Lecture 3: Corpus Data Informs Theory

The Role of Annotation in Linguistic Theory

- Semantic annotation is critical for robust language understanding:

  Summarization, question answering, inference

- Annotation schemata should focus on a single coherent theme:

  Different linguistic phenomena should be annotated separately over the same corpus

- Annotations must be consistent with each other:

  Unification and merging of multiple annotation is necessary

GL Theory and Corpus Pattern Analysis

- In order to understand what is actually going on with the theory matching data, we need to design a particular approach to doing corpus analysis, not just bag of words analysis. This is CPA.
- CPA points out what is plausible and implausible about typing judgments.
- We compare each case with theoretical case from lecture 2.
- How does typing as selection help in determining context for WSD?
- Conclusions: Maybe dot objects are too pervasive. What's the most computationally efficient way to encode transformations reflecting data?
  - Accommodation instead of subtyping
  - Relational typing instead of dot object typing
  - GL theory of composition informs clustering and selectional modeling
  - Selection encode context

COURSE OUTLINE

- Monday: Framing the Problem:
  What is Compositionality?
  Generative Lexicon as a Theory of Selection
- Thursday: Corpus Data on Semantic Transformations
  Lexical Sets and Corpus Pattern Analysis
- Friday: Extending and Enriching the Model of
  Generative Lexicon

Methodology of Empirically-Grounded Semantics

- Annotation scheme: assumes a given feature set.
- Feature set: encodes specific structural descriptions and properties of the input data.
- Structural descriptions: theoretically-informed attributes derived from empirical observations over the data.

The Model-Annotate-Test Paradigm

Enriching Compositionality

If all you have for composition is function application, then you need to create as many lexical entries for an expression as there are environments it appears in. (Weak Compositionality)

Two ways to overcome this:

(1) Type Shifting Rules: Partee-Rooth MG, CG, HPSG.

(2) Type Coercion Operations: GL, Hendriks, Moens and Steedman

Maintaining Compositionality

- Generative Mechanisms of Argument Selection:
  - Selection
  - Accommodation
  - Coercion:
    (i) Introduction
    (ii) Exploitation
- Qualia-based Type Structure:
  - Natural,
  - Artifactual,
  - Complex.

Generative Mechanisms of Argument Selection

- Pure Selection: The type a function requires is directly satisfied by the argument.
- Accommodation: The type a function requires is inherited by the argument.
- Coercion: The type a function requires is imposed on the argument type. This is accomplished by either:
  - Exploitation: selecting part of the argument's type structure to satisfy the function's typing;
  - Introduction: wrapping the argument with the type the function requires.

Type Coercion

– Exploitation: selecting part of the argument's type structure to satisfy the function's typing;

– Introduction: wrapping the argument with the type the function requires.

Two Kinds of Coercion in Language

- Domain-shifting: The domain of interpretation of the argument is shifted;
- Domain-preserving: The argument is coerced but remains within the general domain of interpretation.

Domain-Shifting Coercion

- Entity shifts to event:
  I enjoyed the beer
- Event shifts to interval:
  before the party started. . .
- Entity shifts to proposition:
  I doubt John.

Domain-Preserving Coercion

– Count-mass shifting: There's chicken in the soup.
– NP Raising: Mary and every child came.

Types and Composition of Local Contexts

Compositionality mediated through richer selectional mechanisms:

		VERB TYPE	
COMPOSITION	Natural	Artifactual	Complex
Selection	die(x)	fix(x,y)	read(x,y)
Accommodation	wipe(x,hand)	spill(beer)	burn(x,book)
Coercion	enjoy(rock)	spoil(water)	read(x,joke)

Lecture 3: Corpus Data Informs Theory

That's all well and good, but...

Corpus Data on spoil (Patrick Hanks, p.c.)

In both BNC and Associated Press, over 80% of Direct Objects of spoil are Events. Typically, they are Events that one would expect to enjoy. The implicature is that, by spoiling an Event, one kills the enjoyability of it. One might say that spoil is a causative antonym of enjoy.

The lexical set of direct objects of spoil include:

fun, enjoyment, magic, pleasure, holiday, party, Christmas, birthday, dinner, evening, morning, day, half-hour, event, occasion, view, performance, opera, game, match, ...

Compositional Selection in Prepositions

Types of Locations

- Natural Location: defined by 3-D coordinates
- Artifactual Location: defined by Telic on Natural
- Complex Location: defined by coherence relation with Physical Entity

Natural Locations

From the abstraction of spatial coordinates, there are entities which have spatial denotations without entity extention. e_{NL} is structured as a join semi-lattice, $\langle e_{NL}, \sqsubseteq \rangle$;

(33)a. $point$, $spot$, $position$, $area$

b. $space$, sky

Artifactual Locations: e_{AL}

(34)a. $x : e_{NL} \otimes_T \tau$

b. $g \vdash x : e_{NL} \otimes_T \tau \ =_{df} \ g \vdash x : e_{AL}$

c. $g \vdash P : e_{NL} \otimes_T \tau \rightarrow \underline{t} \ =_{df} \ g \vdash P : e_{AL} \rightarrow \underline{t}$

Examples of types in e_{AL}.

(35)a. *seat*: $loc \otimes_T sit$

b. *home*: $loc \otimes_T live\_in$

Complex Locations: e_{CL}

(36)a. $g \vdash x : \sigma \bullet \tau \ =_{df} \ g \vdash x : e_{CL}$

b. $g \vdash P : (\sigma \bullet \tau) \rightarrow \underline{t} \ =_{df} \ g \vdash P : e_C \rightarrow \underline{t})$

Examples of types in e_{CL}.

(37)a. *door*: $phys \bullet loc \otimes_T walk\_through$

b. *window*: $phys \bullet loc \otimes_T see\_through$

Closer Look at the Data

Consider the physical objects from $\mathcal{E}$:

1. Natural Types (No Selection):
   rock, tree, tiger
   We'll meet up with you at the tigers.
2. Artifactual Types (Partial Selection):
   blackboard, computer, table, bar, sink, stove,
   garage_1, station, park, museum, restaurant
3. Complex Types (Selection): door, window, room,
   pool

Non-selecting Artifactual Entities

1. train, chair, phone, garage$_2$, kitchen, sofa, bed
2. But...
   on the sofa, in bed, on the phone, ...

