Corpus Annotation: Framework and Exercises

Eduard Hovy
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
USA
hovy@isi.edu
http://www.isi.edu/~hovy

Julia Lavid
Departamento de Filologia Inglesa
Universidad Complutense de Madrid
28040 Madrid
Spain
lavid@filol.ucm.es
http://www.ucm.es/info/atg/webpages/lavid/julia-webpage.html

Acknowledgments

- For the OntoNotes materials, and for learning about annotation, we thank
  - Martha Palmer and colleagues, U of Colorado at Boulder
  - Ralph Weischedel and Lance Ramshaw, BBN
  - Mitch Marcus and colleagues, U of Pennsylvania
  - Robert Belvin and the annotation team at ISI
  - Ann Houston, Grammarsmith
- For an earlier project involving annotation, we thank the IAMTC project:
  - Bonnie Dorr and Rebecca Green, U of Maryland
  - David Farwell and Stephen Helmreich, New Mexico State U
  - Teruko Mitamura and Lori Levin, CMU
  - Owen Rambow and Advaith Siddharth, Columbia U
  - Florence Reeder and Keith Jones, MITRE
- For funding, we thank DARPA and the NSF

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

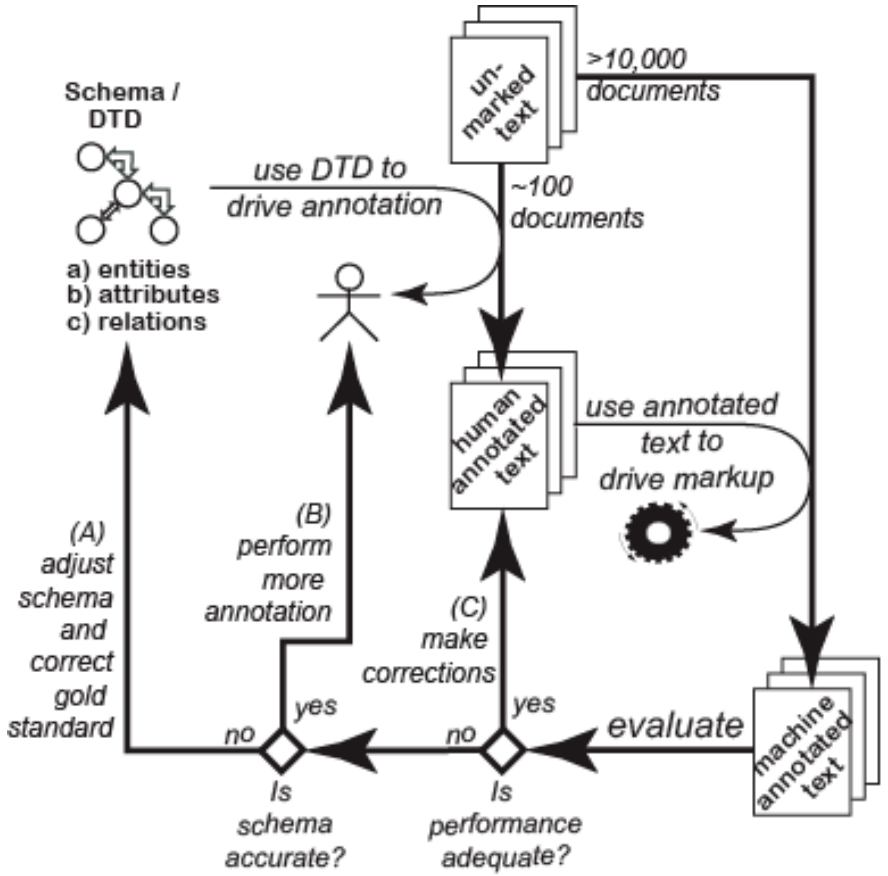
Are we entering an era of corpus building?

- The ‘statistics revolution’ in speech and NL processing is now complete:
  - Most people see speech and NL processing as a notation rewrite problem:
    - Speech → text, Italian → Chinese, sentence → parse tree → case frame, long text → short text…
  - Everyone uses machine learning to learn the rewriting ‘rules’
  - Everyone agrees creating rewriting rules by hand is infeasible for most transformations — the phenomena are too complex
- Results:
  - A new hunger for annotated corpora
  - A new class of researcher: the Annotation Expert
- BUT: How rigorous is Annotation as a ‘science’?

Ex: Annotation for Info Extraction

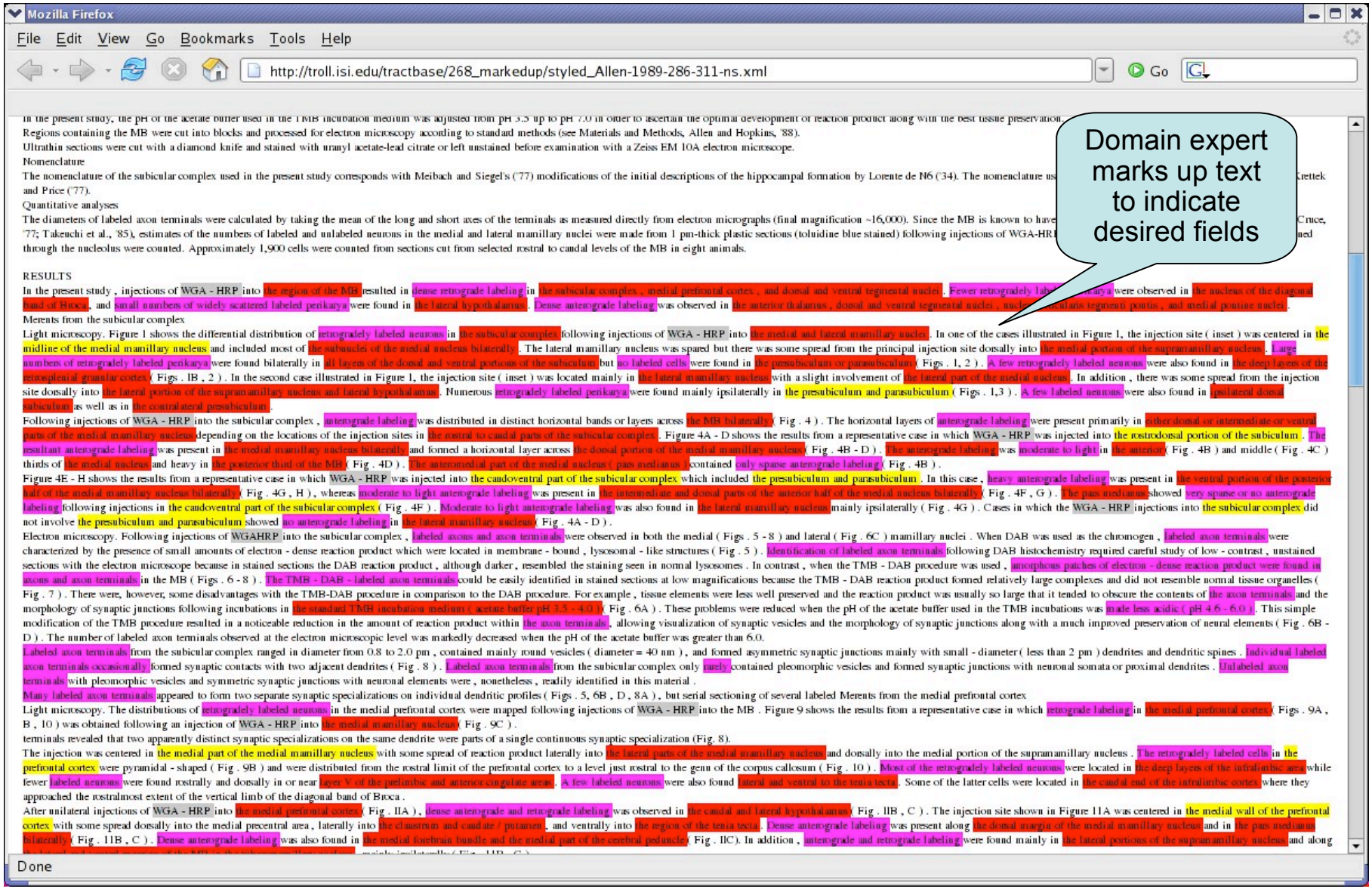
- Task: Identify desired information in free-form text and:
  - either extract info and put in database
  - or mark occurrence in text
- Examples: organization names, types and symptoms of disease, people's opinions about products, etc.
- As the items to extract become more complex, defining what exactly to extract becomes harder: move from pre-specified (hard-coded) rules to automated learning…
- …and this requires annotation…

- What is the role of annotation?
- How to define the IE, and how to determine acceptability of annotation for IE?

