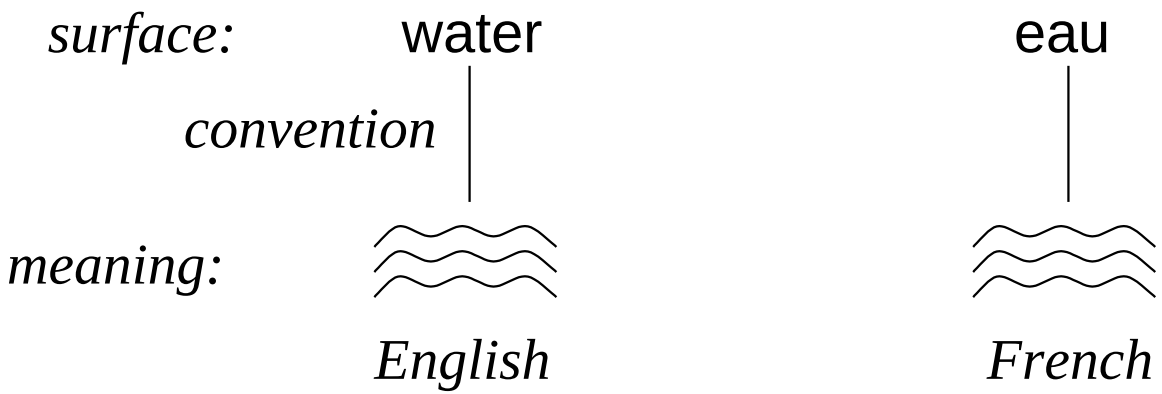
1.1.1 INTRODUCTION

The analysis of natural languages and of natural language communication has long been an interdisciplinary enterprise involving

linguistics,
philosophy,
psychology,
physiology,
neurology,
sociology,
mathematics,
and computer science.

Let’s start with something really basic, namely the structure of words.

1.1.2 BASIC STRUCTURE OF A WORD

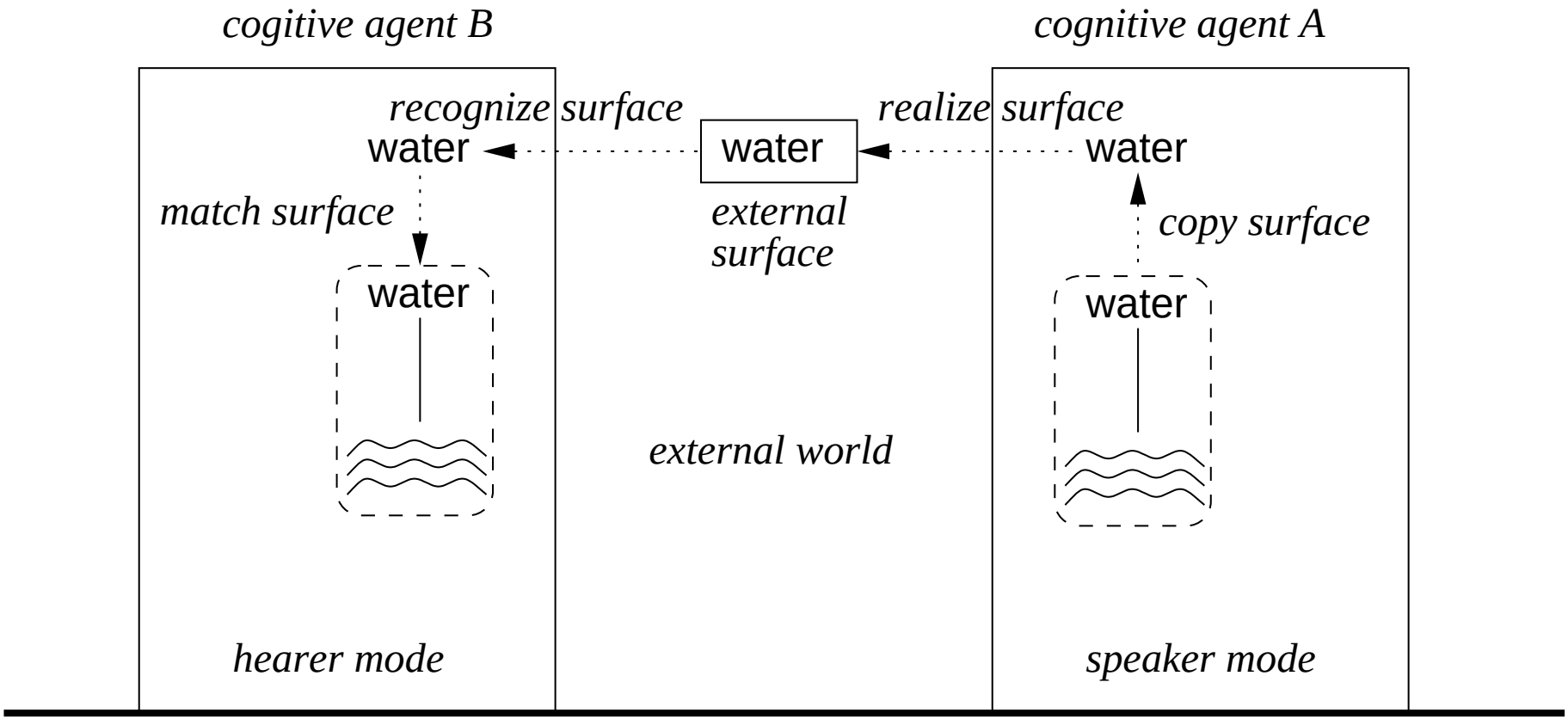


1.1.3 TASKS OF LEARNING THE WORDS OF A FOREIGN LANGUAGE

- learning to recognize and produce the foreign surfaces in the modalities of spoken and written language, and

- learning the conventional connections between these foreign surfaces and familiar meanings.

1.1.4 PRODUCTION AND RECOGNITION OF A WORD



1.1.5 THE FIRST MECHANISM OF COMMUNICATION (MOC-1) (Hausser 2009b)

Natural language communication relies on modality-dependent external surfaces which have neither meaning nor any grammatical property.

1.1.6 DIFFERENT MODALITIES

- vision (handwritten or printed signs)
- audition (spoken signs)
- signing (signs gestured by the hearing impaired)
- tactile sense (Braille for the blind)

in addition there are the modalities of taste, smell, temperature, etc., which are not used for language communication

1.1.7 EXAMPLES OF MODALITY-FREE CODING

- Neurological coding in natural agents
- Computational coding like 7 bit ASCII in artificial agents

1.1.8 FUNCTIONAL MODEL OF NATURAL LANGUAGE COMMUNICATION

A functional model of natural language communication requires

1. a set of cognitive agents each with (i) a body, (ii) external interfaces for recognition and action, and (iii) a memory for the storage and processing of content,

2. a set of external language surfaces which can be recognized and produced by these agents by means of their external interfaces using pattern matching,

3. a set of agent-internal (cognitive) surface-meaning pairs stored in memory, whereby the internal surfaces correspond to the external ones, and

4. an agent-internal algorithm which constructs complex meanings from elementary ones by establishing semantic relations between them.

1.1.9 COMMUNICATION WITHOUT A NATURAL LANGUAGE

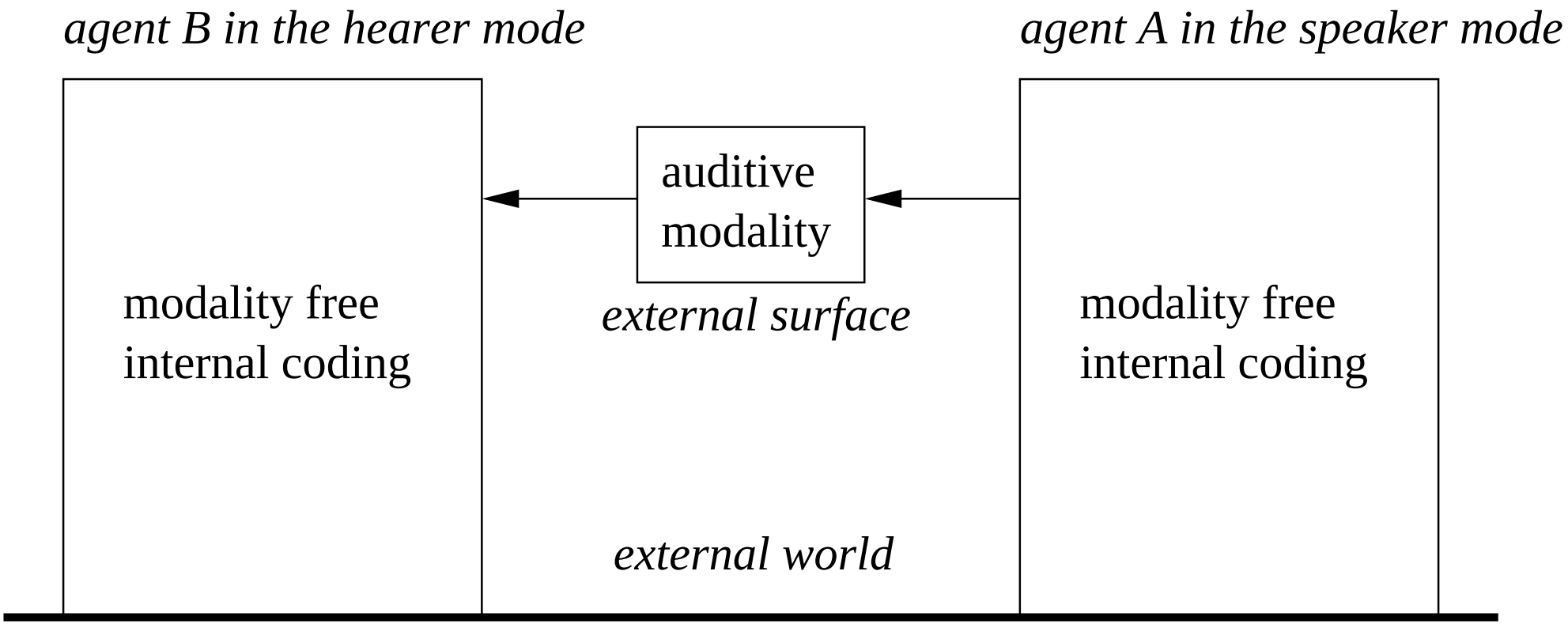
- endocrinic messaging by means of hormones,
- exocrinic messaging by means of pheromones, for example in ants, and
- the use of samples, for example in bees communicating a source of pollen.

1.1.10 ADVANTAGES FOLLOWING FROM MOC-1

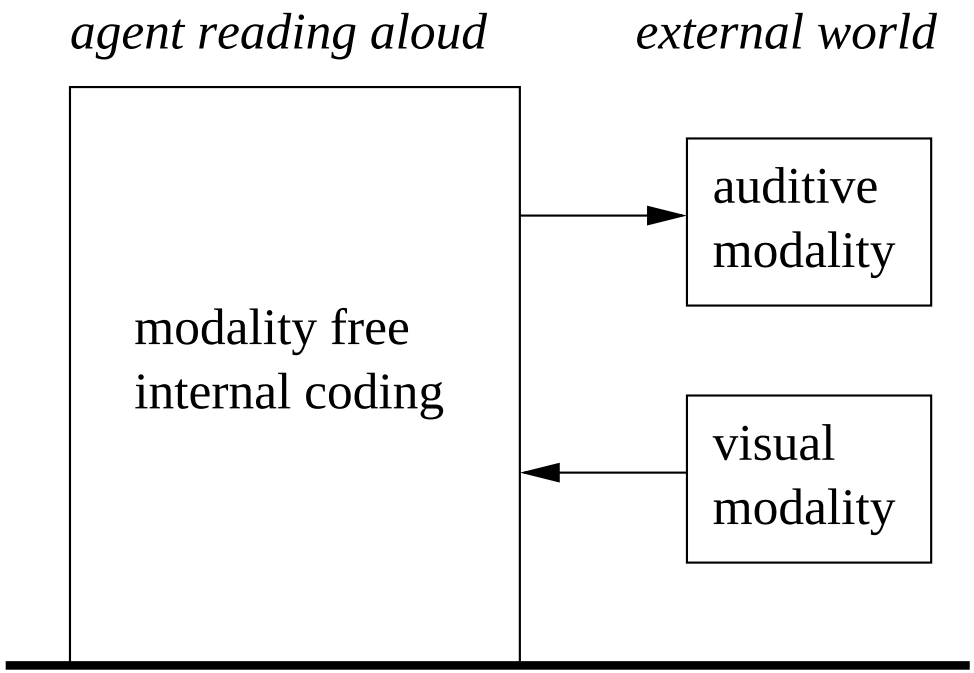
1. The modality of an external surface imposes no restriction on the kind of meaning which may be attached to the internal counterpart of this surface.
2. The external surfaces are much more suitable for (i) transfer and (ii) long-term storage than the associated meanings.

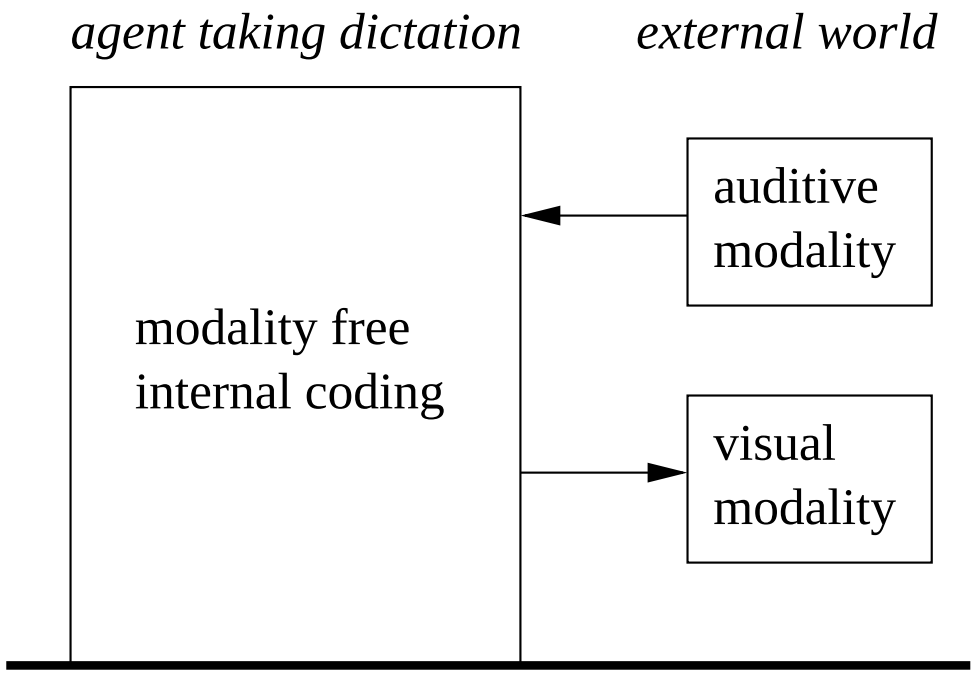
1.2 Modality Transfer in the Speaker and the Hearer Mode

