Lexical knowledge representation and natural language processing

James Pustejovsky

Branimir Boguraev

Morphosyntax-Semantics Interface
in Lexicalist Theories SS 2003

aims of the semantic lexicon

- create a lexical model which
  - can do disambiguation and complement selection without using word sense enumeration
  - can account for new word contexts and creative use of language
- overall goal: significant improvement of natural language processing systems in **computational lexical semantics**

semantic lexicon methods and basic formalisms

- lexical formalism makes use of a knowledge representation framework
- disambiguation becomes part of the ***semantic analysis*** procedure
- different word senses result directly from interaction of mutual compatible roles in the lexical entry - ***compositional approach***

lexical ambiguity
word sense enumeration

1) a fast car
2) a fast typist
3) a fast book
4) a fast game

lexical ambiguity
word sense enumeration

1) a fast car
2) a fast typist
3) a fast book
4) a fast game

fast(1): moving quickly
fast(2): performing some act quickly
fast(3): doing something requiring a short space of time
fast(4): involving rapid motion

- additional selection restrictions

lexical ambiguity
word sense enumeration

1) a fast car
2) a fast typist
3) a fast book
4) a fast game
5) a fast motorway

fast(1): moving quickly
fast(2): performing some act quickly
fast(3): doing something requiring a short space of time
fast(4): involving rapid motion

- additional selection restrictions
- enumeration → incompleteness

lexical ambiguity - a semantical approach

- consider the following examples:
  - a fast car
  - a fast typist
  - a fast book
  - a fast motorway
- the lexical sign of the noun contains semantic information which can modified by the adjective
- more powerful mechanism than word sense enumeration

ambiguity and compositionality new lexical entries

- lexical entry encodes a range of representative aspects of lexical meaning
- mutual compatible roles force a special interpretation of a word in a specific context
  - more complex process than matching of features
  - mechanisms for composition on phrasal level
- compositional approach revises the more conventional view which treats verbs as functions and nouns as arguments

lexical signs
four levels of representations

- Argument structure:
  - mapping from a word to a function
  - number and type of arguments
- Event structure:
  - event type for a verb or a phrase
  - state (S), process (P) and transition (T)
  - information about the event and event composition

lexical signs
four levels of representations

- Qualia structure
  - defines essential attributes of objects, events, and relations
  - specifies an argument structure for nouns and nominal phrases
- Lexical inheritance structure
  - defines relations to other words in the lexicon
  - provides a link to general world knowledge

Qualia structure

- system of relations that characterizes the semantics of nominals
- specifies four aspects of word meaning:
  - ***Constitutive role -*** the relation between an object and its constituent parts
  - ***Formal role -*** that which distinguishes it within a larger domain
  - ***Telic role -*** its purpose and function
  - ***Agentive role*** - factors involved in its origin or „bringing it about“

Qualia structure – an example from the lexicon

door (x,y)

CONST = **aperture(y)**

FORMAL = **physobj(x)**

TELIC = **walk_through(P,w,y)**

AGENTIVE = **artifact(x)**

Qualia structure
the role approach

- roles determine a minimal semantic description of a word
- approach enriches the semantics of nominal types to „spread the semantic load“ more evenly through the lexicon
- possibility to create an account for new word senses arising in syntactic composition
- motivated by linguistic phenomena

some examples

- They walked through the **door**.
  She will paint the **door** red.
- Black smoke filled the **fireplace**.
  The **fireplace** is covered with soot.

