

Association for Machine Translation in the Americas

AMTA - 2006
CONFERENCE

TUTORIAL ON

A Gentle Introduction to
Ontologies

Presenter:

Eduard Hovy

Information Sciences Institute,

USC

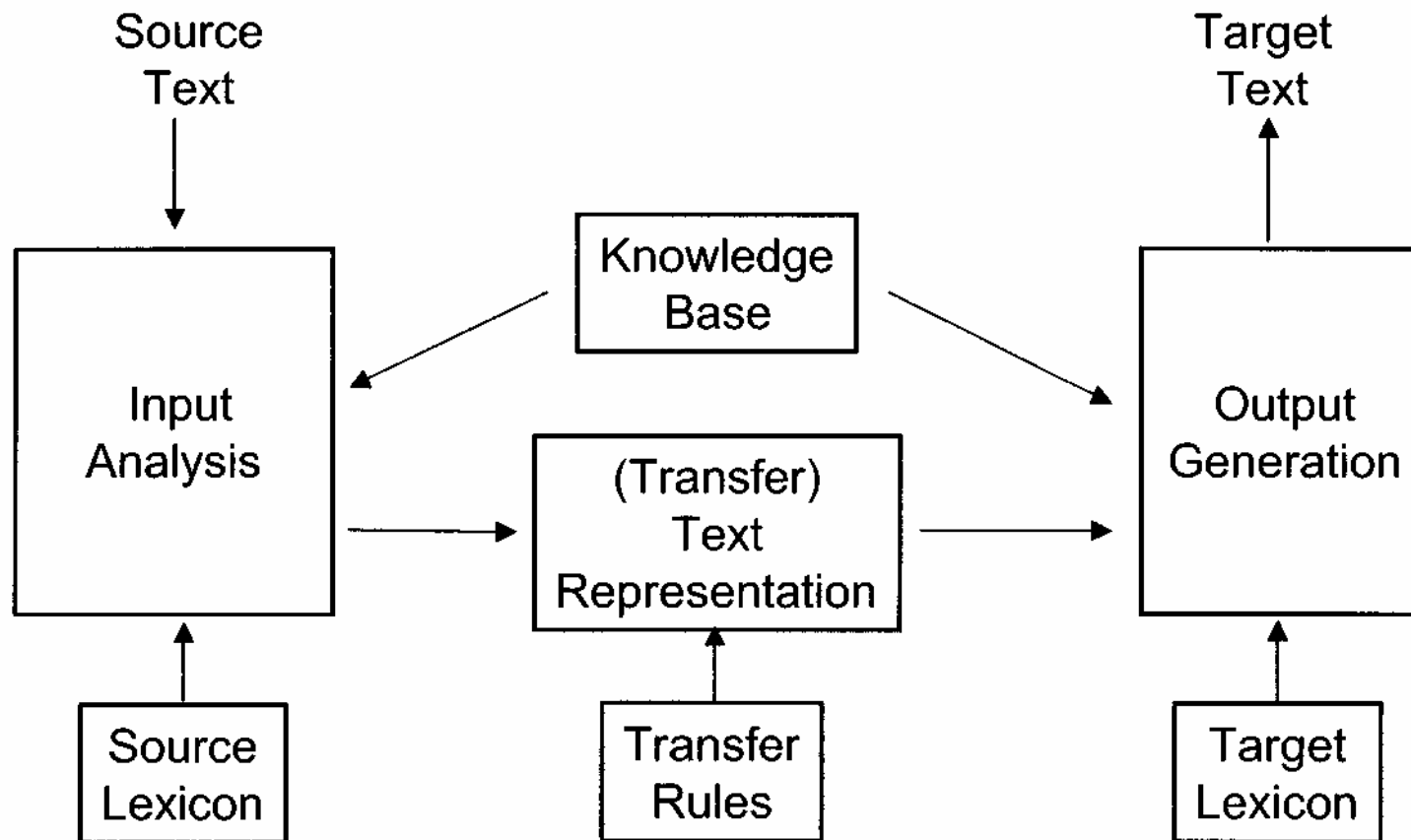
BOSTON MARRIOTT CAMBRIDGE
CAMBRIDGE, MA

8 - 12 AUGUST 2006

Outline

1. Ontologies for MT
2. Ontological Semantics
3. Ontology Construction: The Truth problem
4. Ontology content
 - Comparing Upper Models
 - Middle Models: Words, senses, and concepts
5. Conclusion

Components of MT systems



Some definitions

dream green ideas furiously colorless

- **Syntax:** rules of grammar

colorless green ideas dream furiously

- **Semantics:** meaning

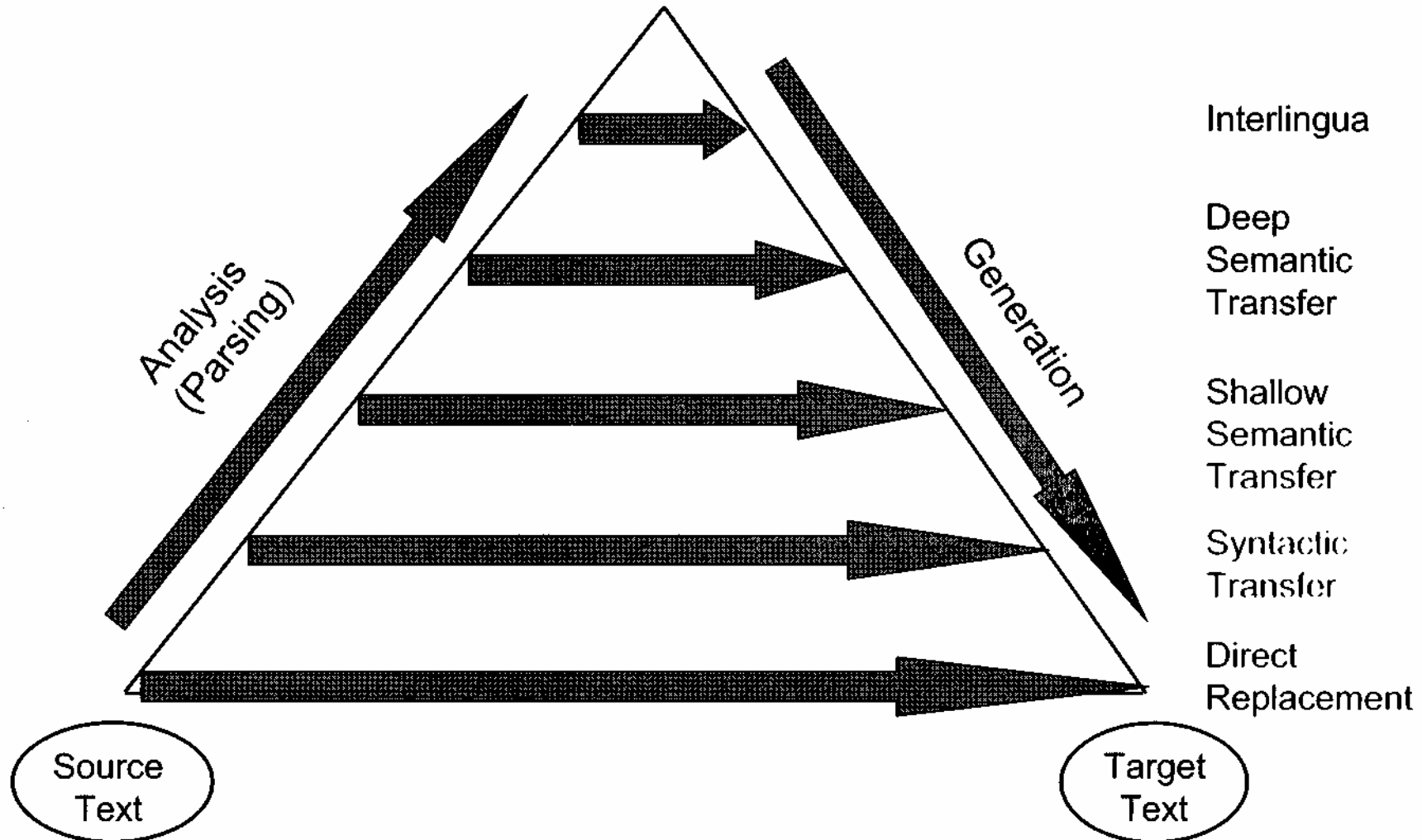
tired old men dream peacefully

- **Lexical semantics:** meaning of each word alone
- **Compositional hypothesis:** can compose meanings to build larger meaning
- **Pragmatics:** meanings from context: situation and interpersonal (author–reader) relationships

Outline

1. Ontologies for MT
2. Ontological Semantics
3. Ontology Construction: The Truth problem
4. Ontology content
 - Comparing Upper Models
 - Middle Models: Words, senses, and concepts
5. Conclusion

Increasing depth of analysis: The Vauquois Triangle



Problems with shallow approaches

- Direct replacement:
 - Keep source language word order:
De student zal de man zien
The student will the man see
- Syntactic transfer:
 - No meaning—just correct grammar
[AUX OBJ VB] → [AUX VB OBJ]
- Shallow semantic transfer:
 - Very little / simple meaning transfer; no pragmatics

E: I like singing

G: Ich singe gern (= I sing gladly/likingly)

E: I am hungry

G: Ich habe hunger (= I have hunger)

E: I have a headache

J: My head hurts

} Need adequate
representations
of meaning

Interlinguas

- For all transfer systems, need $2n.(n-1)$ rules for n languages
- For Interlingual systems, need only $2n$ sets of rules ($\rightarrow IL \rightarrow$)
- Interlingua is the 'deep' semantic notation of the meaning (the idea) behind the text:
 - An Interlingua is a system of symbols and notation to represent the meaning(s) of (linguistic) communications with the following features:
 - language-independent
 - formally well-defined
 - expressive to arbitrary level of semantics
 - non-redundant
- Three parts of an interlingua:
 1. Symbols — 'conceptual lexicon' for meaning aspects:
 - open-class terms: 'concepts'
 - closed-class terms: features and value sets
 2. Formalism — notation and syntax:
 - rules for composing symbols legally
 - rules for writing correctly (where the parentheses go, etc.)
 3. Substrate — knowledge representation system:
 - how concepts are stored; how property inheritance is implemented; etc.