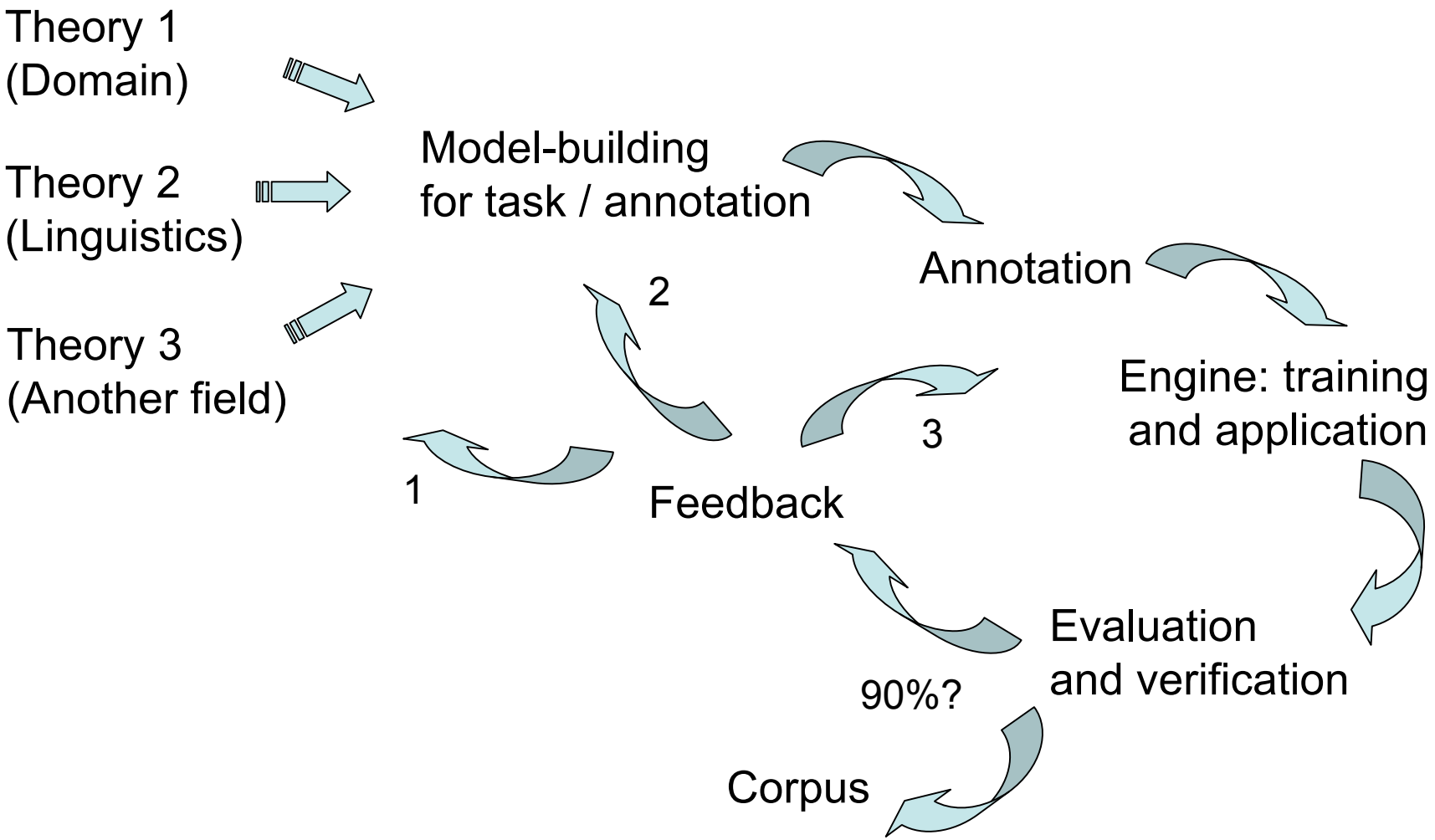


Biomed text markup

(Burns, Feng, Hovy 06; 07)



The generic annotation pipeline

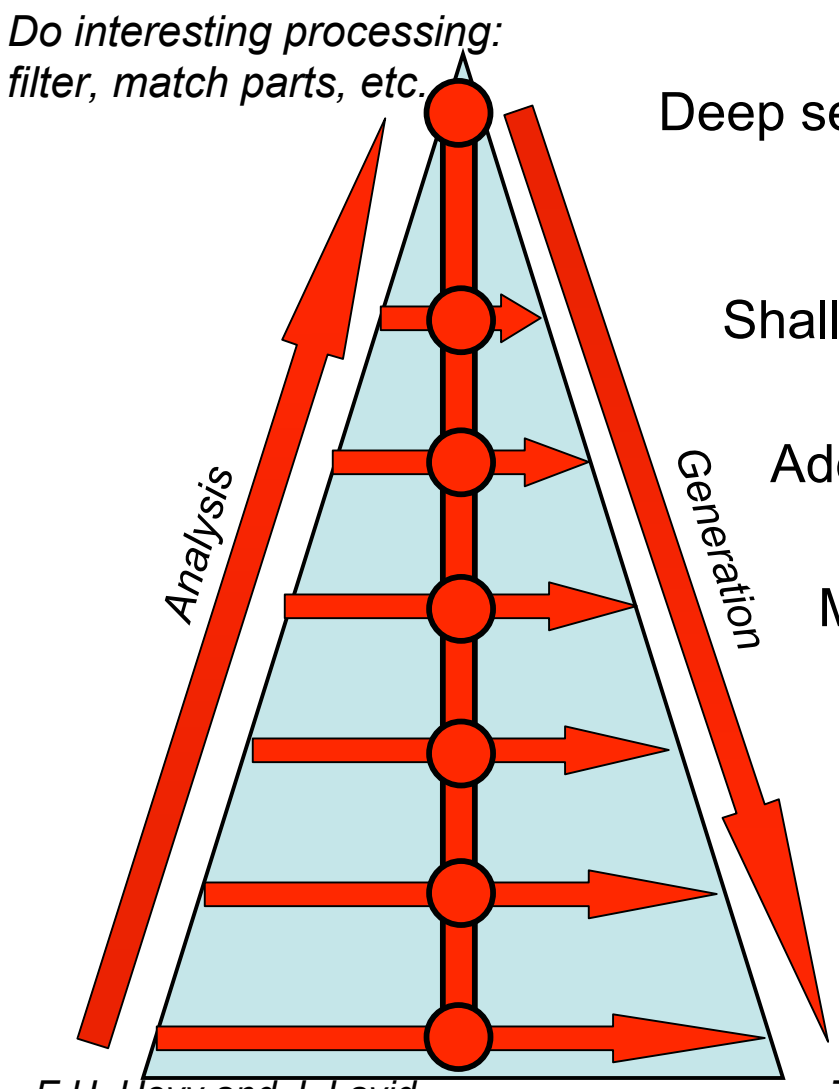


Two reasons to annotate

- **Traditional goal**: Fundamental belief that domain semantics is useful:
  - for reasoning in / studying the domain
  - to help improve NLP
- Methodologies: Transform pure text into interpreted/ extracted/marked-up text
  - Old methodology: manually-built rules for transformations
  - New methodology: machine learning of transformations
    1. Have humans manually annotate texts with transformation info
    2. Train computers on the corpus to do the same job

- **Additional goal**: Use annotation as mechanism to **test aspects of the theory** of domain semantics empirically — actual theory formation as well

NLP at increasing depths



Deep semantics: ?