  → relation between the ***Formal*** and the ***Constitutive qualia***

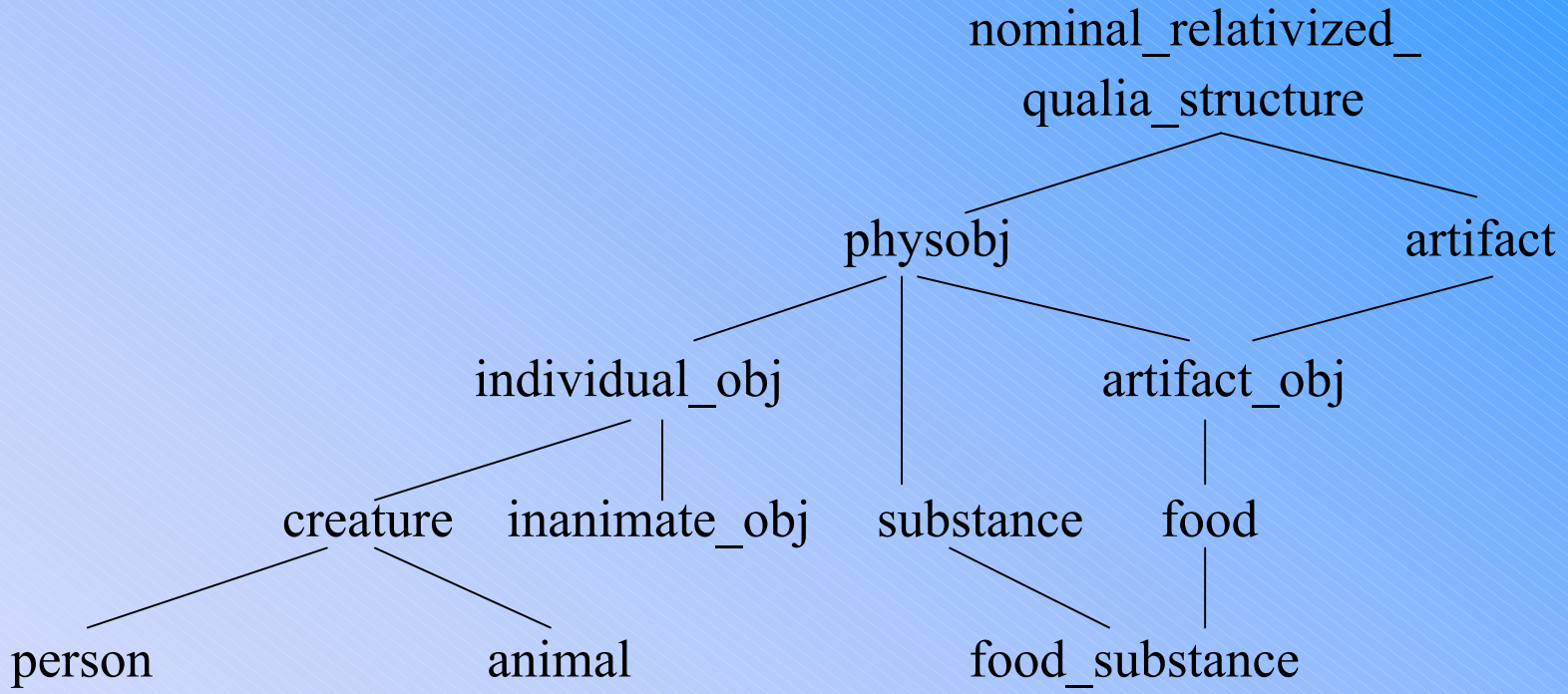
some examples

- They walked through the **door**.
  She will paint the **door** red.