Interlingua example

Source sentence:

“In the following cases, TV sets may overheat.”

Interlingua ‘sentence’:

(*E-OVERHEAT
(MOOD DEC)
(MODAL POSSIBILITY)
(CONDITION
 (*O-CASE
 (NUMBER PLUR)
 (REFERENCE DEFINITE)
 (ATTRIBUTE (*P-FOLLOWING))))
(THEME
 (*O-TELEVISION
 (NUMBER PLUR))))

Source: [KANT system, CMU: Mitamura, Nyberg, Carbonell et al.]

Shallow and deep semantics

- **She sold him the book from her**

(X1 :act **Sell** :agent She :patient (X1a :type ...))

Which roles?

(X2a :act **Transfer** :agent She :patient (X2c :type Book) :recip He)

(X2b :act **Transfer** :agent He :patient (X2d :type Money) :recip She)

How define states and state changes?

- **He has a headache / He gets a headache**

(X3a :prop **Headache** :patient He) (...?..)

How handle relations?

(X4c :type Head :owner He) :state -3)

(X4b

How handle negation?

How handle comparatives?

- **Though it's not perfect, democracy is the best system**

(X4 :type **Contrast** :arg1 (X4a ...?..) :arg2 (X4b ...?..))

More phenomena of semantics

Somewhat easier

Bracketing (scope) of predications
Word sense selection (incl. copula)
NP structure: genitives, modifiers...
Concepts: ontology definition
Concept structure (incl. frames and thematic roles)
Coreference (entities and events)
Pronoun classification (ref, bound, event, generic, other)
Identification of events
Temporal relations (incl. discourse and aspect)
Manner relations
Spatial relations
Direct quotation and reported speech

More difficult

Quantifier phrases and numerical expressions
Comparatives
Coordination
Information structure (theme/rheme)
Focus
Discourse structure
Other adverbials (epistemic modals, evidentials)
Identification of propositions (modality)
Pragmatics/speech acts
Polarity/negation
Presuppositions
Metaphors

Outline

1. Ontologies for MT
2. Ontological Semantics
3. Ontology Construction: The Truth problem
4. Ontology content
 - Comparing Upper Models
 - Middle Models: Words, senses, and concepts
5. Conclusion

What is an ontology?

- “An ontology is the specification of a conceptualization”
— Gruber, 1993
- “A specific artifact designed with the purpose of expressing the intended meaning of a (shared) vocabulary” — Guarino, 2003
- “In philosophy, **ontology** (from the Greek $\omega\nu$ = *being* and $\lambda\omicron\gamma\omicron\sigma$ = *word/speech*) is the most fundamental branch of metaphysics. It is the study of being or existence as well as the basic categories thereof — trying to find out what entities and what types of entities exist. Ontology has strong implications for the conceptions of reality.” — Wikipedia
- “Ontology” dates to 17th century; meta-physics back to Aristotle

Related fields

Libraries and such:
Classification codes
(Dewey, NAICS, SIC, etc.)
⇒ classification / indexing

Philosophy
⇒ understanding the world

Mathematical Logic / AI
⇒ automated reasoning

ontology

Domains:
Biology, Zoology, Medicine, etc.
⇒ development of theories

Linguistics / NLP / MT
⇒ language applications

Databases:
Metadata schemas
⇒ def and management of data

What's inside an ontology?

- **Concepts:** represent a conceptualization; the class of all the examples of that event or entity
- **Relations:** represent a relationship between concepts
- **Axioms:** express a necessary fact holding between concepts and relationships
- **Instances:** represent a specific individual

*happiness,
children*

*colour-of,
location-of*

*if X is mortal
then X will die
one day*

John Lennon

T-Box

A-Box

...but what about Beethoven's 9th symphony?

The OL layer cake

$\forall x, y (sufferFrom(x, y) \rightarrow ill(x))$

Rules

cure(dom:DOCTOR,range:DISEASE)

Relations

is_a(DOCTOR,PERSON)

Concept Hierarchies

DISEASE:=<I,E,L>

Concepts

{disease,illness}

Synonyms

disease, illness, hospital

Terms

(Expressed in Ontology Learning challenge problem, 2005: Buitelaar et al.)

Content building steps

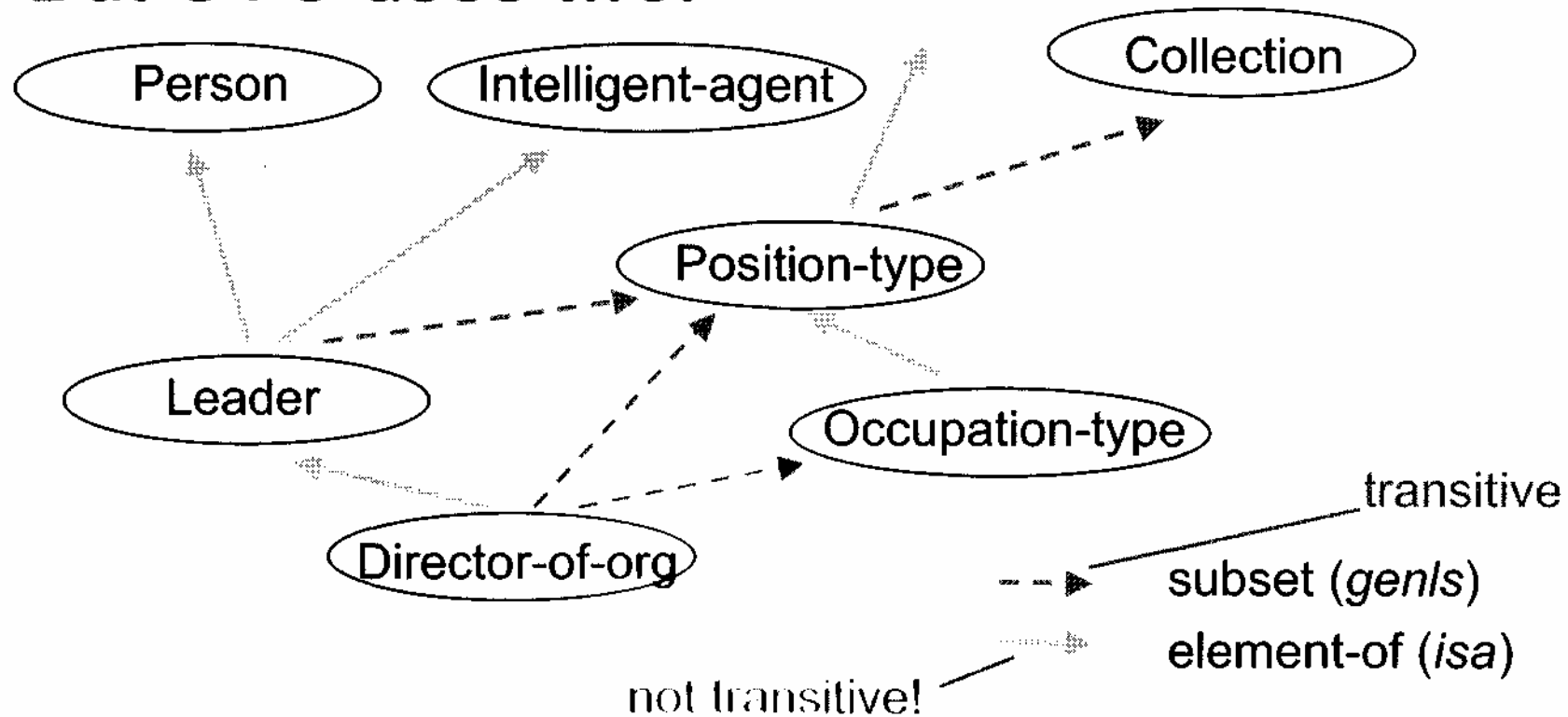
- List **terms** that denote the entities, events, qualities, relationships, etc. in the domain
 - Link them using one or more **relations**:
 - *structuring* relations (subsumption, others)
 - *definitional* relations
 - *additional info* relations
 - Define **axioms** and properties
 - rules that specify what must be true about what
 - Provide **additional information** resources:
 - lexicons, glossaries, documentation, etc.
- Terminology
'ontology'
(e.g.,
WordNet)
- 'True'
ontology
- External
resources

The main relation: Subsumption

(also called *a-kind-of*, *IS-A...*)

Most people use only ISA

But CYC uses two:



Reasoning in ontologies

- **Axioms** and inference
 - Rules that allow one / the KR system to add knowledge automatically
- **Relation/slot filler requirements** that enable reasoning
 - Conceptual type (*:color* filled by a *Color*)
 - Number (*:number-of-wheels* of *Car* = < 4)
 - Value on scale (*:happiness-degree* $\in [-1,+1]$)
- **Property inheritance**
 - Properties defined for superordinate concepts inherited by subordinate ones (overridden?)
 - Property requirements also inherited (or not?)
- **Automated classification**
 - Given a concept or instance with properties but no specified *isa* subsumer(s), a classifier finds where it belongs in the ontology (the most general location)

OWL example: simple axiom

- A number restriction on a slot:
 - *Every human has two parents*
- Plus an assertion:
 - *Socrates is a human*
- Allows an inference:
 - *Therefore Socrates has two parents*

OWL representation
(from John Sowa)

“Every human has two parents”:

```
<owl:Class rdf:about="#Human">
```

```
<rdfs:subClassOf><owl:restriction>
```

```
<owl:cardinality
```

```
  rdf:datatype="&xsd:nonNegativeInteger">2
```

```
</owl:cardinality>
```

```
<owl:onProperty rdf:resource="#Parent">
```

```
</owl:restriction></rdfs:subClassOf></owl:Class>
```

we're talking
about Humans:

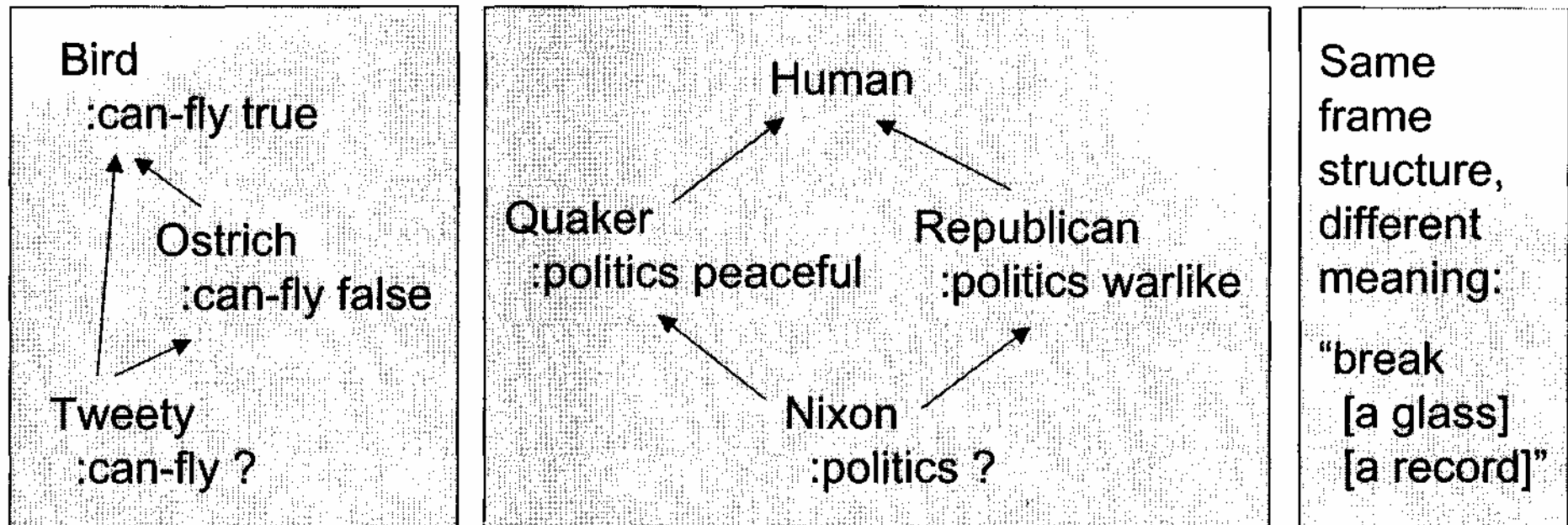
we're making a
restriction:

the cardinality
is 2...

...on the Parents

Tweety and the Nixon Diamond

- Problems with values inherited from (multiple) parents
- Defeasible reasoning; various logics that operate over the ontology's symbols and structure



Cognitive Scientists and Linguists don't care
AI people unhappy

Linguists happy
Cognitive Scientists and AI people unhappy

Axiomatizations and axiom sets

- Upper levels—some creation efforts:
 - Time (Allen 80s; Hobbs et al. 05–; Pustejovsky et al. TIME-ML 04–05)
 - Space
 - Meronymy (Guarino et al. 00)
- Middle zone—not many; too much work!
 - But: Extended WordNet, a large axiom set derived automatically from WordNet glosses (Moldovan et al.)
- Domain ontologies:
 - Axioms are typical here; they *define* the structure and behaviour of the domain

Outline

1. Ontologies for MT
2. Ontological Semantics
3. Ontology Construction: The Truth problem
4. Ontology content
 - Comparing Upper Models
 - Middle Models: Words, senses, and concepts
5. Conclusion

Now we're ready for an exercise...

Create your own ontology of the following 33 words:

apples, beans, beef, bread, breakfast cereal, cabbage, cake, carrots, cauliflower, cheese, cookies, cucumbers, custard, eggs, ground beef, herring, mushrooms, peaches, peas, pickled herring, pies, plums, pork, potatoes, pudding, rice, sausages, scrambled eggs, toast, tomatoes, veal, wheat, yoghurt

What did you do?

- The easy part:
 - *vegetables, fruits, meats...* but what of *tomatoes*? Is your experience right, or are the biologists right?
- Or, what about:
 - *starches, proteins, greens...* this is what's **inside**; is this a better organization? Should you have both?
- The harder part:
 - *Eggs and scrambled eggs; milk and cheese; pies...* Methods of preparation — define somewhere else, and then somehow apply this to the basic foodstuffs?
- What is right?
 - If I organized them by **color** or **size**, would that be wrong? By **sweetness**? What if I were diabetic?

Decisions when ontologizing

1. Should you create the term?
 2. Where should the term go relative to the other terms? — *species*
 3. What is special/unique/different about this term? — *differentium/ae*
- *How do you know you're right?*
 - *How do you decide between two or more alternatives?*

Five styles of truth

1. Abstraction and feature combination: *the philosophers*
 2. Intuitive distinctions: *the cognitive scientists*
 3. Inference-based organization: *the computational people*
 4. Cross-linguistic phenomena: *the linguists*
 5. Domain analysis: *the domain specialists*
- Taxonomic clarity: *everyone*



Example philosopher: Aristotle

- The gold plated KR approach:

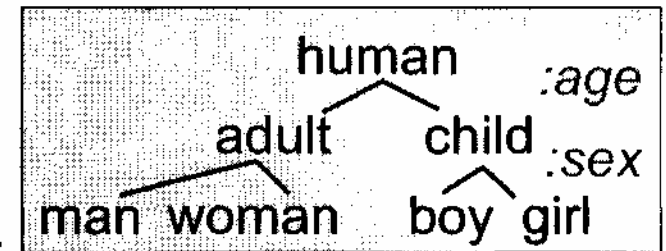
- Find a primitive concept—undefined
- Specialize it in various ways by adding various differentiae
- Define these differentiae elsewhere in the ontology
- Don't confuse **definitional** aspects with mere **properties!**

- An *apple* is-a *fruit* with essential differentium *XXX* and with properties *:colour=red, :size=tennis-ball-sized...*

human

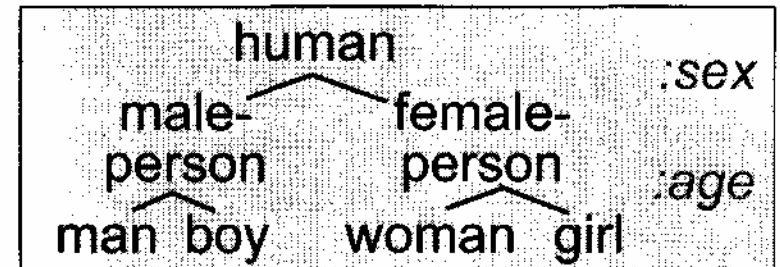
man, woman
:sex

:sex, :animacy...



- Problems:

- What are the differentiae?
- How do you order them?



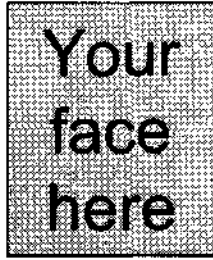


Example cognitive scientist: Rosch

- Functional purpose of classes: “provide maximum information with the least cognitive effort”
- Established experimental paradigms for determining subjects’ ratings of how good an example of a category a member is judged to be
- **Basic Level** categories:
 - A basic category is the largest class of which we can form a fairly concrete image, like *chair* or *ball*. These are the first classifications that children make
 - Superordinate categories are collections of basic categories: *furniture* includes chairs, lamps, desks, beds, etc.; *toys* include balls, dolls, furry animals. No one object clearly represents them
 - Subordinate categories represent divisions of basic classes (*deck chairs, bar stools, teddy bears, school desks*)

The problem of categories: The Prototype Theory view

- Traditional theory: people categorize using the common features of the members (differentiae)
- Rosch observations:
 - (1) When people categorize, they cannot tell you what features they are using — often don't know the differentiae!
 - (2) When people categorize, they usually find some members of categories more “typical” (“better”) than others (e.g., a *robin* is a better member of the category *Bird* than an *ostrich*)
 - (3) When people categorize, they categorize more typical members more quickly than less typical ones
- Rosch suggestion:
 - Create ‘star structure’ of prototypes rather than (or in addition to?) a subsumption hierarchy with differentiae



Example computationalist: you?

- Have you ever written a program and created several data types?
 - Typical MT termsets:
 - Part of speech tags
 - Syntactic categories
 - Named entity categories (Person, Organization, Numerical-expression, Location, Time-expression, etc.)
- You define a small set of categories
- Your program does different things with the different categories
- Have you ever made the set of termsets explicit? — your first ontology

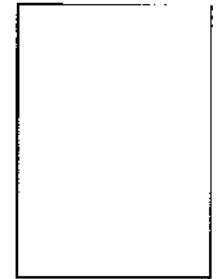


Example (cognitive) linguist: Lakoff

- Create classes according to the way one (or more) language(s) behave(s):
 - Classes of noun
 - Classes of verb
- Do they make conceptual categories? How do we judge?
- E.g., Dyirbal noun categorization:
 - Class I: *human males + storms, rainbow* (from myths) + *fish* (and so also fishing tools) + *moon* (husband of the sun) + ...
 - Class II: *human females + birds* (myth: because they have female spirits) + *sun* (wife of the moon) + *fire* (associated with sun) + *hot things* (experienced like fire) + ...
 - Class III: *edible plants*
 - Class IV: the rest
- E.g., Hopi time categorization



WordNet: Miller and Fellbaum



- Cognitive scientists at Princeton University
- Word hierarchy built by hand during 1980s, using dictionaries and manual insight
- Approx. 110,000 nodes at present:
 - Synonym, antonym, part-of links; examples; frequencies
 - WordNets for other languages: EuroWordNet (Vossen et al.): Dutch, Italian, Spanish, English
 - Global WordNet: see <http://www.globalwordnet.org/>
 - Hierarchy info:
 - Noun hierarchy depth ~12
 - Verb hierarchy depth ~3
 - Adjective/adverb not in hierarchy, but in star structure
 - Almost no top-level structure
- Freely available: <http://wordnet.princeton.edu/>
- Extensively used in CompLing, but not very useful yet
 - Except: definitions converted to axioms and used for theorem proving in automated QA (Moldovan et al.)