Shallow semantics: frames

Adding more: semantic features

Medium changes: syntax

Adding info: POS tags, etc.

Small changes: demorphing, etc.

Direct: simple replacement

Shallow and deep semantics

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

(X1 :act Sell :agent She :patient (X1a :type

(X2a :act Transfer :agent She :patient (X2c :type Book) :recip He)
(X2b :act Transfer :agent He :patient (X2d :type Money) :recip She)

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

(X3a :prop Headache :patient He) (…?…)

(X4c :type Head :owner He) :state -3)
(X4b

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

(X4 :type Contrast :arg1 (X4a …?…) :arg2 (X4b …?…))

Which symbols?

Which roles?

How define states and state changes?

How handle relations?

How handle negation?

How handle comparatives?

Some phenomena to annotate

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)
- Opinions and subjectivity
- Pragmatics/speech acts
- Polarity/negation
- Presuppositions
- Metaphors

Annotation project desiderata

- Annotation must be:
  - **Fast**… to produce enough material
  - **Consistent**… enough to support learning
  - **Deep**… enough to be interesting

- Thus, need:
  - Simple **procedure** and good **interface**
  - Several people for **cross-checking**
  - Careful attention to the source **theory**!

Annotation as a science

- Increased need for corpora and for annotation raises new questions:
  - **What kinds/aspects of 'domain semantics' to annotate?** …it's hardly an uncontroversial notion…
  - Which corpora? How much?
  - Which computational tools to apply once annotation is 'complete'? When *is* it complete?
  - How to manage the whole process?
- Results:
  - A new hunger for annotated corpora
  - A new class of researcher: the Annotation Expert

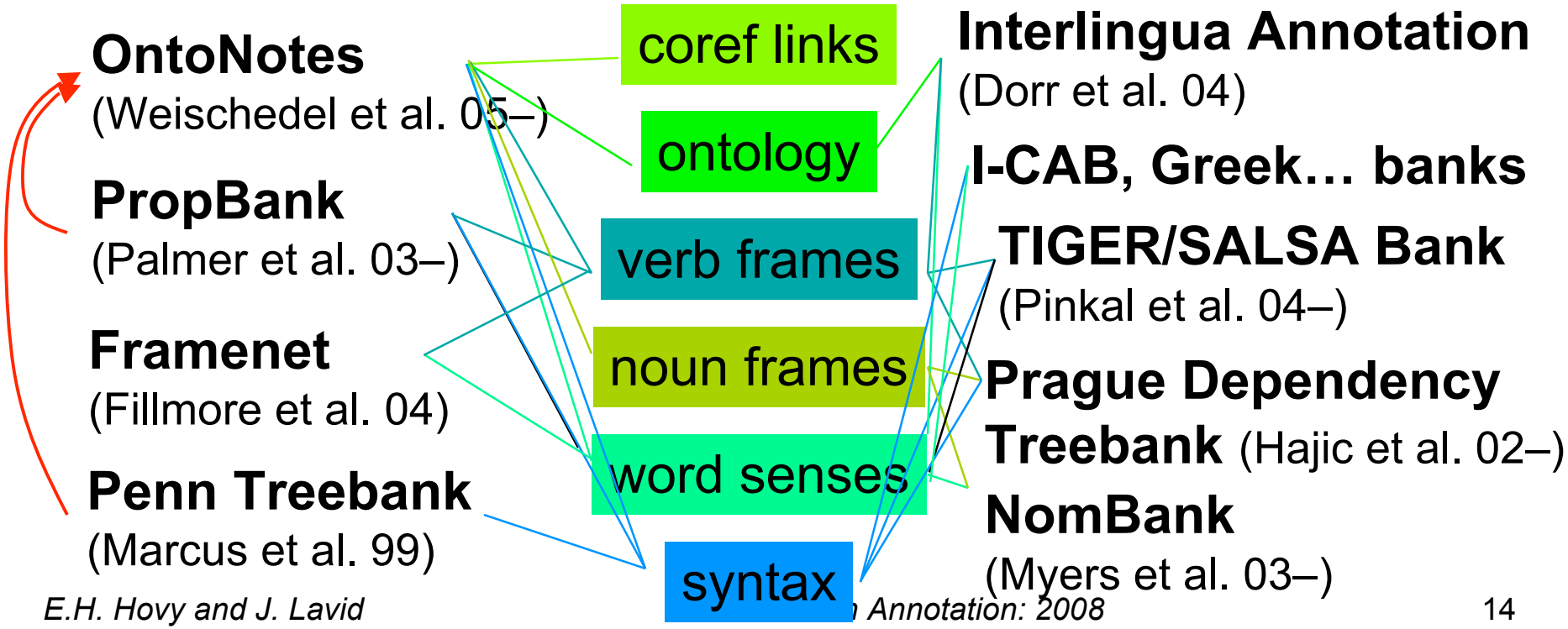
- Need to systematize annotation process — BUT: How rigorous is Annotation as a 'science'?

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Semantic annotation projects

- Goal: corpus of pairs (sentence + semantic rep)
- Process: humans add information to sentences (and their parses)
- Recent projects:



Other recent annotation projects

- US:
  - Time-ML (Pustejovsky et al.)
  - MPQA: subjectivity / ‘opinion’ (Wiebe et al.)
- EU:
  - Several annotation projects
- Japan:
  - Two ministries (MIC & METI) planning next 8 years’ NLP research — annotation important role
  - MIC theme: Universal communication (knowledge construction and multimedia integration, input and output)

OntoNotes goals

- Goal: In 4 years, annotate corpora of 1 mill words of English, Chinese, and Arabic text:
  - Manually provide semantic symbols for nouns and verbs
  - Manually connect sentence structure in verb and noun frames
  - Manually link anaphoric references
  - Manually construct supporting ontology of senses

- Even so, many words untouched!:

  Results of automated annotation by system trained on OntoNotes corpus:

The Bush **administration** (WN-Poly **ON-Poly**) had **heralded** (WN-Poly False) the Gaza **pullout** (WN-Poly False) as a big **step** (WN-Poly ON-Mono) on the **road** (WN-Poly ON-Mono) **map** (WN-Poly False) to a separate Palestinian **state** (WN-Poly **ON-Poly**) that Bush **hopes** (WN-Poly ON-Mono) to **see** (WN-Poly **ON-Poly**) by the **time** (WN-Poly False) he **leaves** (WN-Poly False) **office** (WN-Poly False) but a Netanyahu **victory** (WN-Mono False) would **steer** (WN-Poly False) Israel away from such **moves** (WN-Poly **ON-Poly**) .

