Knowledge Based Machine Translation

Cristina Vertan

University of Hamburg • Informatics Department

Natural Language Systems Group

WWW: http://nats-www.informatik.uni-hamburg.de/~cri/

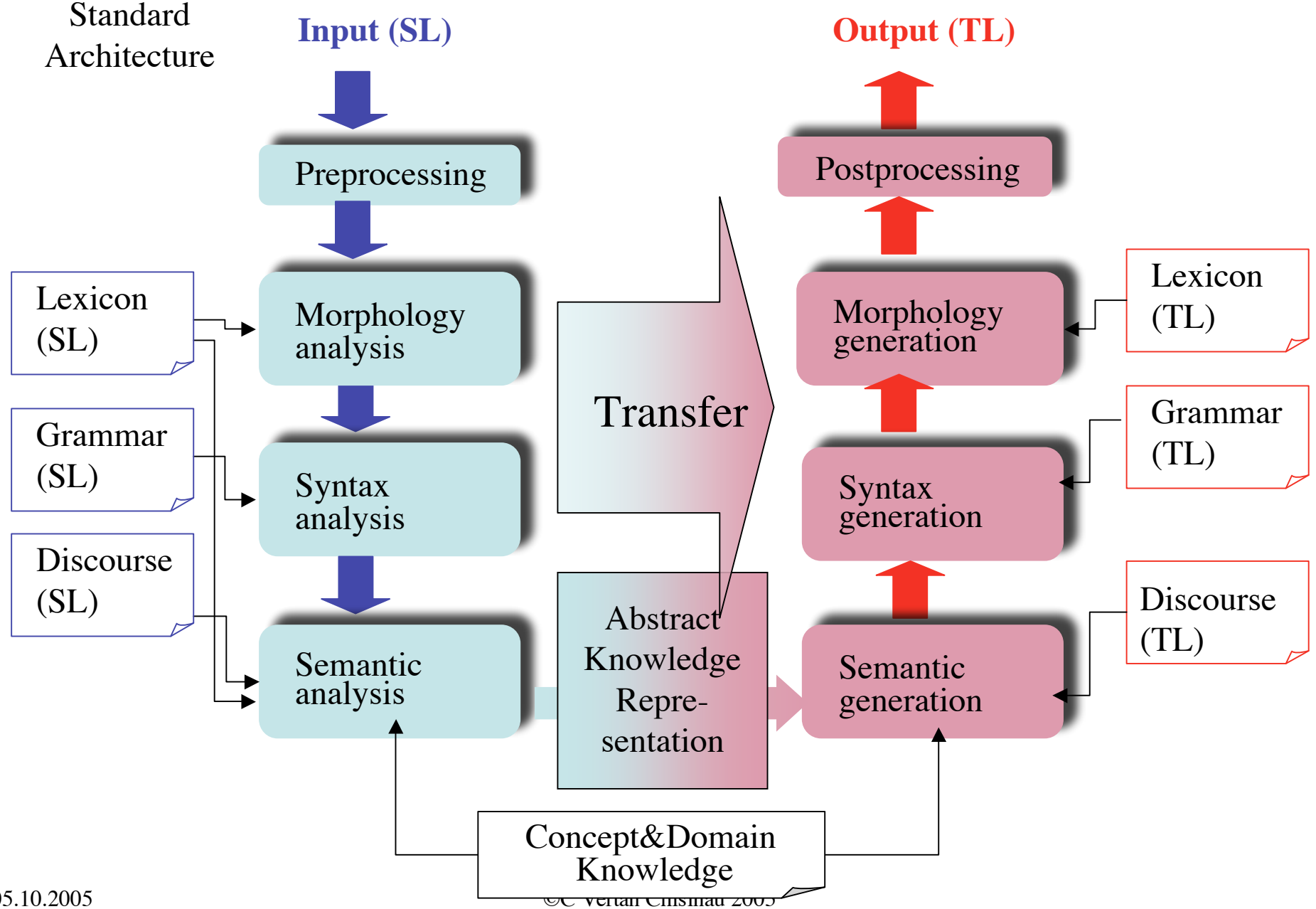
E-Mail: vertan@informatik.uni-hamburg.de

Most part of the slides from this section I owe with explicit permission from Walther v. Hahn