Dot Objects with Functions

Consider the objects from $\mathcal{C}$:
school, work, hospital

1. Stage-level: at (the) school
2. Individual-level: in school, in the army

Events as Containers

Consider the events from $\mathcal{R}$:

1. Symmetric:
   party, conference, workshop, meeting, battle, breakfast
2. Asymmetric:
   lecture, talk, concert

Degree of Involvement

Symmetric event in the container:

(38) a. John is at a meeting
b. Mary is at an appointment.

Asymmetric event in the container:

(39) a. John is at a lecture. (he's not giving it).
b. * John is at his lecture.
c. John is at a concert. (He's not performing).

The Selective Force of Locative AT

(40) a. Any Locative Type from Entity Domain:
b. Some physical objects from Entity Domain:
c. Some Events from Relation Domain:

The Semantics of Locative AT

(41) a. Locative Relation is proximity along horizontal dimension.

b. Telic property of the location or object is exploited.

(42)a. $x : e_{NL} \otimes_T \tau$

b. $g \vdash x : e_{NL} \otimes_T \tau \ =_{df} \ g \vdash x : e_{AL}$

c. $g \vdash P : e_{NL} \otimes_T \tau \rightarrow \underline{t} \ =_{df} \ g \vdash P : e_{AL} \rightarrow \underline{t}$

Artifactual Locative Relations

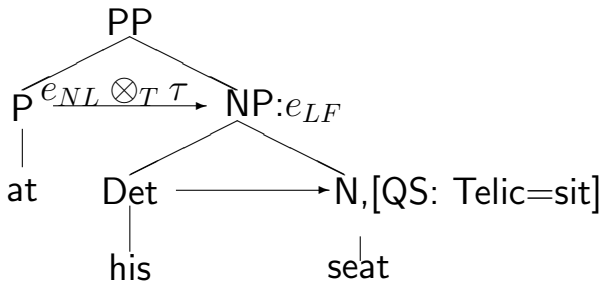
(43) *at*: $e_{AL} \rightarrow (e \rightarrow \underline{t})$

Locative Selection

Location Types:

at his seat

(44)



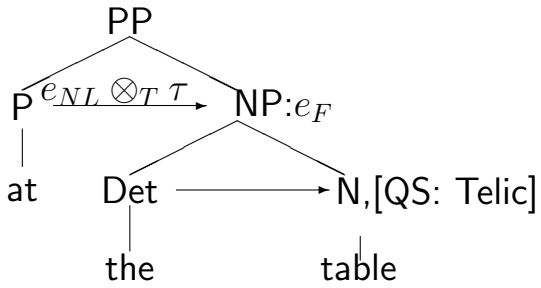
(45) $\lambda x \lambda e \exists y[loc(x, y) \wedge sit(e, x, y) \wedge seat(y)]$

Artifactual Locative Coercion

Objects are coerced to Locations

at the table

(46)



(47) $\Theta[phys \sqsubseteq loc] : phys \rightarrow loc$

(48) $\lambda x \lambda e(\iota y)[loc(x, y) \wedge Telic(e, x, y) \wedge table(y)]$

Violations of Selectional Constraints

– at the chair: locative relation is violated.

– at the tree: Artifactual (Telic) constraint is violated.

Catalan Locatives (p.c. Roser Saurì)

(49) On són les claus?
where are-3pl the keys?

(50) Són a la cadira de lentrada.
Are-3pl at the chair of the hall.

(51) al despatx/cuina /menjador
in-the office /kitchen /dinning room

(52) al calaix.
in-the drawer.

(53) a /sobre la taula.
at/over the table.

(54) where is-3sg the cat?

(55) El gat sobre la taula.
∗El gat a la taula.
The cat is on the table.

(56) where is-3sg the cup?

(57) sobre la taula.
a la taula.
on the table.

Qualia Selection and Default Arguments

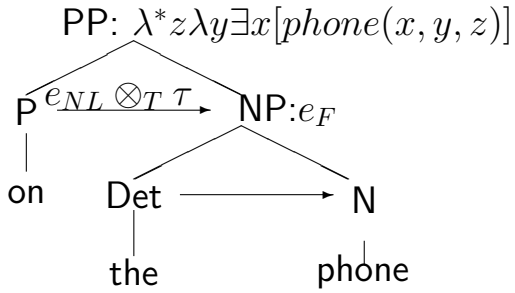
(58) És al telfon (, parlant amb la Maria).
Is-IND.LEVEL at-the phone (, speaking with the-SG-FEM Mary)
He is on the phone.

(59) Està parlant per telèfon (amb la Maria).
Is-STAGE.LEVEL speaking for phone (with the-SG-FEM Mary)
He is speaking through/by the phone .
*Est per telfon.

Qualia Selection and Default Arguments

(60)
$$\left[\begin{array}{l}\textbf{phone} \\ \text{QS} = \left[\begin{array}{l}\text{F} = phys(x) \\ \text{T} = [communicate(e,y,z) \wedge \textbf{with}(e,x)]\end{array}\right]\end{array}\right]$$

(61)



on the phone with Mary

Classifier Systems and Coercion

(Data from David Wilkins (2000))

(62)a. *thipe*: flying, fleshy creatures;

b. *yerre*: ants;

c. *arne*: ligneous plants;

d. *name*: long grasses;

e. *pwerte*: rock related entities.

Classifier Systems and Coercion

(63)a. *kere*: game animals, meat creatures;

b. *merne*: edible foods from plants;

c. *arne*: artifact, usable thing;

d. *tyape*: edible grubs.

(64)a. *kere aherre*: kangaroo as food;

b. *merne langwe*: edible food from bush banana;

c. *pwerte athere*: a grinding stone

Type Distinctions

(65) a. SPECIFIC NOUN: sortal classification, a Natural type;

b. GENERIC NOUN: a Artifactual type;

c. CLASSIFIER CONSTRUCTION: the instantiation and binding of the qualia role from the Artifactual type onto the Natural Type.

Natural vs. Functional

(66) *Iwerre-ke anwerne aherre arunthe-∅ are-ke.*

way/path-DAT 1plERG kangaroo many-ACC see-pc

"On the way we saw some kangaroos."

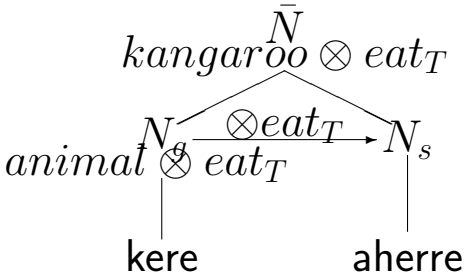
(67) *the imarte arratye kere aherre-∅ arlkwe-tye.lhe-me-le.*

1sgERG then truly meat kangaroo-ACC eat-GO&DO-npp-SS

'When I got there I ate some kangaroo meat."