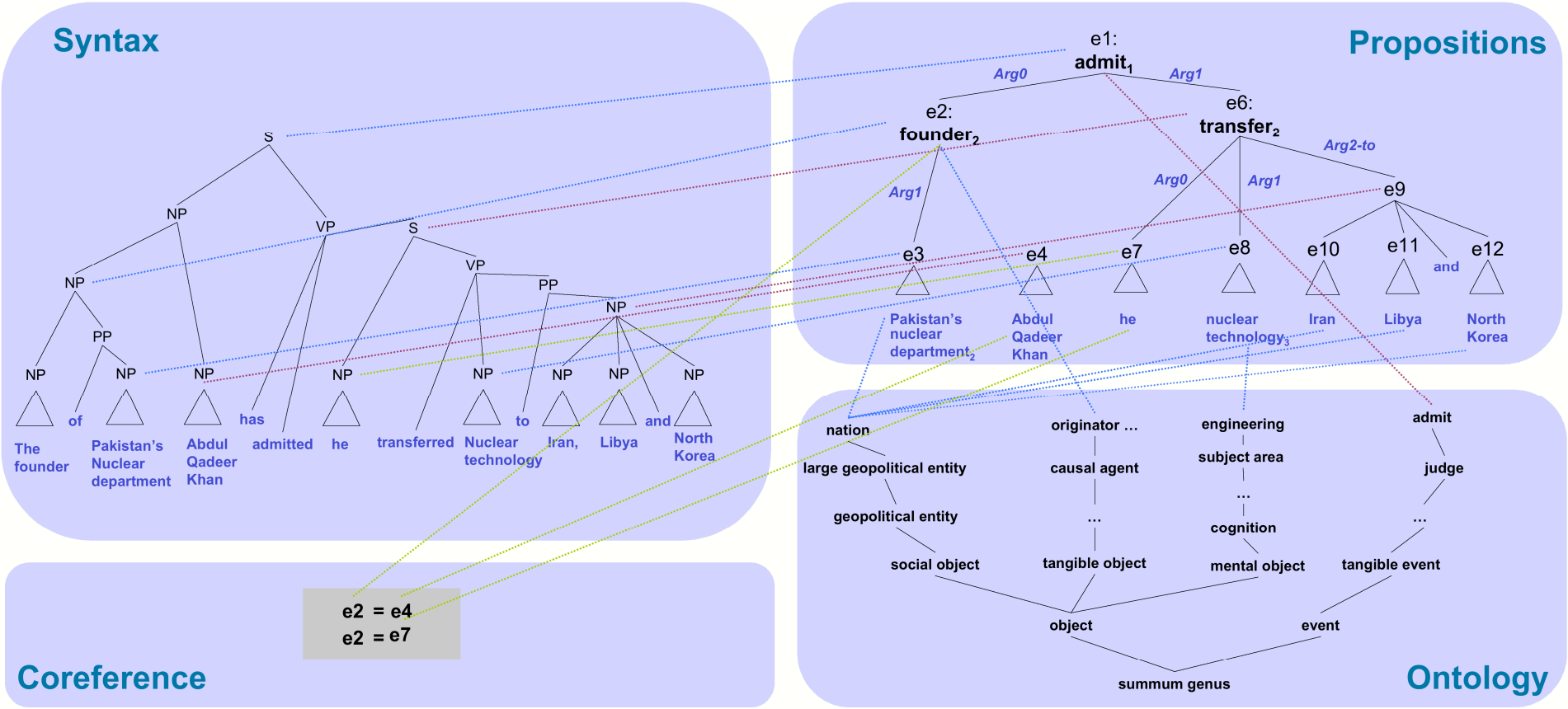
The Israeli **generals** (WN-Poly ON-Mono) **said** (WN-Poly **ON-Poly**) that if the **situation** (WN-Poly ON-Mono) did not **improve** (WN-Poly ON-Mono) by Sunday Israel would **impose** (WN-Poly ON-Mono) `` more restrictive and thorough **security** (WN-Poly False) **measures** (WN-Poly False) '' at other Gaza **crossing** (WN-Poly ON-Mono) **points** (WN-Poly **ON-Poly**) that Israel **controls** (WN-Poly **ON-Poly**), **according** (WN-Poly False) to **notes** (WN-Poly False) of the **meeting** (WN-Poly False) **obtained** (WN-Poly ON-Mono) by the New York Times.

Why an Onto-Bank?

- We focus on only the very simplest, 'literal', semantics
- We believe that using even OntoNotes's literal semantics can improve performance on GALE tasks:
  - MT:
    - Prefer translations in which target sentence pred-arg structures are fully connected and properly align with source sentence structures — **proposition structure**
  - Distillation:
    - Match correct sense of ambiguous words to query — **semantic word sense**
    - Obtain more accurate query term expansion — **semantic word sense and ontology-based inference**
    - Resolve pronouns and nominal mentions for more complete response creation — **coreference**
    - Find semantic redundancy and overlaps in retrieved fragments — **coref, semantic word sense, ontology-based inference**

OntoNotes content

(Slide by L. Ramshaw, et al.)



The founder of Pakistan's nuclear department, Abdul Qadeer Khan, has admitted he transferred nuclear technology to Iran, Libya, and North Korea.

Example of result

(Slide by M. Marcus, R. Weischedel, et al.)

3@wsj/00/wsj_0020.mrg@wsj: Mrs. Hills said many of the 25 countries that she placed under varying degrees of scrutiny have made `` genuine progress '' on this touchy issue .

In various formats…

Propositions

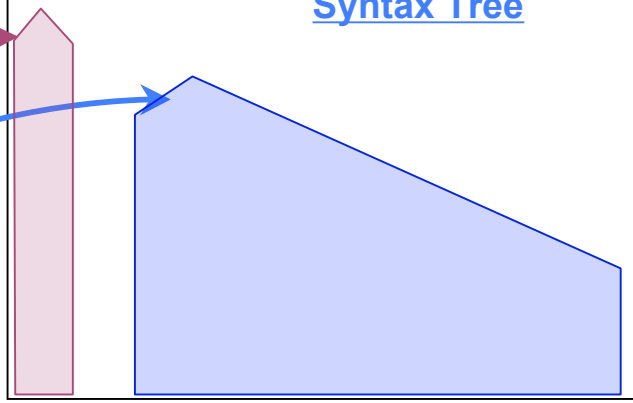
predicate : say
pb sense : 01
on sense : 1

ARG0: Mrs. Hills [10]
ARG1: many of the 25 countries that she placed under varying degrees of scrutiny have made `` genuine progress '' on this touchy issue

predicate : make
pb sense : 03
on sense : None

ARG0: many of the 25 countries that she placed under varying degrees of scrutiny
ARG1: “ genuine progress '' on this touchy issue

Syntax Tree



Coreference chains

ID=10; TYPE=IDENT
Sentence 1: U.S. Trade Representative Carla Hills
Sentence 3: Mrs. Hills
Sentence 3: she
Sentence 4: She
Sentence 6: Hills