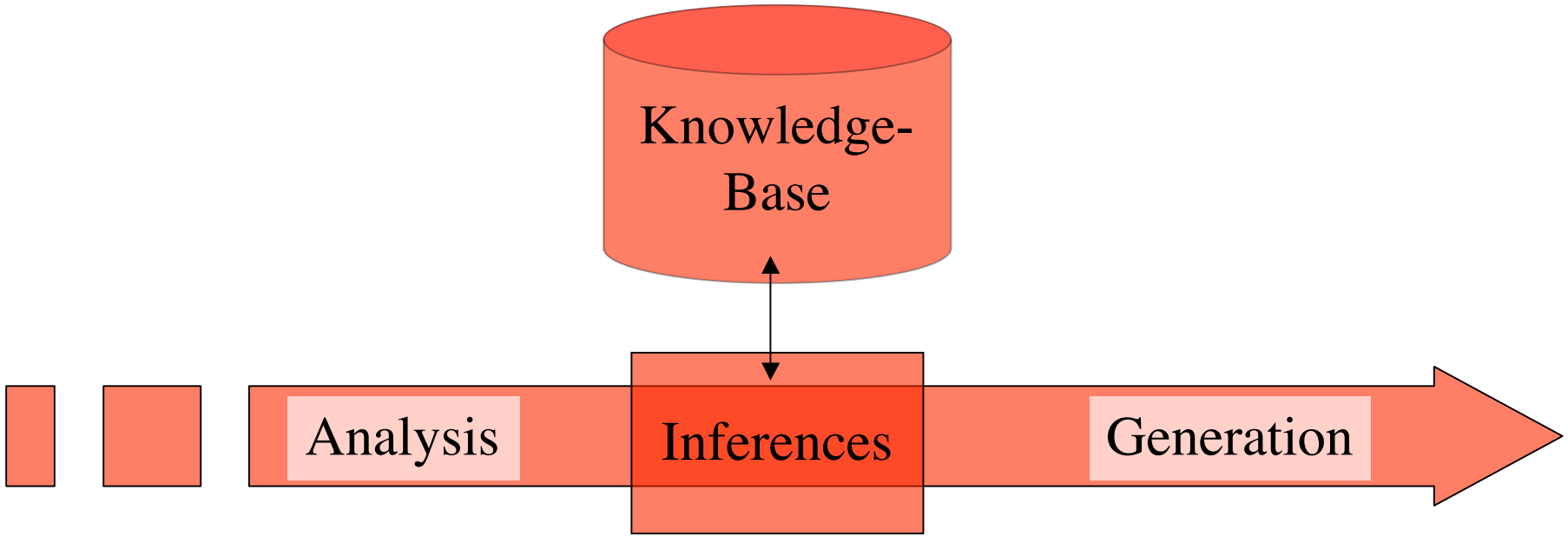
Outline

- What is knowledge?
- Why knowledge in Machine Translation?
- Linguistic examples
- 3 Examples of systems
  - KBMT 1989
  - DBR-MAT 1999
  - Verbmobil 2001
- The Semantic Web Idea

Standard Architecture



Derivation of Solutions from General Knowledge



Knowledge Based Processing

```
?Grammatical function of "dog"
     "dog" sub-concept of animal,
     "animal" sub-concept of
living
     "living" = agent of "bite"
!~"dog" = grammatical subject
```

Terminological Knowledge = Knowledge-based System?

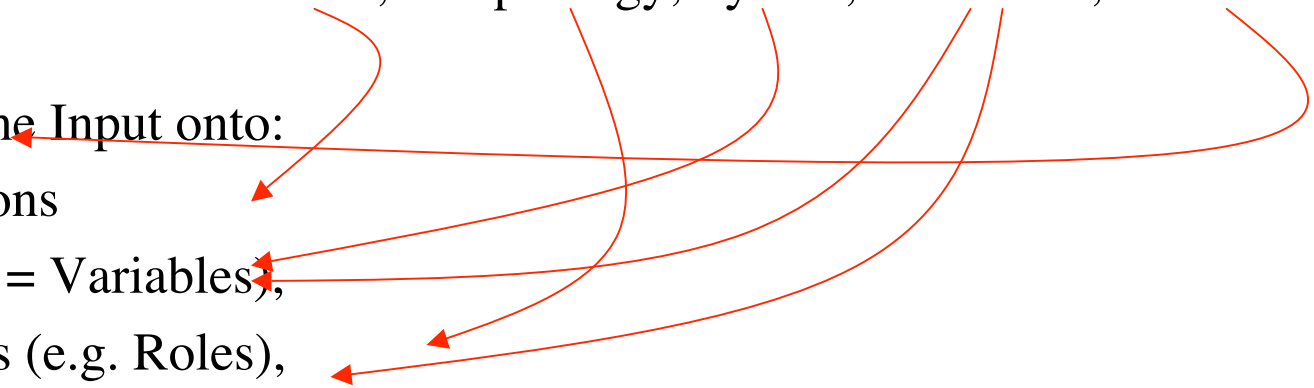
- In a sense, systems that use terminological material (in a systematic order according to the domain), can be called knowledge based systems.
- However, the ontological knowledge (conceptual ordering) of the field is not declarative, but implicit in the ordering of the terminology (nomenclature). The ontology is not visible directly.

"Real" Knowledge in MT

Structural Transfer with Interlingua

Tasks:

- Content analysis from the Lexicon, Morphology, Syntax, Semantics, Pragmatics
- = Mapping of the Input onto:
  - Presuppositions
  - Objects (e.g. = Variables),
  - Relationships (e.g. Roles),
  - Quantifiers (e.g.= Negation, Number)
- Consistency check (e.g. Presupposition check)
- Semantic extraction
- Reordering of results in the generation phase



Interlingua Expression

```
[    request (referent (_5747))
     presuppose (exists (_4340))  ]
     some (_4340)
[    unique (_4407)
     single (_4407)
     instance (_4407, person)
     propval (Person, _4407, sex, male)
     [    some (_4725)
          [    unique (_5033)
               single (_5033)
               instance (_5033, project)
               propval (project, _5033, name, str
                                          (LOKI))  ]
          instance (_4725, leading)
          propval(leading, _4725, theta, _5033)
          propval(leading, _4725, alpha, _4407)
          topic(_4407)      ]
]
```

Interlingua systems construct a deep meaning representation by using additional knowledge of some sort, at least about what sentences are about.

Presupposition part of an Interlingua expression

Knowledge Layers

Knowledge used in an MT system may be

- Conceptual knowledge ("ontology", "upper model")
- World knowledge (chemical laws, e.g.)
- Factual knowledge (situational knowledge) about the actual state of affairs

- Some examples:

Conceptual Knowledge

$\forall b$ Bear (b) $\Rightarrow$ Animal (b) *"Bears are animals"*

$\forall b$ Bear (b) $\Leftrightarrow$ SpeciesOf (b) = Ursidae

"Bears belong to the species of Ursidae"

$\forall x,y$ part of (x,y) & PhysicalThing (y) $\Rightarrow$ PhysicalThing (x) *"Parts of physical objects are physical objects"*

World Knowledge

$\forall$ x PhysThing (x) $\Rightarrow \exists$s size(x) = s

"All physical things have a size"

Bear (Pooh) *"Pooh is a bear"*

T (Area (Poland, SqMiles (233000)), AD1426)

"In 1426 Poland had the size of 239.000 square meters"

Situation Knowledge

In (Pooh, LivingRoom3)

"Pooh is in the living room"

Female (Speaker23)

"The current speaker is female"

(most examples from Russell/Norvig)

Do We Need Knowledge in MT at all?

- Whatever a MT system does, it needs at least some of the before mentioned knowledge layers. If the functionality of the system is not very demanding, some parts of the conceptual knowledge can be described in the lexicon, e.g. under semantic features, sub-categorization, roles or constraints.

- But any system will run into difficulties …

Disambiguation

source: situation knowledge

"I will go to my office in an hour"

→ "Ich gehe zu meinem Büro in eine Stunde"
→ "Ich fahre zu meinem Büro in eine Stunde"
→ " Ich fliege zu meinem Büro in eine Stunde"

(engl.→ germ.)

dependent on *how far away* my office is.