Classifier Construction

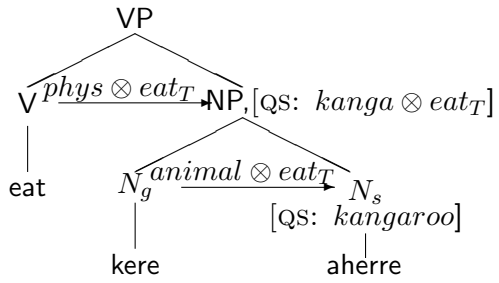
(68)



$$\begin{bmatrix} \textbf{see} \\ \text{CAT} = \textbf{verb} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = animal \\ \text{ARG2} = phys \end{bmatrix} \end{bmatrix}$$

Artifactual Selection with Classifier Construction

$$\begin{bmatrix} \textbf{eat} \\ \text{CAT} = \textbf{verb} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = animal \\ \text{ARG2} = phys \otimes eat_T \end{bmatrix} \end{bmatrix}$$



$$\Theta[kangaroo \sqsubseteq phys] : kangaroo \rightarrow phys$$

Accounting for Agency

1. Selection: x *assassinated/murdered* y
2. Accommodation: x *rolled down the hill*
3. Coercion: x *flies to Boston*

– Human is a complex type of rational animal.

– human: $anim \otimes_{\mathsf{A,T}} (\mathrm{E,E'})$

(69) a. The child /storm / tree killed the teacher.

b. The child /*storm / *tree murdered the teacher.

(70) a. *kill*: $anim \rightarrow (e_N \rightarrow t)$

b. *murder*: $anim \rightarrow (human \rightarrow t)$

Selection of Agency

John murdered Mary.

1. murder: $\lambda x[\mathit{murder}(\mathsf{x},\mathsf{m})]$,
   $\langle m\colon anim, x\colon anim \otimes_{\mathsf{A,T}} (\mathrm{E},\mathrm{E}')\rangle$
2. john: $anim \otimes_{\mathsf{A,T}} (\mathrm{E},\mathrm{E}')$
3. $\exists e[\mathsf{murder}(e, j, m)]$, Intentional Act

Accommodation of Agency

John killed Mary (intentionally).

Co-Composition

Classic Co-composition cases:

(71)a. John baked a potato.

b. John baked a cake.

(72)a. The bottle is floating in the river.

b. The bottle floated under the bridge.

(73)
$$\begin{bmatrix} \textbf{float} \\ \text{ARGSTR} = \left[\text{ARG1} = \boxed{1}\left[\textbf{physobj} \right] \right] \\ \text{EVENTSTR} = \left[\text{E}_1 = \textbf{e}_1\textbf{:state} \right] \\ \text{QUALIA} = \left[\text{AGENTIVE} = \textbf{float}(\textbf{e}_1, \boxed{1}) \right] \end{bmatrix}$$

(74)
$$\begin{bmatrix} \textbf{into the cave} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} & = \boxed{1}\begin{bmatrix}\textbf{physobj}\end{bmatrix} \\ \text{ARG2} & = \boxed{2}\begin{bmatrix}\textbf{the\_cave}\end{bmatrix} \end{bmatrix} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E}_1 = \textbf{e}_1\textbf{:process} \\ \text{E}_2 = \textbf{e}_2\textbf{:state} \\ \text{RESTR} = <_\propto \\ \text{HEAD} = \textbf{e}_2 \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORMAL} = \textbf{at}(\textbf{e}_2, \ \boxed{1}\ , \ \boxed{2}\ ) \\ \text{AGENTIVE} = \textbf{move}(\textbf{e}_1, \ \boxed{1}\ ) \end{bmatrix} \end{bmatrix}$$

(75)
$$\lambda x \lambda e_1 \exists e_2 [move(e_1, x) \wedge \circ(e_1, e_2) \wedge float(e_2, x)]$$
$$\Rightarrow \textit{while floating}$$

(76)
$$\begin{bmatrix} \textbf{float into the cave} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} & = \boxed{1}\begin{bmatrix}\textbf{physobj}\end{bmatrix} \\ \text{ARG2} & = \boxed{2}\begin{bmatrix}\textbf{the\_cave}\end{bmatrix} \end{bmatrix} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E}_1 = \textbf{e}_1\textbf{:state} \\ \text{E}_2 = \textbf{e}_2\textbf{:process} \\ \text{E}_3 = \textbf{e}_3\textbf{:state} \\ \text{RESTR} = <_\propto (e_2, e_3) \ , \circ_\propto (e_1, e_2) \\ \text{HEAD} = \textbf{e}_3 \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORMAL} = \textbf{at}(\textbf{e}_3, \ \boxed{1}\ , \ \boxed{2}\ ) \\ \text{AGENTIVE} = \textbf{move}(\textbf{e}_2, \ \boxed{1}\ ), \textbf{float}(\textbf{e}_1, \ \boxed{1}\ ) \end{bmatrix} \end{bmatrix}$$

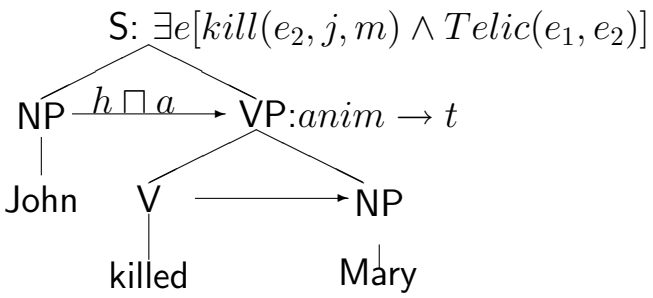
$$\begin{bmatrix} \textbf{kill} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E}_0 = \textbf{e}_0\textbf{:state} \\ \text{E}_1 = \textbf{e}_1\textbf{:process} \\ \text{E}_2 = \textbf{e}_2\textbf{:state} \\ \text{RESTR} = <_\propto \\ \text{HEAD} = \textbf{e}_1 \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = \boxed{1} \begin{bmatrix} \textbf{ind} \\ \text{FORMAL} = \textbf{physobj} \end{bmatrix} \\ \text{ARG2} = \boxed{2} \begin{bmatrix} \textbf{animate\_ind} \\ \text{FORMAL} = \textbf{physobj} \end{bmatrix} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \textbf{cause-lcp} \\ \text{FORMAL} = \textbf{dead}(\textbf{e}_2, \boxed{2}) \\ \text{AGENTIVE} = \textbf{kill\_act}(\textbf{e}_1, \boxed{1}, \boxed{2}) \\ \text{PRECOND} = \neg\textbf{dead}(\textbf{e}_0, \boxed{2}) \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \textbf{kill} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E}_0 = \textbf{e}_0\textbf{:state} \\ \text{E}_1 = \textbf{e}_1\textbf{:process} \\ \text{E}_2 = \textbf{e}_2\textbf{:state} \\ \text{RESTR} = <_\propto \\ \text{HEAD} = \textbf{e}_1 \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = \boxed{1} \begin{bmatrix} \textbf{ind} \\ \text{FORMAL} = \textbf{physobj} \end{bmatrix} \\ \text{ARG2} = \boxed{2} \begin{bmatrix} \textbf{animate\_ind} \\ \text{FORMAL} = \textbf{physobj} \end{bmatrix} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \textbf{cause-lcp} \\ \text{FORMAL} = \textbf{dead}(\textbf{e}_2, \boxed{2}) \\ \text{AGENTIVE} = \textbf{kill\_act}(\textbf{e}_1, \boxed{1}, \boxed{2}) \\ \text{TELIC} = \textbf{P}(\textbf{e}_3, \boxed{1}) \\ \text{PRECOND} = \neg\textbf{dead}(\textbf{e}_0, \boxed{2}) \end{bmatrix} \end{bmatrix}$$