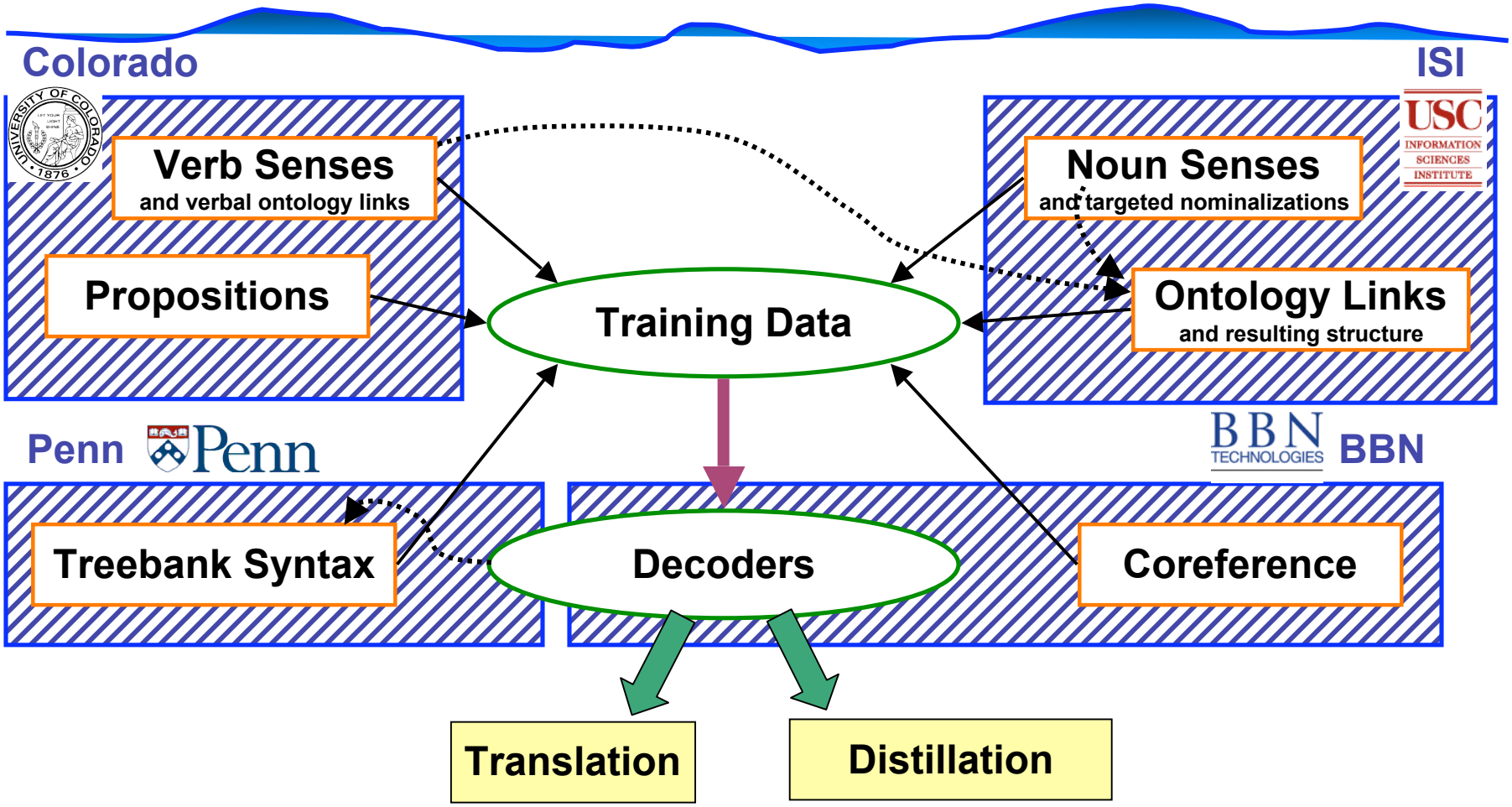
Omega ontology for senses

Say.A.1.1.1: DEF “…” EXS “…” FEATS … POOL [State.A.1.2 Declare.A.1.4…]

Say.A.1.1.2: DEF “…” EXS “…” POOL […]

Project structure & parts

(Slide by M. Marcus, R. Weischedel, et al.)



- Syntactic structure
- Predicate/argument structure
- Disambiguated nouns and verbs
- Coreference links
- Ontology
- Decoders

Syntax layer

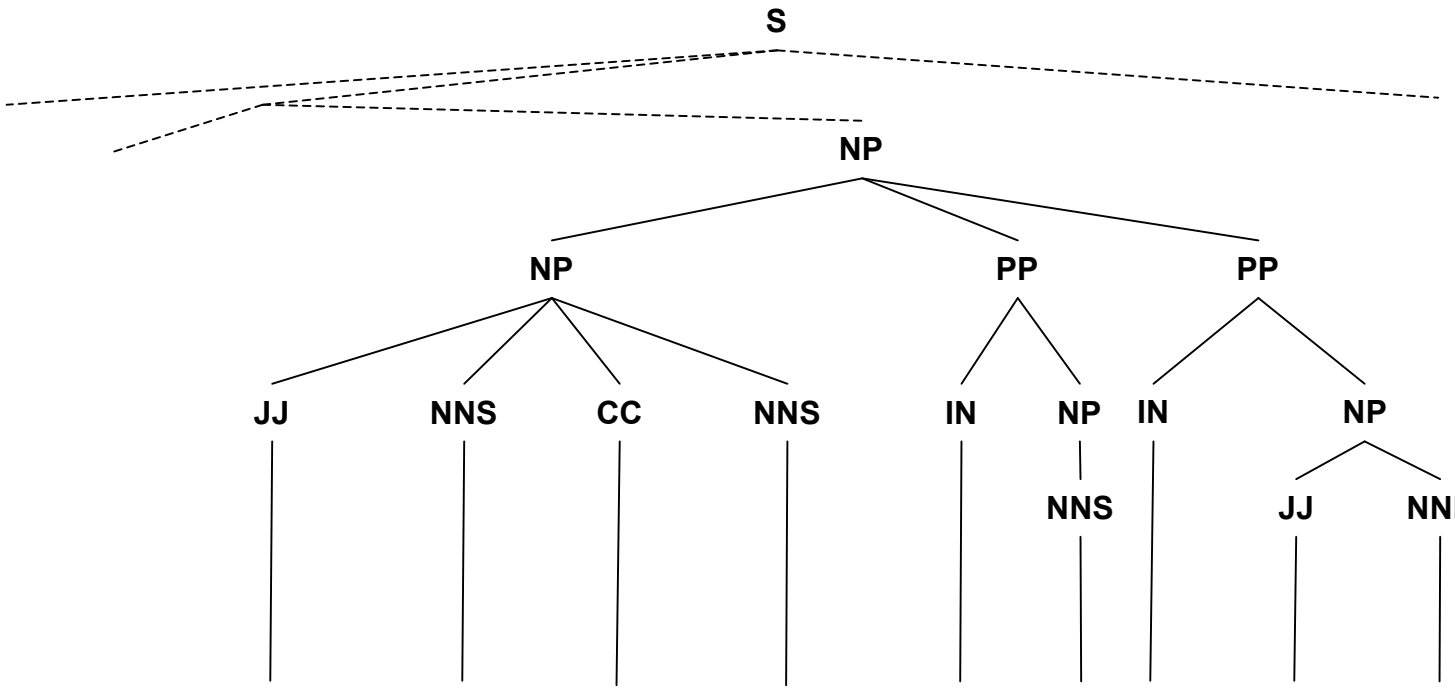
Penn Treebank structure

(Slide by S. Pradhan, BBN)

Identifies meaningful phrases in the text

Lays out the structure of how they are related

Concerns about the pace of the Vienna talks -- which are aimed at the destruction of some 100,000 weapons , as well as major reductions and realignments of troops in central Europe – also are being registered at the Pentagon .



... major reductions and realignments of troops in central Europe – ...

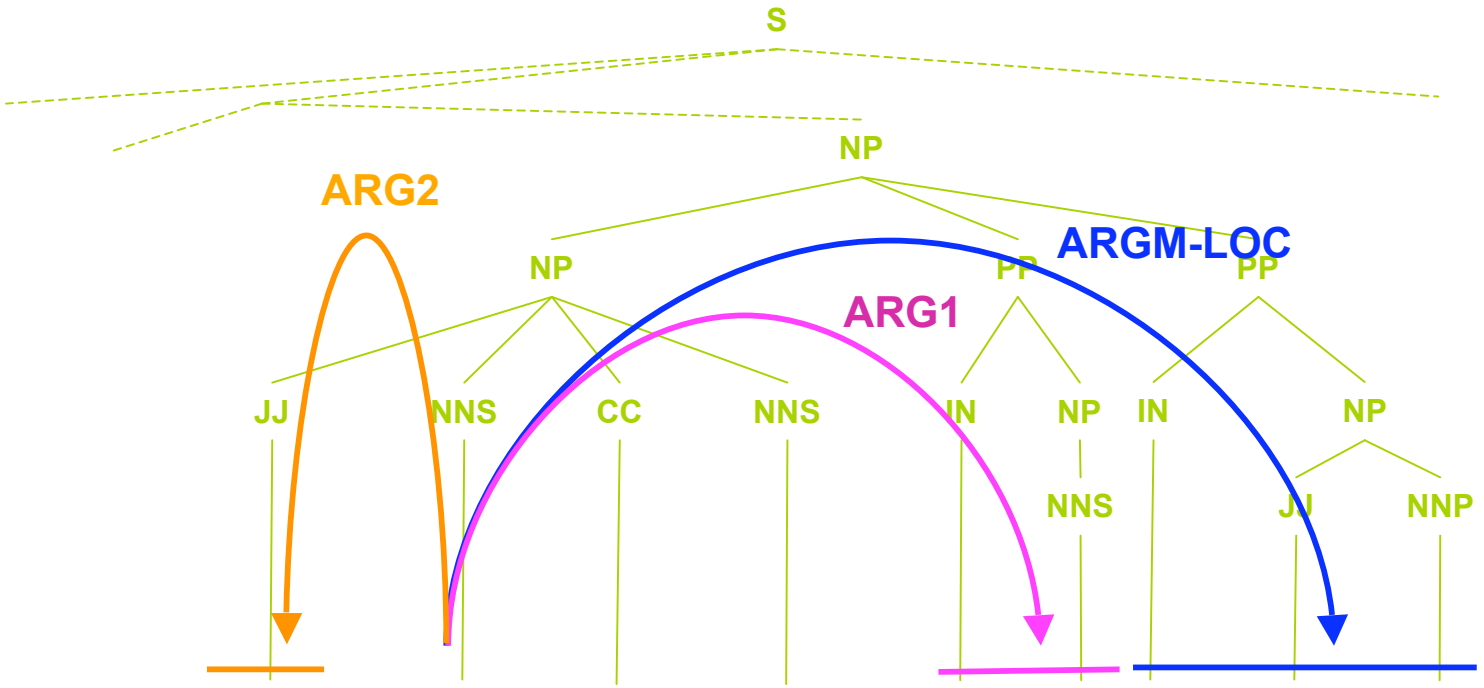
Propositional structure

Propbank structure

(Slide by S. Pradhan, BBN)