1.2.1 INTERAGENT COMMUNICATION USING SPEECH



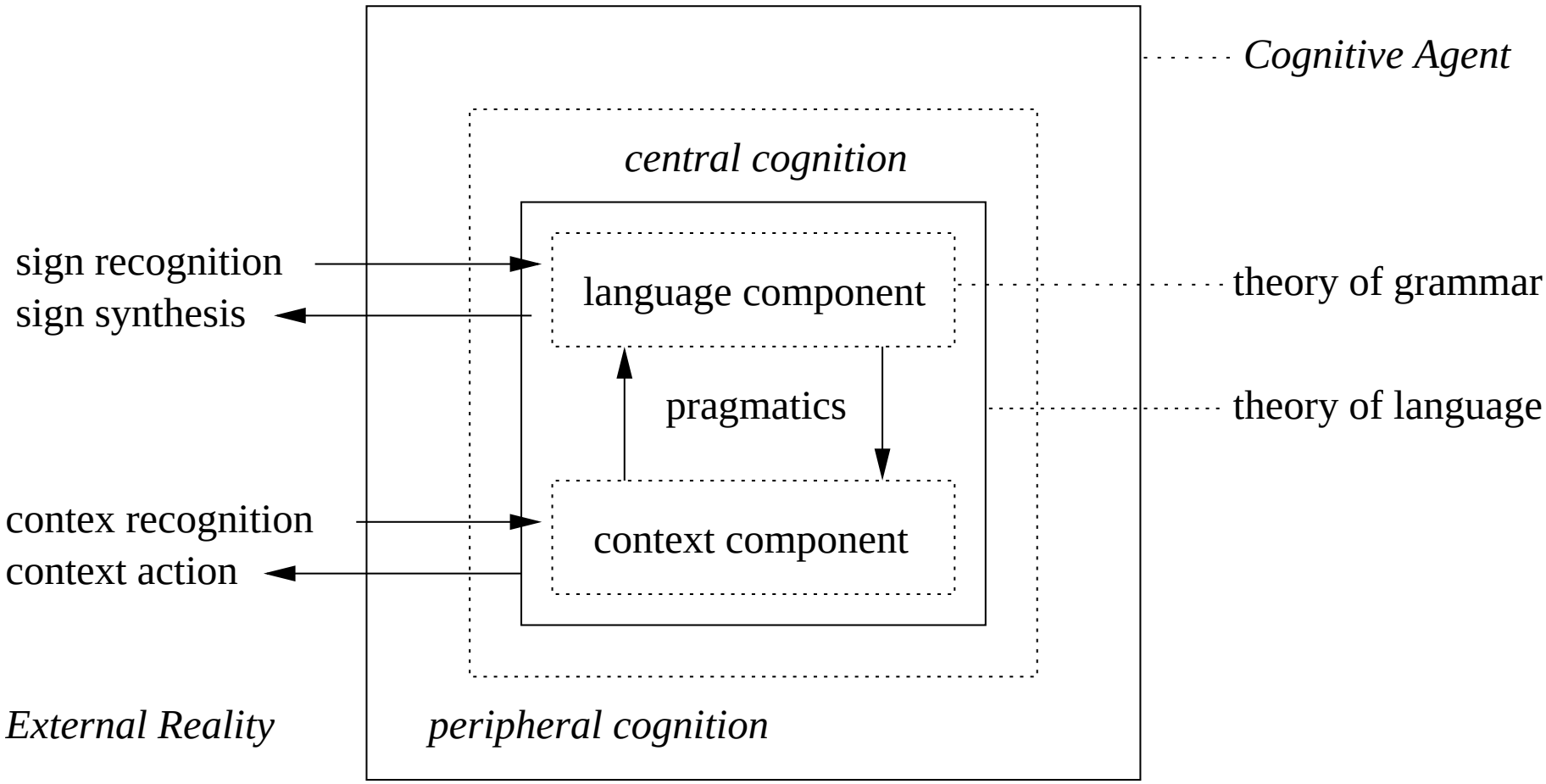
1.2.2 TWO KINDS OF MODALITY CONVERSION





1.3 The Context Component

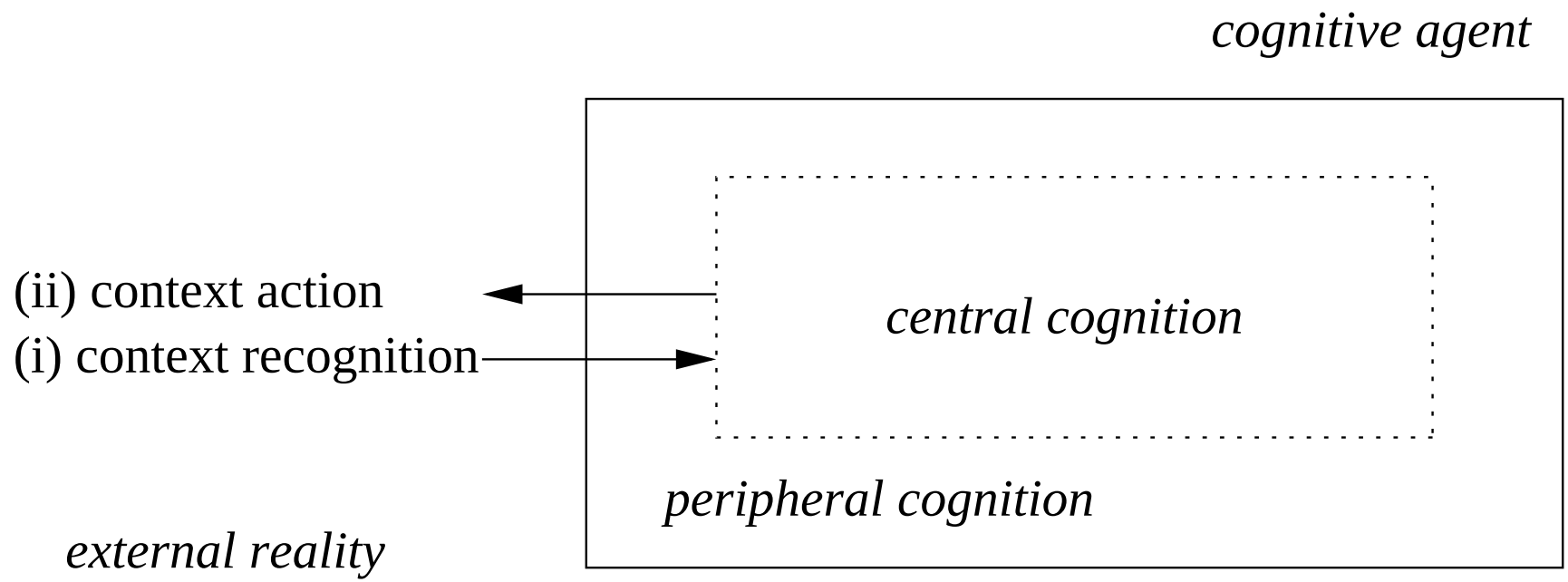
1.3.1 CONTEXT AS PART OF A COGNITIVE AGENT WITH LANGUAGE



In phylogeny and ontogeny, the evolution of the context level precedes language acquisition.

phylogeny: development of a species
ontogeny: development of a member of the species

1.3.2 CONTEXT AS A COGNITIVE AGENT WITHOUT LANGUAGE



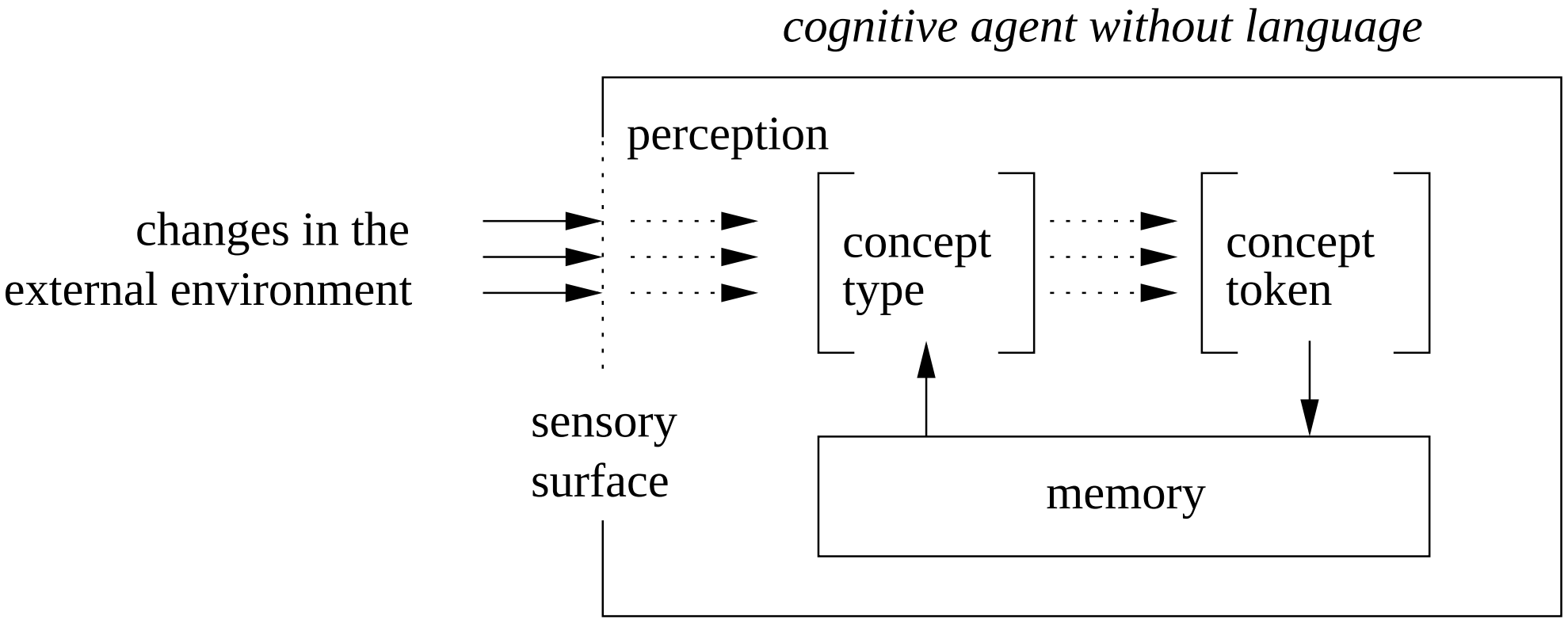
What are cognitive abilities which occur before language and are re-used by language?

1.3.3 TYPE AND TOKEN OF THE CONCEPT *square*

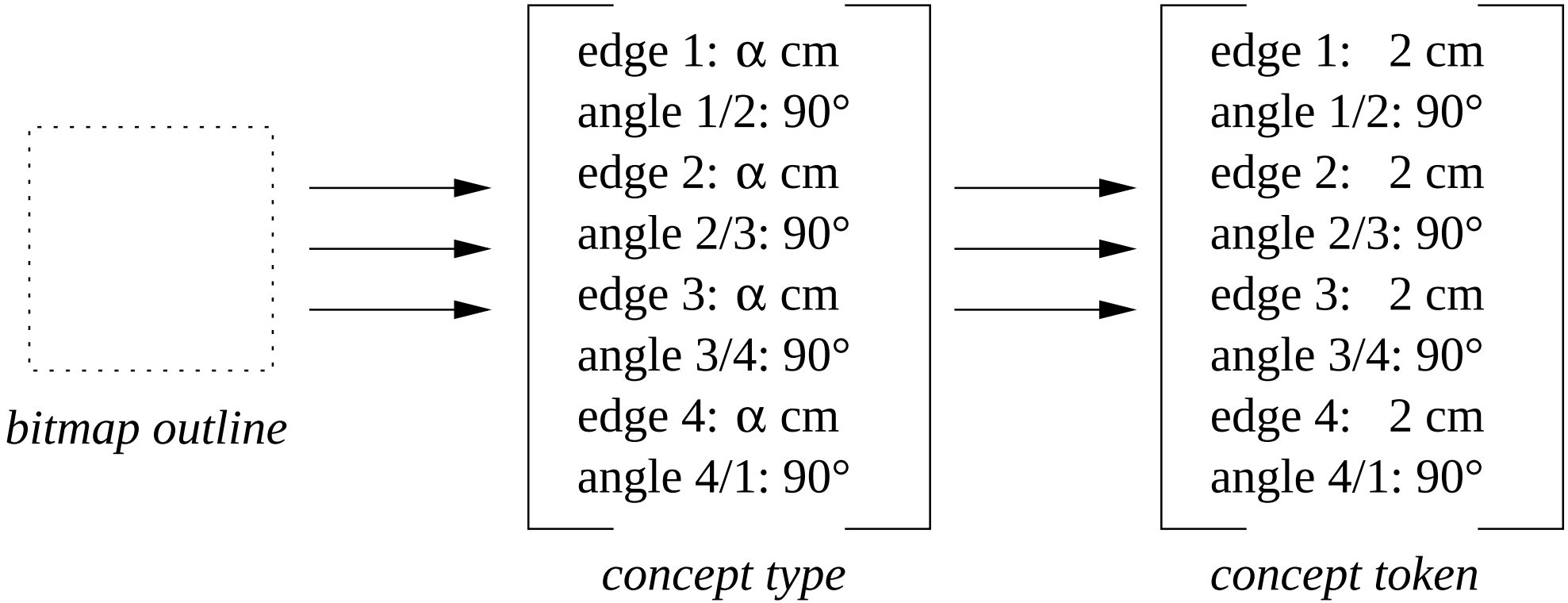
type	*token*
edge 1: α	edge 1: 2 cm
angle 1/2: 90°	angle 1/2: 90°
edge 2: α	edge 2: 2 cm
angle 2/3: 90°	angle 2/3: 90°
edge 3: α	edge 3: 2 cm
angle 3/4: 90°	angle 3/4: 90°
edge 4: α	edge 4: 2 cm
angle 4/1: 90°	angle 4/1: 90°

The type-token distinction was introduced by the American philosopher and logician C. S. PEIRCE (1839–1914).

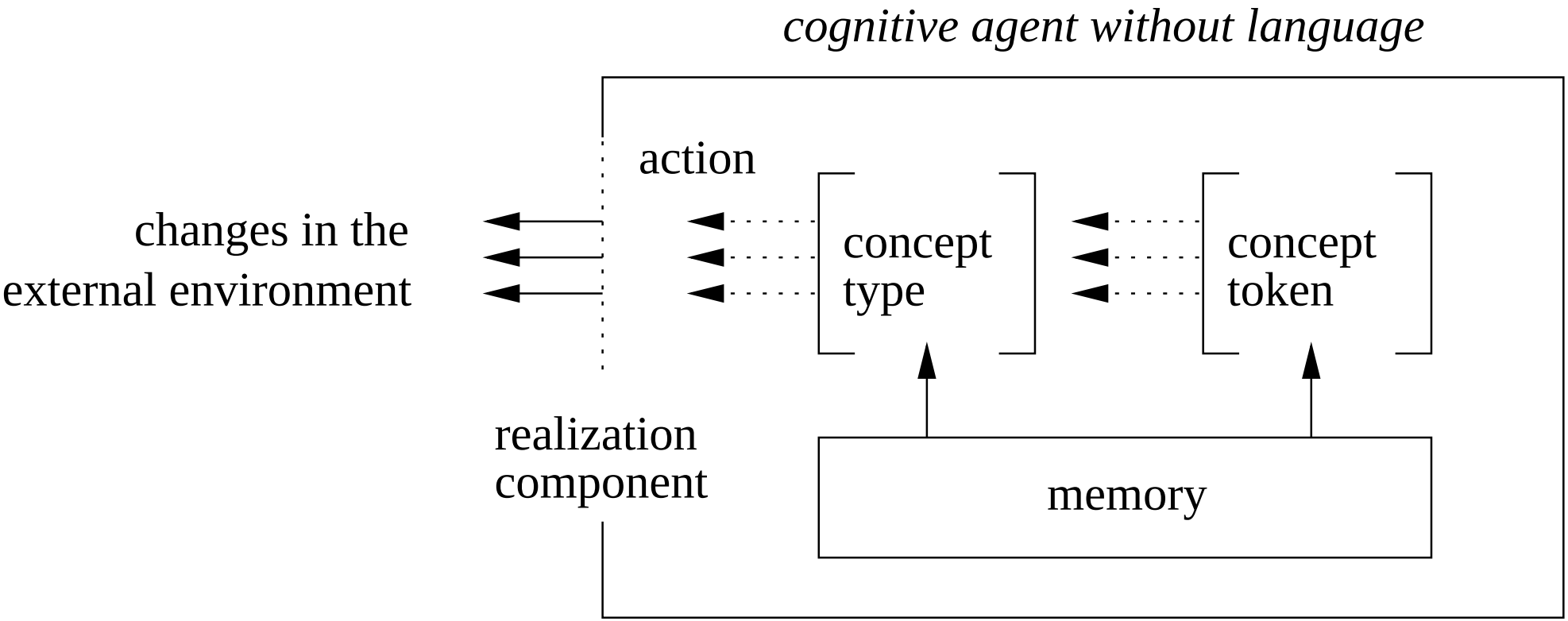
1.3.4 CONCEPT TYPE AND CONCEPT TOKEN IN CONTEXTUAL RECOGNITION



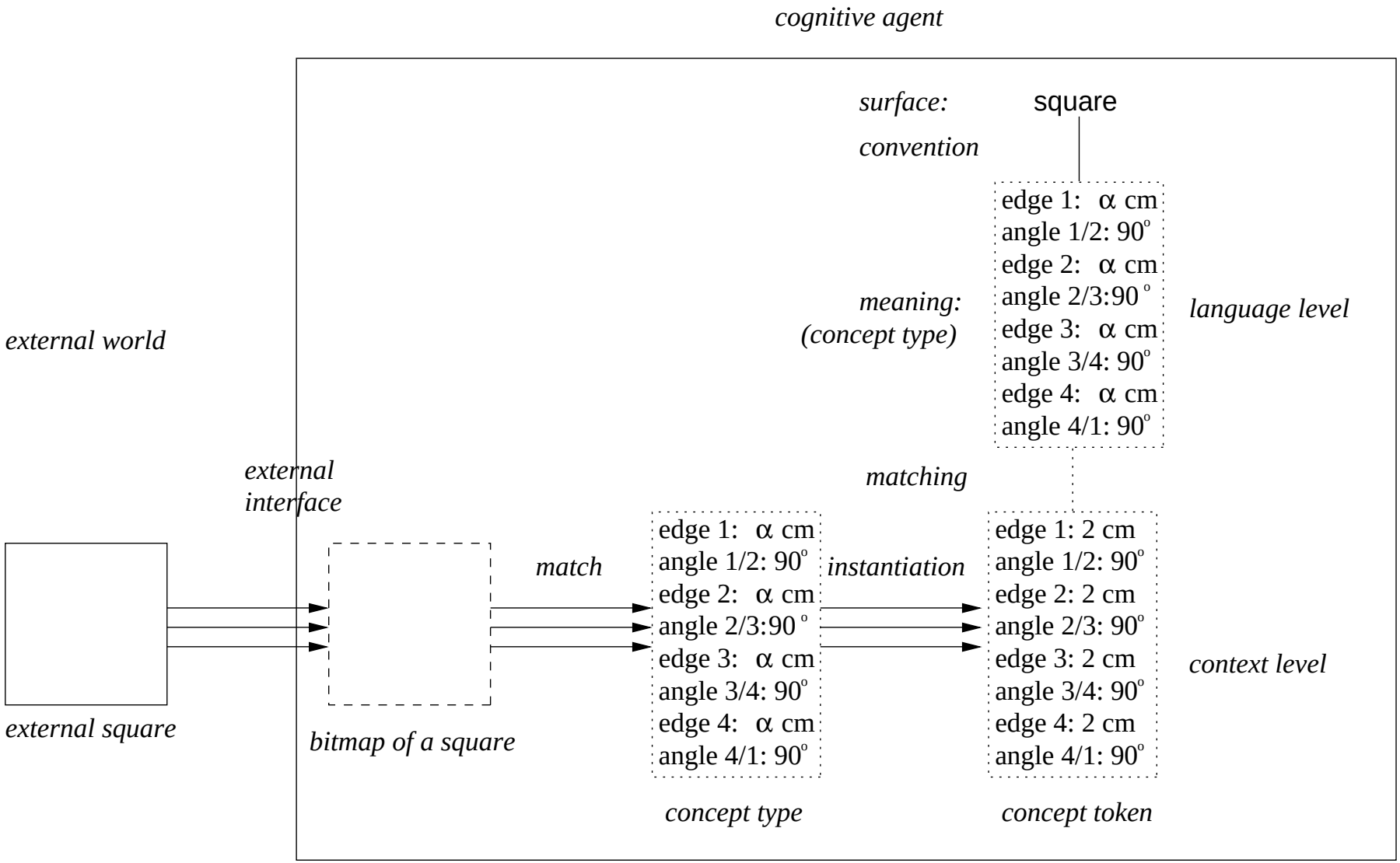
1.3.5 CONCEPT TYPE AND CONCEPT TOKEN IN RECOGNIZING A SQUARE