Accommodation of Agency

1. kill: λx[*kill*(x,m)], $\langle m : anim, x : anim \rangle$
2. john: $anim \otimes_{\mathsf{A,T}} (\mathrm{E},\mathrm{E}')$
3. Agent Accommodation: λx[*kill*(x,m)],
   $\langle m : anim, x : anim \otimes_{\mathsf{A,T}} (\mathrm{E},\mathrm{E}') \rangle$
4. Function Application:
5. $\exists e[\mathsf{kill}(e, j, m)]$

(77)



(78) a. John killed the flowers accidently / intentionally.
b. John/the rock rolled down the hill.
c. John cooled off with an iced latte.

(79) a. John gave Mary a book.
b. John gave Mary a shower.
c. John gave the plants a spray.

Coercion of Agency

(80)a. We painted$_{R(i,j)}$ our house last summer.
We$_i$/They$_j$ used Benjamin Moore paints.
They$_j$/*We$_i$ even worked in the heat of the day.

b. I dry-cleaned$_{R(i,j)}$ my shirts before I left on the trip.
They$_j$/*I$_i$ stained the sleave, though.

e. I washed$_{R(i,j)}$ my car yesterday.
They$_j$/*I$_i$ waxed the exterior too.

(81)a. Lufthansa flies to Boston.

b. McDonalds has served 1 trillion burgers.

Corpus Data on Selection: believe [__ S+-fin]

• inf clause	1996	0.6
• ing clause	18	1.9
• that clause	13974	2.6
• wh clause	486	0.5

Corpus Data on Selection: believe [__ NP]

• luck	73	33.05
• ear	48	22.14
• story	73	20.67
• word	95	18.9
• eye	74	14.78
• hype	6	14.16
• myth	12	14.07
• truth	19	13.39
• it	8	12.91
• lie	10	12.57
• opposite	7	12.22
• tale	13	12.16
• nonsense	7	11.62
• propaganda	7	11.17

Concordance for believe [__ NP]:

31 percent said they'd believe the newspaper, primarily because they had "more

He seems to have made the mistake of believing his own propaganda .

Politicians are always at their most vulnerable when they believe their own propaganda .

They weren't quite so stupid as to believe wholly their own propaganda .

The trouble with the hon. Gentleman is that he believes his own propaganda .

The trouble is , the media is able to influence the public and unfortunately influential people in the trade union and labour movements , and maybe they believe the propaganda that socialism is dead and respond accordingly .

PropBank: doubt

Predicate *doubt*:

Frames file for 'doubt' based on sentences in financial subcorpus. No access to verbnet. Comparison with 'believe'.

Roleset doubt.01 "doubt, disbelieve":

Roles:

Arg0:*disbeliever*

Arg1:*disbelief*

Examples:

sentential disbelief (-)

```
Although takeover experts said they doubted Mr. Steinberg will
make a bid by himself...
```

Arg0: they
REL: doubted
Arg1: Mr. Steinberg will make a bid by himself

As usual, leave 'that' complementizers out of the Arg1.

nominal disbelief (-)

```
John doubted Mary.
```

Arg0: John
REL: doubted
Arg1: Mary

Corpus Data on Selection: doubt [__ NP]

•	ability	40	28.79
•	validity	16	27.09
•	sincerity	8	23.7
•	sanity	6	20.83
•	existence	13	18.5
•	correctness	4	16.9
•	accuracy	7	16.7
•	thomases	2	15.99
•	wisdom	6	15.53
•	viability	4	14.28
•	truth	7	13.22
•	authenticity	3	12.83
•	word 1	4	12.75

•	feasibility	3	12.42
•	suitability	3	12.2
•	veracity	2	12.11
•	strength	7	11.84
•	seriousness	3	11.64
•	faith	5	10.75
•	value	9	10.17
•	presupposition	2	9.74
•	possibility	5	9.72
•	claim	6	9.6
•	Sebastian	2	9.48
•	commitment	5	9.4

Corpus Data on Artifactual Selection: repair [__ NP]

• damage	107	42.92
• roof	16	20.31
• covenant	9	18.38
• fence	10	18.1
• gutter	5	15.89
• ravages	4	15.82
• hernium	4	15.6
• car	23	15.39
• shoe	10	15.04
• leak	5	15.01
• bridge	10	14.03
• crack	6	14.02
• fencing	4	13.91
• wall	14	13.77
• puncture	3	13.54
• building	16	13.52

pipe	7	12.92
saddlery	2	12.79
ligament	3	11.85
road	13	12.24

Corpus Data on Complex Selection: read [__ NP]

• book	772	43.31
• newspaper	205	35.76
• bible	82	34.24
• papers	144	32.61
• article	156	31.89
• letter	226	30.44
• poem	85	29.39
• novel	88	28.57
• paper	175	28.54
• text	112	26.93
• passage	82	26.89
• story	148	26.03
• comic	26	25.41
magazine	85	25.38
script	37	24.37
poetry	46	24.12
report	180	23.37
page	89	23.25
paragraph	38	22.92
word	162	21.85

Corpus Data on Propositional Selection: tell [__ NP]

• story	1293	51.85
• truth	602	49.55
• lie	254	45.4
• tale	275	41.0
• reporter	170	38.53
• inquest	82	34.16
• court	639	33.72
• Reuter	44	33.62
• conference	288	30.81
• fib	18	30.49
• joke	94	28.63

Corpus Data on Polysemous Alternating Verbs: open:

Before Bramble could answer , the door **opened** and another stranger entered

As he hesitated the door **opened** and Gilbert Forbes came out in a rush ,

Dressing-room doors **opened** , voices questioned , feet clattered on

and when the door **opened** again he started violently and spilled

It turned . He pulled . The door **opened** . He looked out . The corridor dusky .