Similarly, in Romanian a merge → either walk or drive

Anaphora antecedents
source: concepts/facts

"She took the ice cream from the fridge and ate it"

→ "Ella ha tomando la tarta del refrigerador y la ha comido" (not "el")

(engl. → esp.)

Pronouns

source: facts

"They go"

→"Ei merg" / "Ele merg" / "Ei si ele merg"

(engl. → rom.)

Dependent on whether the set of persons ("they") contains

men or woman or both

Lexical Specification

source: concept hierarchy/facts

"cousin" → "cousin" or "cousine"

(engl. → fr.)

"Uncle" → "farbror" or "morbror"

(germ. → dan.)

3 Examples of the use of knowledge in MT

- KBMT 1989
- DBR-MAT 1999
- Verbmobil 2001

Example I: The **KBMT** project

Assumptions behind KBMT:

One "functionally complete" meaning representation can serve for translations to a number of languages,

no total representation of human understanding of a text is necessary for machine translation,

applicable to relatively unambiguous, e.g. technical documents.

Basic Components of a KBMT System

An ontology of concepts ("domain model", "ontology")

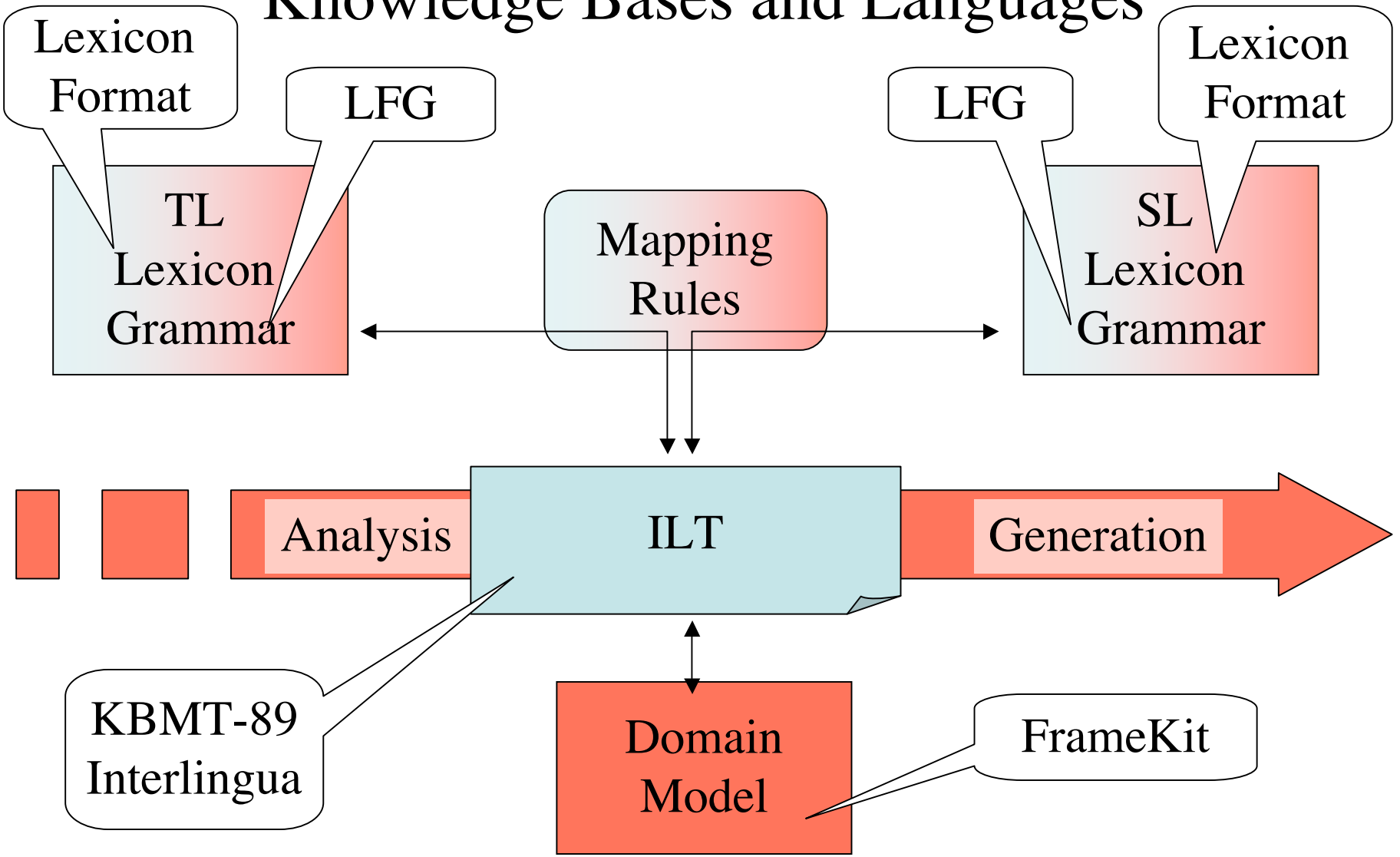
Source language (SL) lexicon and grammar for the analysis process

Target language (TL) lexicon and grammar for the generation processes

Mapping rules between the Interlingua and SL/TL syntax.

Technical aim of KBMT-89: Bidirectional translations for single sentences of English and Japanese in the domain of PC manuals.