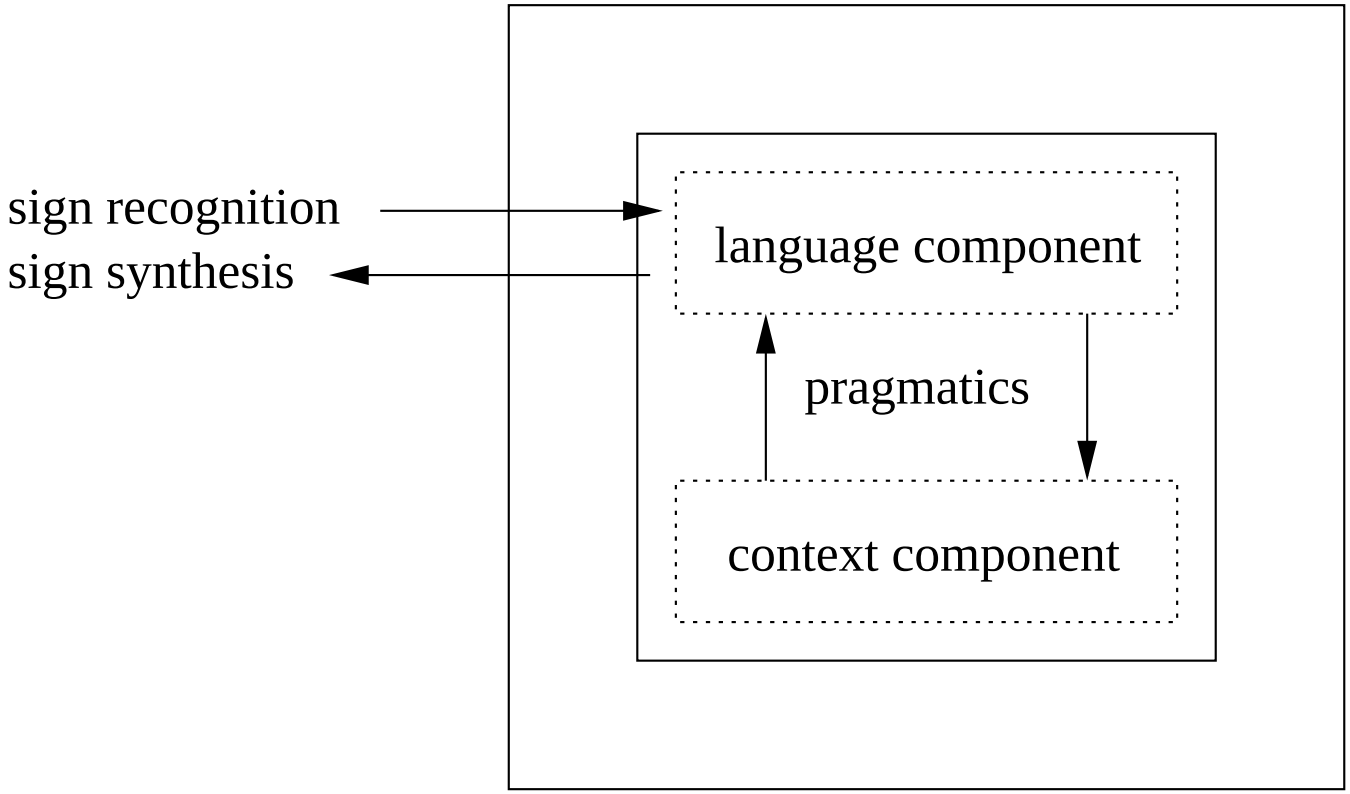
1.3.6 CONCEPT TYPES AND CONCEPT TOKENS IN ACTION



1.3.7 CONCEPT TYPES AT THE CONTEXT AND LANGUAGE LEVEL



1.3.8 USE OF EXTERNAL INTERFACES IN MEDIATED REFERENCE



1.3.9 DISADVANTAGES OF NOT HAVING CONTEXTUAL INTERFACES

1. *The conceptual core of language meanings remains undefined.*
   Most basic concepts originate in agents without language as recognition and action procedures of their contextual interfaces, and are re-used as the core of language meanings in agents with language.Therefore, agents with language but without contextual interfaces use meanings which are void of a conceptual core – though the relations between the concepts, represented by place holder words, may still be defined, both absolutely (for example in the *is-a* or *is-part-of* hierarchies) and episodically.

2. *The coherence or incoherence of content cannot be judged autonomously.*
   The coherence of stored content originates in the coherence of the external world.Therefore, only agents with contextual interfaces are able to relate content 'imported' by means of language to the data of their own experience. An agent without contextual interfaces, in contrast, has nothing but imported data – which is why the responsibility for their coherence lies solely with the users who store the data in the agent.

2. Day Two: Reference

2.1 Four Different Approaches to Reference

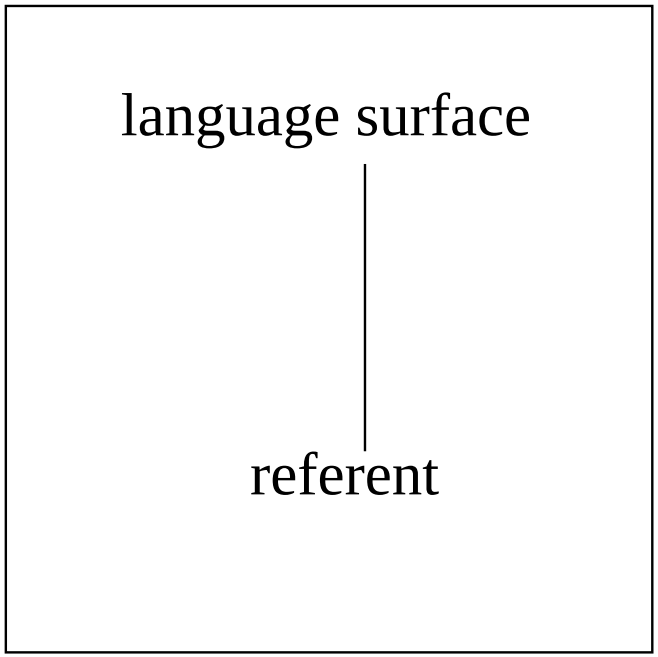
The notion of *reference* is generally defined as “a relation between language and the world.”

2.1.1 FIRST QUESTION: -SENSE OR +SENSE?

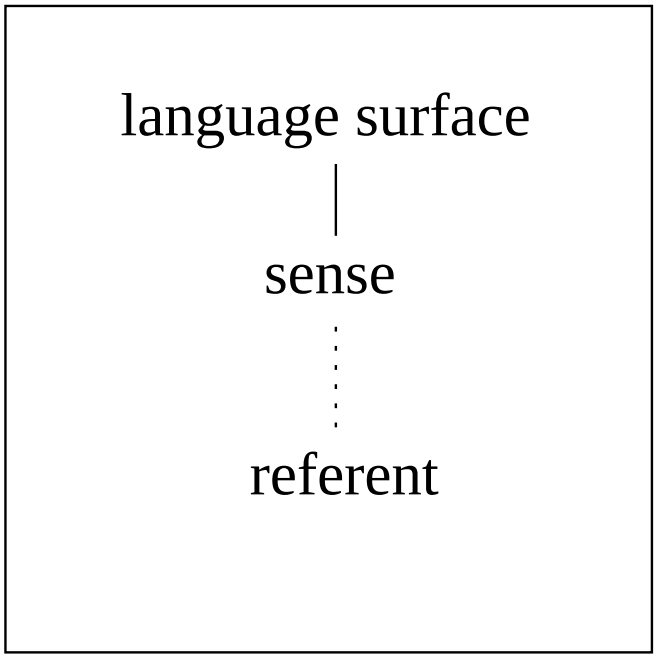
Should the relation of reference be defined directly between the language surface and the referent or should there be an intermediate level of sense (meaning)?

2.1.2 CHARACTERIZING THE ± SENSE DISTINCTION SCHEMATICALLY

−− sense



+ sense

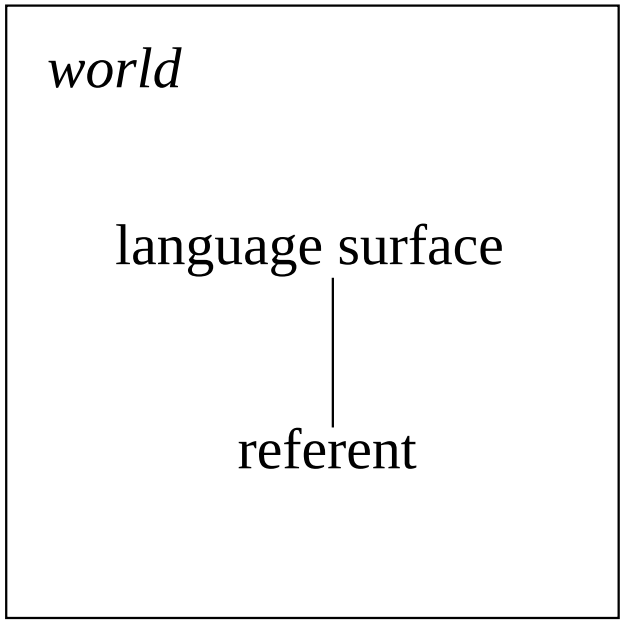


2.1.3 -CONSTRUCTIVE OR +CONSTRUCTIVE?

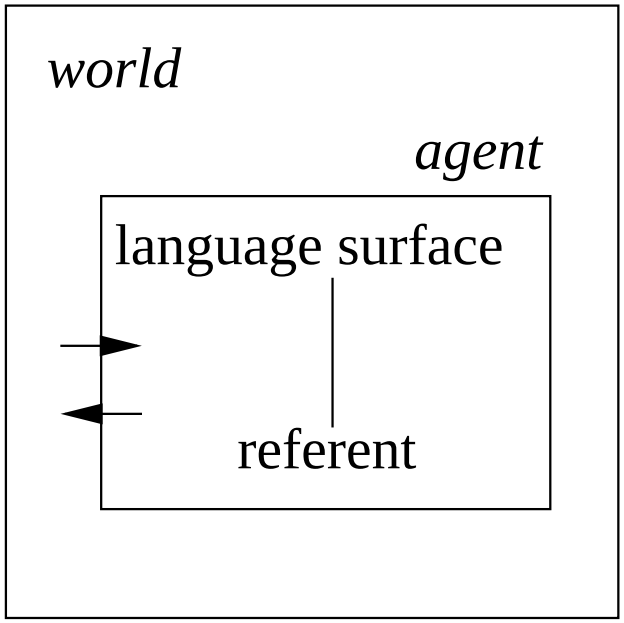
Should the relation of reference be defined between the language surface and the referent “out there in the world” or should it be reconstructed as something cognitive inside the agent?

2.1.4 CHARACTERIZING THE DISTINCTION ± CONSTRUCTIVE DISTINCTION SCHEMATICALLY

−− constructive



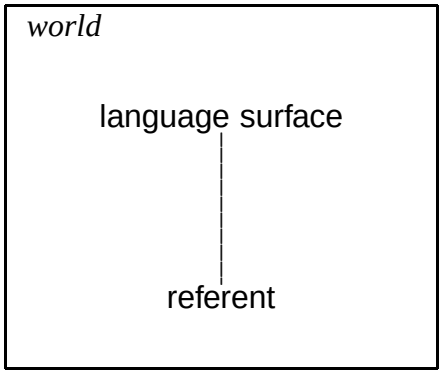
+ constructive



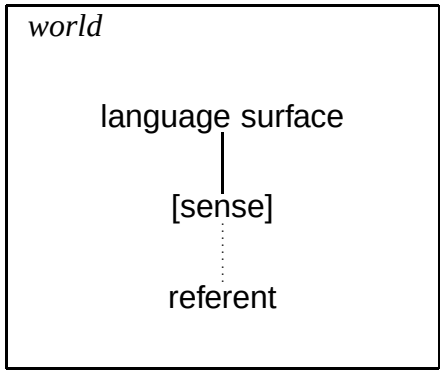
The combination of these binary features results in four different approaches, namely (i) [–sense, –constructive], (ii) [+sense, –constructive], (iii) [–sense, +constructive], and (iv) [+sense, +constructive]. All four approaches have been adopted in the history of semantics.

2.1.5 ONTOLOGIES OF SEMANTIC INTERPRETATION

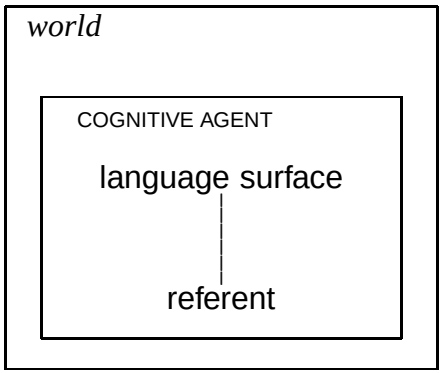
i [–sense, –constructive]
Russell, Carnap, Quine, Montague



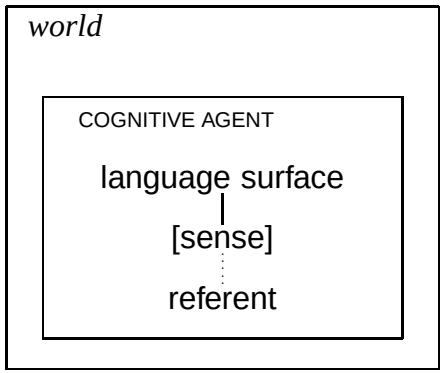
ii [+sense, –constructive]
Frege



iii [–sense, +constructive]
Newell & Simon, Winograd, Shank



iv [+sense, +constructive]
Anderson, CURIOUS, SLIM-machine



2.2 Problems with Approaches to Reference which are not [+sense]

2.2.1 THE [-SENSE, -CONSTRUCTIVE] APPROACH

Thrived in the scholasticism of the Middle Ages. Associated with William of Occam.

Defines reference as a direct relation between language surface and the referent, both out there in the world (extensional approach).

Motivation: only concrete surfaces and concrete referents are allowed in order to arrive at a reliable characterization of *TRUTH*.

Two basic rules of inference:

Substitutivity of Identicals (SI)
Existential Generalization (EG)

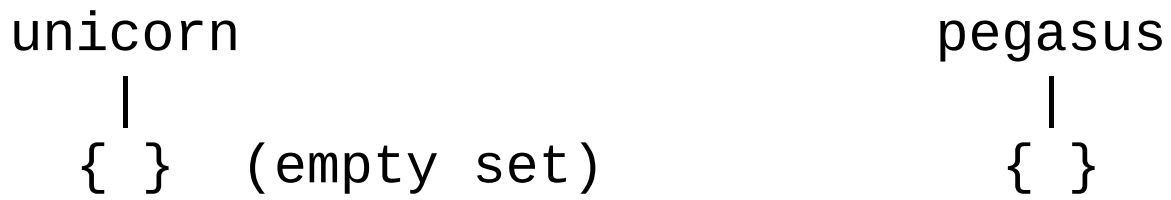
2.2.2 SUBSTITUTIVITY OF IDENTICALS (SI)

If f(a) and a=b, then f(b)

For example, if *Richard is sleeping*, formally sleep(Richard), is true and
Richard is the Prince of Burgundy, formally Richard = Prince of Burgundy, is true,
then it follows that *the prince of Burgundy is sleeping*,
formally sleep(Prince of Burgundy) is true.

2.2.3 PROBLEM WITH NON-REFERRING EXPRESSIONS

Mythical beings like unicorns and Pegasus do not exist. Formally:



Therefore, it follows from *Julia is looking for a unicorn* via SI that *Julia is looking for Pegasus*, which is regarded as a false inference by the logicians.

2.2.4 EXISTENTIAL GENERALIZATION (EG)

If f(a,b) is true, then it is true that a exists and b exists.

For example, if *Julia found a unicorn*, formally find (Julia, unicorn), is true
then *Julia exists* and *unicorn exists* is true.

2.2.5 PROBLEM WITH INTENSIONAL CONTEXTS

1) Julia finds a unicorn. $>$ A unicorn exists.
2) Julia seeks a unicorn. $\not>$ A unicorn exists.

The premises in these two examples have exactly the same syntactic structure, namely F(a,b). The only difference consists in the choice of the verb. Yet in (1) the truth of the premise implies the truth of the consequent, in accordance with the rule of existential generalization, while in (2) this implication does not hold.

How can a relation be established between a subject and an object if the object does not exist? How can Julia seeks a unicorn be grammatically well-formed, meaningful, and even true under realistic circumstances?

2.3 Frege's [+sense] approach to Reference

2.3.1 THE [+SENSE, –CONSTRUCTIVE] APPROACH

Pioneert by Gottlob Frege (1892) to solve the [–sense, –constructive] problems (intensional approach). Continued by William van Orman Quine (Quine 1960) and Richard Montague (Montague 1970). Continues to treat reference as an external relation between the language signs and their referents "out there in the world" ([–constructive])

Basic idea: distinction between
even and *uneven* contexts (Frege),
transparent and *opaque contexts* (Quine), or
extensional and *intensional contexts* (Montague).

Solution to the first problem: Existential generalization (EG) holds only in *even* (*transparent*, *extensional*) contexts.

Solution to the second problem: Substitutivity of Identicals (IG) holds in *even* (*transparent*, *extensional*) contexts if the extensions are the same, and in *uneven* (*opaque*, *intensional*) contexts if the senses (meanings, intensions) are the same.

2.4 Problem of [–constructive] approaches in general

Linguistic externalism
must be metalanguage based

Possible world semantics

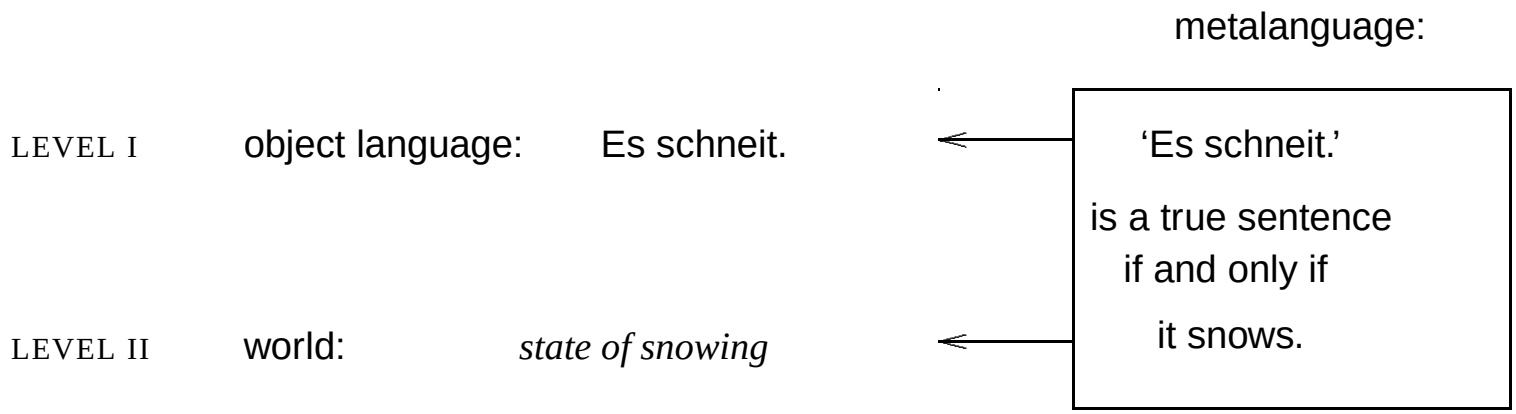
2.4.1 SCHEMA OF TARSKI'S T-CONDITION

T: x is a true sentence if and only if p.