Tells who did what to whom…for both verbs and nouns

Concerns about the pace of the Vienna talks -- which are aimed at the destruction of some 100,000 weapons , as well as major reductions and realignments of troops in central Europe -- also are being registered at the Pentagon .



... major reductions and realignments of troops in central Europe – ...

Predicate frames

Propbank frames

(Slide by S. Pradhan, BBN)

Predicate frames define the meanings of the numbered arguments

Concerns about the pace of the Vienna talks -- which are aimed at the destruction of some 100,000 weapons , as well as major reductions and realignments of troops in central Europe -- also are being registered at the Pentagon .

reduction

reduce.01 – Make less

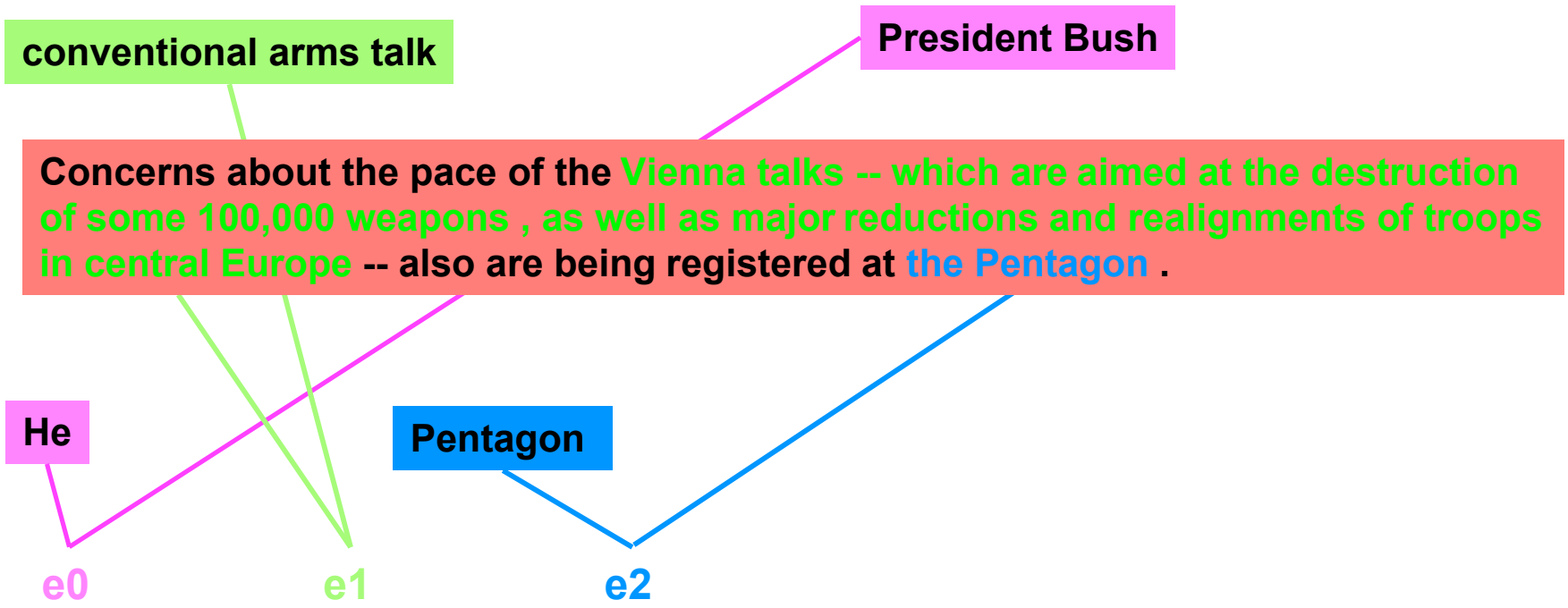
ARG0 – Agent	-
ARG1 – Thing falling →	the troops
ARG2 – Amount fallen →	major
ARG3 – Starting point	-
ARG4 – Ending point	-

Coreference

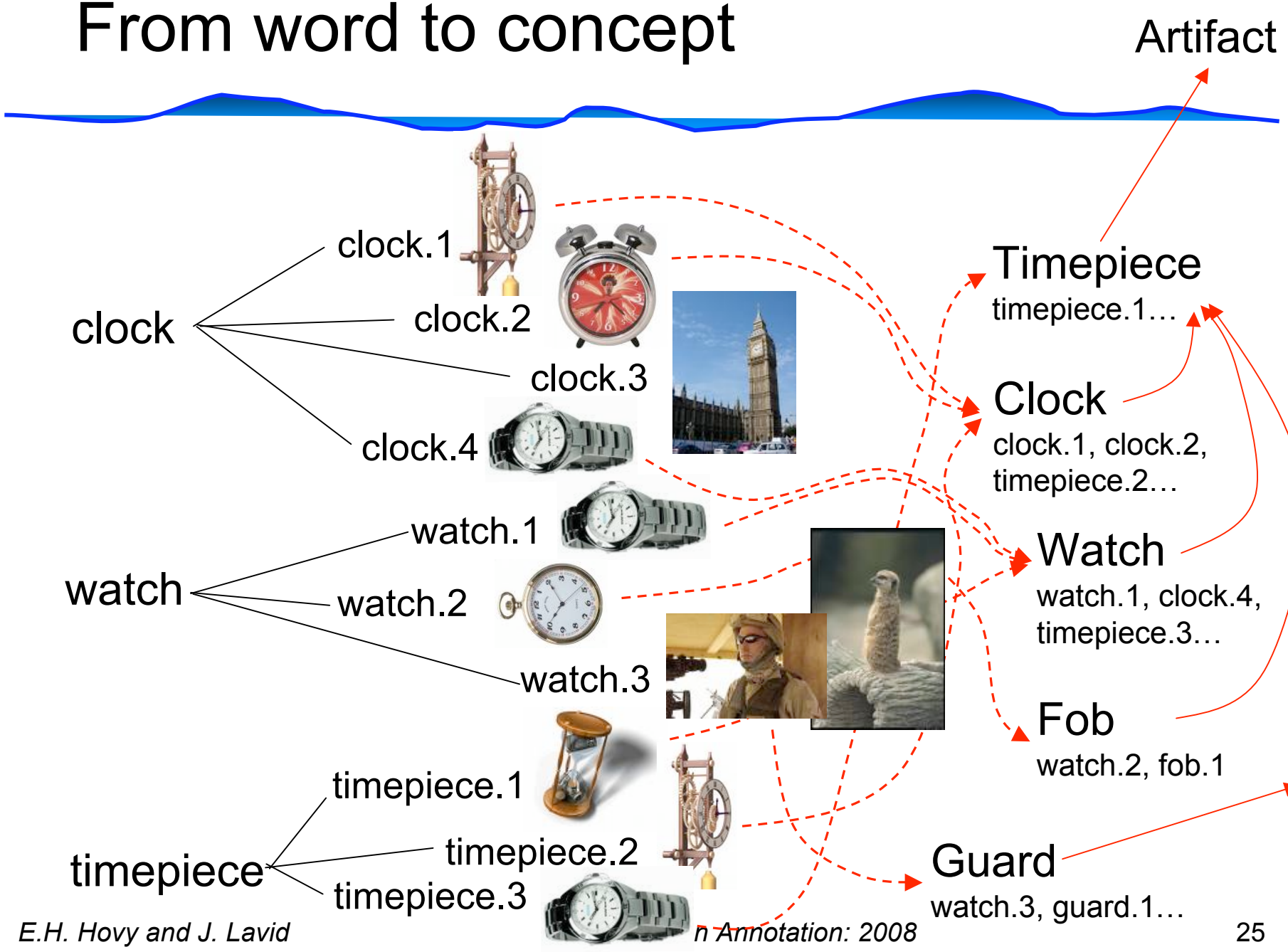
(Slide by S. Pradhan, BBN)