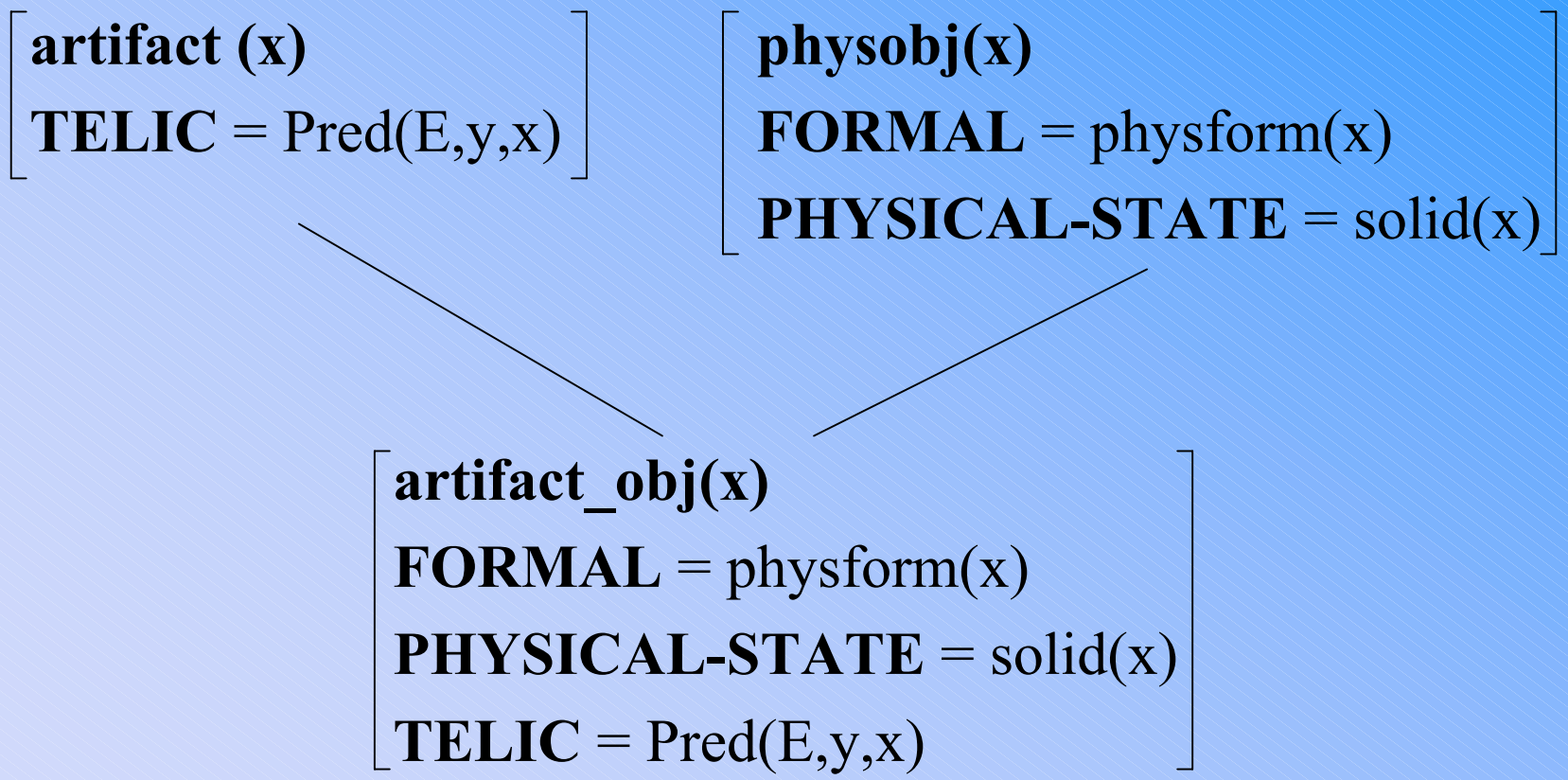
  →relation between the ***Formal*** and the ***Constitutive qualia***

$$\begin{bmatrix} \textbf{door (x,y)} \\ \text{CONST} = \textbf{aperture(y)} \\ \text{FORMAL} = \textbf{physobj(x)} \\ \text{TELIC} = \textbf{walk\_through(P,w,y)} \\ \text{AGENTIVE} = \textbf{artifact(x)} \end{bmatrix}$$

qualia structure
an example type hierachy



inheritance in the type hierarchy



formal aspects of the type hierarchy

- constraints as restrictions on PATR-II-type feature structures
- well formedeness conditions:
  - completeness
  - well typedness
  - uniqueness

Lexical ambiguity and compositionality

- a fast car
- a fast typist
- a fast book
- a fast motorway

„fast“ modifies the
Telic role of the noun
which is of type ***event***

Lexical ambiguity and compositionality

- a fast car
- a fast typist
- a fast book
- a fast motorway

„fast“ modifies the Telic role of the noun which is of type ***event***

$$\begin{bmatrix} \text{car (x)} \\ \text{CONST} = \{\text{body,engine,...}\} \\ \text{FORMAL} = \text{physobj(x)} \\ \text{TELIC} = \text{drive(P,y,x)} \\ \text{AGENTIVE} = \text{artifact(x)} \end{bmatrix}$$

Lexical ambiguity and compositionality

- a fast car
- a fast typist
- a fast book
- a fast motorway

„fast“ modifies the Telic role of the noun which is of type ***event***

$$\begin{bmatrix} \text{car (x)} \\ \text{CONST} = \{\text{body,engine,...}\} \\ \text{FORMAL} = \text{physobj(x)} \\ \text{TELIC} = \text{drive(P,y,x)} \\ \qquad\qquad \& \text{ fast (P)} \\ \text{AGENTIVE} = \text{artifact(x)} \end{bmatrix}$$

Lexical ambiguity

- different meanings of „fast“ can be derived from only one lexical entry
- no word sense enumeration
- new word meaning can be created eg.