2.4.2 INSTANTIATION OF TARSKI'S T-CONDITION

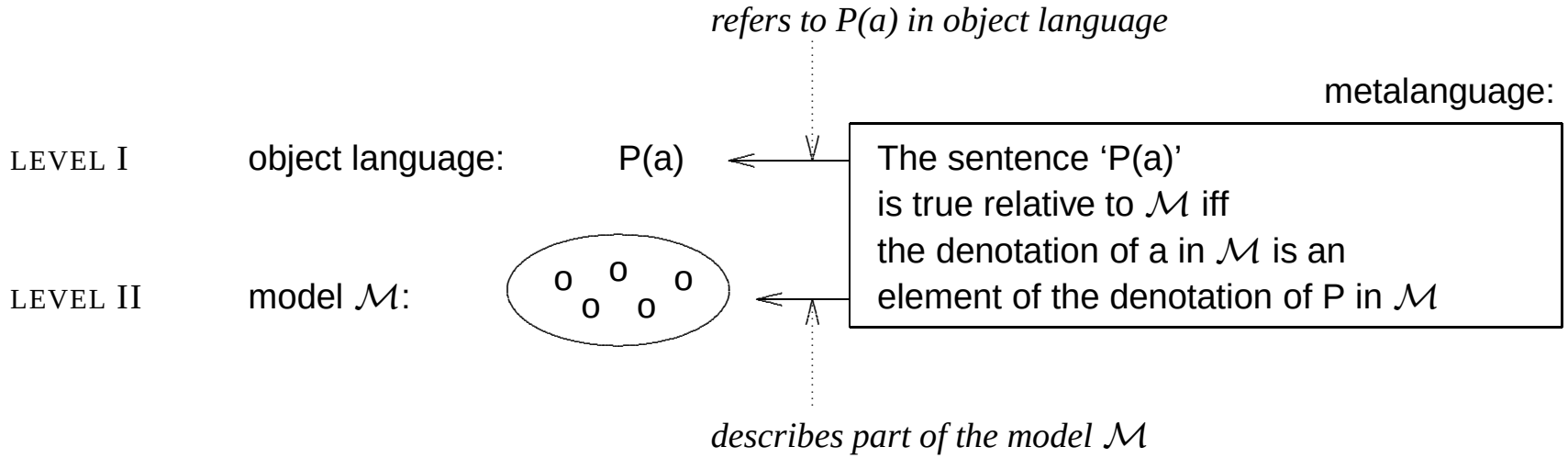
'Es schneit' is a true sentence if and only if it snows.

2.4.3 RELATION BETWEEN OBJECT AND METALANGUAGE



Crucial importance of verification (FoCL'99, Sect. 19.3).

2.4.4 T-CONDITION IN A LOGICAL DEFINITION



2.4.5 THE APPEAL TO IMMEDIATE OBVIOUSNESS IN MATHEMATICS

En l'un les principes sont palpables mais éloignés de l'usage commun de sorte qu'on a peine à tourner late tête de ce côte-la, manque d'habitude : mais pour peu qu'on l'y tourne, on voit les principes à peine; et il faudrait avoir tout à fait l'esprit faux pour mal raisonner sur des principes si gros qu'il est presque impossible qu'ils échappent.

[In [the mathematical mind] the principles are obvious, but remote from ordinary use, such that one has difficulty to turn to them for lack of habit : but as soon as one turns to them, one can see the principles in full; and it would take a thoroughly unsound mind to reason falsely on the basis of principles which are so obvious that they can hardly be missed.]

B. PASCAL (1623 -1662), *Pensées*, 1951:340

2.5 Metalanguage-based versus procedural semantics

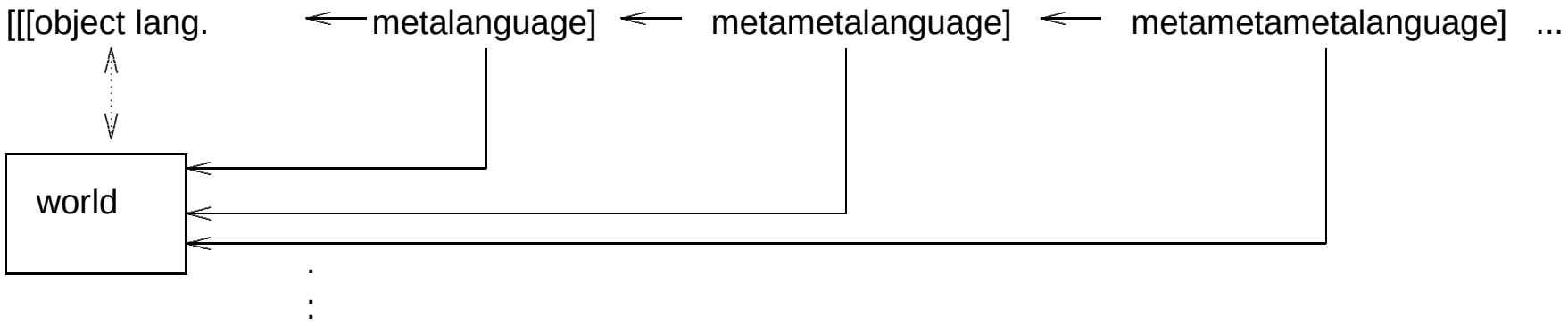
2.5.1 EXAMPLE OF A VACUOUS T-CONDITION

'A is red' is a true sentence if and only if A is red.

2.5.2 IMPROVED T-CONDITION FOR red

'A is red' is a true sentence if and only if A refracts light in the electromagnetic frequency interval between α and β.

2.5.3 HIERARCHY OF METALANGUAGES



2.5.4 AUTONOMY FROM THE METALANGUAGE

Autonomy from the metalanguage does not mean that computers would be limited to uninterpreted, purely syntactic deduction systems, but rather that Tarski's method of semantic interpretation is not the only one possible. Instead of assigning semantic representations to an object language by means of a metalanguage, computers use an operational method in which the notions of the programming language are realized automatically as machine operations.

2.5.5 EXAMPLE OF AUTONOMY FROM METALANGUAGE

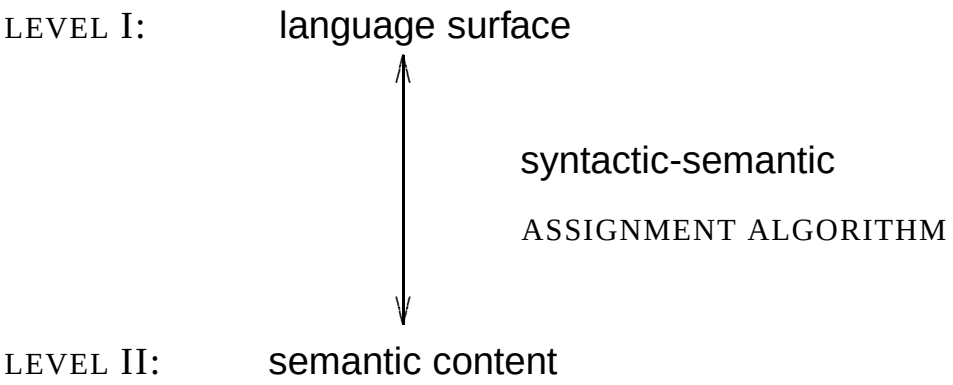
There is no problem to provide an adequate metalanguage definition for the rules of basic addition, multiplication, etc. However, the road from such a metalanguage definition to a working calculator is quite long and in the end the calculator will function mechanically – without any reference to these metalanguage definitions and without any need to understand the metalanguage.

2.5.6 PROGRAMMING LOGICAL SYSTEMS

There exist many logical calculi which have not been and never will be realized as computer programs. The reason is that their metalanguage translations contain parts which may be considered immediately obvious by their designers (e.g., quantification over infinite sets of possible worlds in modal logic), but which are nevertheless unsuitable to be realized as empirically meaningful mechanical procedures.

2.6 Basic structure of semantic interpretation

2.6.1 THE 2-LEVEL STRUCTURE OF SEMANTIC INTERPRETATION



2.6.2 THE FUNCTION OF SEMANTIC INTERPRETATION

For purposes of transmission and storage, semantic content is coded into surfaces of language (representation). When needed, the content may be decoded by analyzing the surface (reconstruction).

The expressive power of semantically interpreted languages resides in the fact that representing and reconstructing are realized *automatically*: a semantically interpreted language may be used correctly without the user having to be conscious of these procedures, or even having to know or understand their details.

2.7 Logical, programming, and natural languages

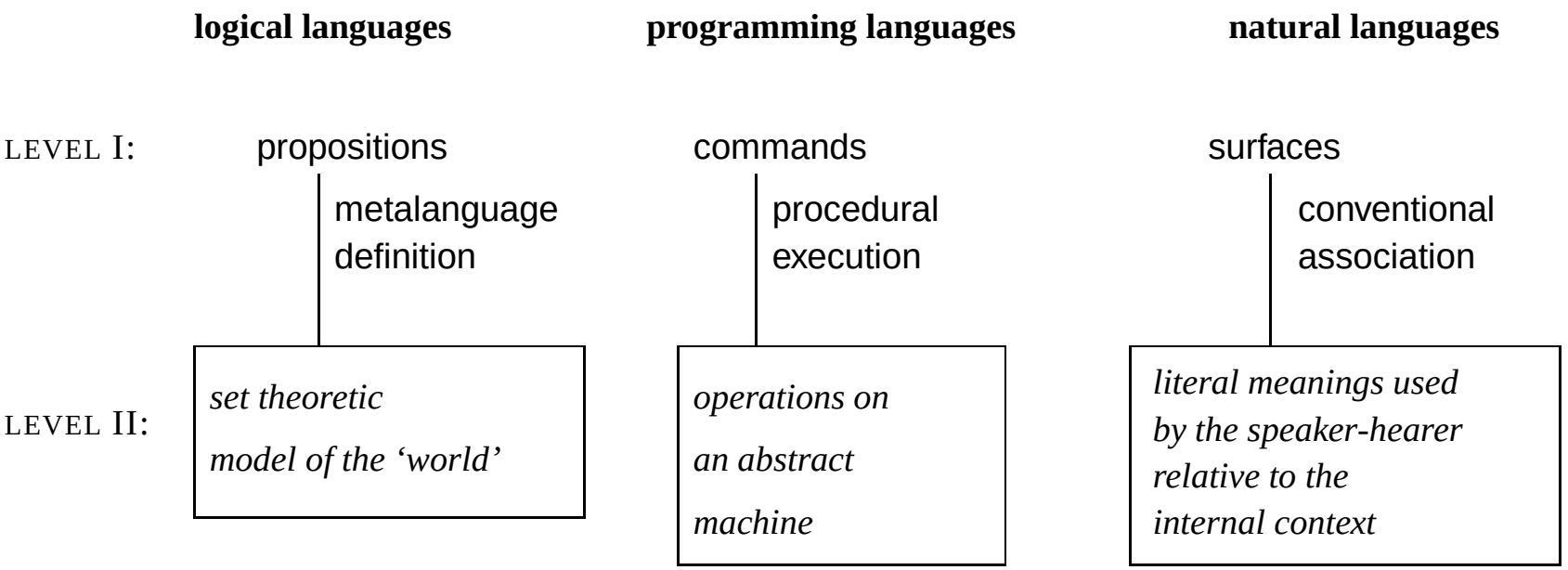
2.7.1 THREE DIFFERENT TYPES OF SEMANTIC SYSTEMS

1. *Logical languages*
   Designed to determine the truth value of arbitrary propositions relative to arbitrary models. The correlation between the two levels is based on *metalanguage definitions*.

2. *Programming languages*
   Designed to simplify the interaction with computers and the development of software. The correlation between the two levels is based on the *procedural execution* on an abstract machine, usually implemented electronically.

3. *Natural languages*
   Preexisting in the language community, they are analyzed syntactically by reconstructing the combinatorics of their surfaces. The associated semantic representations have to be deduced via the general principles of natural communication. The correlation between the two levels is based on *conventional association*.

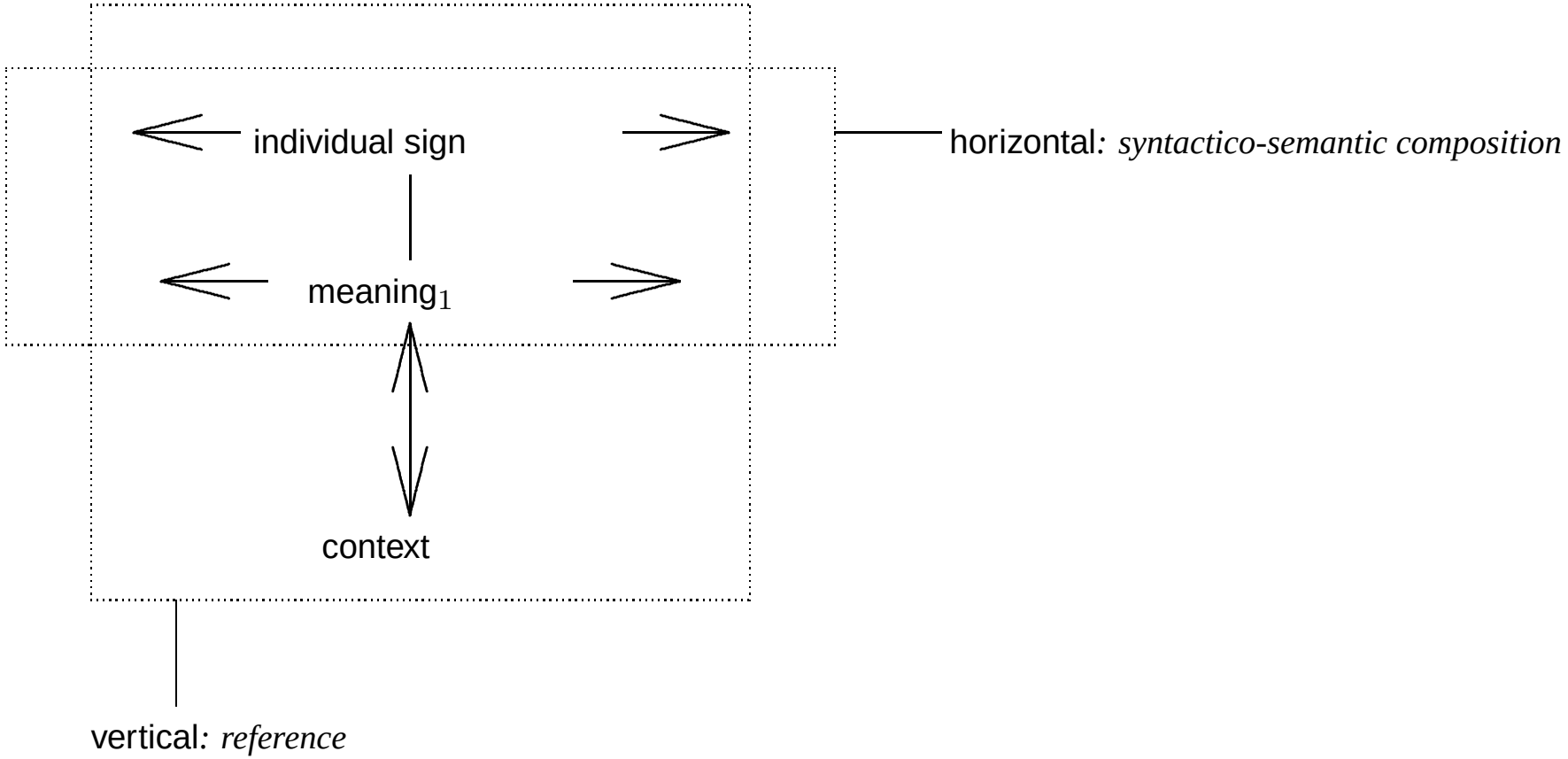
2.7.2 THREE TYPES OF SEMANTIC INTERPRETATION



3. Day Three: Semantic Relations

3.1 Traditional Notions of Grammar

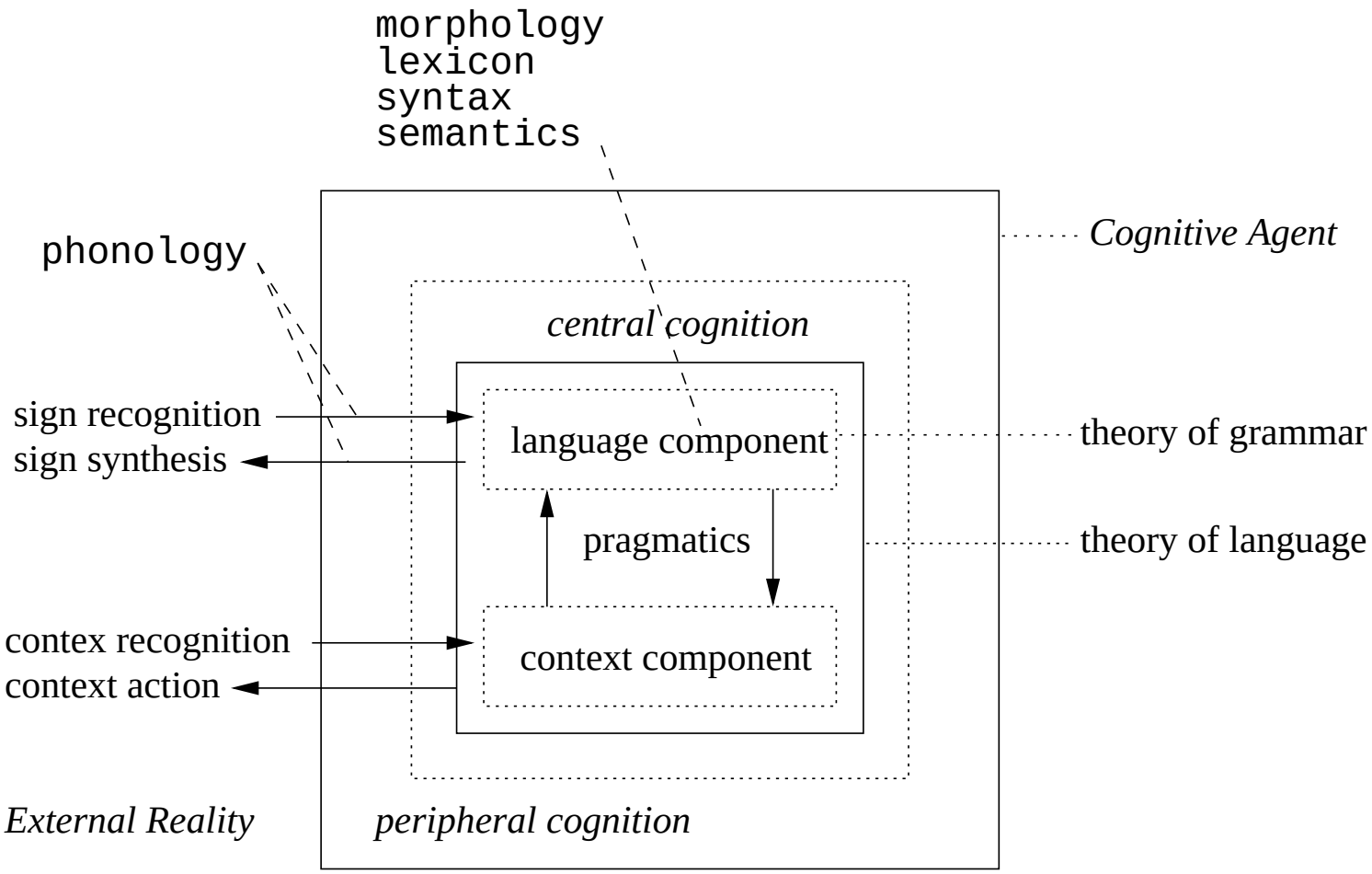
3.1.1 HORIZONTAL AND VERTICAL ASPECT OF ANALYSIS