Knowledge Bases and Languages



FrameKit Representation of *"Save the document"*

$$\begin{bmatrix} instance\_of: & save \\ isa: & physical\_event \\ id: & save\_43 \\ agent: & user \\ patient: & \begin{bmatrix} instance\_of & document \\ isa & separable\_entity \\ id: & document\_72 \\ reference: & definite \end{bmatrix} \end{bmatrix}$$

KB: Frames with linguistic and non-linguistic knowledge

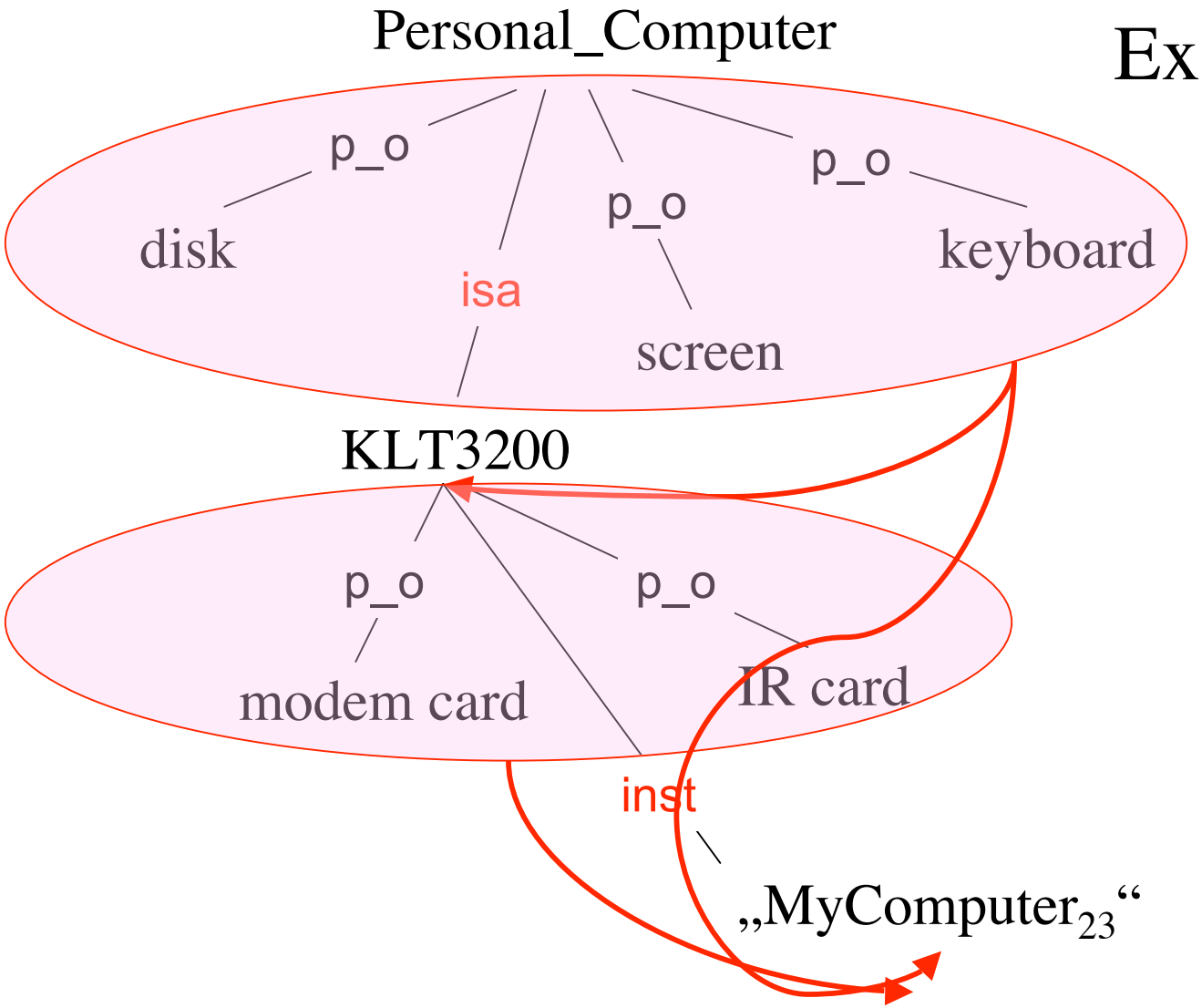
The KBMT-89 ontology contains

```
Objects
  Events
  Properties of objects or events
  Relations
  Attributes
```

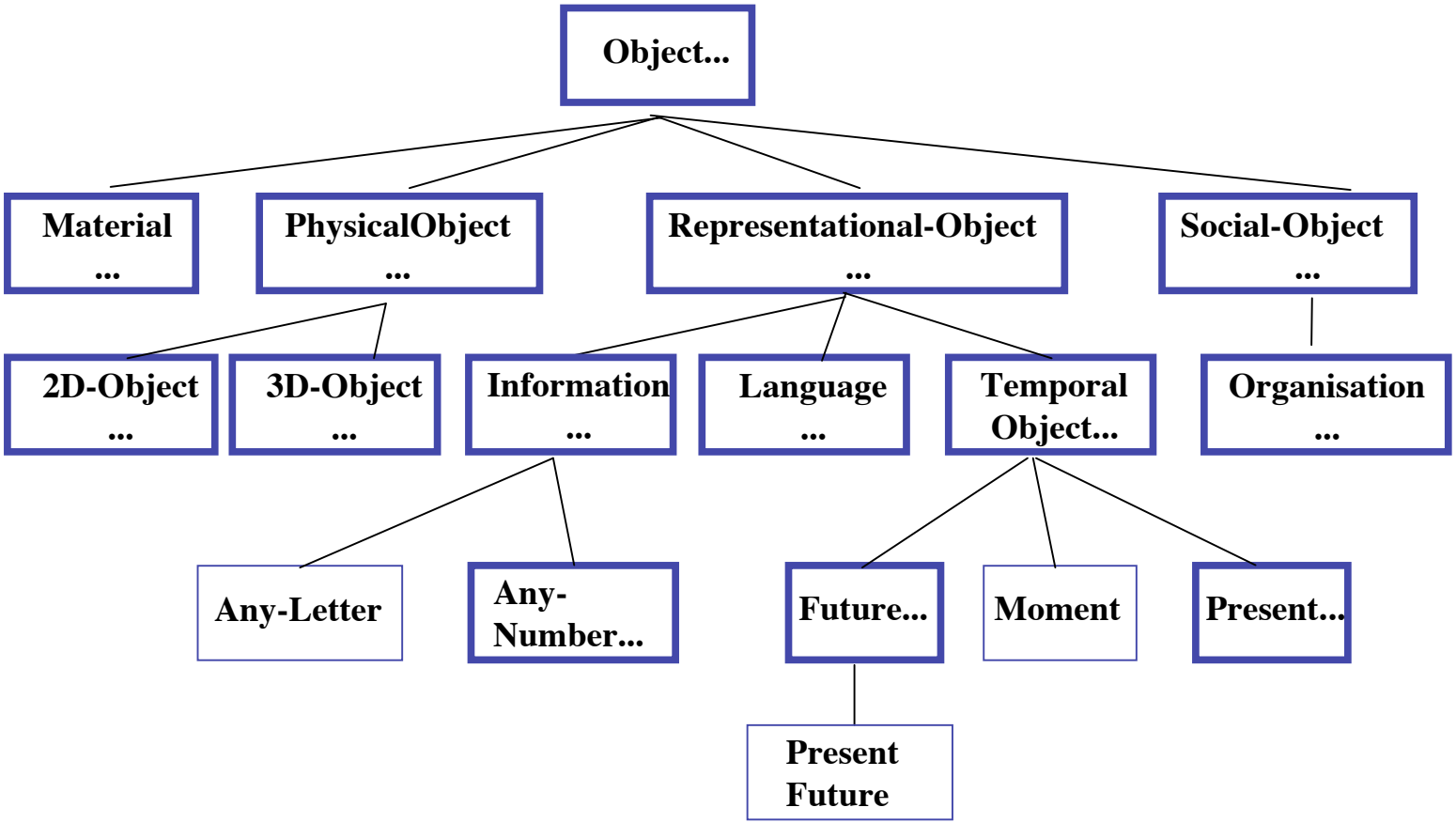
Concepts are linked to others by relations. Each concept has attributes which specify value sets. Value sets contain only literals (i.e. no concepts).