Identifies different mentions of the same entity within a document – especially links definite NPs, referring noun phrases, and pronouns to their antecedents

Two coref types tagged – *Identity* and *Attributive*



From word to concept



Word senses and Ontology

Omega ontology

Meaning of nouns and verbs are specified using a catalog of possible senses, with semantic features

Synonymous senses are pooled into 'concepts', pooling features

Concepts linked into Omega Ontology under Upper Model

Concerns about the pace of the Vienna talks -- which are aimed at the destruction of some 100,000 weapons , as well as major reductions and realignments of troops in central Europe -- also are being registered at the Pentagon .

Word Sense

aim

1. **Point or direct object, weapon, at something ...**
2. **Wish, propose, or intend to achieve something**

Word Sense

propose

1. **Suggest a course of action**
2. **Ask someone to get married**
3. **Request funds or other support for a project**
4. **State a hypothesis**

Sense Pool (concept): [aim2+propose1]

Why an ontology?

- Current HLT systems depend on impoverished text models:
  - Bags of words, ngram word sequences, syntactic structure
- OntoNotes provides a (very slightly) deeper and more semantic (meaning-based) representation that:
  - Resolves meaning ambiguity of words in terms of senses
  - Connects the word senses to an ontology of symbols
  - The ontology symbols are organized in semantic clusters
  - The symbols also contain features
- **Why not just senses?**
- For more effective HLT systems, it may be useful to exploit the symbols' organization and features
  - Applications (Information Extraction, Question Answering, Summarization…) and tasks (entailment, semantic analysis for learning by reading, etc.) all use inference
  - Ontology may support limited inference for **term expansion, term substitution, term matching, structure matching**, etc.

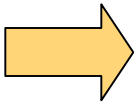
Four major subtasks

- How do you go from

  The founder of Pakistan's nuclear department, Abdul Qadeer Khan, has admitted he transferred nuclear technology to Iran, Libya, and North Korea

  to

  P1: :type Person3 :name "Abdul Qadeer Khan"
  P2: :type Person3 :gender male
  P3: :type Know-How4
  P4: :type Nation2 :name "Iran"
  P5: :type Nation2 :name "Libya"
  P6: :type Nation2 :name "N. Korea"
  X0: :act Admit1 :speaker P1 :saying X2
  X1: :act Transfer2 :agent P2 :patient P3 :dest (P4 P5 P6)
  coref P1 P2

  instances

  semantic symbols

  frame structure

  coref links

  sense groups

- Tasks:
  1. Create word senses for words
  2. Annotate sentences with the senses
  3. Annotate sentences for co-reference
  4. Group senses and insert into Omega ontology, as concepts

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Annotation: The 7 core questions

1. **Preparation**
   - Choosing the corpus — which corpus? What are the political and social ramifications?
   - How to achieve balance, representativeness, and timeliness? What does it even mean?
2. **'Instantiating' the theory**
   - Creating the annotation choices — how to remain faithful to the theory?
   - Writing the manual: this is non-trivial
   - Testing for stability
3. **Interface design**
   - Building the interfaces. How to ensure speed and avoid bias?
4. **The annotators**
   - Choosing the annotators — what background? How many?
   - How to avoid overtraining? And undertraining? How to even know?
5. **Annotation procedure**
   - How to design the exact procedure? How to avoid biasing annotators?
   - Reconciliation and adjudication processes among annotators
6. **Validation**
   - Measuring inter-annotator agreement — which measures?
   - What feedback to step 2? What if the theory (or its instantiation) 'adjusts'?
7. **Delivery**
   - Wrapping the result — in what form?
   - Licensing, maintenance, and distribution

Q1. Prep: Choosing the corpus

- Corpus collections are worth their weight in gold
  - Should be unencumbered by copyright
  - Should be available to whole community
- Value:
  - Easy-to-procure training material for algorithm development
  - Standardized results for comparison/evaluation
- Choose carefully—the future will build on your work!
  - (When to re-use something?—Today, we're stuck with WSJ…)
- Important sources of raw and processed text and speech:
  - ELRA (European Language Resources Association)
    www.elra.info
  - LDC (Linguistic Data Consortium)
    www.ldc.upenn.edu/

Q1. Prep: Choosing the corpus

- **Technical issues**: *Balance, representativeness, and timeliness*
  - When is a corpus representative? —"stock" in WSJ is *never* the soup base
    - We need a methodology of 'principled' corpus construction for representativeness (even BNC process rather ad hoc)
  - How to balance genre, era, domain?
    - Effect of (expected) usage of corpus
    - See (Kilgarriff and Grefenstette, CL 2003)
  - Experts: corpus linguists or domain specialists
- **Social, political, funding issues**:
  - How do you ensure agreement / complementarity with others? Should you bother?
  - How do you choose which phenomena to annotate? Need high payoff…
  - How do you convince funders to invest in the effort?

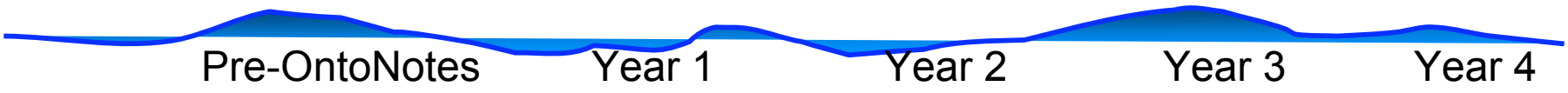
OntoNotes decisions

- Year 1: started with what was available
  - Penn Treebank, already present, allowed immediate proposition and sense annotation
  - Problem: just *Wall Street Journal*: all news, very skewed sense distributions
- Year 2:
  - English: balance by adding transcripts of broadcast news
  - Chinese: start with newspaper text
- Later years:
  - English, then Chinese: add transcripts of tv/radio discussion, then add blogs, online discussion
  - Add Arabic: newspaper text
- Questions:
  - How much parallel text across languages?
  - How much text in specialized domains?
  - How much additional to redress imbalances in word senses?
  - etc.

OntoNotes corpus growth

OntoNotes Release	Genres	Languages	Release Date
1.0	Newswire	Eng & Chi	2007-03
2.0	Broadcast News	Eng & Chi	2007-11
3.0	Broadcast Conversation	English	2008-11
		Chinese	2009-04
	Newswire	Arabic	2008-11
4.0	Newsgroups & Weblogs	English	2009-11
	Newsgroups	Chinese	2010-04
	Newswire	Arabic	2009-11
5.0	Conversational Telephone Speech	English	2010-11
	Weblogs	Chinese	2011-04
	Newswire	Arabic	2010-11

Corpus delivery by year



	Eng	Chi	Eng	Chi	Eng	Chi	Eng	Chi	Eng	Chi
Selection	NW 300	NW 250	BN 200	BN 300	BC 200	BC 150	WL NG	NG		
Treebank										
PropBank										
Word Sense										
Ontology										
Coref										
Delivery										

ON 1.0

ON 2.0

ON 3.0

ON 4.0

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Q2: Instantiating the theory