Example linguist: Matthiessen

- Penman NL generation system (ISI, 1979–1997, with Bill Mann and others); KPML (various; 1995–; John Bateman):
 - Systemic-functional Linguistics grammar and system
 - Penman Upper Model: taxonomy (network) of approx. 300 terms
 - Input representation terms defined in Domain Model; connected to Upper Model
 - For NLG, many grammar decisions determined by very general categories capturing English structure and word behavior:
 - Nouns / verbs (of various types) / adjectives
 - Count nouns / mass nouns
 - Tenses etc.
- Upper Model nodes represent conceptual-grammatical categories: at interface of language and world

Domain specialist examples

- Computational / expert systems:
 - **Protégé Ontologies Library**: Stanford University's collection of 18 influential ontologies (<http://protege.stanford.edu/ontologies/ontologies.html>)
 - **OntoSelect**: over 700 ontologies in various domains (<http://views.dfki.de/Ontologies/>)
- Medical:
 - **UMLS: Metathesaurus** (over 1 mill biomedical concepts and 5 mill concept names from over 100 controlled vocabularies and classifications (some in multiple languages) used in patient records, administrative health data, bibliographic and full-text databases, expert systems), the Semantic Network (*isa* for type hierarchy; *physically related*, *spatially related*, *temporally related*, *functionally related*, *conceptually related*), and the SPECIALIST lexicon (<http://www.nlm.nih.gov/research/umls/>)
- Industrial etc.:
 - **NAICS** (North American Industry Classification System): numerical classifications of construction, agriculture, technology, wholesale, retail, industry, etc., (<http://www.census.gov/epcd/www/naics.html>)



Domain specialists

- Is a *dolphin* a *mammal* or a *fish*?
- Is a *steelhead trout* a *salmon* or not?
- When is someone *Jewish*?

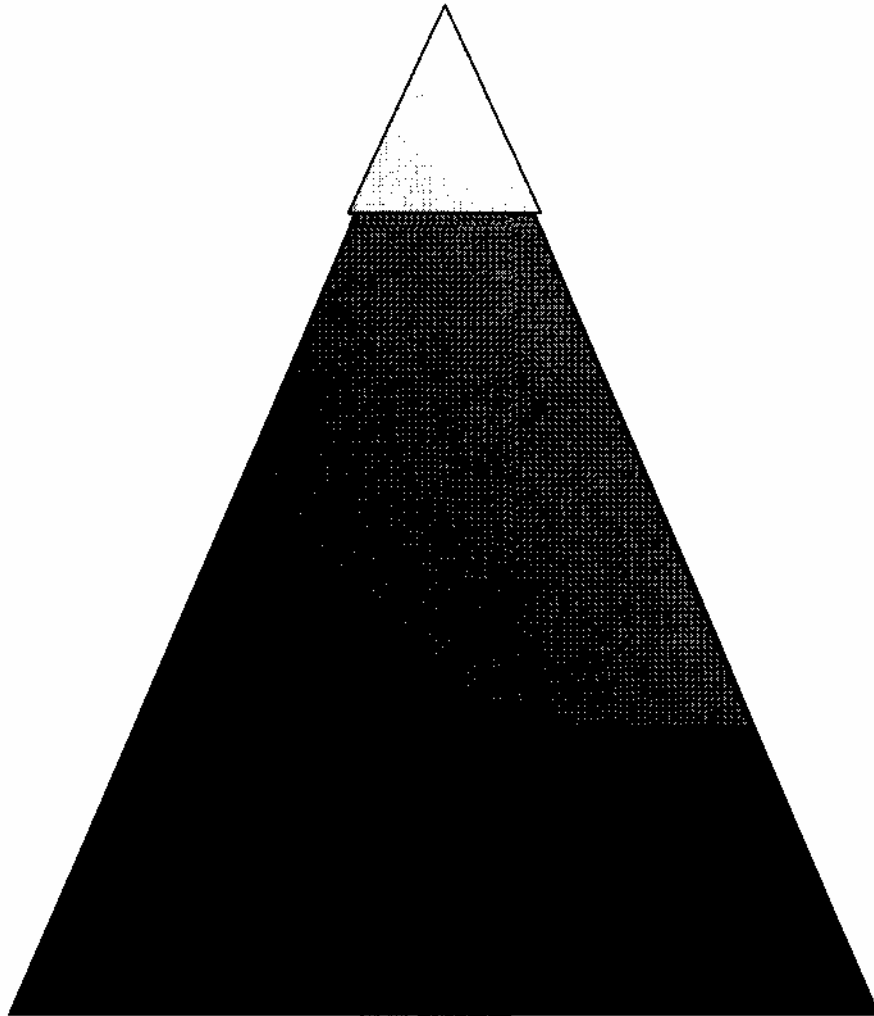
- Which features are the determinate ones? Why? Who decides?

There is no **authority**: it can be tradition, the law, social consensus, or simply ad hoc purpose-driven. The point is to know which you adopt and to **be careful and consistent**.

Outline

1. Ontologies for MT
2. Ontological Semantics
3. Ontology Construction: The Truth problem
4. Ontology content
 - Comparing Upper Models
 - Middle Models: Words, senses, and concepts
5. Conclusion

The 'zones' of ontologies



- **Upper Model:**
 - Between 100 and 500
 - Most abstract generalizations
 - Used for overall organization
 - Very general axioms
 - Not really lexicalized
 - Built by theoreticians: philosophy, AI, KR
- **Middle Model:**
 - Between 500 and 100,000
 - The world in general
 - Used for NLP: IR, QA, etc.
 - Usually not axiomatized: too much work!
 - Built by cognitive scientists and linguists
- **Domain Model:**
 - Between 200 and 2,000
 - Specific domain concepts
 - Used mainly for domain reasoning
 - Often highly axiomatized
 - Built by domain experts and system builders, often jointly

Questions

1. What do you include in the Upper Model, and what in the Middle Model?
2. Where is the boundary?
3. How 'primitive' are your concepts?
What granularity should you use?

Parsimonious vs profligate

Parsimonious

- Few symbols
- Easy to see conceptual relatedness
- Easy to define and run inferences
- Hard to compose complex meanings

There is no 'correct' position: what you choose depends on how much inference you need vs how complex your domain is

Profligate

- Many symbols
- Hard to determine conceptual relatedness
- Hard work to define inferences
- No need to compose complex meanings
- Easy to fall into the trap of semantics-by-capitalization (or '*wishful mnemonics*': McDermott: Artificial Intelligence Meets Natural Stupidity, 1981)



CYC

- Creator: Lenat (CYCorp, Austin Texas); since 1990s
- CYC: largest and richest ontology; millions of axioms
- ResearchCYC: 25571 concs
- RCYC (which was translated into RDF) omits all second order concept expressions, (for example functional operator expression) and so it has a lot of missing supers
- R-CYC in Omega:
 - Lexical items not in Omega yet
 - Missing supers require dummy root "Protocol Root" which now has 95 arbitrary child concepts

Mammoth - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back

Address http://omega.isi.edu:8888/index

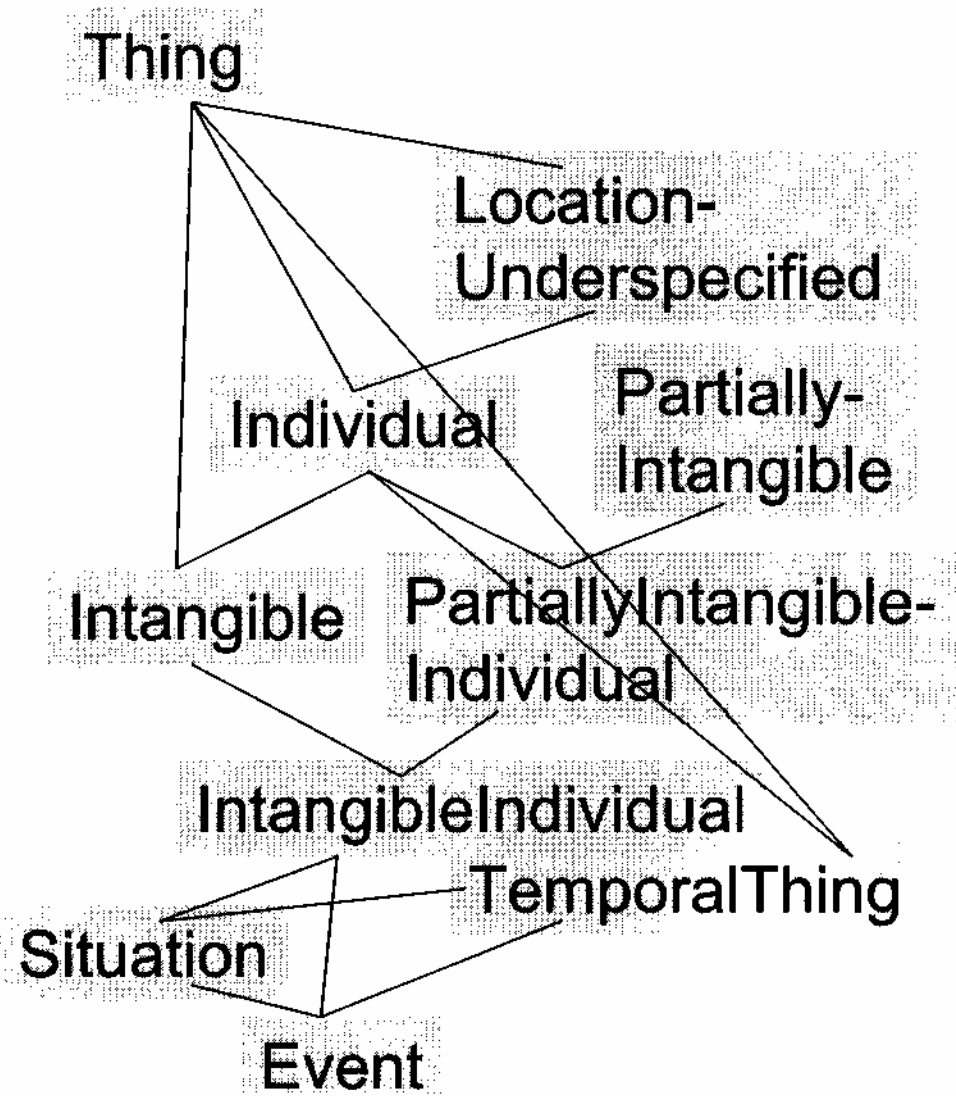
Find: word Language(s):

Concept: **CYC Protocol Root**

Direct Subclass:

- VariedOrderCollection**
- Unaccounted**
- TimeIndependentCollection**
- TimeDependentCollection**
- Thing**
- ThingTypeByCapability**
- TheTerm**
- TheCOASpecification**
- TestingConstant**
- TestQueryConstant**
- TermPhrasesConstraint**
- TangibleObjectTypeByGenericDamageNonScalar**
- TKBTemplateIndividualAffiliation**
- System**
 - FunctionalSystem**
 - EducationalSystem**
 - Ecosystem**
 - EconomicSystem**
 - ClassificationSystem**
 - AudioSystem**
- SubLTemplate**
- SubLExpression**
- SocialStatusCollectionType**
- SetOrCollectionTypeByCardinality**
- SecurityClearance**
- ScriptTypeByStructuralOrderFeature**
- SKSExternalTermDenotingFunction**

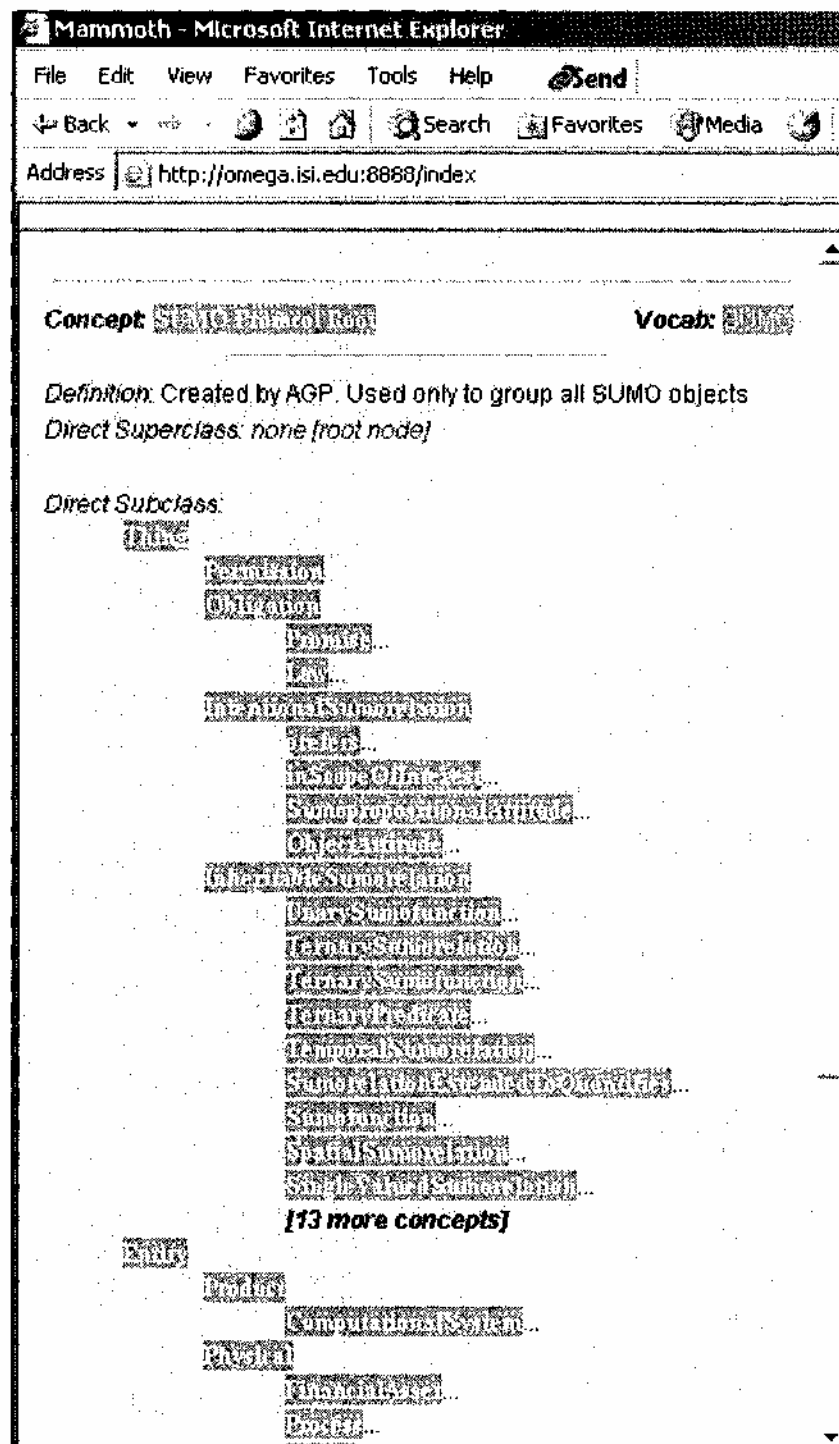
Top of ResearchCYC



- Largest and most developed ontology
- Principally aimed at AI: inference-heavy and NL-light
- Very tangled network; hard to understand and use unless you absorb the CYC's philosophy / methodology
- Full CYC, for sale: over 1M axioms
- Has been tried by many research groups in the past; successful adoption rate low

SUMO

- Creator: Pease and consortium; recent (USA)
- Suggested as standard: 'Suggested Upper Model Ontology'
- 1653 concepts
- No lexical items
- Adopts more traditional KR-style / Description Logic approach, with lots of internal reasoning mechanism constructs
- More hierarchical than RCYC or DOLCE, with few uplinks pointing to expressions
- Omega version not complete and a little 'buggy'



Some top parts of SUMO

- Entity
 - Physical
 - Object
 - Process
 - Financial Asset
 - Abstract
 - Quantity
 - Graph
 - Attribute
 - Thing
 - Permission
 - Obligation
 - InheritableSumorelation
 - IntentionalSumorelation
 - BinarySumoRelation
 - CaseRole
 - ...
 - Product
 - Computational System...
- Caserole
 - Agent-rel
 - Patient
 - Result
 - Resource
 - ResourceUsed
 - Instrument
 - ComputerRunning
 - StandardOutputDevice
 - ...
 - DataProcessed
 - Experiencer
 - Origin
 - Destination
 - Direction
 - Path

CYC *Event* and SUMO *Process*

Definition: An important specialization of *Situation* and thus also of *IntangibleIndividual* and *TemporallyExistingThing* (qq.v). Each instance of *Event* is a dynamic situation in which the state of the world changes; each instance is something one would say happens. Events are intangible because they are changes per se, not tangible objects that effect and undergo changes. Notable specializations of *Event* include *Event-Localized*, *PhysicalEvent*, *Action*, and *GeneralizedTransfer*. *Events* should not be confused with *TimeIntervals* (q.v.). The temporal bounds of events are delineated by time intervals, but in contrast to many events time intervals have no spatial location or extent.

Definition: Intuitively, the class of things that happen and have temporal parts or stages. Examples include extended events like a football match or a race, actions like *Searching* and *Reading*, and biological processes. The formal definition is: anything that lasts for a time but is not an *Object*. Note that a *Process* may have participants 'inside' it which are *Objects*, such as the players in a football match. In a 4D ontology, a *Process* is something whose spatiotemporal extent is thought of as dividing into temporal stages roughly perpendicular to the time-axis.

Outline

1. Ontologies for MT
2. Ontological Semantics
3. Ontology Construction: The Truth problem
4. Ontology content
 - Comparing Upper Models
 - Middle Models: Words, senses, and concepts
5. Conclusion

Toward building a Middle Model

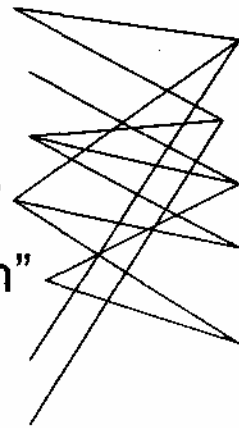
- You can start with words (terms in the domain, in the dictionary, etc.)...
- ...but you have to bring together synonyms and separate out word senses...
- And then you have to group similar meaning clusters for later ontology inheritance and inference...
- What does this all look like?

From lexemes to concepts

Lexical space

- Words
- Monolingual

- “drive”
- “steer”
- “fahren”
- “steuern”
- “besturen”
- “rijden”
- “drijven”
- ...



Sense space

- Word senses
- Multilingual

- Drive1
- Drive2
- Drive3
- ...
- Manage

Concept space

- Concepts
- Interlingual (?)

?

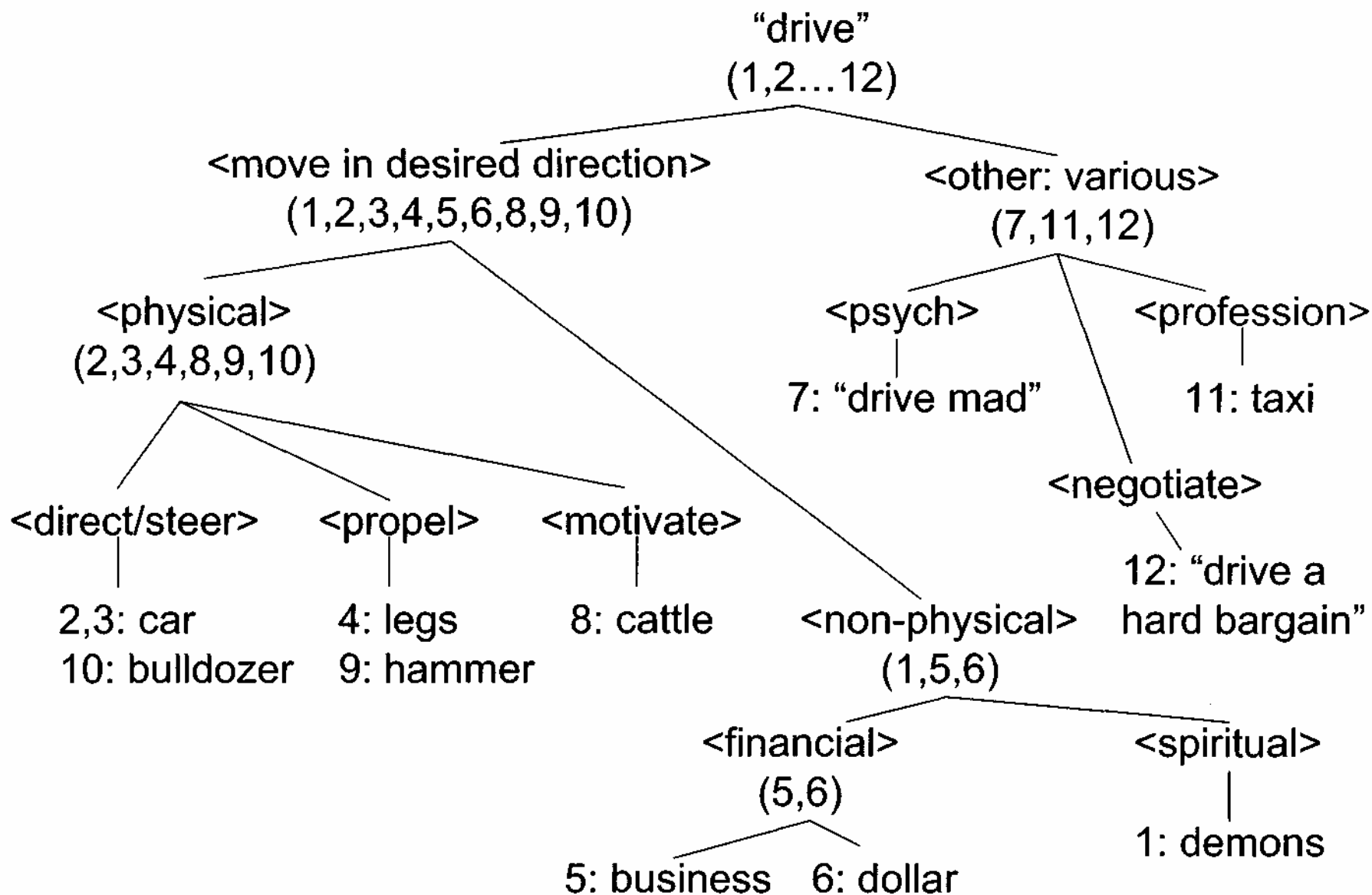
- How many concepts?
- How related to senses?