Frames support inheritance mechanisms:

Inheritance Example



Fragment of the ontology of KBMT-89



(Concepts in bold boxes = further sub-concepts are omitted for clarity)

ILT for "Get the diagnostics diskette from the back of this manual"

```
·  [*RECEIVE
     (AGENT *READER)
     (THEME [*DISKETTE
          (NUMBER SINGULAR)
          (PURPOSE [*DIAGNOSE
               (NUMBER SINGULAR)])
               (REFERENCE DEFINITE)])
           (SOURCE [*BACK-OF-3D
               (NUMBER SINGULAR)
               (REFERENCE DEFINITE)
               (PART-OF [*MANUAL
                    (NUMBER SINGULAR)
                    (REFERENCE DEFINITE)
                    (NEAR +)])])
           (TENSE PRESENT)
                    (MOOD IMPERATIVE)
                      (CLAUSAL-MARK +)
                      (NUMBER-BULLET [*ANY-NUMBER
                         (CARDINALITY 1)])]
```

Example II: The **Verbmobil** System

Transfer system for English, Japanese and German

Spontaneous speech input, on-line processing near real-time.

Sentence representation: VIT (Verbmobil Interface Term) for necessary linguistic and non-linguistic information.

no representation of semantic contents or pragmatic sense in VITs

VITs represent **dialogue acts** (--> syntactic disambiguation, repair of gaps in speech hypotheses, and

prosody as indicator for structural boundaries, particle interpretation and sentence mood.

```
·  Vit( vitID(sid(104,a,en,10,800,1,en,y,semantics),        %SegmentI
       [word(he,1,[l126]),                                  %WHG string
          word(is,2,[]),
                  word(coming,3,[l127]),
                  word(at,4,[l136]),
                  word(the,5,[l128]),
                  word(beginning,6,[l135]),
                  word(of,7,[l135]),
                  word("August",8,[l134])]),
       index(l138,l125,i35),                                %Index)
       [beginning(l135,i37),                                %Conditions
        arg3(l135,i37,i38),
        come(l127,i35),
        arg1(l127,i35,i36),
        decl(l137,h43),
        pron(l126,i36),
        at(l136,i35,i37),
        mofy(l134,i38,aug),
        def(l128,i37,h42,h41),
        udef(l131,i38,h45,h44)],
       [in_g(l126,l125), in_g(l137,l138),                   %Constraints
        in_g(l127,l125), in_g(l128,l130),
        in_g(l131,l133), in_g(l134,l132),
        in_g(l135,l129), in_g(l136,l125),
        leq(l125,h41), leq(l125,h43),
        leq(l129,h42), leq(l129,h44),
        leq(l130,h43), leq(l132,h45),
        leq(l133,h43)],
       [s_sort(i35,situation),                              %Sorts
        s_sort(i37,time),
        s_sort(i38,time),
       [dialog_act(l125,inform),                            %Discourse
        dir(l136,no),
        prontypel136,third,std()],
       [cas(i36,nom),                                       %Syntax
        gend(i36,masc),
        num(i36,sg), num(i37,sg), num(i38,sg),
        pcase(l135,i39,of)],
       [ta_aspect(i35,progr),                               %Tense and Aspect
        ta_mood(i35,ind),
        ta_perf(i35,nonperf),
        ta_tense(i35,pres),]
       [pros_accent(l135,progr)],                           %Prosody
·  )
```

"He is coming at the beginning of August"