But midway through the afternoon the door **opened** . Pike came in. x

xThe bedroom door **opened** and she rushed in . ` Want anything

The door **opened** and there she stood . She was wearing a

they sang as the back door **opened** and Nick came in , a bottle of wine in

but then the door **opened** . The policeman smiled showing large flashy

then the door **opened** . A Bengali girl , absurdly young , stood

The door **opened** and Sheila came in . ` What are

still searching for them as the front door **opened** and Herr Nordern came in.

Corpus Data on Complex Types: lunch (as Obj)

eat	93	42.49
cook	34	34.46
serve	44	28.44
skip	9	23.41
finish	21	22.58
enjoy	25	21.97
prepare	21	20.66
attend	15	18.54
miss	12	16.96
take	48	15.47
provide	26	15.21
bring	21	15.06
get	40	14.98
include	12	10.89

buy	14	14.21
arrange	8	13.18
want	19	12.69
host	4	12.17
organise	6	11.1
cancel	4	11.08
order	6	10.74
spoil	3	9.72
share	6	9.75

Corpus Data on Complex Types: lecture (as Obj)

attend	75	38.84
deliver	65	38.02
give	226	35.18
entitle	12	19.41
organise	9	14.38
present	13	14.16
sponsor	5	12.55
illustrate	7	12.44
finish	7	11.81
include	13	11.4
organize	5	11.21
publish	8	10.99
prepare	7	10.52
get	22	9.82

record	6	9.73
hold	12	9.55
arrange	5	9.46
read	6	8.59
write	8	8.54
begin	6	6.4

Corpus Data on Complex Types: seminar (as Obj)

• attend	65	39.64
• organise	56	38.75
• hold	88	32.76
• host	7	18.77
• entitle	9	18.08
• run	19	17.09
• convene	5	16.94
• chair	6	16.83
• arrange	9	15.72
• sponsor	6	15.5
• conduct	8	14.93
• address	7	13.84
• give	24	12.71

plan	7	11.98
design	5	8.84
present	5	8.4
aim	6	11.87
follow	6	7.15

Corpus Data on Complex Types: appointment (as Obj)

make	454	35.11
announce	71	30.09
terminate	20	27.2
confirm	35	24.53
approve	31	24.52
arrange	32	24.26
cancel	16	22.16
keep	55	20.42
accept	32	19.64
get	89	18.58
secure	17	18.21
relinquish	7	17.67
book	9	16.21
include	30	15.47
ratify	6	15.32

hold	36	15.49
follow	30	14.69
welcome	11	14.5
recommend	11	14.06
receive	20	13.23
block	7	12.81
oppose	7	12.01
veto	5	15.44
miss	9	11.83

Corpus Data on Complex Types: book (as Subj)

•	contain	119	30.89	consist	16	15.28
•	deal	51	24.3	devote	11	14.97
•	cover	48	19.9	trace	11	14.7
•	include	58	18.85	reveal	20	14.66
•	review	19	18.62	concentrate	15	14.59
•	lie	28	18.4	explain	24	14.58
•	provide	70	17.69	chronicle	6	18.06
•	publish	30	17.37	describe	28	13.93
•	show	65	17.07			
•	appear	37	17.01			
•	bargain	6	16.28			
•	help	37	15.54			

Corpus Data on Complex Types: book (as Obj)

read	772	53.51
write	933	50.44
publish	416	44.21
balance	76	32.65
buy	187	29.16
entitle	66	27.96
borrow	43	24.94
illustrate	65	24.38
close	76	22.84
produce	146	22.66
research	26	22.34
open	100	22.05
rewrite	16	21.69
sell	92	21.25
print	34	20.74
recommend	44	20.17
get	301	20.15
Compile	23	19.81

dedicate	23	19.53
ban	27	18.52
purchase	28	18.2
consult	22	17.52
finish	38	17.37
edit	18	17.27

Complex Types: book (modified by Adjective)

• concerned	61	34.79
• available	65	31.21
• useful	20	22.67
• full	30	21.91
• enjoyable	8	20.99
• readable	5	19.09
• interesting	13	18.45
• unreadable	3	15.46
• relevant	9	14.78
• complete	9	14.59
• ready	9	14.45
• up to date	4	14.29
• valuable	6	13.73

good	18	13.2
popular	8	13.03
encyclopaedic	2	12.69
blasphemous	2	12.53
open	9	11.67
invaluable	3	11.56
impressive	4	11.2
supposed	5	11.03

Complex Type Structure is Exploited Differently in Different Grammatical Positions

- Book in Subject position exploits the information type
- Book in Object position exploits the physical type

Computational Lexical Resources

- Structured according to theoretical model of particular linguistic phenomena
- Need empirical grounding in corpus data
- This lecture discusses
  - Modifying the underlying theory for the KB from corpora
  - contexualizing a lexical knowledge base to corpus data

Computational Lexical Resources

WordNet	• Synsets • Inter-synset relations • Sentence frames
Levin Verb Classes	• Enumerative senses
PropBank	• Verb-specific semantic roles
FrameNet	• Frame Theory • Constellations of selectional possibilities
Generative Lexicon	• Aims to account for context • Underspecified representation • Difficult to scale

Contextualizing Lexical Resources

- Brandeis Semantic Ontology (BSO)
  - Lexical KB based on Generative Lexicon principles
  - Consists of an ontology and a dictionary
  - Follows specification adopted by EU-sponsored SIMPLE project
- Corpus Pattern Analysis (CPA)
  - Semi-automated corpus analysis methodology
  - Derives from analysis of large corpora for lexicographic purposes (e.g. Cobuild dictionary)
  - Identifies typical context elements responsible for activating word senses of a target words
  - Creates inventory of word senses for the target word