3.1.2 TRADITIONAL COMPONENTS OF GRAMMAR

- *Phonology*: Science of language sounds (modality-dependent)
- *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

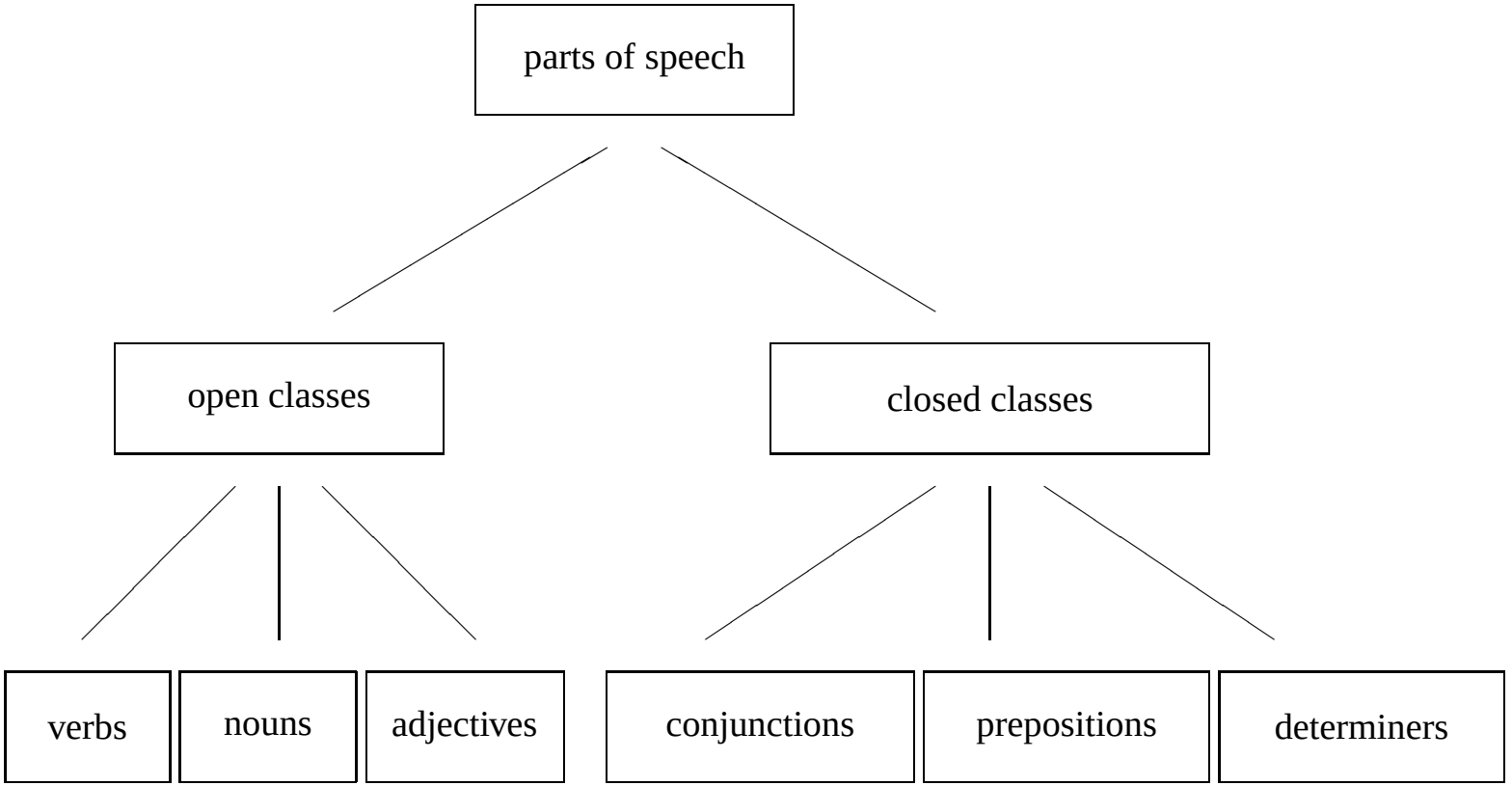
3.1.3 LOCATING THE COMPONENTS OF GRAMMAR IN COGNITION



3.1.4 TRADITIONAL PARTS OF SPEECH

- *verbs*, e.g., walk, read, give, help, teach, . . .
- *nouns*, e.g., book, table, woman, messenger, arena, . . .
- *adjectives*, e.g., quick, good, low, . . .
- *conjunctions*, e.g., and, or, because, after, . . .
- *prepositions*, e.g., in, on, over, under, before, . . .
- *determiners*, e.g., a, the, every, some, all, any, . . .
- *particles*, e.g., only, already, just. . .

3.1.5 CLASSIFICATION OF THE PARTS OF SPEECH INTO OPEN AND CLOSED CLASSES



3.1.6 COMPARISON OF THE OPEN AND THE CLOSED CLASSES

- The open classes comprise several 10 000 elements, while the closed classes contain only a few hundred words.

- The morphological processes of inflection, derivation, and composition are productive in the open classes, but not in the closed classes.

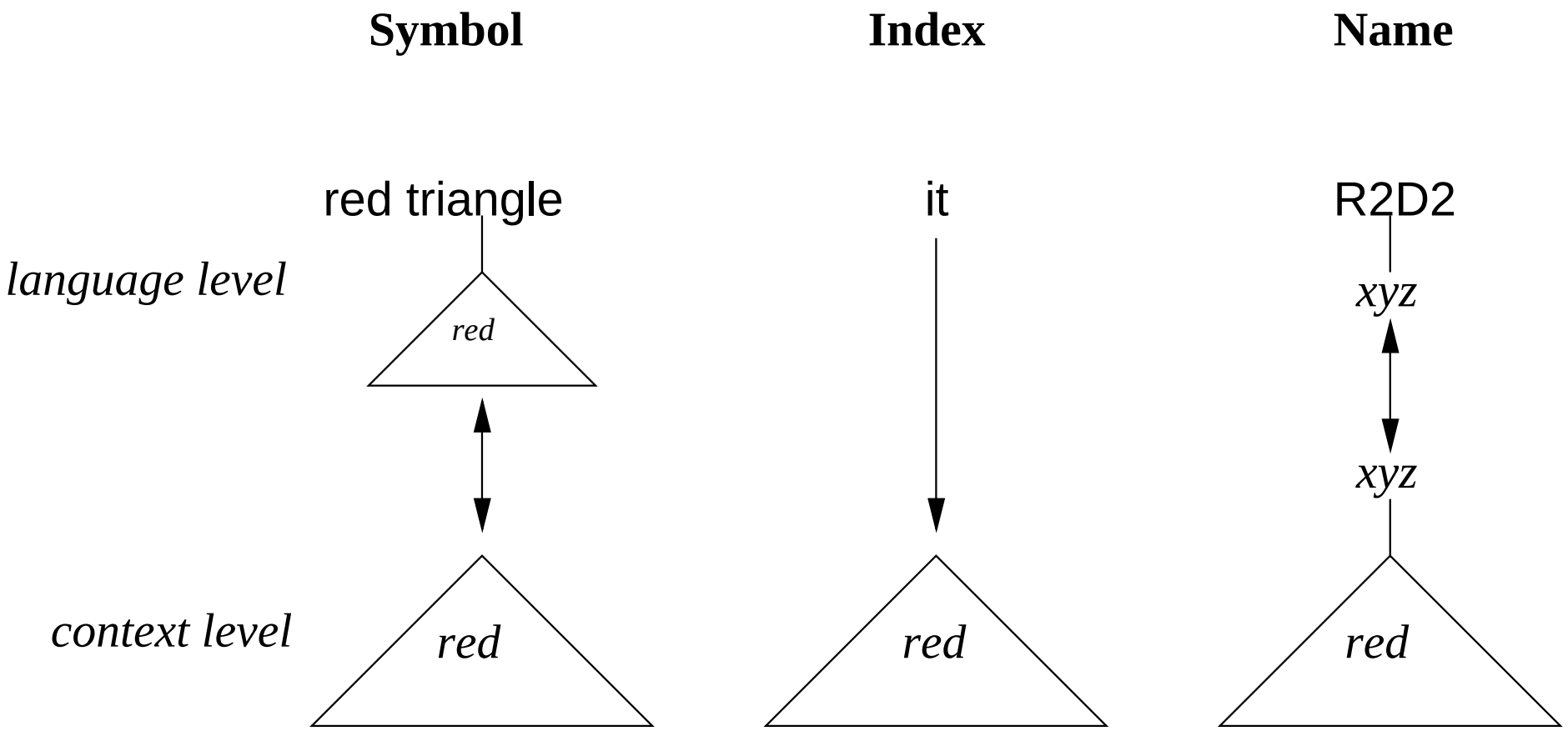
- In the open classes, the use of words is constantly changing, with new ones entering and obsolete ones leaving the current language, while the closed classes do not show a comparable fluctuation.

3.1.7 RELATING THE PARTS OF SPEECH TO THE KINDS OF SIGNS

The elements of the open classes are also called *content words*, while the elements of the closed classes are also called *function words*. In this distinction, however, the sign kind must be taken into consideration besides the category.

This is because only the *symbols* among the nouns, verbs, and adjective-adverbials are content words in the proper sense. *Indices*, on the other hand, e.g. the personal pronouns he, she, it etc., are considered function words even though they are of the category noun. Indexical adverbs like here or now do not even inflect, forming no comparatives and superlatives. The sign type *name* is also a special case among the nouns.

3.1.8 COMPARING ICONIC, INDEXICAL, AND NAME-BASED REFERENCE



3.2 Compositionality

3.2.1 FREGEAN PRINCIPLE

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

3.2.2 DIFFERENT COMPOSITION, SAME PARTS

The dog bites the man.
The man bites the dog.

3.2.3 SAME COMPOSITION, DIFFERENT PARTS

The dog is chasing a cat.
The dog is chasing a squirrel.

3.2.4 THE FREGEAN PRINCIPLE APPLIES TO *analyzed surfaces*

Syntactic ambiguity in an unanalyzed surface: Suzy saw the man with a telescope

analyzed surface 1:

Suzy$_{subject\ noun}$
saw$_{two-place\ verb}$
the man$_{object\ noun}$
with the telescope$_{adnominal\ adjective}$ $\Rightarrow$ *the man has the telescope*

analyzed surface 2:

Suzy$_{subject\ noun}$
saw$_{two-place\ verb}$
the man$_{object\ noun}$
with the telescope$_{adverbial\ adjective}$ $\Rightarrow$ *Suzy has the telescope*

3.2.5 THE FREGEAN PRINCIPLE APPLIES TO THE *literal meaning*

3.2.6 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.

3.2.7 PRAGMATIC AMBIGUITY: USING THE SAME SIGN RELATIVE TO DIFFERENT CONTEXTS OF USE

That's beautiful weather!

uttered relative to a context of a gorgeous summer day: *the weather is beautiful.*
(literal use, language level and context level are in a relation of correspondence)

uttered relative to a context of a miserable winter day: *the weather is awful.*
(ironice use, language level and context level are in a relation of contrast)

Conclusion: If the Fregean Principle is applied to (i) analyzed surfaces and (ii) their literal meaning$_1$ there are neither syntactic nor pragmatic ambiguities.

3.3 Parts of Speech at the Elementary, Phrasal, and Clausal level

3.3.1 PARTS OF SPEECH AT THE PHRASAL LEVEL

1. *noun*:

Julia	slept	*elementary*
the pretty young girl	slept	*phrasal*

The phrasal noun consists of a determiner (function word), two elementary adjectives, and an elementary noun, the latter being content words.

2. *verb*:

Julia	slept	*elementary*
Julia	could have been sleeping	*phrasal*

The phrasal verb consists of (forms of) a modal verb, two auxiliaries, and a main verb.

3. *adjective*:

Julia slept	well	*elementary*
Julia slept	for five minutes	*phrasal*

The phrasal adjective for five minutes consists of a preposition, a determiner, and a noun, the latter being a content word.

3.3.2 PARTS OF SPEECH AT THE CLAUSAL LEVEL

1. *noun*:

Julia	pleased her mother.	*elementary*
the pretty young girl	pleased her mother.	*phrasal*
that Julia slept	pleased her mother.	*clausal*

2. *adnominal adjective*:

the	black	dog	barked	*elementary*
the dog	with the bone		barked	*phrasal*
the dog	which Mary saw		barked	*clausal*

3. *adverbial adjective*:

Recently	Mary smiled.	*elementary*
After her nap	Mary smiled.	*phrasal*
When Fido barked	Mary smiled.	*clausal*

3.3.3 Principles of syntactic-semantic well-formedness

Agreement, valency, and word order.

3.3.4 Violation of agreement

* Every girl want a coke. (should be wants)

3.3.5 Violation of valency

* John gave Mary. (should be John gave Mary something)

3.3.6 Violation of word order

* A book read John. (should be John read a book or A book John read.)

3.4 Semantic Relations

Basic task of a grammar formalism: characterizing the *semantic relations* between elementary, phrasal, and clausal parts of speech.

3.4.1 THE TWO KINDS OF SEMANTIC RELATIONS IN NATURAL LANGUAGE

functor-argument structure:

JULIA$_{argument}$ SLEPT$_{functor}$ SOUNDLY$_{modifier}$

The arguments of functors are obligatory, the modifiers of functors or arguments are optional.

coordination:

JULIA$_{conjunct}$ SUSANNE$_{conjunct}$ AND MARY$_{conjunct}$

Coordination combines elements of the same part of speech, while functor-argument structure combines elements of different parts of speech.

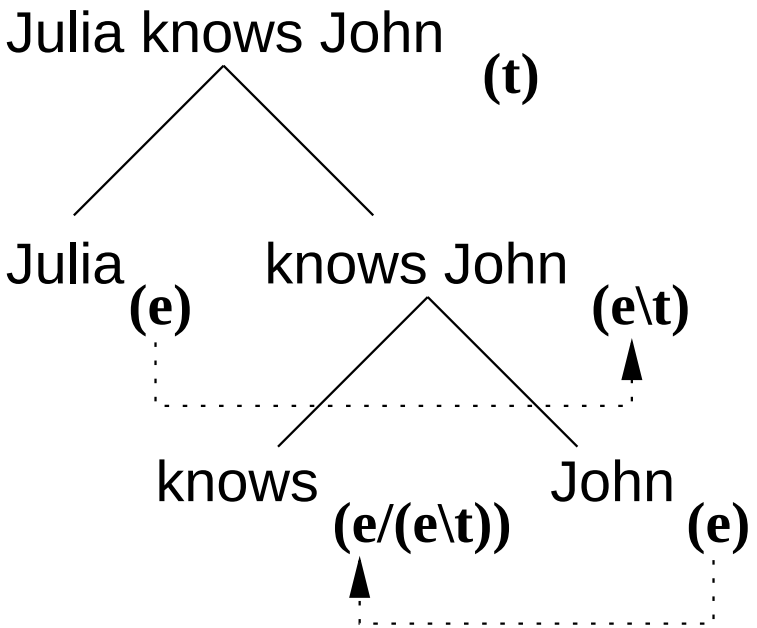
3.5 Three Sign-oriented Grammar Formalisms

3.5.1 BOTTOM-UP CATEGORIAL ANALYSIS OF Julia knows John.

Categorial Grammar was invented by Leśniewski (1929) and Ajdukiewicz (1935), and first applied to natural language by Bar Hillel (1953):

rule **2:** $\beta_{(X)} * \alpha_{(X\backslash Y)} \longrightarrow \beta\alpha_{(Y)}$

rule **1:** $\alpha_{(X/Y)} * \beta_{(X)} \longrightarrow \alpha\beta_{(Y)}$



Characterization of semantic relations (functor-argument structure) in terms of canceling an argument position in the functor.

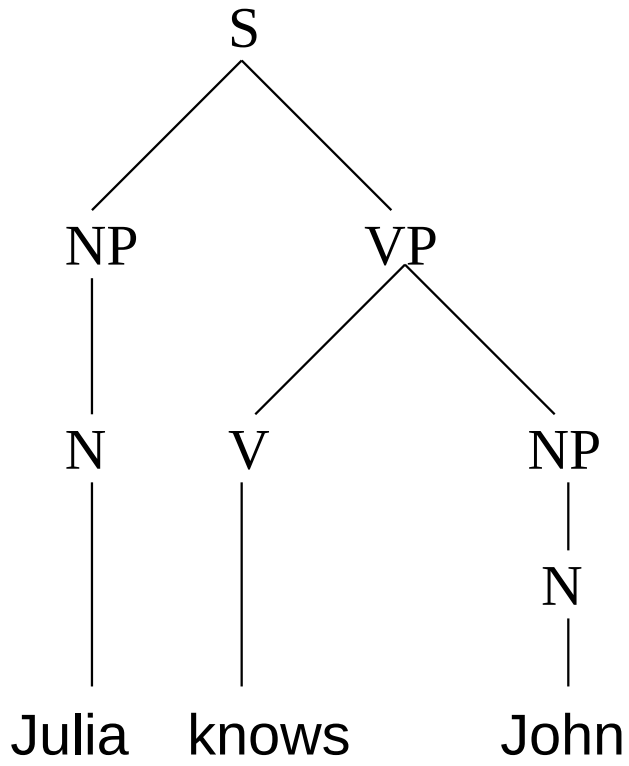
3.5.2 TOP-DOWN PHRASE-STRUCTURE ANALYSIS Julia knows John.