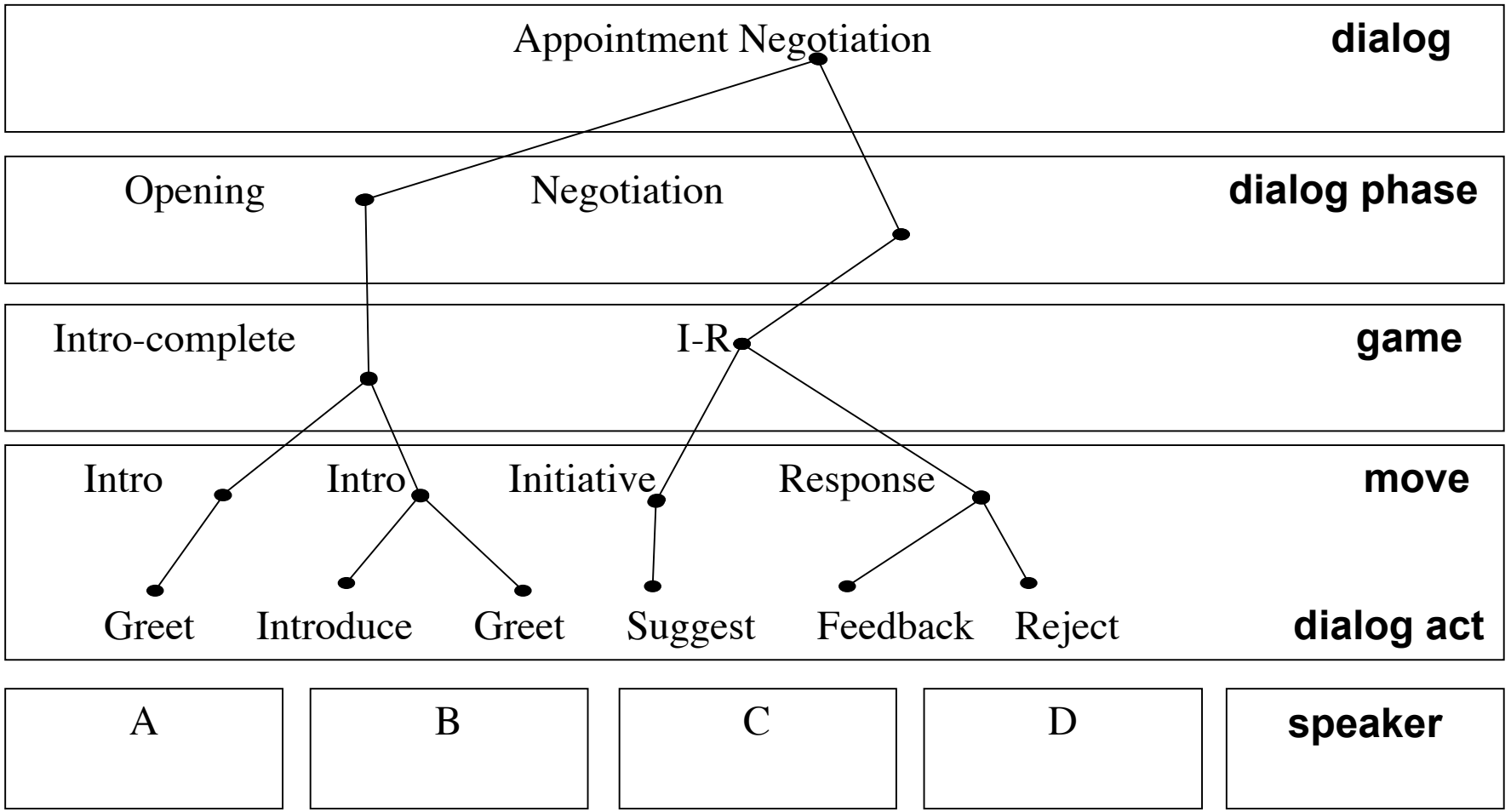
Multilingual Semantic Data Base

Base form of words,
Lexical semantic decomposition,
Interlingual representation, if available,
Semantic class (nominals, quantifiers, verbs, modifiers, etc.),
Syntactic valency with mappings onto grammatical functions and thematic roles (linking),
Ontological sorts, e.g.:

- abstract
  - property, field, info-content, institution, symbol
- space-time
  - temporal
    - situation
      - meeting_sit, communicate_sit, action_sit, …)
    - time
  - entity
    - object
      - agentive, thing
    - location

Selectional restrictions on arguments.