Brandeis Semantic Ontology

- Based on Generative Lexicon principles, in line with SIMPLE specifications
- Coverage at present (approximate figures)
  - Type lattice
    - 3500 type nodes total
    - Entity types, Event types, Property types
    - Event and property types cover events + relations
  - Lexical coverage
    - 40,000 lexical entries (multiple senses)
    - 6,000 collocational entries
    - 29,000 nouns, 5,000 verbs, 6,000 adjectives
    - adverbs, prepositions, numerals, pronouns, determiners

Generative Lexicon

- Lexical items endowed with structure that aims to account for compositionality of meaning
- Four levels of lexical information:
  - Lexical typing structure
  - Argument structure: *specifies predicate's arguments*
  - Event structure: *specifies event type and subevents*
  - Qualia structure (4 basic roles)

    FORMAL *object's basic type, 'isa' relation*

    CONSTITUTIVE *object's constituent parts*

    TELIC *object's purpose or function*

    AGENTIVE *object's origin, how it came into being*
- Lexical inheritance is typed and follows qualia links
- ***sandwich***(x)

  FORMAL = **physform**(x)

  CONST = {bread, …}

  TELIC = **eat**(P, w, x)

  AGENTIVE = **make**(z, x)

BSO Structure

- Events, Entities, Properties
- Qualia are defined for Entity types
- Arguments are specified for Event types
- Type inheritance principles
  - Inheritance is typed; a simple type may inherit its qualia from different supertypes
  - Inheritance for Entities follows qualia links
  - Inheritance for Events mirrors argument type inheritance

BSO entry for “beer”

BSO 0 other senses

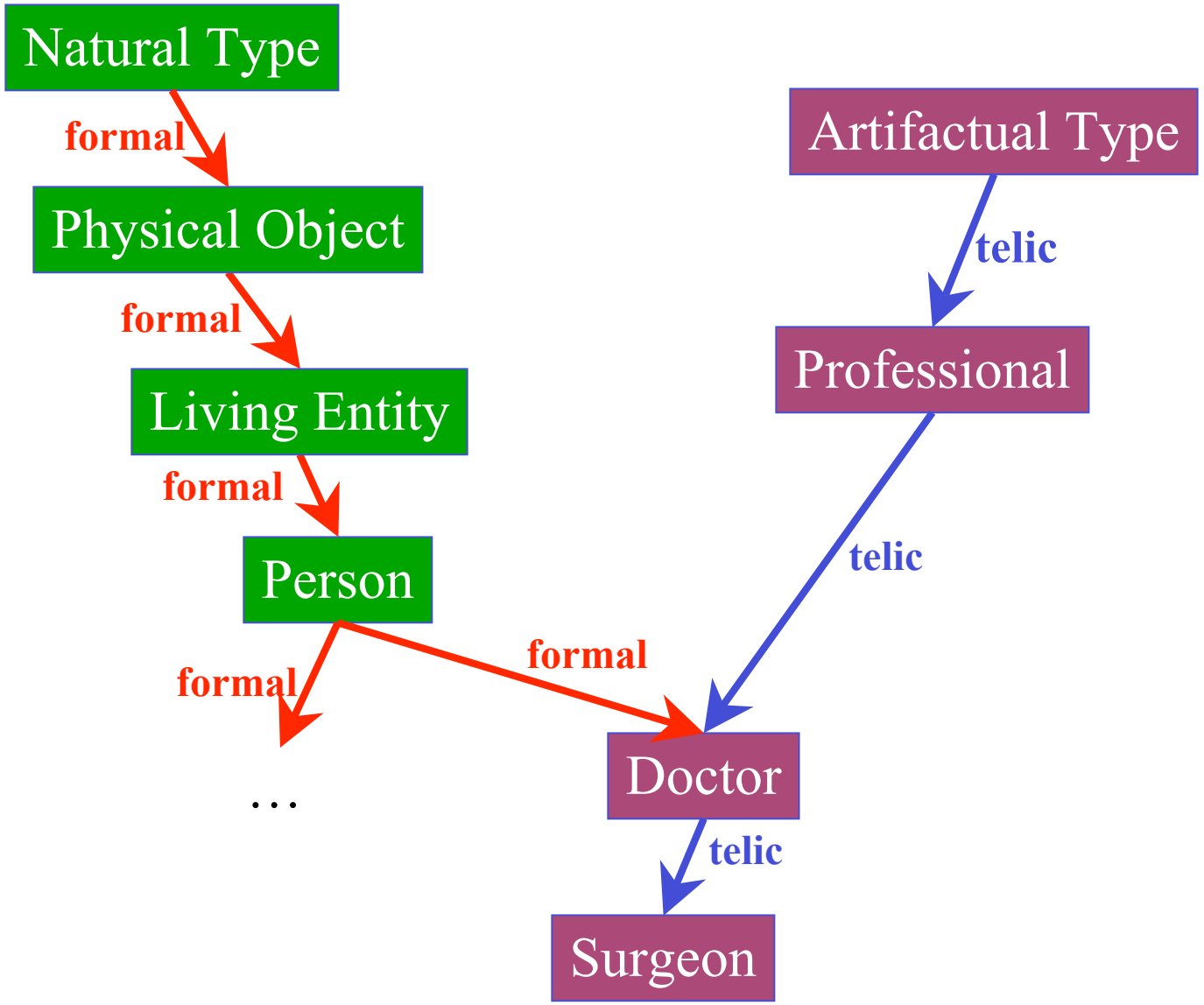
LEMMA: beer
POS: noun
Type: Beer
Inherited Type: Alcoholic Beverage
Has Elements: Alcohol

Qualia:
Indirect Telic: Drink Activity
Instrumental Telic: Event
Indirect Agentive: Create Material Entity Activity
Constitutive: Alcohol

Entity Hierarchy

- Natural types
  - Inherit formal quale of supertype
- Artifactual types
  - Inherit telic quale of supertype
  - Formal quale is inherited through **formal mapping**
- Complex types
  - “dot types” (e.g. building, book, lecture)
  - very shallow hierarchy
  - inherit from two or three functional and/or natural types
- Lexical items for Entity types
  - multiple senses
  - inherit qualia specifications from their type
  - may link to a more specific type in their quale

Type Inheritance for Naturals and Artifactuals



Event Hierarchy

- Shallow, corpus-driven
- Arguments are specified by links to Entity types
- Event typing depends on argument typing
- Lexical items for Event types
  - multiple senses
  - inherit argument type specifications from their type
  - may link to a more specific argument type
  - may add a type specification for an argument missing from its type

Corpus Pattern Analysis

Pustejovsky and Hanks (2001)
Pustejovsky, Hanks, and Rumshisky (2004),
Hanks and Pustejovsky (2005),
Rumshisky, Hanks, Havarsi, and Pustejovsky (2006)

Corpus Pattern Analysis (CPA)

Corpus Pattern Analysis (CPA) is a corpus analytic and automated induction technique that:

1. Identifies the typical syntagmatic patterns for each word and determines discriminant context features.
2. Catalogues semantic types of arguments that are relevant for distinguishing between different senses.
3. Creates an inventory of syntactic and lexical realizations for relevant semantic types.