  „a fast motorway“:

[TELIC: travel (P,cars) & on (P,x) & fast (P)]

→fast selects a complement of type TELIC

example: advantages of the type hierarachy

sentence „Mary finished her sandwich.“

sandwich(x)
CONST = {bread,...}
FORMAL = physobj(x)
TELIC = eat(P,w,x)
AGENTIVE = artifact(x)

- process is an event which is part of the Telic role value
- each artifact has a Telic role

→ compositional combination is possible

Qualia structure – complement selection

- Mary woke John up.
- The cup of coffee woke John up.
- John‘s drinking the cup of coffee woke him up.

- Mary enjoyed the movie.
- Mary enjoyed watching the movie.

- syntactic differences, but complements are all events of some sort
- word sense enumeration is unintuitive

→complement selection by ***type coercion***

Qualia structure – type coercion

Type coercion: Convert the argument to the type which is expected by the function.

Aliases $\Sigma\alpha$ of an expression α allow polymorphical treatment of α and allow changes of type and denotation

α of type a can be coerced to type b if there is an σ in $\Sigma\alpha$ so that σ(a) is of type b

Qualia structure – type coercion an example

adjective ***fast*** wants to modify a noun e.g. ***motorway***

the adjective needs an argument of type **event** but the noun is of type **artifact_object**

because ***motorway*** is an artifact it has an TELIC value with some event in it

→ type of ***motorway*** can be coerced to event using an alias σ in Σ_α

[TELIC: travel (P,cars) & on (P,x) & fast (P)]

type coercion and functional application

functional application: Is α of type <b,a>, and is β of type b, then α(β) is of type a.

functional application with type coercion:

If α is of type <b,a>, and β is of type c, then

1) α(β) is of type a, if c=b
2) α(σ(β)) is of type a, if there is a $\sigma \in \Sigma_\beta$ where σ(β) is of type b
3) type error otherwise

Qualia structure – subtypes und type paths

- ***type path*** in a type hierarchy:
  - If a is a type, then is [a] a type path
  - If a and b are type paths with b ≤ a, then [a b] is a type path

  where ≤ is the order in the type hierachy
- extended ***type definition***:
  - [e] is a type
  - [t] is a type
  - if [a] and [b] are any types, then <[a],[b]> is a type

  where e,t are standard types from Montague

Qualia structure – type coercion an example

- phrase „begin a book“
  - begin expects phrase of type [event]
  - book has type [individual physobj]
  - noun‘s alias permits type coercion of „book“ to type [event]
  - existence of the alias is guaranteed by inheritance in the type hierarchy
  - two candidates associated with „book“

Qualia structure – subtype coercion

- ***subtype coercion***: conversion of the argument to the subtype which is expected by the function

- ***functional application with subtype coercion***:

  If α is of type <[b c],[a]>, and β is of type [b d], then

  1) $\alpha(\beta)$ is of type a, if c=d
  2) $\alpha(\sigma(\beta))$ is of type a, if there is a $\sigma \in \Sigma'_\beta$ with $\sigma(\beta)$ is of type c
  3) type error otherwise