Dialog Act Structure



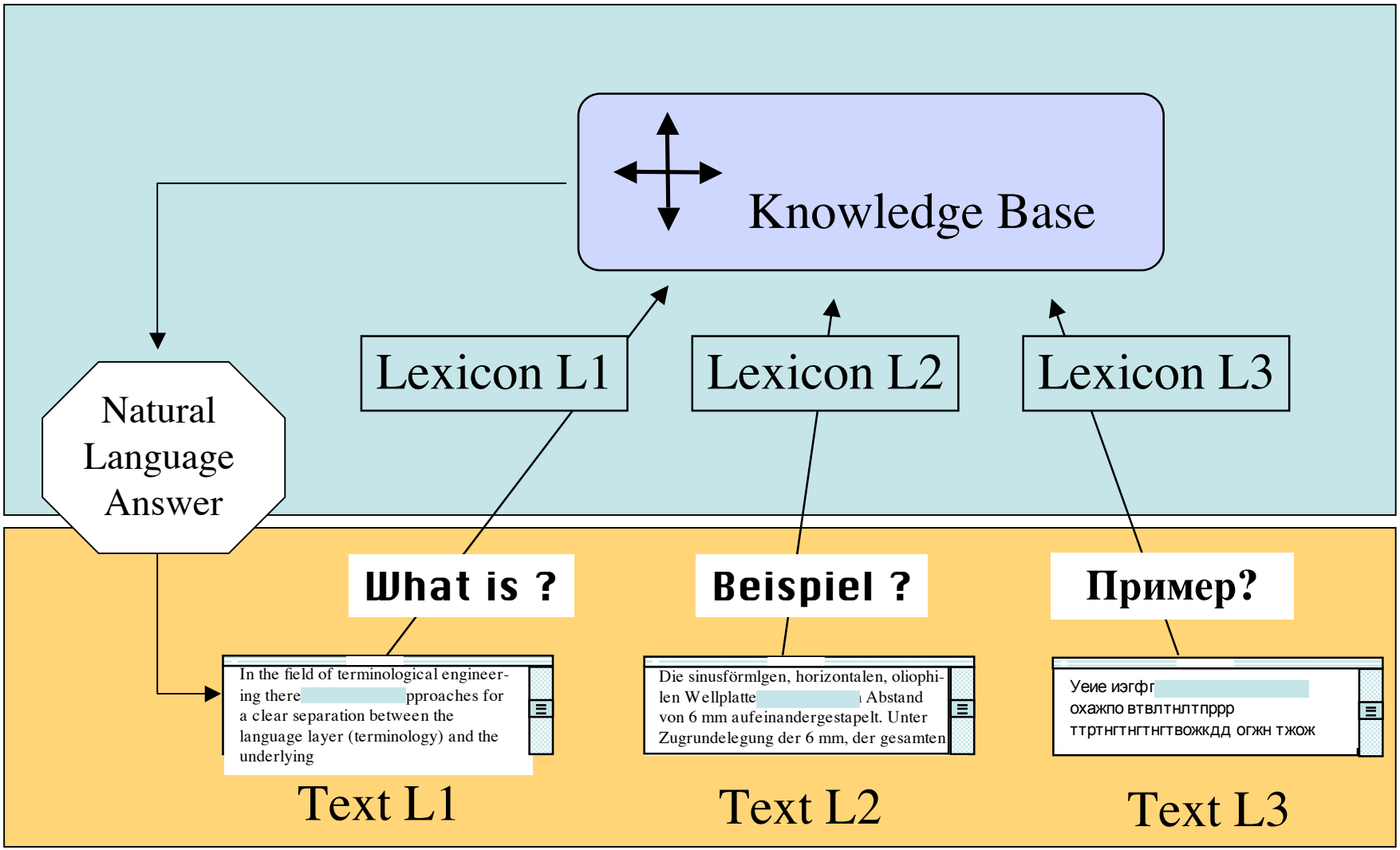
Example III: The DBR-MAT System

Functionality:

Technical translators are supported.

- They can ask conceptual questions concerning,
  - intensional definitions,
  - extensional definitions,
  - characteristics and attributes,
  - hyponyms or meronyms
  - Examples
  - etc.
- ask for pictures,
- recursively ask about the explanations.

The DBR-MAT Paradigm



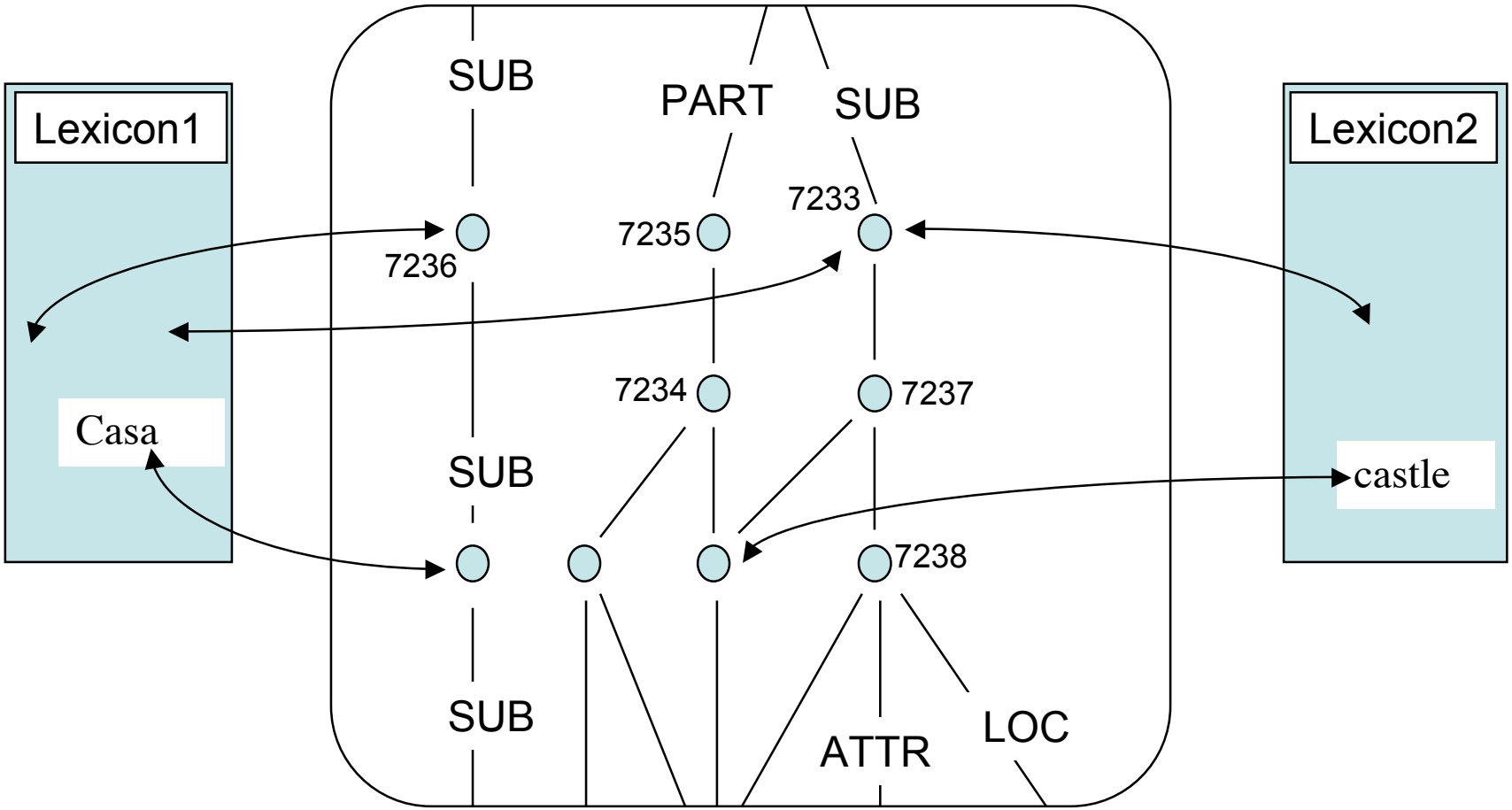
Conceptual Graphs

Examples of complex conceptual relations ("situations")