CPA Components

- Lexical discovery
  - Manual discovery of selection context patterns through corpus analysis
  - Apply this procedure to predicates
- Feature set verification
  - Sorting unseen instances of verb use according to nearest match to identified patterns
  - Similar to conventional WSD
- Automatic pattern acquisition
  - Acquisition of patterns for unanalyzed cases
    - Discriminant feature selection
    - Predicate-based argument clustering

Analyzing Context of Usage

- Consider the word ***treat*:**

  Peter treated Mary badly.
  Peter treated Mary with antibiotics.
  Peter treated Mary with respect.
  Peter treated Mary for her asthma.
  Peter treated Mary to a fancy dinner.
  Peter treated Mary to his views on George W. Bush.
  Peter treated the woodwork with creosote.

Analyzing Context of Usage

- Consider the word ***treat*:**

  Peter treated Mary badly.
  Peter treated Mary with antibiotics.
  Peter treated Mary with respect.
  Peter treated Mary for her asthma.
  Peter treated Mary to a fancy dinner.
  Peter treated Mary to his views on George W. Bush.
  Peter treated the woodwork with creosote.

What features are relevant?

- Extending classification of minor categories

  e.g. adverbials of manner/effect

  - Peter treated Mary rudely.
  - Peter treated Mary effectively.

- Argument structure and contrasting argument types are the most frequent source of meaning differentiation for the predicate

  - The customer will absorb the cost.
  - The customer will absorb this information.

CPA Patterns for “absorb”

The customer will absorb the cost.

Mr. Clinton wanted energy producers to absorb the tax.

PATTERN 1: [[Abstract] | [Person]] absorb [[Asset]]

They quietly absorbed this new information.

Meanwhile, I absorbed a fair amount of management skills.

PATTERN 2: [[Person]] absorb {([QUANT]) [[Abstract= Concept]}

Water easily absorbs heat.

The SO_2 cloud absorbs solar radiation.

PATTERN 3: [[PhysObj] | [Substance]] absorb [[Energy]]

The villagers were far too absorbed in their own affairs.

He became completely absorbed in struggling for survival.

PATTERN 4: [[Person]] {be | become} absorbed {in [[Activity] | [Abstract]}

Argument Typing in CPA

- Lexical discovery produces two pieces of information regarding the arguments:
  - Argument type (shallow type)
  - Semantic subspecification, if any

- For example,

  [[Person=Doctor]] treat [[Person=Patient]] (at | in [[Hospital]])

  [[Person]] fire [[Artifact=Firearm]] (at [[PhysObj]])

  [[TopType]] take {[[Person]]'s mind} {off [[TopType= Bad]]}

BSO Lite

- BSO Lite is a shallow projection of BSO
  - Used in CPA to help identify lexical sets of predicate arguments with semantic types
  - Selected for frequently contributing to existing CPA patterns
- 65 Shallow Types
  - Abstract, Asset, Animate, Artifact, Document, HumanGroup, Information, Institution, Location, Person, PhysObj, Process, Substance, Surface, TimePeriod, etc.
- BSO Lite has been used to improve WSD for a subset of Senseval-3 verbs

Semantic Subspecification

- An interpretation assigned to the argument (in underspecified cases)
- A unifying semantic feature for the **lexical set** (i.e. the lexical items found in that argument position)

[[Person=Doctor]] treat [[Person=Patient]] (at | in [[Hospital]])

[[Person]] fire [[Artifact=Firearm]] (at [[PhysObj]])

[[TopType]] take {[[Person]]'s mind} {off [[TopType= Bad]]}

Semantic Subspecification

- Lexical sets
  - predicate-based groupings of similarly typed lexical elements that typically fill a given argument slot of the target predicate
  - [[Person]] fire [[Artifact=Firearm]] (at [[PhysObj]])

  Firearm (object argspec for ***fire***)
  - gun, rifle, Kalashnikov, pistol, revolver, MK17
- also Properties (e.g. Bad) and Roles (semantic role, e.g. Beneficiary)
  - [[TopType]] take {[[Person]]'s mind} {off [[TopType= Bad]]}

  Bad (iobj argspec for ***take***)
  - problem, troubles, depression, nausea, dizziness, anxiety, tragedy, crime, disillusionment, pain

Lexical Discovery

- Initial inventory of relevant features is created
- Initial inventory of semantic types
- Lexical sets
  - Uncover semantic features that contribute to predicate disambiguation
  - In BSO, correspond to Subtype (usually a functional subtype)
  - Place additional restrictions on semantic type of the argument
  - Populated through type-filtered cluster analysis, in each argument position of the target lemma

CPA Pattern Elements

- Syntactic Parsing
  - Phrase-level parsing (clause roles)
- Shallow Semantic Typing
  - Generic semantic features
    - shallow types from BSO Lite
- Minor Category Parsing
  - Adverbial Phrases, Locatives, Purpose Clauses, Rationale Clauses, Temporal Adjuncts, etc.
- Subphrasal Syntactic Cue Recognition
  - Genitives, partitives, bare plural/determiner distinction, infinitivals, negatives, past participles, etc.