Phrase Structure grammar (PS-grammar) was proposed by Chomsky (1957, 1965), based on a formal system by Post (1936):

rule **1: S** ⟶ **NP VP**

rule **2: NP** ⟶ **N**
rule **3: VP** ⟶ **V NP**

rule **4: N** ⟶ Julia John,
rule **5: V** ⟶ knows



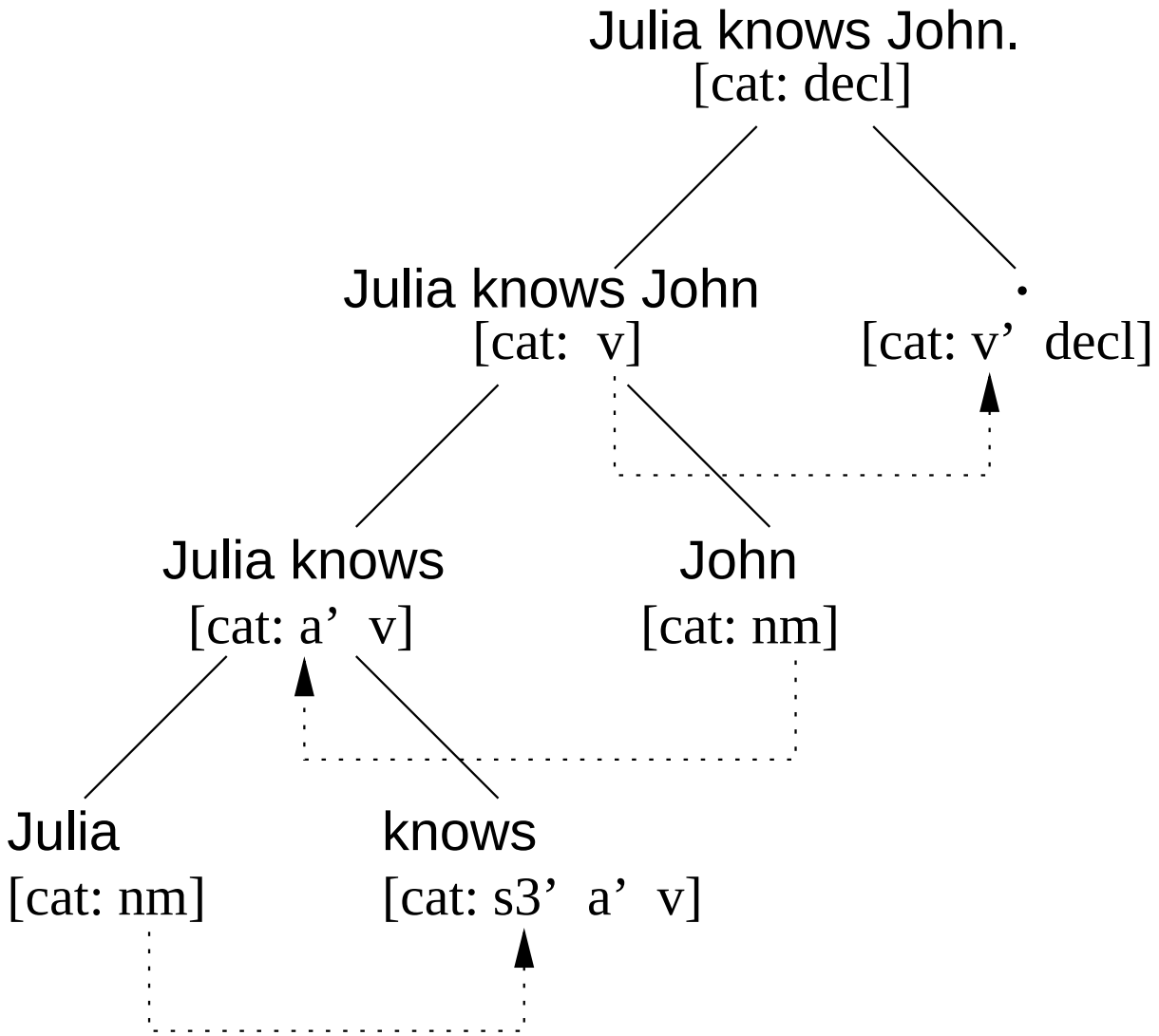
Characterization of semantic relations (functor-argument structure) in terms of substituting a node by two (or three) “daughters”.

3.5.3 TIME-LINEAR BOTTOM-UP NEWCAT DERIVATION OF Julia knows John.

3 S+IP { }
[cat: VT] [VT' SM]
cancel VT

2 FV+NP {S+IP}
[cat: NP' X VT] [NP]
cancel NP

1 Nom+FV {FV+NP}
[cat: NP] [cat: NP' X VT]
cancel NP



3.6 From a Sign-Oriented to an Agent-Oriented Approach

3.6.1 THE PROBLEM WITH SIGN-ORIENTED APPROACHES

What to do with their grammatical analyses?

3.6.2 HOW THIS QUESTION IS ANSWERED BY THE AGENT-ORIENTED APPROACH OF DBS

The analysis of language expressions (theory of grammar) must be embedded into a functional model of how communicating with natural language works (theory of language).

3.6.3 SCIENTIFIC NECESSITY OF AN AGENT-ORIENTED APPROACH

An agent-oriented approach is essential for a scientific understanding of natural language, because the general structure of natural language is determined by its function, and the function of language is communication (FoCL'99, 4.5.3).

This is in concord with Darwin's theory of evolution in which anatomy, for example, will be structured according to functions associated with use.

3.6.4 COMPUTATIONAL DESIGN DECISIONS

1. What should be the *data structure* (abstract data type) of the content items stored in memory?

2. What should be the *algorithm* to read items of content into (hearer mode) and out of (speaker mode) the agent's memory?

3. What should be the *data base schema* according to which the content items are stored in and retrieved from memory?

4. What should be the functional flow?

3.7 Verification

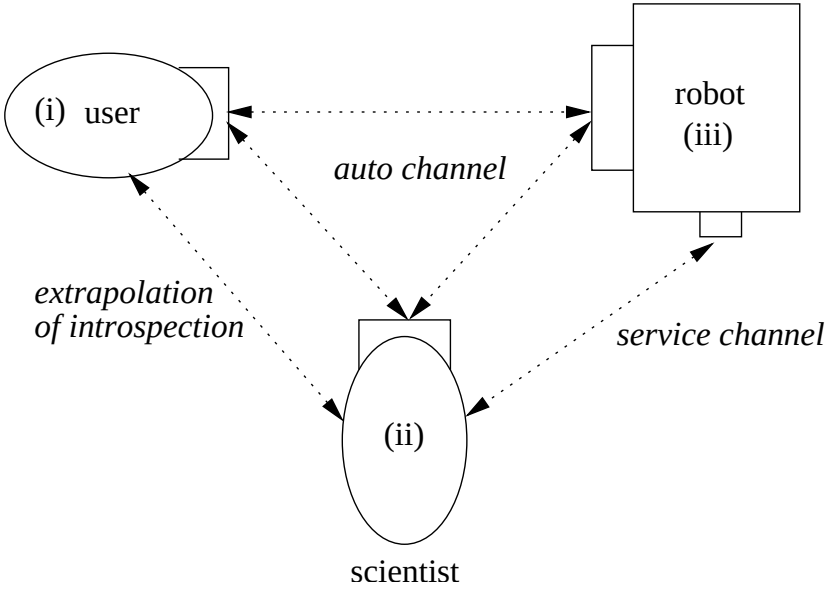
3.7.1 VERIFICATION OF AN AGENT-ORIENTED APPROACH

Like any scientific theory, the DBS mechanism of natural language communication must be *verified*. For this, the single most straightforward method is implementing the theory computationally as a talking robot. This method of verification is distinct from the repeatability of experiments in the natural sciences, and may serve as a unifying standard for the social sciences.

3.7.2 CONSTELLATIONS PROVIDING DATA FOR A SCIENTIFIC RECONSTRUCTION OF COMMUNICATION

1. Interaction between (i) the user and (iii) the robot
2. Interaction between (i) the user and (ii) the scientist
3. Interaction between (ii) the scientist and (iii) the robot

3.7.3 INTERACTION BETWEEN USER, ROBOT, AND SCIENTIST



3.7.4 DATA CHANNELS OF COMMUNICATIVE INTERACTION

1. The *auto-channel* processes input automatically and produces output autonomously, at the context as well as the language level. In natural cognitive agents, i.e., the user and the scientist, the auto-channel is present from the very beginning in its full functionality. In artificial agents, in contrast, the auto-channel must be reconstructed – and it is the goal of Database Semantics to reconstruct it as realistically as possible.

2. The *extrapolation of introspection* is a specialization of the auto-channel and results from the scientists' effort to improve man–machine communication by taking the view of the human user. This is possible because the scientist and the user are natural agents.

3. The *service channel* is designed by the scientist for the observation and control of the artificial agent. It allows direct access to the robot's cognition because its cognitive architecture and functioning is a construct which in principle may be understood completely by the scientist.

3.7.5 WHY THE ARGUMENTS AGAINST MENTALISM OR PSYCHOLOGISM DO NOT HOLD AGAINST DBS

A computational model of natural language communication requires a reconstruction of cognition. The usual arguments against mentalist or psychologist approaches do not apply because of the specific method of verification available to DBS.

3.7.6 THE EQUATION PRINCIPLE OF DATABASE SEMANTICS (NLC'06, 1.3.1)

1. The more realistic the reconstruction of cognition, the better the functioning of the model.
2. The better the functioning of the model, the more realistic the reconstruction of cognition.

3.7.7 PRACTICAL ADVANTAGES OF A COMPUTATIONAL RECONSTRUCTION OF NL COMMUNICATION

Given that natural language communication is a real and objective procedure, it is a legitimate scientific goal to model this procedure as a theory of how natural language communication works. Such a theory is not only of academic interest, but is also the foundation of free human-machine communication in natural language. The practical implications of having machines which can freely communicate in natural language are enormous: instead of having to program the machines we could simply talk with them.

4. Day Four: The Cycle of Natural Language Communication

4.1 The Data Structure of Database Semantics

4.1.1 Automatic word form recognition: matching an unanalyzed surface onto a key

unanalyzed word form surface: water

matching

lexical entry: [water (lexical description)]

4.1.2 Format of Analyzed Word Forms: the lexical noun proplets water and eau

	English	*French*	
surface	sur: water	sur: eau	*lexical features*
core attribute	noun: *water*	noun: *water*	
category	cat: sn	cat: sn	
semantic property	sem: mass	sem: mass	
modifier(s)	mdr:	mdr:	*continuation features*
functor	fnc:	fnc:	
next conjunct	nc:	nc:	
previous conjunct	pc:	pc:	
identity	idy:	idy:	*bookkeeping features*
proposition number	prn	prn	

4.1.3 Conceptual requirements on the Data Structure

Straightforward realization of the hearer mode, the think mode, and the speaker mode.

4.1.4 Technical requirements on the Data Structure

1. easy coding of lexical details

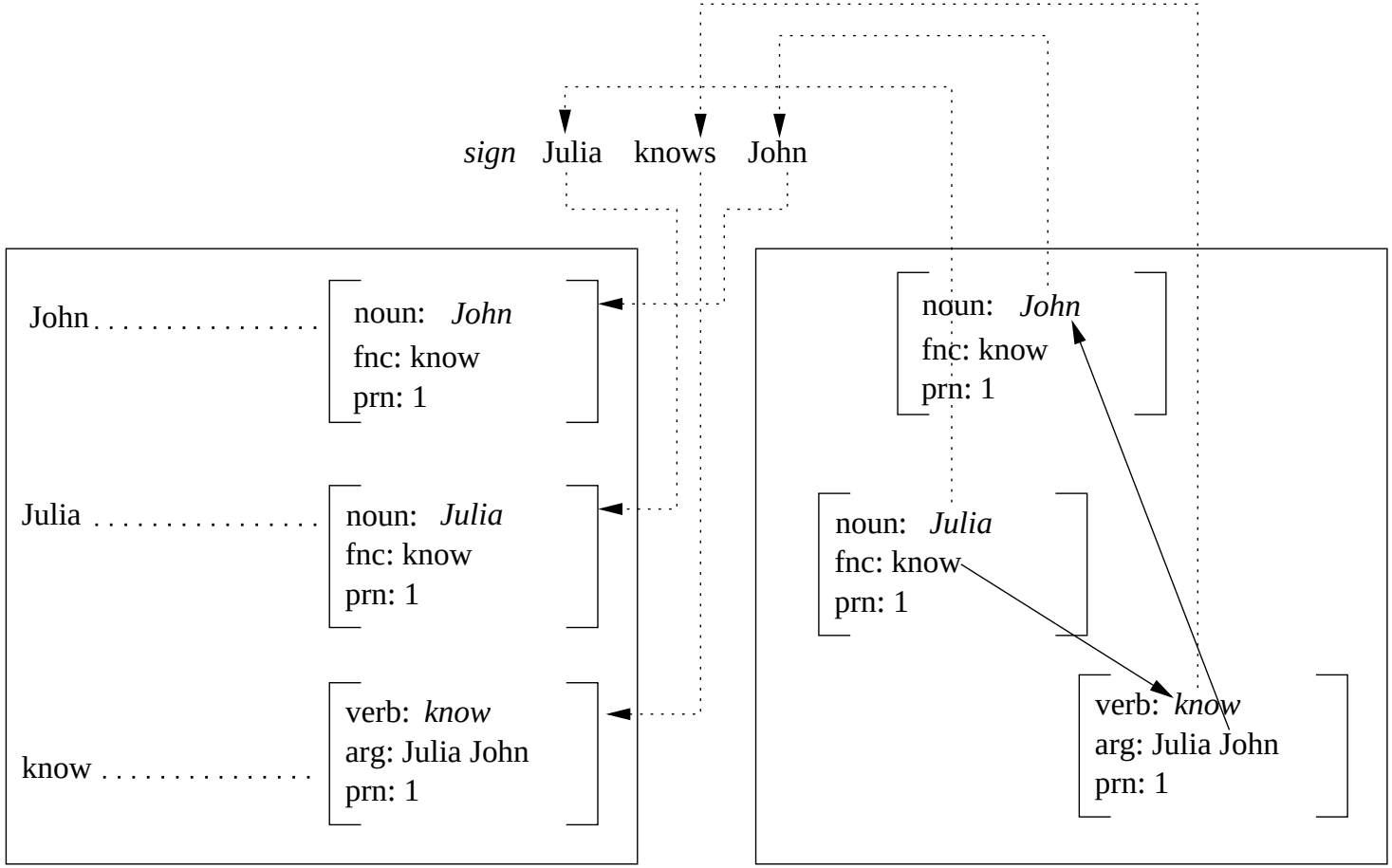
2. easy coding of semantic relations

3. suitability for a computationally straightforward matching
   (a) between rule patterns and language proplets
   (b) between language level and context level

4. suitability for storage and retrieval in a database: order-free coding of relations between basic items

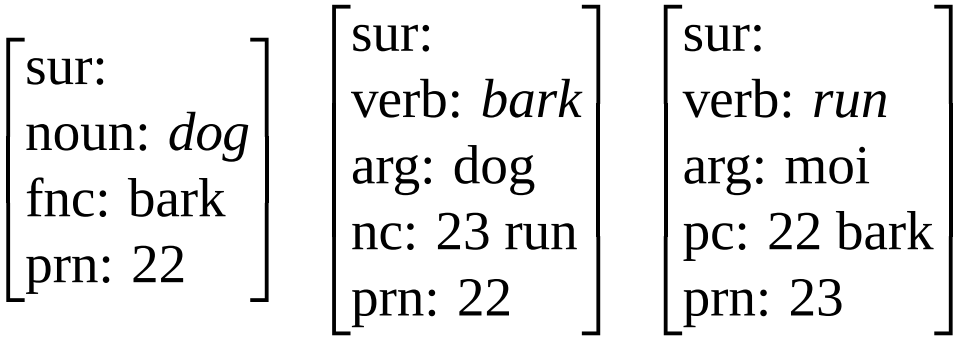
4.2 The Cycle of NL Communication

4.2.1 TRANSFER OF CONTENT FROM THE SPEAKER TO THE HEARER

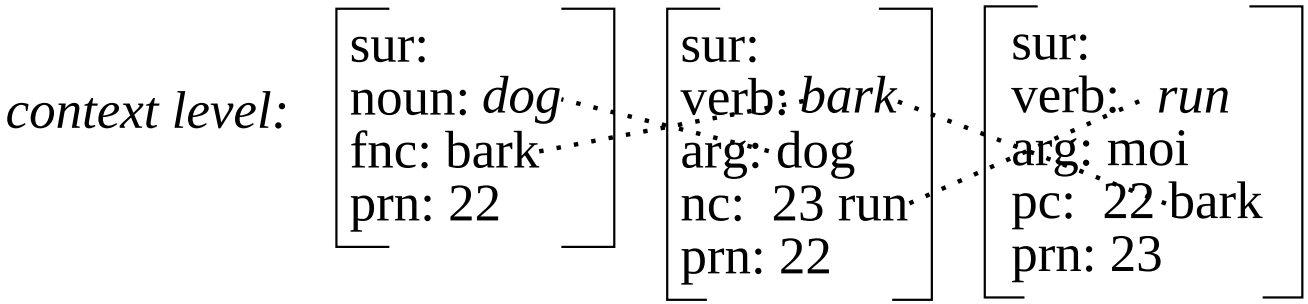


hearer: key−word−based storage *speaker: retrieval−based navigation*

4.2.2 THE CODING OF CONTENT: *dog barks. (I) run.*



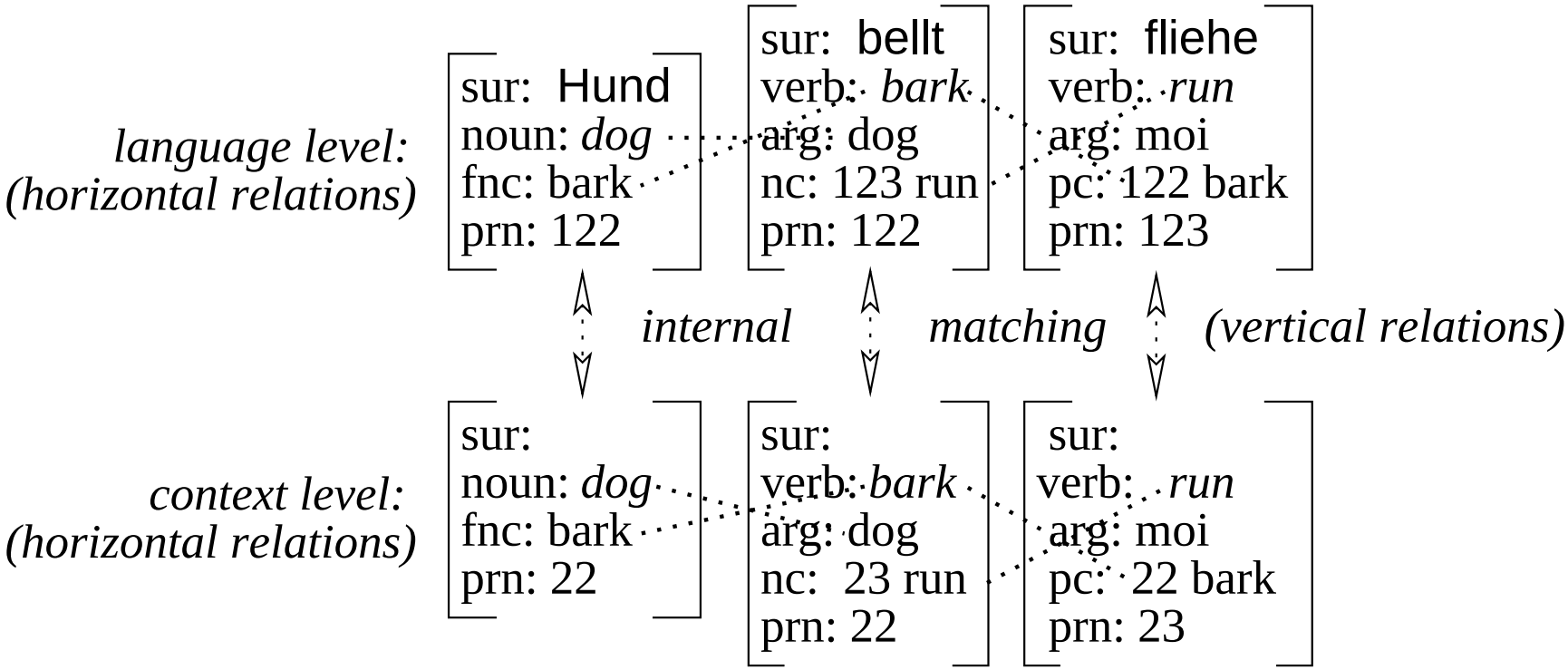
4.2.3 CODING OF RELATIONS BETWEEN CONCEPTS VIA ATTRIBUTE-VALUE PAIRS IN PROPLETS