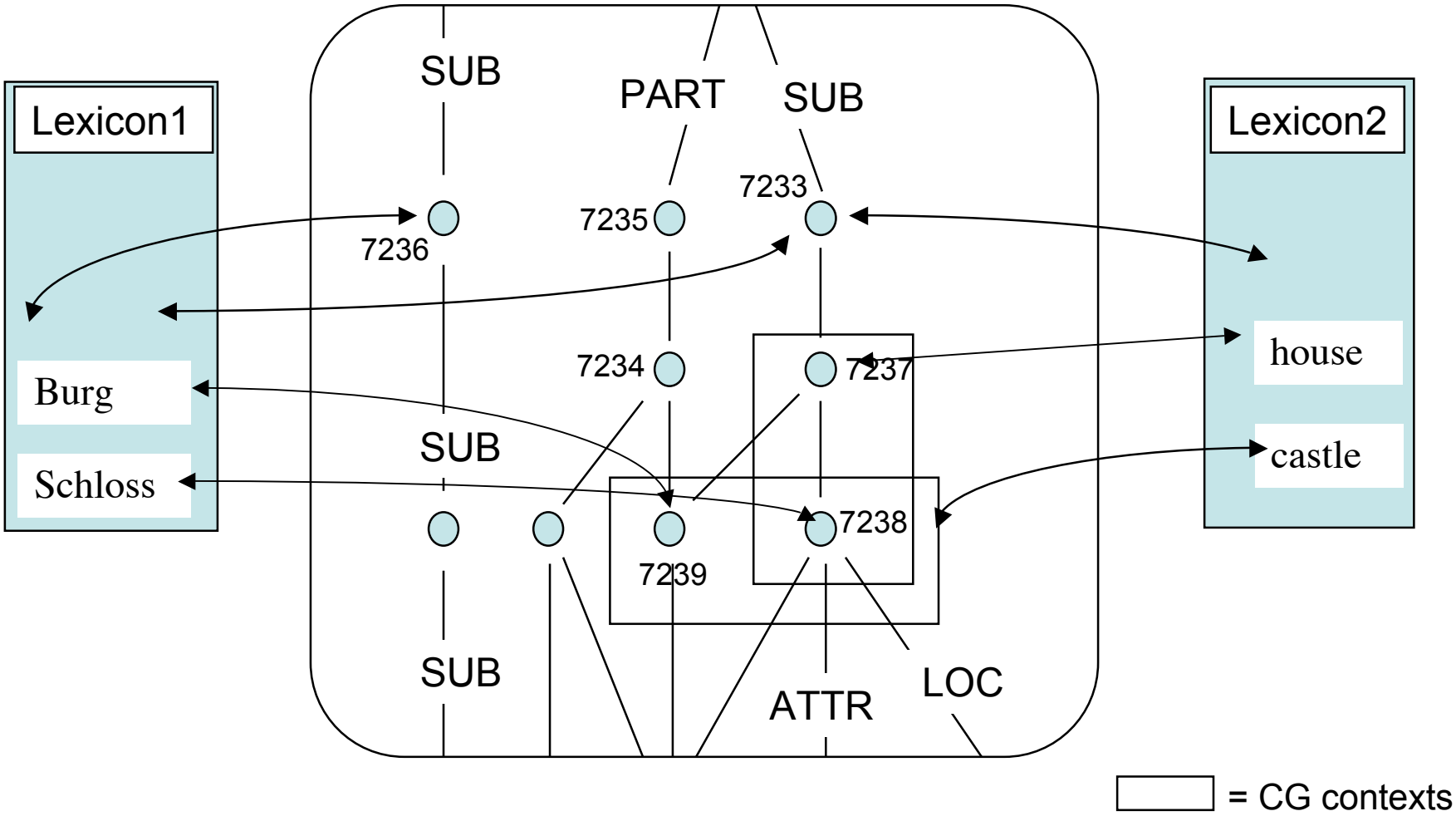
```
[SITUATION: [OIL FRAGMENT: {*}] →  (IN) →
[WATER: {*}]
          →  (CHAR) →  [PHYSICAL STATE:
             disj{MEMBRANE, DROPS, COLLOID,
                  EMULSION, SOLUTION}].
```

```
[SITUATION:
          [WASTE WATER: {*}] → (CONTAIN) →
          [OIL FRAGMENT: {*}]
          → (ATTR) → [FLOATING]
               → (ATTR) → [ROUGHLY
DISPERSED]
          → (PTNT) → [PRECIPITATION]].
```

Connecting words and concepts



Multiple inheritance for differences in conceptualizing



Knowledge Base

The knowledge base of DBR-MAT contains the following objects:

KB Objects	Example
Concepts	[OIL SEPARATOR]
Individuals	the separator C334
Conceptual relations	(part_of)
Contexts (situations)	[SITUATION: [WASTE WATER: {*}] -> …
A type hierarchy	[SEPARATOR] ⇓ [OIL SEPARATOR]

Traversing rules

Submenu	***Item***	***Evaluated Conceptual Relations***	***Inheritance***
What is?		Types of… - All + ATTR, Char, PART_OF	✓
Types of -	All	Superconcepts + subconcepts + sister concepts	
	General	All superconcepts from the hierarchy	
	Concrete	All subconcepts from the hierarchy	
	Similar	All sister concepts from the hierarchy	
Characteristics -	All	Attributes + Who + Object + How + Where	✓
	Attributes	ATTR + CHAR	✓
	Who	AGNT	
	Object	OBJ + PTNT	
	How	INST	
	Where	LOC + DEST + FROM + IN + TO	
More…		All remaining relations	
Examples		Individual concepts	
Want All		All mentioned above, without duplicates	✓

Critical problems of knowledge-based systems are still

1. The huge effort to build up knowledge bases,
2. A practical definition of the size (“ coverage”) of the knowledge base, and
3. The choice of the representation language and its necessary logical/formal properties.

Advantage of Knowledge Bases

- Using knowledge bases the developer definitely knows, what is represented where, although he cannot predict, what can be derived with the inference rules. With implicit and procedural (local) representations there is no method to check multiple representation.
- Declarative knowledge sources, are global, can be maintained in isolation, can be exchanged and may be used in other inference machines or grammars. It even can be used in other systems than translation systems.

The Semantic Web Idea

- This basic idea of declarative and modularized knowledge has become very important since the famous paper of Berners-Lee in 2001 on the "semantic web".
- The Semantic Web (according to the definition of the W3C consortium) is
- *"the abstract representation of data on the World Wide Web, based on the RDF standards and other standards to be defined. It is being developed by the W3C, in collaboration with a large number of researchers and industrial partners.*

The Semantic Web Idea

- *The semantic web will serve as one (or several) ontology(ies) to which all WWW objects refer and which can be used consequently for web operations like data mining, information extraction, summarization, etc*"
- and translation!
- The first obvious result of the semantic web activities for translation is, that widely accepted ontologies of specific domains can be used as knowledge bases for machine translation. This solves point (1) to (3) of the above mentioned list of problems.