CPA Database

- Database of hand-constructed CPA patterns
  - about 100 verbs with varying degrees of polysemy
  - about 900 patterns

- Similar to what we have seen so far
  - CPA patterns for ***treat*** :

[[Person]] treat [[Person]] (to [[Event]])

[[Person 1]] treat [[Person 2]] [Adv[Manner]]

  - CPA patterns for ***assemble*** :

[[Person]] assemble [[Artifact]]

[PLURAL[Person]] | [[Human Group]] assemble (in [[Location]])

Corpus Pattern Acquisition

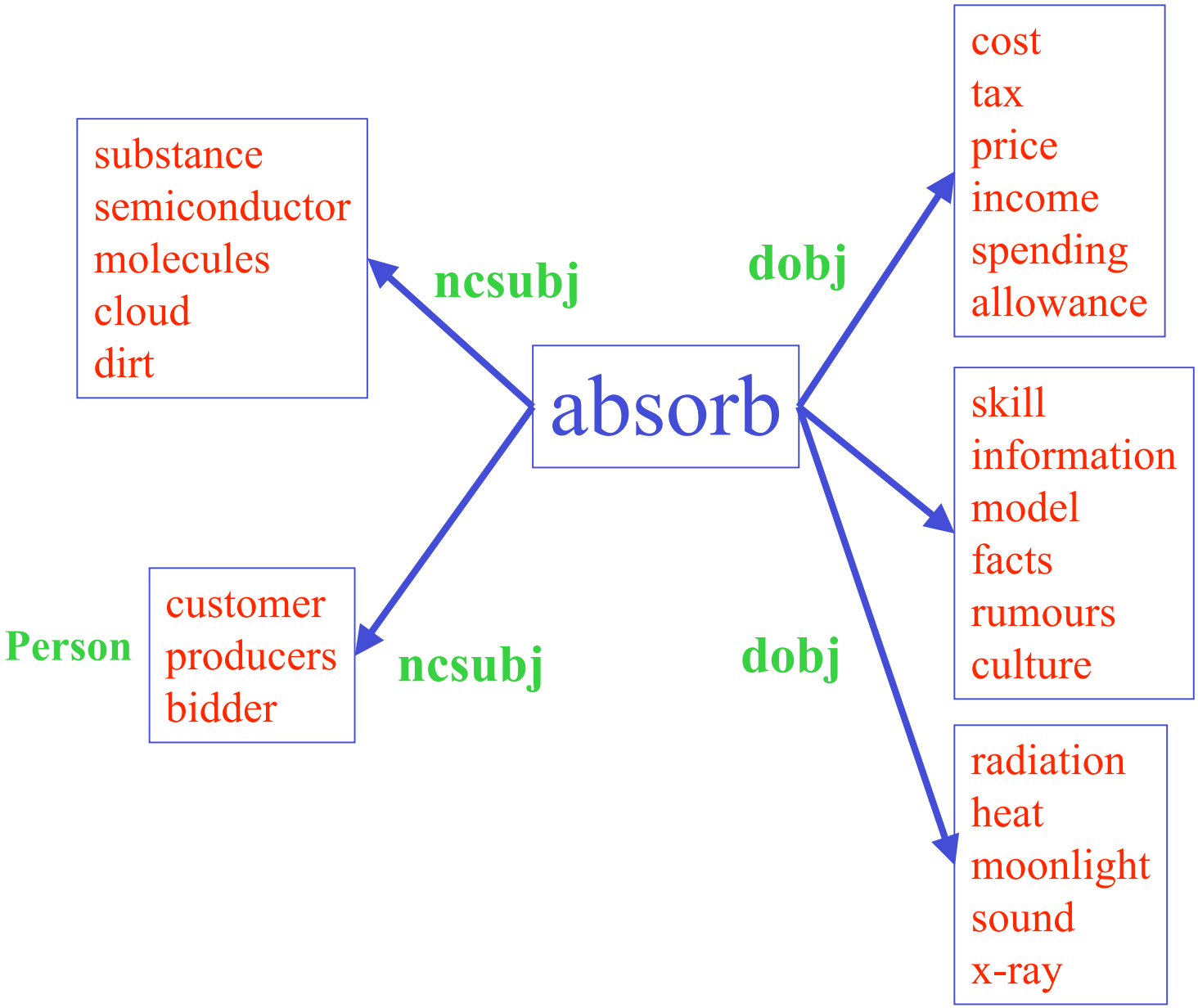
- Acquisition of patterns for unanalyzed cases
  - Discriminant feature selection
  - Predicate argument clustering
- Bootstrapping
  - initial set of context features
  - lexical sets for initial set of verbs

  (from lexical discovery)

Corpus Pattern Acquisition

- Feature selection
  - Use grammatical relation features
    - parse-derived
    - similar to the kind used in by D. Lin, P. Pantel, A. Kilgarriff's Word Sketch Engine, etc.
    - argument typing with BSO Lite
  - WSD experiments conducted as a part of feature verification process
- Used to produce
  - CPA patterns and
  - lexical sets implicated in those patterns through clustering-for-sense-discrimination on predicates.
- Semi-automatic

Corpus Pattern Acquisition



Contextualizing BSO with CPA

- CPA => BSO
  - BSO undergoes verification with respect to corpus information as recorded in CPA context patterns
    - Entity hierarchy: verification restructuring
    - Event hierarchy: enriching argument specification
- BSO => CPA:
  - Keep track of what BSO types capture relevant type distinctions in CPA
    - Refine semantic features (BSO Lite)

Contextualizing BSO Entity Hierarchy

- Nouns denoting entities are grouped together and typed according to their tendency to co-occur in the same argument slot in relation to verbs
- Goal is to substantiate existing hierarchy and restructure where necessary by verifying lexical extensions of each type
- Use sense discrimination by context elements as a criterion
- Semantic type is retained if it carries a semantic feature that verifiably contributes to producing actual sense distinctions in predicates (as observed in corpora)

Contextualizing BSO Event Hierarchy

- Enrich verb argument specification through semantic typing information from CPA patterns
  - Event arguments linked to appropriate Entity types
  - “Direct “ sense disambiguation based on argument type information
- Event typing linked to argument typing

BSO entry for “fire”

CPA Patterns for “fire” (selected)

I. DISCHARGE A GUN AT A TARGET (60%)

1. [[Person]] fire [[Artifact=Firearm]] (at [[PhysObj]])

2. [[Person]] fire [[Artifact=Projectile]] (off) (from [[Artifact=Firearm]]) (at [[PhysObj]] | [Adv[Direction]])

3. [[Artifact=Firearm]] fire [NO OBJ] (at [[PhysObj]] | on [[HumanGroup]] | [Adv[Direction]])

II. DISMISS AN EMPLOYEE (11%)

5. [[Person 1]] fire [[Person 2]] (for [[Action=Bad]])

Contextualized BSO entry for “fire”

Conclusion

- Lexical Typing is Structured Lexical Decomposition
- The Predicate has Structure:
  - Qualia Structure
  - Argument Structure
  - Event Structure
- Context is encoded by strong typing
- Distinction between selection, coercion, and exploitation
- Opposition Structure can be encoded in the predicate's type as a gate.

Conclusions from Today's Lecture

- As a methodology, allows to adept an ontology to specific domain/task, using a specialized corpus
- Sense distinctions not supported by corpus evidence deleted from the type system.
- Type cohesion with respect to corpus evidence.

The End

Thank You!