4.2.4 LANGUAGE PROPLETS REPRESENTING *dog barks. (I) run.*

sur: Hund	sur: bellt	sur: fliehe
noun: *dog*	verb: *bark*	verb: *run*
fnc: bark	arg: dog	arg: moi
prn: 122	nc: 123 run	pc: 122 bark
	prn: 122	prn: 123

4.2.5 IMPACT OF (HORIZONTAL) INTERPROPLET RELATIONS ON (VERTICAL) MATCHING



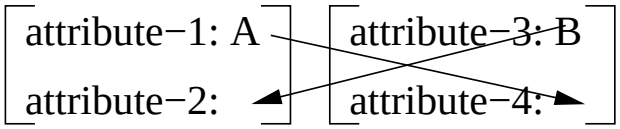
4.2.6 KEYS FOR LEXICAL LOOKUP IN THE SPEAKER AND THE HEARER MODE

sur: Hund	← *key for lexical lookup in the hearer mode*
noun: *dog*	← *key for lexical lookup in the speaker mode*
fnc:	
prn	

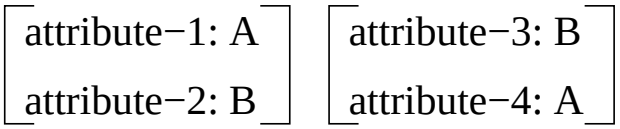
4.3 Coding semantic relations

4.3.1 SCHEMA OF BIDIRECTIONAL POINTERING

cross−copying

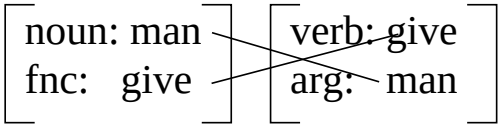


result

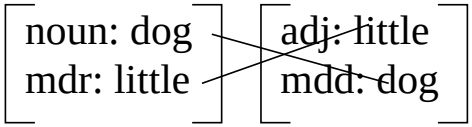


4.3.2 BASIC FUNCTOR-ARGUMENT AND COORDINATION STRUCTURES

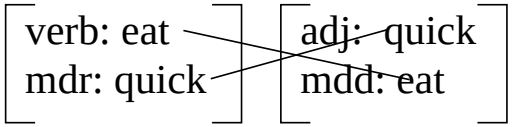
(i) noun−verb



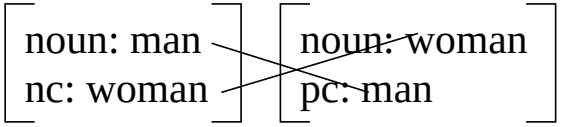
(ii) noun−adnominal



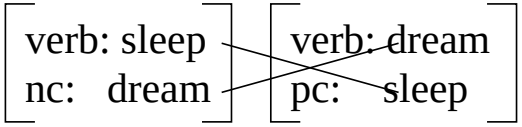
(iii) verb−adverbial



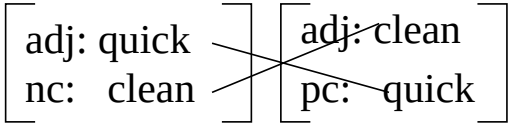
(iv) noun coordination



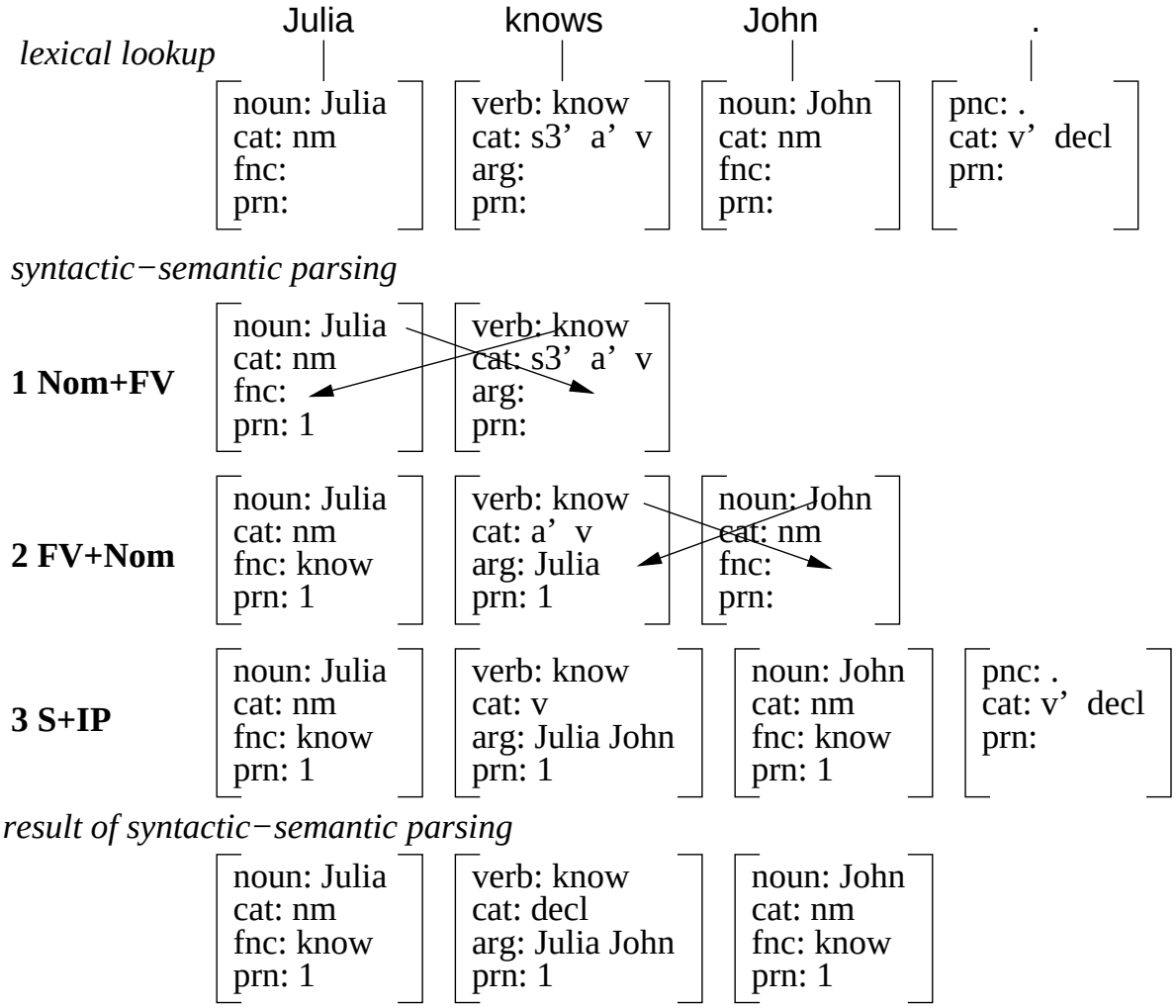
(v) verb coordination



(vi) adjective coordination



4.3.3 HEARER-MODE DERIVATION IN DBS

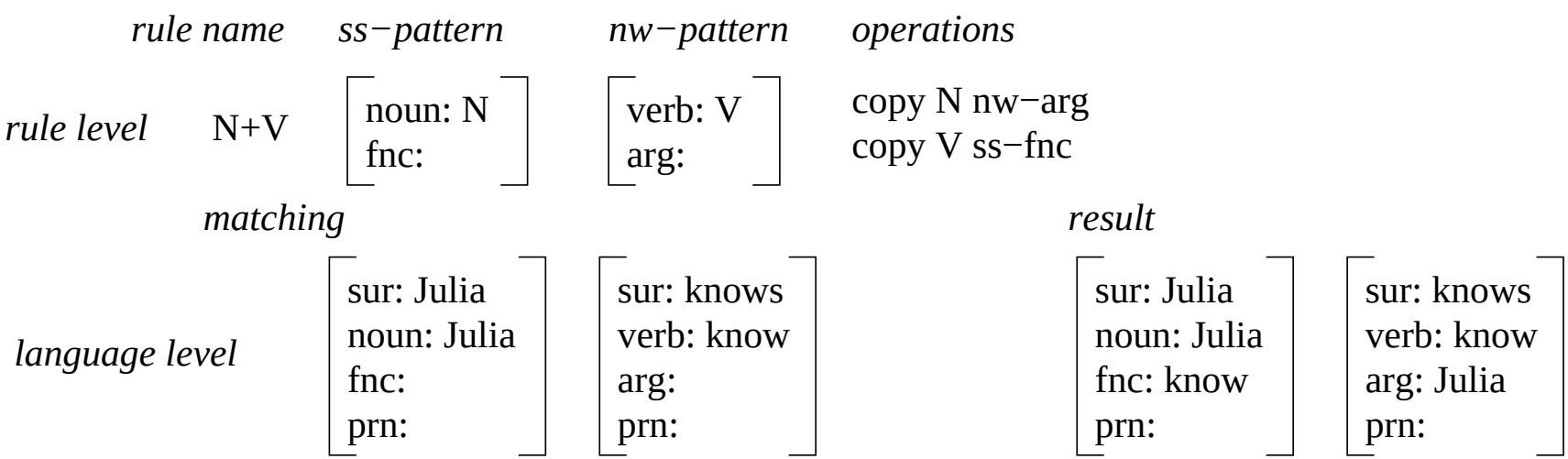


4.3.4 Methods for establishing semantic relations in the hearer mode

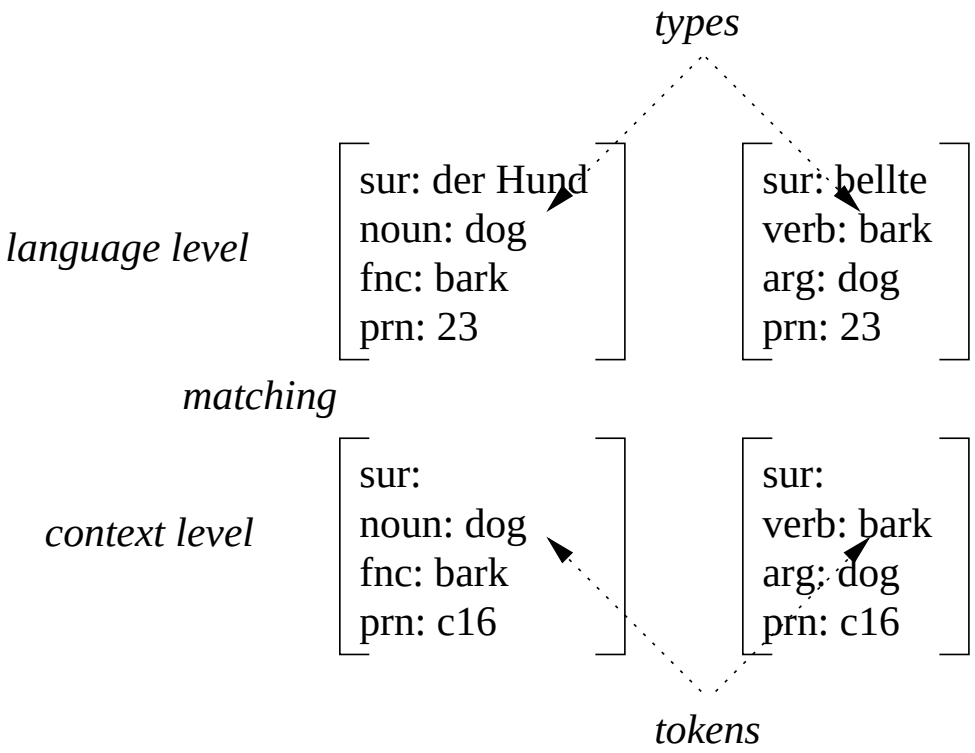
- bidirectional copying
- simultaneous substitution
- deletion (absorption)

4.4 Matching proplets

4.4.1 SCHEMATIC ILLUSTRATION OF MATCHING IN A RULE APPLICATION



4.4.2 SCHEMATIC MATCHING BETWEEN LANGUAGE AND CONTEXT LEVEL



The role of attributes at the two levels

The role of corresponding values at the two levels

The role of the type/token relation between values at the two levels

4.5 Storage in a Database

4.5.1 CONTENT OF Julia knows John.

noun: John	noun: Julia	verb: know
cat: nm	cat: nm	cat: decl
fnc: know	fnc: know	arg: Julia John
prn: 1	prn:1	prn: 1

order-free, here using the alphabetical order of the core values

4.5.2 STORAGE OF PROPLETS IN A WORD BANK

owner records	*member records*	
...	...	
[noun: John]	...	[noun: John cat: nm fnc: know prn: 1]
...	...	
[noun: Julia]	...	[noun: Julia cat: nm fnc: know prn: 1]
...	...	
[verb: know]	...	[verb: know cat: decl arg: Julia John prn: 1]
...	...	

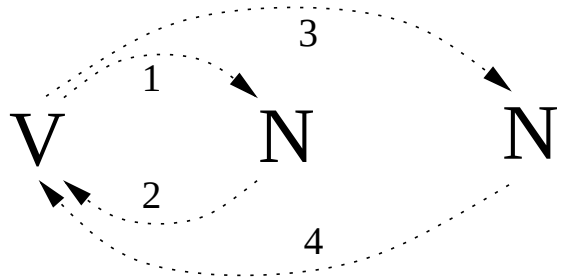
4.6 Database Semantics: Think Mode

A Word Bank goes beyond a classic network database because the semantic relations between proplets provide something like a railroad system. Using the retrieval mechanism of the data base, an autonomous navigation moves an imaginary focus point from one proplet to the next.

4.6.1 CONTENT AS A CONSTELLATION

verb: know	noun: Julia	noun: John
cat: decl	cat: nm	cat: nm
arg: Julia John	fnc: know	fnc: know
prn: 1	prn: 1	prn: 1
V	**N**	**N**

4.6.2 NAVIGATING THROUGH A CONSTELLATION



4.6.3 CONDITIONS ON A NAVIGATION

1. A navigation is a *shortest* route to traverse
2. *all* proplets in the constellation such that
3. each successor proplet is *accessible* from the current proplet.

4.6.4 CONTRIBUTIONS OF NAVIGATION TO LANGUAGE PRODUCTION

1. core values
2. parts of speech
3. semantic relations
4. traversal sequence
5. ingredients of perspective

The navigation is powered by an LA-think grammar and serves to selectively activate content in the Word Bank (conceptualization in the speaker mode).