Qualia structure – functional application with subtype coercion – an application

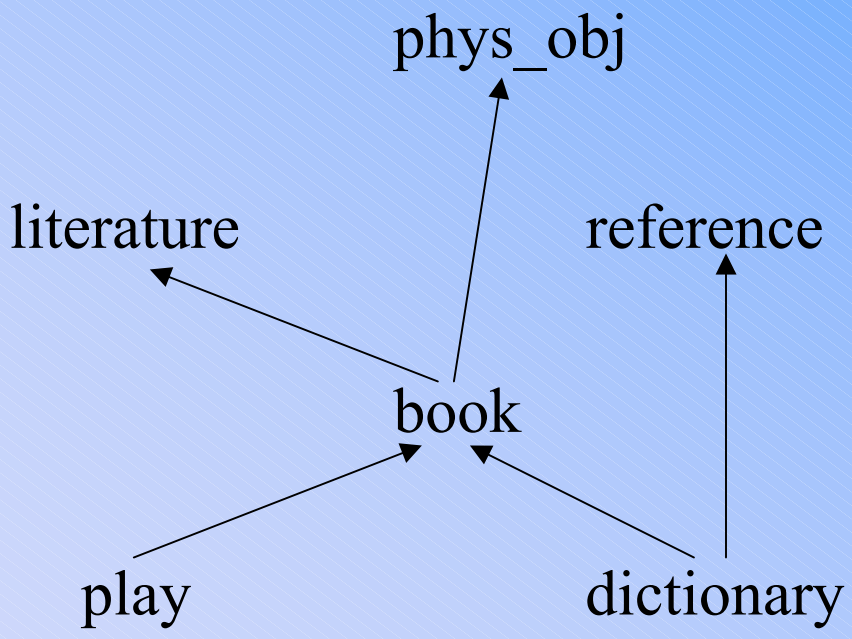
example: a fast car

- generic type of adjectives <[N],[N]>
- fast can be instantiated with the more specific subtype <[N Telic],[N]>
- car has the type <[N artifact]>
- every artifact has a Telic role → existence of an alias σ is guaranteed
- application fast‘(σ(car‘)) is possible

knowledge representation in the semantic lexicon

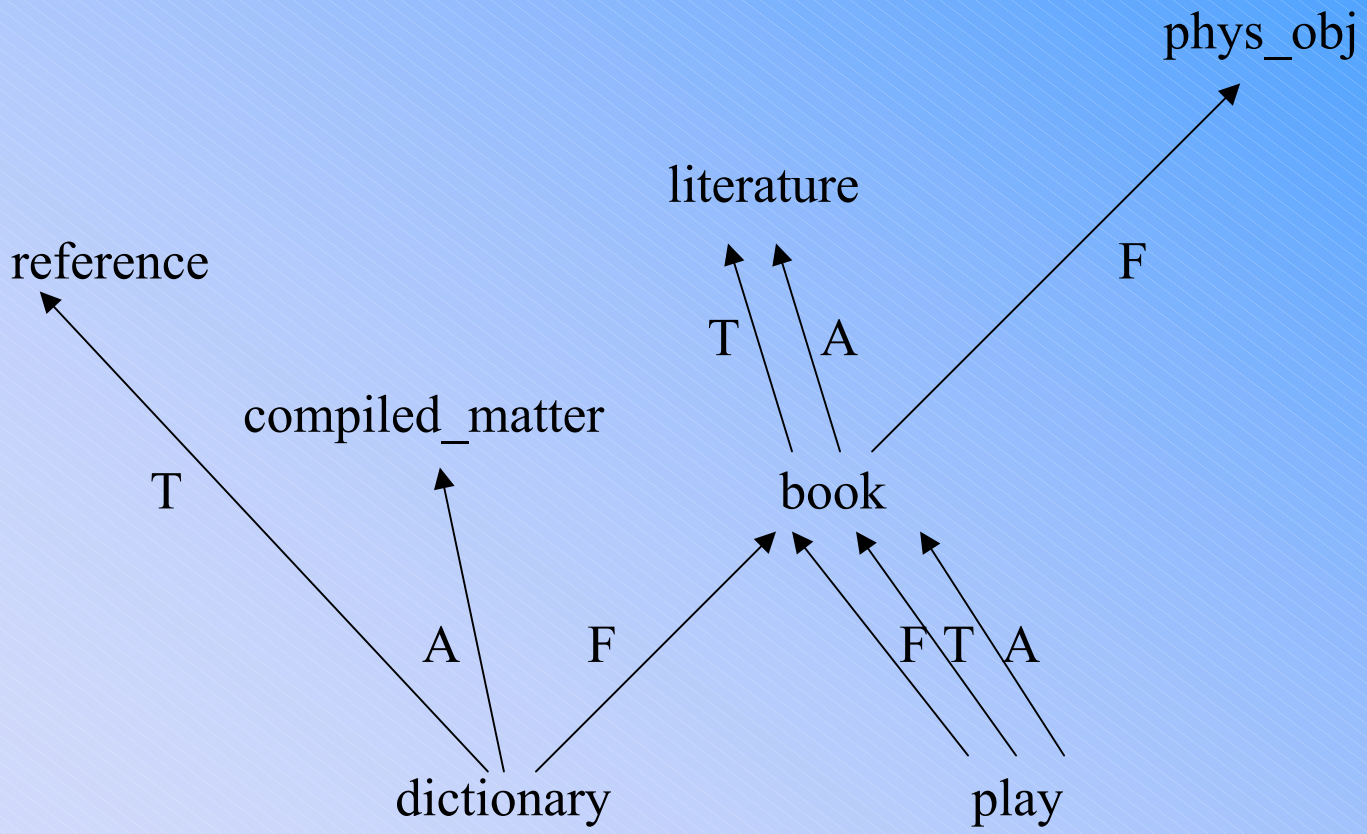
- goal: impose a structure on the lexicon which helps to resolve „hidden“ lexical ambiguity
- simple is_a-hierarchy
- is_a-hierarchy with multiple inheritance to model contextual ambiguities
  - information flow along specification/ generalization edges causes problems
- Qualia structures cause more complex inheritance hierarchies