1. *Drive the demons out of her and teach her to stay away from my husband!!*
2. *Shortly before nine I drove my jalopy to the street facing the Lake and parked the car in shadows.*
3. *He drove carefully in the direction of the brief tour they had taken earlier.*
4. *Her scream split up the silence of the car, accompanied by the rattling of the freight, and then Cappy came off the floor, his legs driving him hard.*
5. *With an untrained local labor pool, many experts believe, that policy could drive businesses from the city.*
6. *Treasury Undersecretary David Mulford defended the Treasury's efforts this fall to drive down the value of the dollar.*
7. *Even today range riders will come upon mummified bodies of men who attempted nothing more difficult than a twenty-mile hike and slowly lost direction, were tortured by the heat, driven mad by the constant and unfulfilled promise of the landscape, and who finally died.*
8. *Cows were kept in backyard barns, and boys were hired to drive them to and from the pasture on the edge of town.*
9. *He had to drive the hammer really hard to get the nail into that plank!*
10. *She learned to drive a bulldozer from her uncle, who was a road maker.*
11. *I used to drive a taxi (for work) before I went to night school.*
12. *Beware—Ralph drives a hard bargain; you will probably lose all your money.*

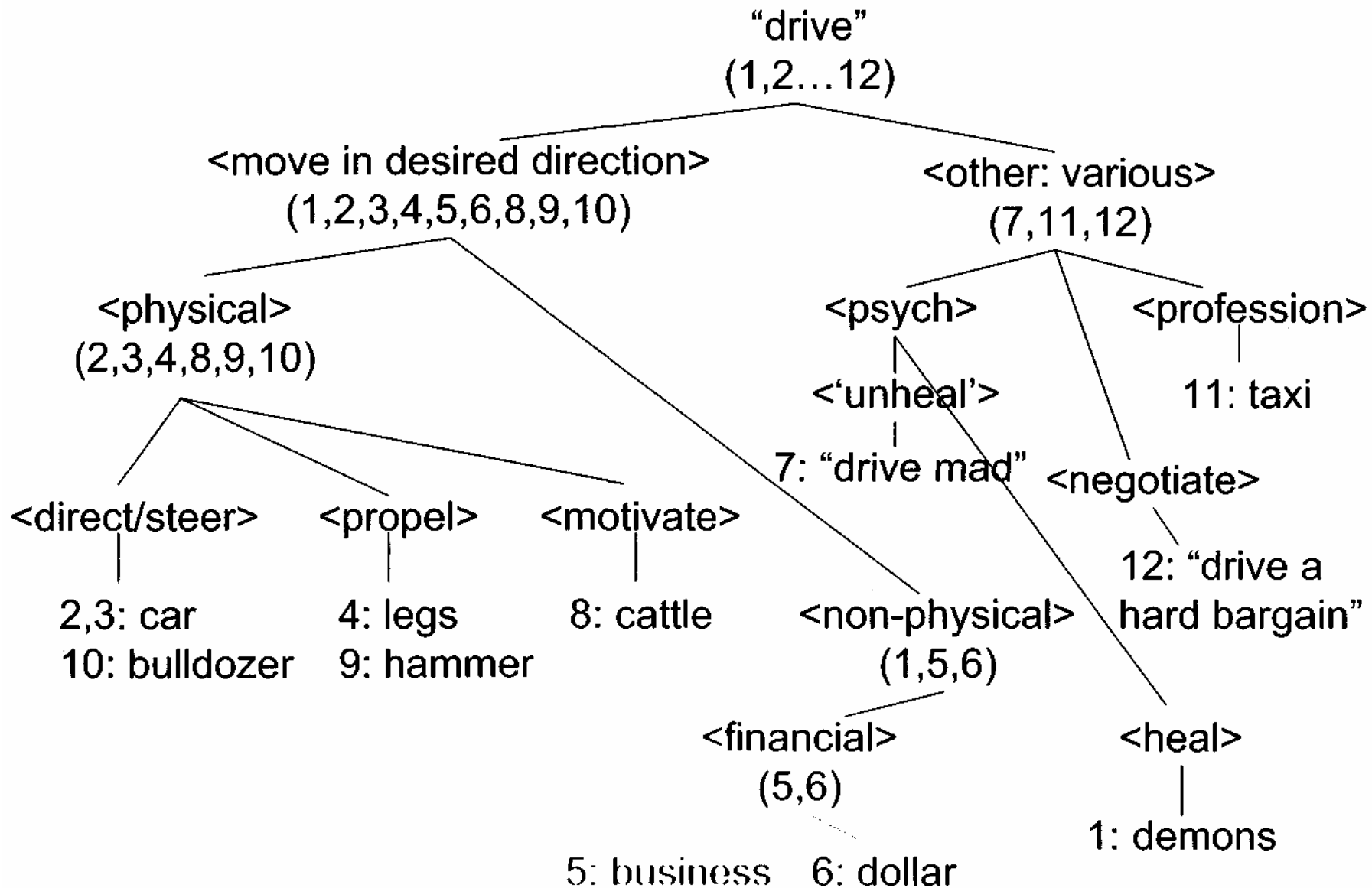
From senses to concepts: Graduated refinement

1. **Initialization:** Given a term (word), collect several dozen sentences containing it. Also collect definitions from various dictionaries
2. Cluster the word's senses into preliminary, loosely similar groups
3. **Differentiation process:** Begin a tree structure with all the groups at the root
4. Considering all the groups, identify the group most different from the others
 1. If you can find one clearly most different group, write down its most important distinction explicitly — this will later become the differentium and be formalized axiomatically
 2. If you cannot find any distinctions by which to further subdivide the group, stop elaborating this branch and continue with some other branch
 3. If you can find several distinctions that subdivide the group in different, but equally valid, ways, also stop elaborating this branch and continue with some other branch
5. Create two new branches in the evolving tree structure, putting the new group under one, and leaving the other groups under the other
6. Repeat from step 4, considering separately the group(s) under each branch
7. **Concept formation:** When all branches have stopped, the ultimate result is a tree of increasingly fine-grained distinctions, which are explicitly listed at each branch point. Each leaf becomes a single concept, not further differentiable in the current task/application/domain. Each distinction must be formalized as an axiom that holds for the branch it is associated with
8. **Insertion into ontology:** Starting from the top, visit each branch point. Do the two branches have approximately the same meaning?
 1. If so, insert them into the ontology at the appropriate point and stop traversing this branch
 2. If not, split the tree and repeat step 8 separately for each branch. Repeat until done

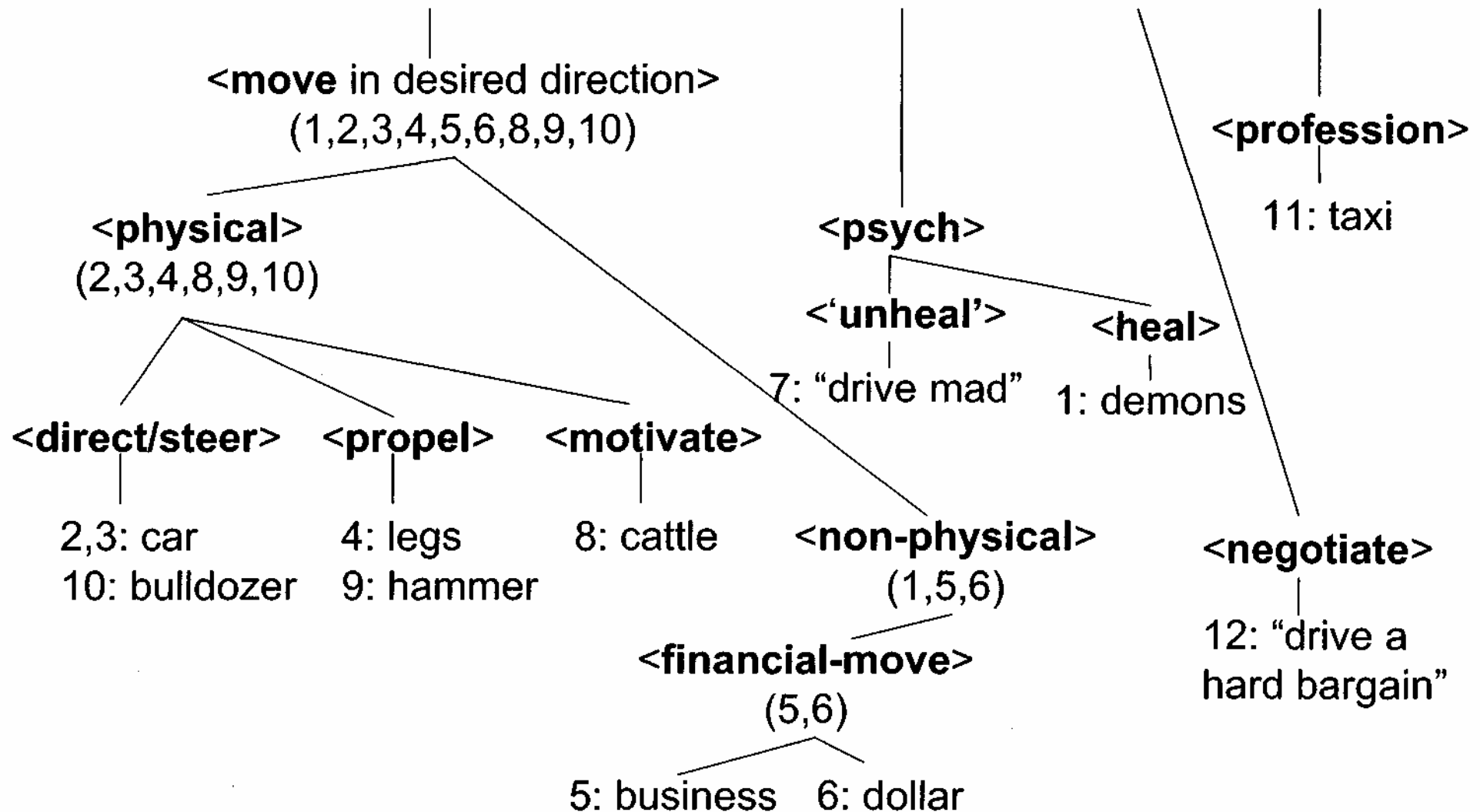
An exercise: “drive”