4.6.5 EXAMPLE OF LA-THINK RULE APPLICATION

i *rule name* ii *rule package*
VNs {NVs}

iii *ss pattern* iv *nw pattern* v *operations*

rule level
$$\begin{bmatrix}\text{verb: } \beta \\ \text{arg: !X } \alpha \text{ Y} \\ \text{prn: k}\end{bmatrix} \quad \begin{bmatrix}\text{noun: } \alpha \\ \text{fnc: } \beta \\ \text{prn: k}\end{bmatrix} \quad \text{output position nw}$$

matching and binding variables *output*

Word Bank level
$$\begin{bmatrix}\text{verb: know} \\ \text{cat: decl} \\ \text{arg: Julia John} \\ \text{prn: 1}\end{bmatrix} \Rightarrow \begin{bmatrix}\text{noun: Julia} \\ \text{cat: nm} \\ \text{fnc: know} \\ \text{prn: 1}\end{bmatrix}$$

4.7 Database Semantics: Speaker Mode

4.7.1 CONTRIBUTIONS OF GRAMMAR

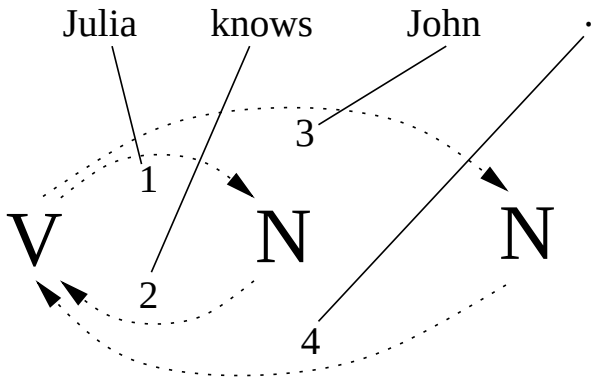
1. language-dependent word order (defined on top of the traversal sequence)

2. language-dependent function word precipitation (utilizing proplet features)

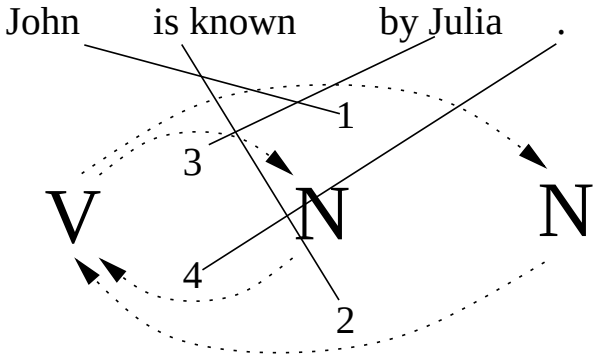
3. selection of word forms (based on proplet features and rules of agreement)

4. lexical selection (driven by the core values of the proplets traversed)

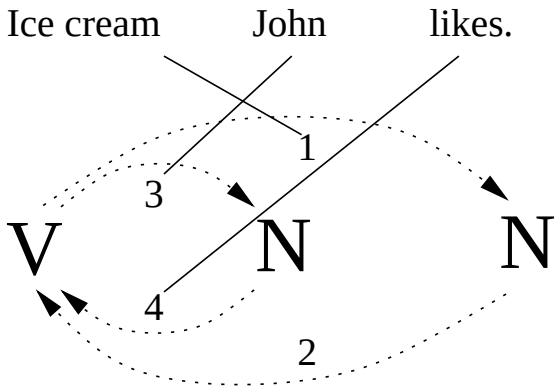
4.7.2 REALIZING A TRANSITIVE SURFACE



4.7.3 REALIZING A PASSIVE SURFACE



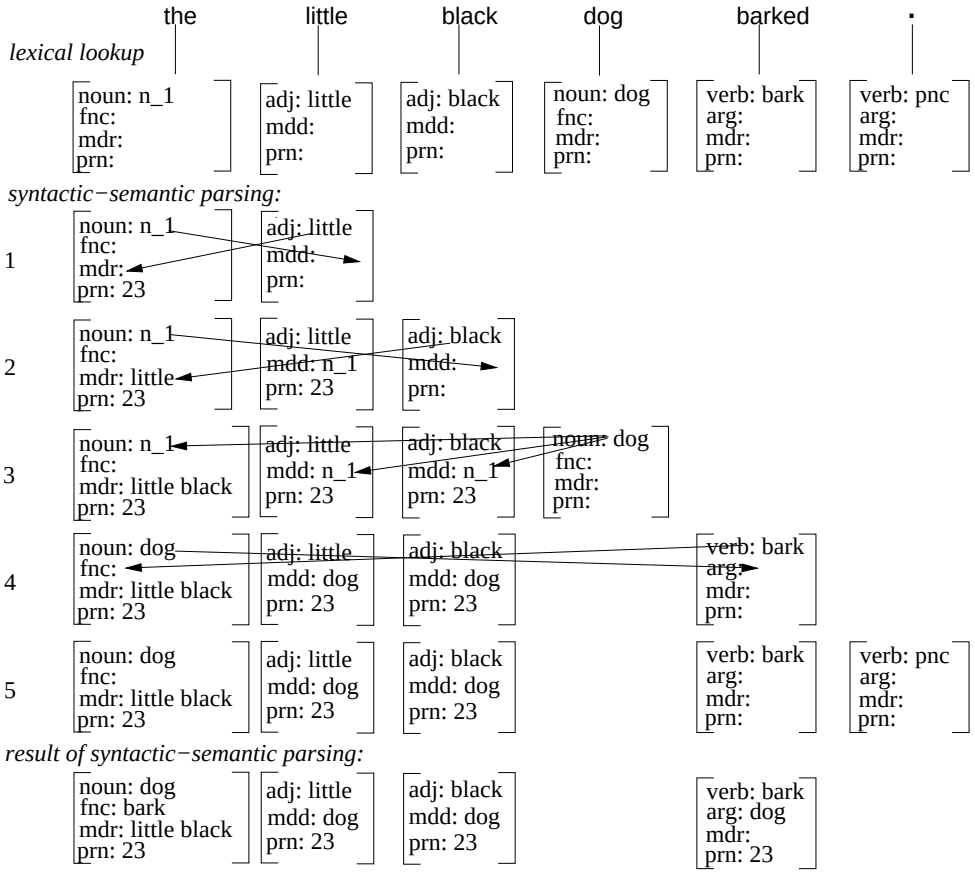
4.7.4 REALIZING A TOPICALIZED OBJECT



5. Day Five: Treating the Phrasal and Clausal Levels

5.1 Establishing Semantic Relations at the Phrasal Level

5.1.1 TIME-LINEAR DERIVATION OF The little black dog barked loudly.



5.1.2 LEXICAL ANALYSIS OF The little black dog barked loudly. (7 PROPLETS, SURFACE ORDER)

sur: the noun: n_1 cat: np sem: def fnc: mdr: idy: prn:	sur: little adj: little cat: adn sem: psv mdd: prn:	sur: black adj: black cat: adn sem: psv mdd: prn:	sur: dog noun: dog cat: def sg sem: count fnc: mdr: prn:	sur: barked verb: bark cat: n′ v arg: mdr: prn:	sur: loudly adj: loud cat: adv sem: psv mdd: prn:	sur: . verb: pnct cat: v′ decl prn:

5.1.3 CONTENT OF The little black dog barked loudly. (5 PROPLETS, ALPHABETICAL ORDER)

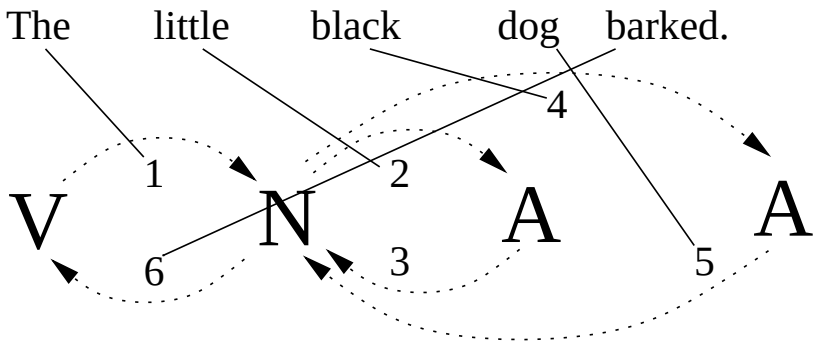
sur: verb: bark cat: decl sem: past arg: dog mdr: loud prn: 23	sur: adj: black cat: adn sem: psv mdd: dog prn: 23	sur: noun: dog cat: def sg sem: count fnc: bark mdr: little black prn: 23	sur: adj: little cat: adn sem: psv mdd: dog prn: 23	sur: adj: loud cat: adv sem: psv mdd: bark prn: 23

5.1.4 Requirements for realizing language-dependent surface in the speaker mode

- language-dependent word order
- language-dependent lexicalization of content words
- language-dependent precipitation of function words
- satisfaction of language-dependent agreement conditions

5.2 Phrasal Level Production in the Speaker Mode

5.2.1 ADNOMINAL MODIFICATION (SPEAKER M.)



1 VNs 2,3 NAr 4,5 NAr 6 NVs

5.2.2 LEXICALIZATION FUNCTIONS

1. lex-d

If one of the following patterns matches an N proplet, then lex-d applied to this proplet produces the associated surface:

pattern	*surface*	*pattern*	*surface*
$\begin{bmatrix}\text{noun:}\alpha\\ \text{sem: indef sg}\end{bmatrix}$	a(n)	$\begin{bmatrix}\text{noun:}\alpha\\ \text{cat: snp}\\ \text{sem: pl exh}\end{bmatrix}$	every
$\begin{bmatrix}\text{noun:}\alpha\\ \text{sem: sel}\end{bmatrix}$	some	$\begin{bmatrix}\text{noun:}\alpha\\ \text{cat: pnp}\\ \text{sem: pl exh}\end{bmatrix}$	all
$\begin{bmatrix}\text{noun:}\alpha\\ \text{sem: def X}\end{bmatrix}$	the		

2. lex-nn

If $\begin{bmatrix}\text{noun: } \alpha \\ \text{cat: snp}\end{bmatrix}$ matches an N proplet, then lex-nn[noun: α] = α.

If $\begin{bmatrix}\text{noun: } \alpha \\ \text{cat: pnp}\end{bmatrix}$ matches an N proplet, then lex-nn [noun: α] = α+s.

3. lex-n

If one of the following patterns matches an N proplet, then lex-n applied to this proplet produces the associated surface:

pattern	*surface*	*pattern*	*surface*
noun: α cat: snp sem: indef sg	a(n) α	noun: α cat: snp sem: pl exh	every α
noun: α cat: snp sem: sel	some α	noun: α cat: pnp sem: sel	some α+s
noun:α cat: pnp sem: pl exh	all α+s	noun:α cat: snp sem: def X	the α
noun:α cat: pnp sem: def X	the α+s		

5.2.3 FORMAL DEFINITION OF LA-SPEAK.E2

ST$_S$ =$_{def}$ {([verb: α] {1 VNs})}

VNs {2 NVs, 3 NAr}

$$\begin{bmatrix} \text{verb: } \beta \\ \text{arg: !X } \alpha \text{ Y} \\ \text{prn: i} \end{bmatrix} \quad \begin{bmatrix} \text{noun: } \alpha \\ \text{mdr: Z} \\ \text{fnc: } \beta \\ \text{prn: i} \end{bmatrix}$$

if adn $\notin$ Z, lex-n [noun: α]
if adn ϵ Z, lex-d [noun: α]
(where adn is an elementary adnominal)

NVs {4 VNs, 5 VVs}

$$\begin{bmatrix} \text{noun: } \alpha \\ \text{fnc: } \beta \\ \text{prn: i} \end{bmatrix} \quad \begin{bmatrix} \text{verb: } \beta \\ \text{arg: X! } \alpha \text{ Y} \\ \text{cat: VT} \\ \text{prn: i} \end{bmatrix}$$

mark α in β-arg
if X = NIL, lex-fv [verb: β]
if Y = NIL, lex-p [verb: β]

NAr {6 NAr, 7 NVs}

$$\begin{bmatrix}\text{noun:}\alpha \\ \text{mdr: !X } \beta \text{ Y} \\ \text{prn: i}\end{bmatrix} \quad \begin{bmatrix}\text{adj: } \beta \\ \text{mdd: } \alpha \\ \text{prn: i}\end{bmatrix}$$

mark β in α-mdr
lex-a [adj: β]
if adn $\notin$ Y, lex-nn [noun: α]

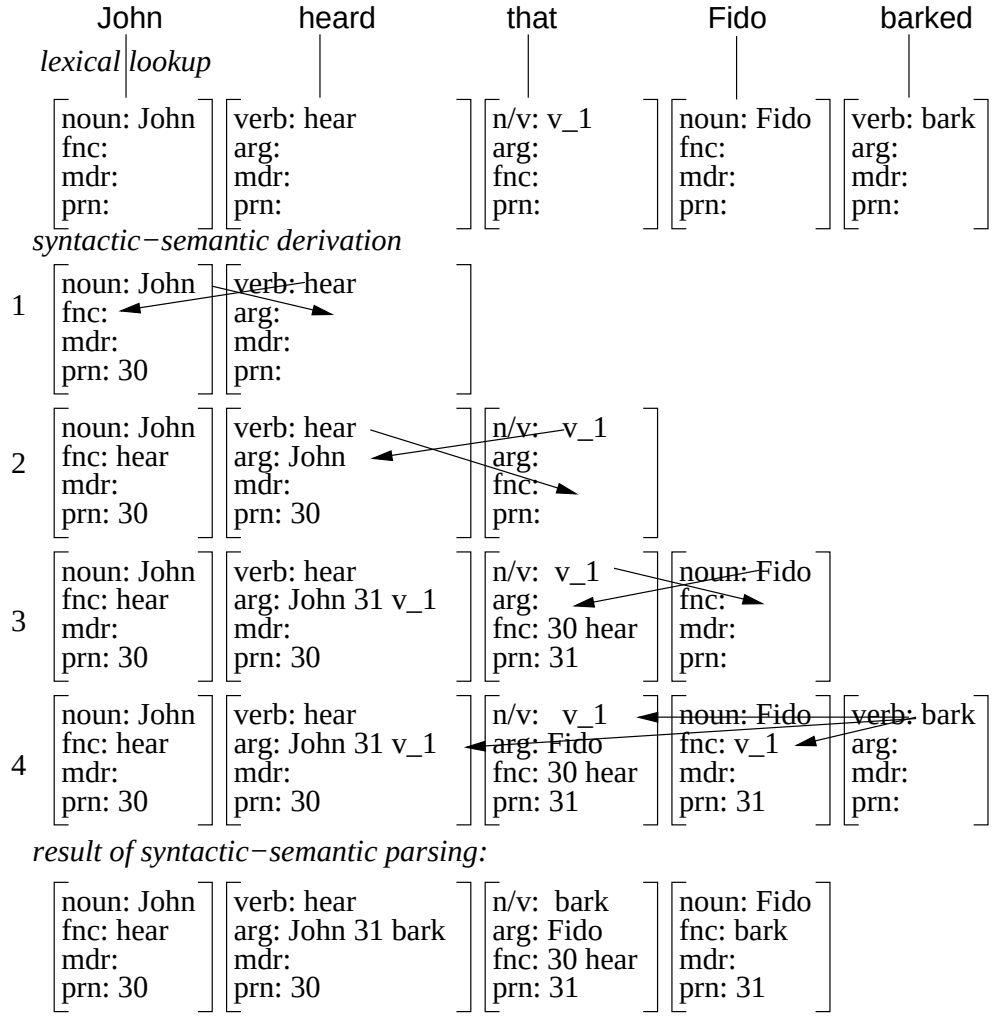
VVs {8 VNs}

$$\begin{bmatrix}\text{verb: } \alpha \\ \text{nc: j } \beta \\ \text{prn: i}\end{bmatrix} \quad \begin{bmatrix}\text{verb: } \beta \\ \text{pc: i } \alpha \\ \text{prn: j}\end{bmatrix}$$

$\mathbf{ST}_F =_{def} \{ (\begin{bmatrix}\text{verb: } \beta \\ \text{arg: X!}\end{bmatrix} \text{rp}_{\text{NVs}})\}$

5.3 Interpretation at the Clausal Level

5.3.1 OBJECT CLAUSE (HEARER MODE)

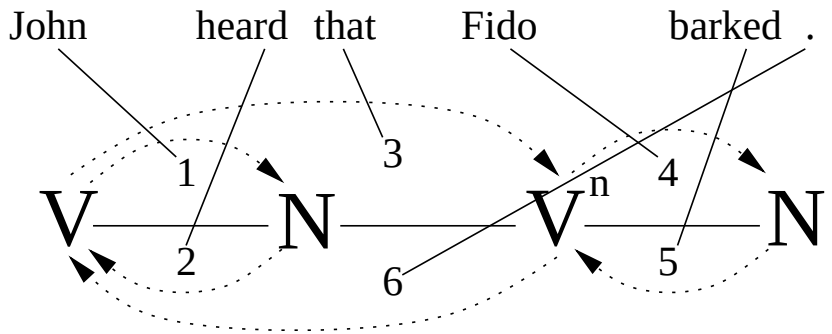


5.3.2 CONTENT OF John heard that Fido barked.

verb: hear	noun: John	n/v: bark	noun: Fido
arg: John 31 bark	fnc: hear	arg: Fido	fnc: bark
mdr:	mdr:	fnc: 30 hear	mdr:
prn:30	prn: 30	prn: 31	prn: 31

5.4 Production at the Clausal Level

5.4.1 OBJECT CLAUSE (SPEAKER MODE)



1 VNs 2 NVs 3 VV^{n}s 4 V^{n}Ns 5 NV^{n}s 6 V^{n}Vs

5.5 Conclusion and References

The time-linear approach of Database Semantics offers an alternative to the hierarchical approaches. This alternative is especially suitable for modeling the cycle of natural language communication, consisting of the hearer mode, the think mode, and the speaker mode.

Hausser, R. (2001) "Database Semantics for natural language," *Artificial Intelligence*, 130.1:27–74 (AIJ'01)

Database Semantics uses the formal algorithm of time-linear LA-grammar, which has been shown to define the first, and so far the only, complexity hierarchy which is orthogonal to the Chomsky hierarchy of regular, context-free, context-sensitive, and r.e. language classes.

Hausser, R. (1992) "Complexity in Left-Associative Grammar," *Theoretical Computer Science*, 106.2:283-308 (TCS'92)

Database Semantics uses the data structure of flat (non-recursive) feature structures, which have been shown here to be suitable for easy coding of lexical analysis and of semantic relations, for a computationally straightforward procedure of matching, and for storage and retrieval in a database.

Database Semantics uses the database schema of a Word Bank, which goes beyond a classic database in that semantic relations are defined between proplets, serving as a railroad system for an autonomous time-linear navigation selectively activating content.

A computationally straightforward and efficient model of how natural language works, even in its first outline, is an essential precondition for achieving human-machine communication, and thus a precondition for a wide range of practical applications, such as the building of talking robots and natural language interaction with the Internet.

For a text-book introduction to computational linguistics taking the general viewpoint of Database Semantics see

Hausser, R. (1999/2001) *Foundations of Computational Linguistics, Human–Computer Communication in Natural Language, 2nd ed.*, pp. 578, Berlin Heidelberg New York: Springer (FoCL'99)

For an application of Database Semantics to the systematic analysis of functor-argument and coordination structures at the elementary, phrasal, and clausal level in English see

Hausser, R. (2006) *A Computational Model of Natural Language Communication*, Berlin

Heidelberg New York: Springer (NLC'06)

A recent paper summarizing the cycle of natural language communication at the elementary, phrasal, and clausal level is

Hausser, R. (2009a) "Modeling Natural Language Communication in Database Semantics," in M. Kirchberg and S. Link (eds.), Proceedings of the APCCM 2009, Australian Computer Science Inc., CIPRIT, Vol. 96

A comparison of the traditional parts of speech and their treatment as core attributes in DBS is presented in

Hausser, R. (2009b) "From Word Form Surfaces to Communication," in H. Kangassalo, Y. Kiyoki, and T. Welzer (eds.), *Information Modelling and Knowledge Bases XXI*. Amsterdam: IOS Press Ohmsha

All papers and the slides of the books are available at

http://www.linguistik.uni-erlangen.de/clue/de/publikationen.html