modelling using an is_a-hierarchy



	play	dictionary
read	√	–
buy	√	√
consult	–	√
begin	√	–

modelling with qualia typed inheritance



denotation of a concept

let P and Q be concepts in our model of lexical organization

- $<Q_1 , P_1 ,... ,P_n>$ is called ***inheritence path***, which can be read as conjunction of ordered pairs
- ***conclusion space*** Φ_q is the set of sequences that lie on an inheritence path <Q,...,P> imposed by the Qualia q
- ***complete conclusion space*** Φ is the set of all conlusion spaces defined for each quale of a concept

denotation of a concept

- Inh(S): set of values inheritable from S

  ϕ
- [[α]]designates the denotation of α with respect to the conclusion space ϕ

  ϕ
- e.g. [[book]] =

λ x [book(x) & Formal(x) = Inh (physobj)
& Telic(x) = Inh (literature)
& Agentive(x) = Inh (literature)]

implications for natural language processing and knowledge representation

- Conflate different word senses in only one lexical meta-entry
  lexical conceptual paradigm (LCP)
  - encoding regularities of word behavior dependend on context
  - reducing the size of the lexicon
  - syntactic information is inheritable between two lexical entries

encoding regularities of word behavior dependend on context

example: product/producer paradigm

qualia can differentiate the different aspects

- The coffee cup is on top of the ***newspaper***.
- The article is in the ***newspaper***.
- The ***newspaper*** attacked the senator from Massachusetts.
- The ***newspaper*** is hoping to fire its editor next month.

inheritance of syntactical information

union (x)
CONST = {entity(y),entity(z)}
FORMAL = entity(x)
AGENTIVE = artifact (x)

merger (x)
CONST = {company(y),company(z)}
FORMAL = company(x)
AGENTIVE = artifact (x)

word – argument realizaion by syntactic schemas

- pattern to realize word-argument structure: **union**

LCP schema (N=union, X=arg1, Y=arg2)
N of X and Y
X‘s N with Y
Y‘s N with X
N between X and Y
N of Z (Z = X + Y)
N between Z

implications for natural language processing and knowledge representation

- shown methods and theories can be adopted by knowledge representation
  - nouns which describe conceptually relations between a physical object and an aperture of some sort:

    window, door, room
  - nouns only referring to the physical object: lid, cap, cover, ...

implications for natural language processing and knowledge representation

- boundary between lexical and commonsense knowledge can be further defined
- generative lexicon with well defined semantic relations
- division of the semantics in different aspects
- semantic information is spread more evenly in the lexicon by a compositional approach with mutual compatible rules
- improved overall robustness of automatic natural language processing concerning lexicon aquisition and language learnability