Deeper semantic “drive”



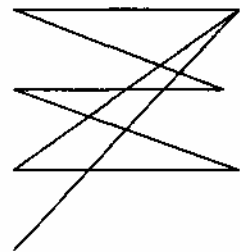
Ontologizing “drive”



From lexemes to concepts

Lexical space

- Words
- Monolingual
- “drive”
- “steer”
- “fahren”
- “rijden”
- ...



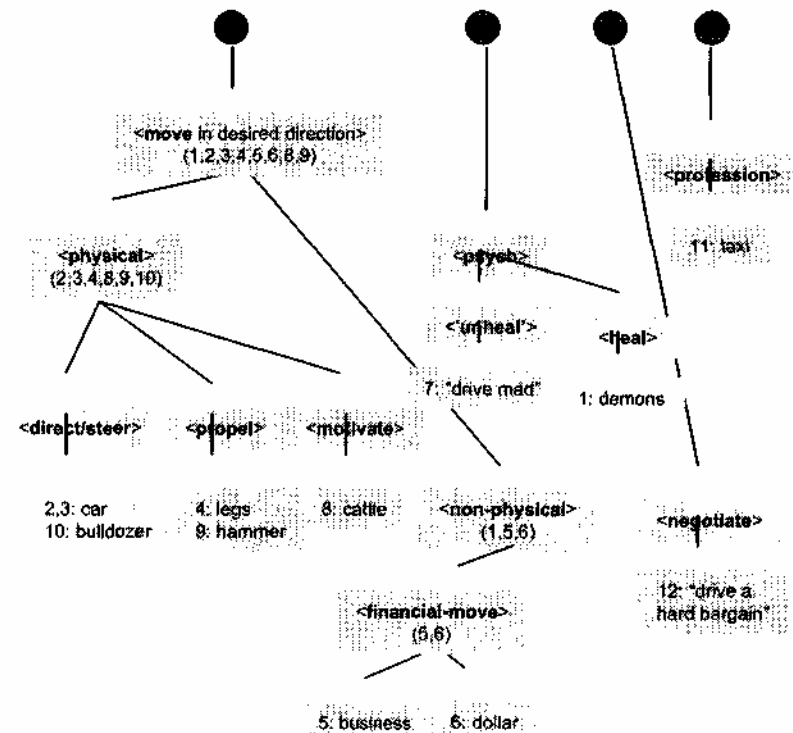
Sense space

- Word senses
- Multilingual
- Drive1
- Drive2
- Drive3
- ...

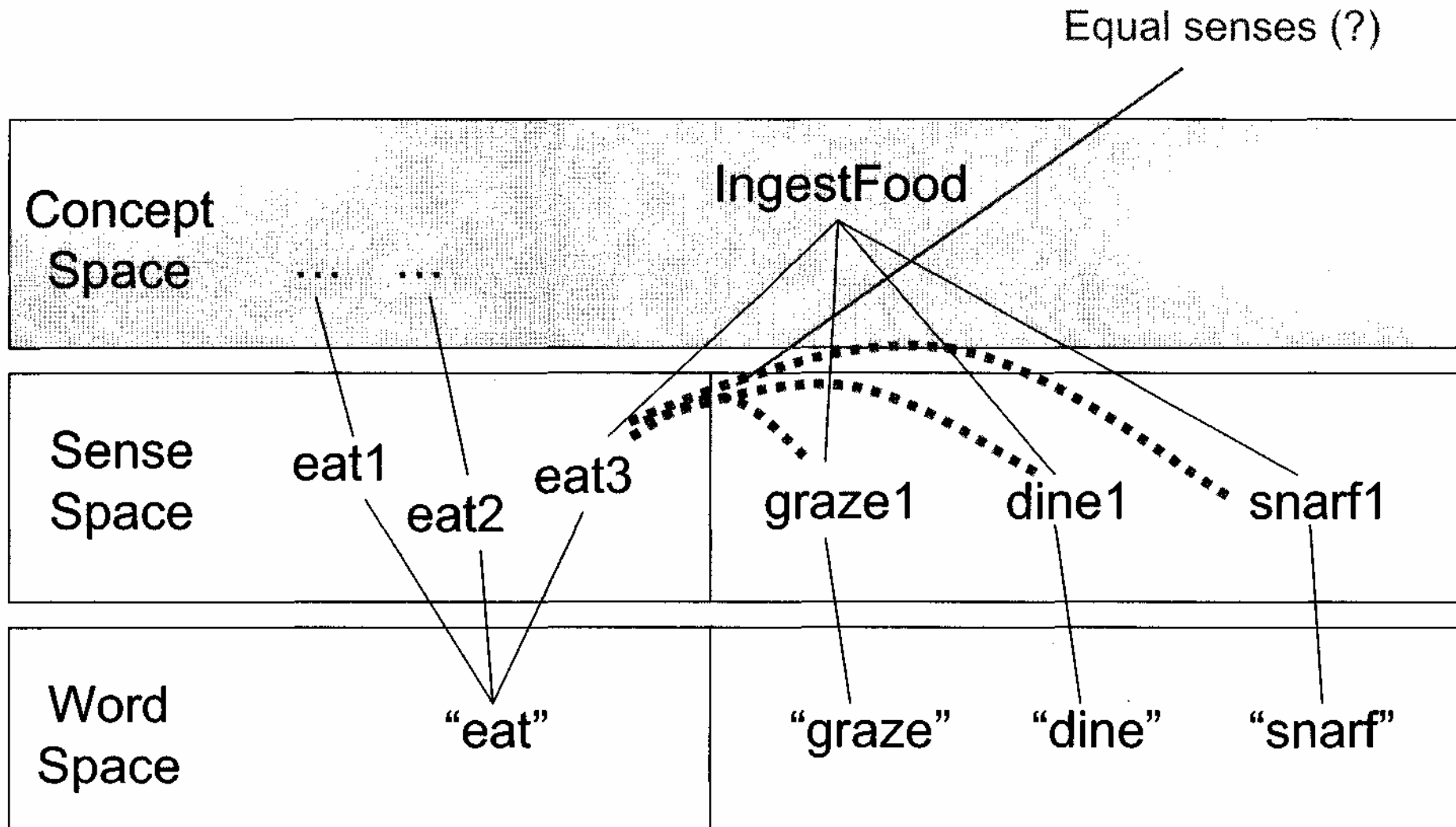
Concept space

- Concepts
- Interlingual (?)

- Graduated granularity: choose
- Generally fewer concepts than senses
- Complex sense-concept mappings

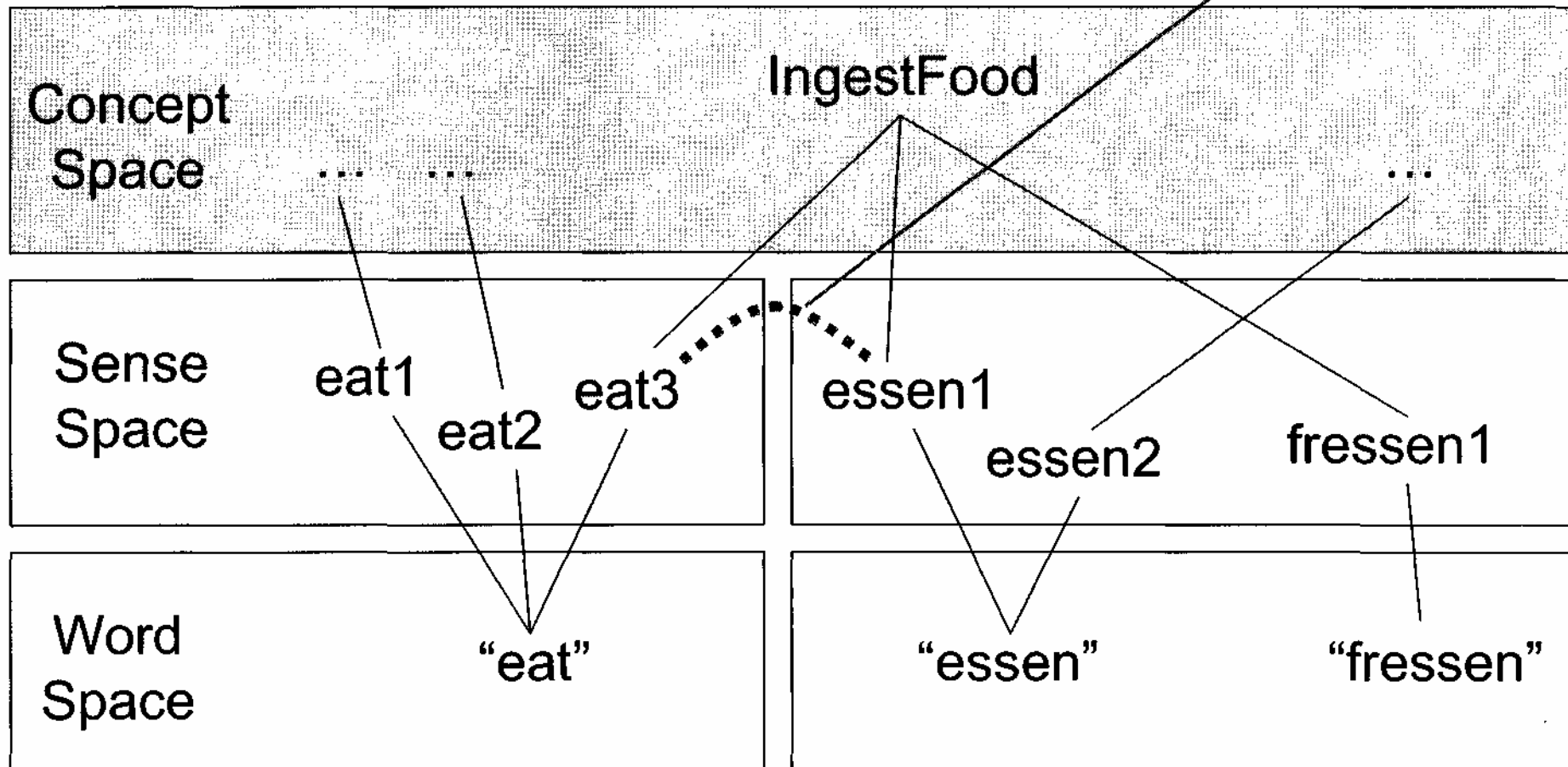


Multilinguality: Word-sense-concept 1



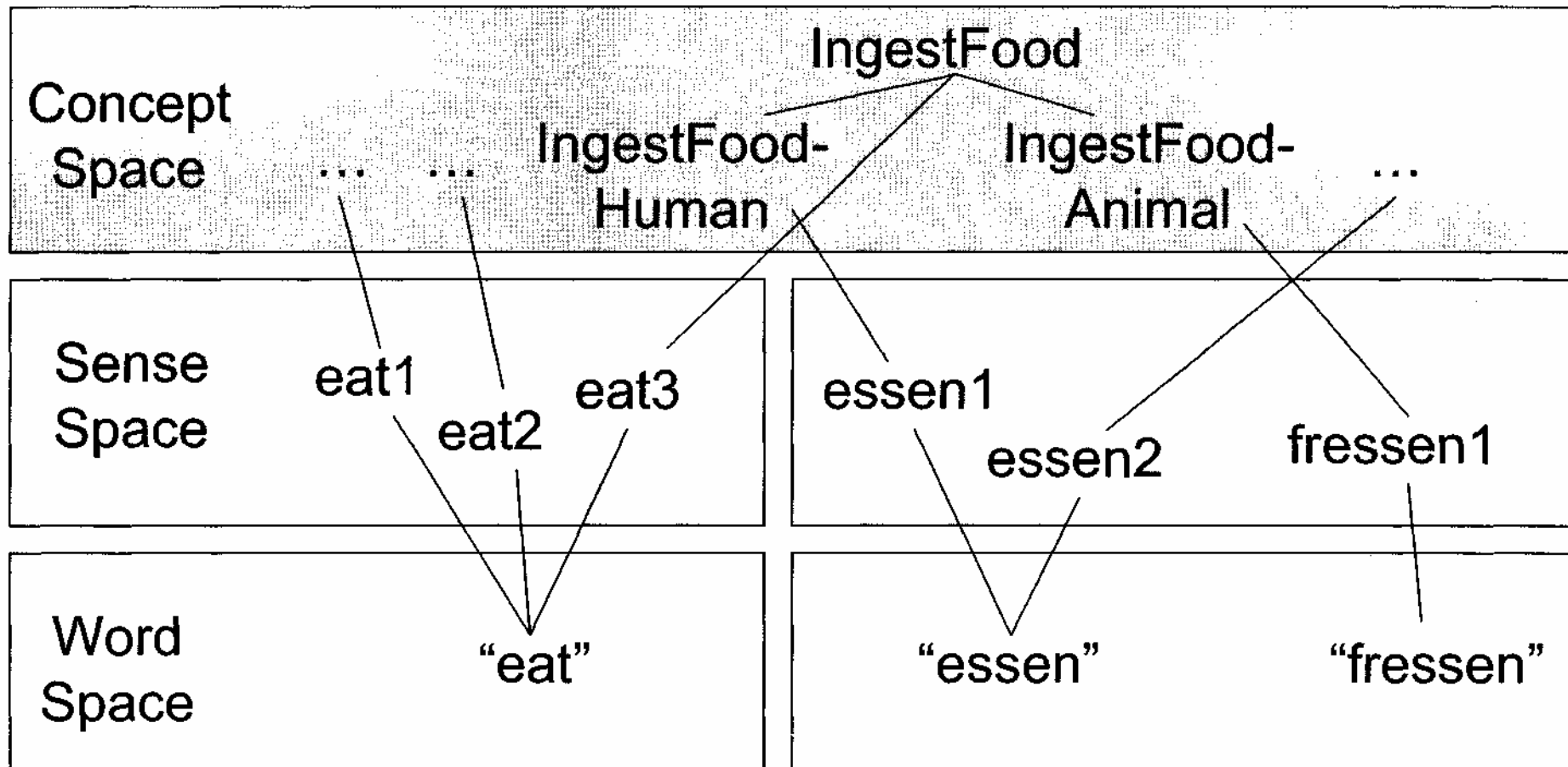
Multilinguality: Word-sense-concept 2

Not necessarily
equal senses!



Multilinguality: Word-sense-concept 3

Other languages may suggest refined conceptualization



Outline

1. Ontologies for MT
2. Ontological Semantics
3. Ontology Construction: The Truth problem
4. Ontology content
 - Comparing Upper Models
 - Middle Models: Words, senses, and concepts
5. Conclusion

Omega content and framework

Goal: one environment for various ontologies and resources

- Concepts: 120,604 Concept/term entries [76 MB]:
 - WordNet (Princeton; Miller & Fellbaum)
 - Mikrokosmos (NMSU; Nirenburg et al.)
 - Penman Upper Model (ISI; Bateman et al.)
 - 25,000+ Noun-noun compounds (ISI; Pantel)
- Lexicon / sense space:
 - 156,142 English words; 33,822 Spanish words
 - 271,243 word senses
- 13,000 frames of verb arg structure with case roles:
 - LCS case roles (Dorr) [6.3MB]
 - PropBank roleframes (Palmer et al.) [5.3MB]
 - Framenet roleframes (Fillmore et al.) [2.8MB]
 - WordNet verb frames (Fellbaum) [1.8MB]
- Associated information (not all complete):
 - WordNet subj domains (Magnini & Cavaglia) [1.2 MB]
 - Various relations: learned from text (ISI; Pantel)
 - TAP domain groupings (Stanford; Guha)
 - SemCor term frequencies [7.5MB]
 - Topic signatures (Basque U; Agirre et al.) [2.7GB]
- Instances [10.1 GB]:
 - 1.1 million persons harvested from text
 - 765,000 facts harvested from text
 - 5.7 million locations from USGS and NGA
- Framework (over 28 million statements of concepts, relations, & instances):
 - Available in PowerLoom
 - Instances in RDF
 - With database/MYSQL
 - Online browser
 - Clustering software
 - Term and ontology alignment software

http://omega.isi.edu/doc/browsers.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Home Search Favorites Media

Address http://omega.isi.edu/doc/browsers.html

Find: Language(s): NAME ANNOT ES EN Namespace(s): O3 PFRM POLCE XSD Style: D TFRM FFRM VNVFRM CVC RCFS


Word: *communicate_{en}*

Senses:

1. join or connect
Concept: *communicate* (VERB.CONTACT)
2. be in verbal contact; interchange information o
Concept: *communicate*<*interact* (VERB.COMMUN
3. transfer to another
Concept: *convey*>*pass* (VERB.POSSESSION)
4. receive Communion, in the Catholic church
Concept: *commune*<*covenant* (VERB.COMMUNIC,
5. administer communion; in church
Concept: *communicate*<*covenant* (VERB.COMMUN
6. transmit thoughts or feelings
Concept: *communicate*>*get* (VERB.COMMUNICAT
7. transmit information
Concept: *pass*<*convey* (VERB.COMMUNICATION)
8. (a) to exchange information, either verbally or r

(b) to exchange information, either verbally or n
of communication. The processes of communicat
primarily are (or focus on) the tangible, multi-part
communication: the interdiscursive medium situat

Mammoth ontology browse



from the Intelligent Systems Division of the USC/Information

This effort expands on the original *SENSUS ontology* develop
translation and its *Ontosaurus Web Browser*, developed by James R. ... and William ...

Our current effort is the *OMEGA successor ontology* by Eduar
information integration and ontology alignment as well as m
the *Omega home page*.

More information on [natural language processing](#) at ISI.

More information on [information source integration](#) at ISI.

Please send comments and suggestions to [Andrew Philpot](#)

Copyright 1995-2004, Information Sciences Institute

All rights reserved

Info selection: source ontologies, frame sources...

Word entry (lex/concept identity, substring match)

Lexical space: word senses

Sense space: frames, etc.

Concept space: hierarchy

Omega browser Mammoth

What would be nice?

- A small number of **(globally) standardized ontologies** and/or core theories of important aspects (time, space, social dynamics, motion, privacy, etc.)
- Solid **theoretical frameworks** for developing ontological notions and theories, and for testing them
- A rich **online world of ontologies, domain models, etc.**, with appropriate ontology creation tools and methodologies
- (Semi-)automated **techniques** for rapidly finding, absorbing, and testing **existing ontologies** for your own applications
- Tools that **automatically create new knowledge bases** on demand, in accord with given ontologies
- **Ontology and knowledge base support technology** that can handle info that may be inconsistent, tenuous, partial, and growing