- Most complex question: What to annotate?
  - Goal: practical task (like IE), theory building (linguistics), or both?
  - Task/theory provides annotation categories/choices
  - Problem: tradeoff between desired detail/sophistication of desired categories and practical attainability of trustworthy annotation
  - General solution: simplify categories to ensure dependable results
  - Problem: How???
- How 'deeply' to instantiate theory?
  - Design rep scheme / formalism very carefully — simple and transparent
  - ? Depends on theory — but also (yes? how much?) on corpus and annotators
  - Do tests first, to determine what is annotatable in practice
- Experts must create:
  - Annotation categories
  - Annotator instruction (coding) manual — very important
  - Experts to build the manual: theoreticians? Or exactly NOT the theoreticians?
- Both must be tested! — Don't 'freeze' the manual too soon
  - Experts annotate a sample set; measure agreements
  - Annotators keep annotating a sample set until stability is achieved

Q2: Instantiating the theory

- Issues:
  - When building the theory, you don't know how many categories there are in the data
  - When addressing a practical task, you don't know how easy it will be to identify all the cases your problem covers
- Likely problems:
  - Categories not exhaustive over phenomena
  - Categories difficult to define / unclear (due to intrinsic ambiguity, or because you rely too much on background knowledge?)
- What you can do:
  - Work in close cycle with annotators, and see week by week what they do
  - Hold weekly discussions with all the annotators
  - Measure the annotator agreement and disagreement (see below)
  - Modify your categories as needed—be led by what is practical
  - Create and constantly update the Annotator Handbook
    - (Penn Treebank Codebook: 300 pages!)
- Measuring stability — measures of agreement: (Lipsitz et al., 1991)
  - Precision (correctness) = P_i = *#correct / N*
  - Entropy (ambiguity, regardless of correctness) = $-\Sigma_i P_i \,.\, ln\, P_i$ (unambig → 0)
  - Odds Ratio (distinguishability of categories) = $\frac{f_{xx} f_{yy}}{f_{xy} f_{yx}}$ (indistinguishable → 0)

Q2: Theory and model

- 'Neutering' the theory: when the theory is controversial, or you cannot obtain stability — you may still be able to annotate, using a more neutral set of terms
  - Ex 1: from Case Roles (*Agent, Patient, Instrument*) to PropBank's roles (*arg0, arg1, argM*) — user chooses desired role labels and maps PropBank roles to them
  - Ex 2: from detailed sense differences to more crude / less detailed ones

- What does this say about the theory, however?

Ensuring trustworthiness/stability

- Problematic issues for OntoNotes:
  1. What sense are there? Are the senses stable/good/clear?
  2. Is the sense annotation trustworthy?
  3. What things should corefer?
  4. Is the coref annotation trustworthy?
- Approach: "**the 90% solution**":
  - Sense granularity and stability: Test with annotators to ensure agreement at 90%+ on real text
  - If not, then **redefine and re-do until 90% agreement** reached
  - Coref stability: only annotate the types of aspects/phenomena for which 90%+ agreement can be achieved

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Exercise 1: Creating senses

- Task: given a word, create its senses
  - Try to make the senses clearly different — the annotators won't agree otherwise!
  - Try to make as many senses as you can — choosing just one or two is not very useful!
  - Remember that the senses will later be put into an ontology — use semantic distinctions, not pragmatic ones

Exercise: Creating senses for “drive”

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.*

Develop your senses here

Annotate according to your senses

Write your sense choices here: _____

Annotate these sentences:

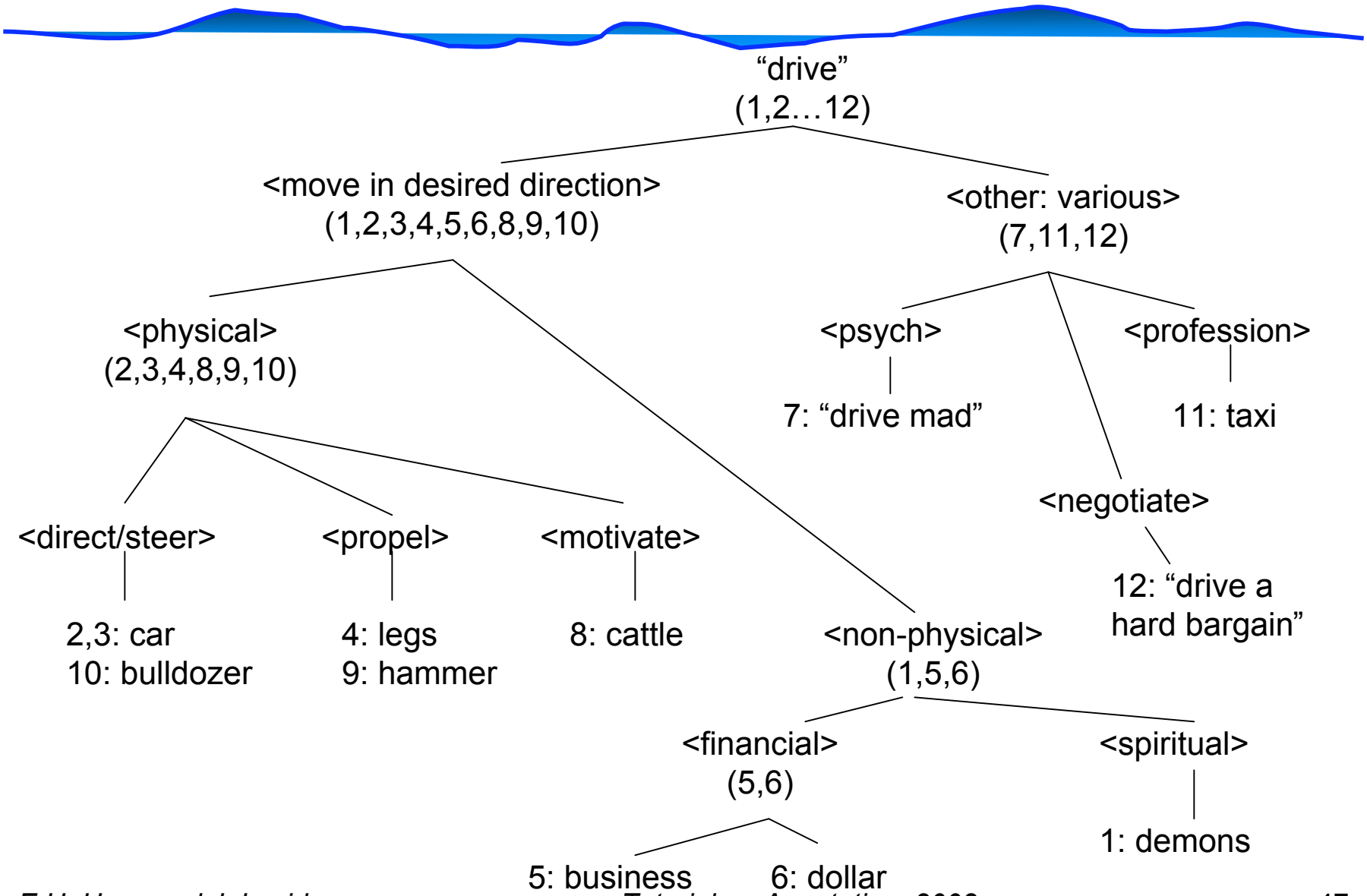
- *Drive* is a short-lived, Emmy Award-nominated television series created by Tim Minear and Ben Queen.
- LaCie is a leading manufacturer of external storage devices including our award winning selection of hard drives.
- Top 10 Scenic Drives: These roads do more than get you there.
- Test drive the 2007 Microsoft Office system programs today!
- Advances in technology have made it very easy for people to drive a clean vehicle.
- Business travel demand will outpace capacity in 2008 and drive rate increases across air, hotel, car rental and meetings.
- In a good golf game, use your club like a catapult to drive your ball straight.
- What we call a life force, drive, urge, compulsion, or impetus, is intimately conjoined with its opposition, that is, its negation, termination, or lack.
- Variables such as the nominal interest rate that drive exchange rate volatility can fluctuate daily.
- Take care that the attack does not drive his defending arm into his opponent's body or head.

- **Your choices**:

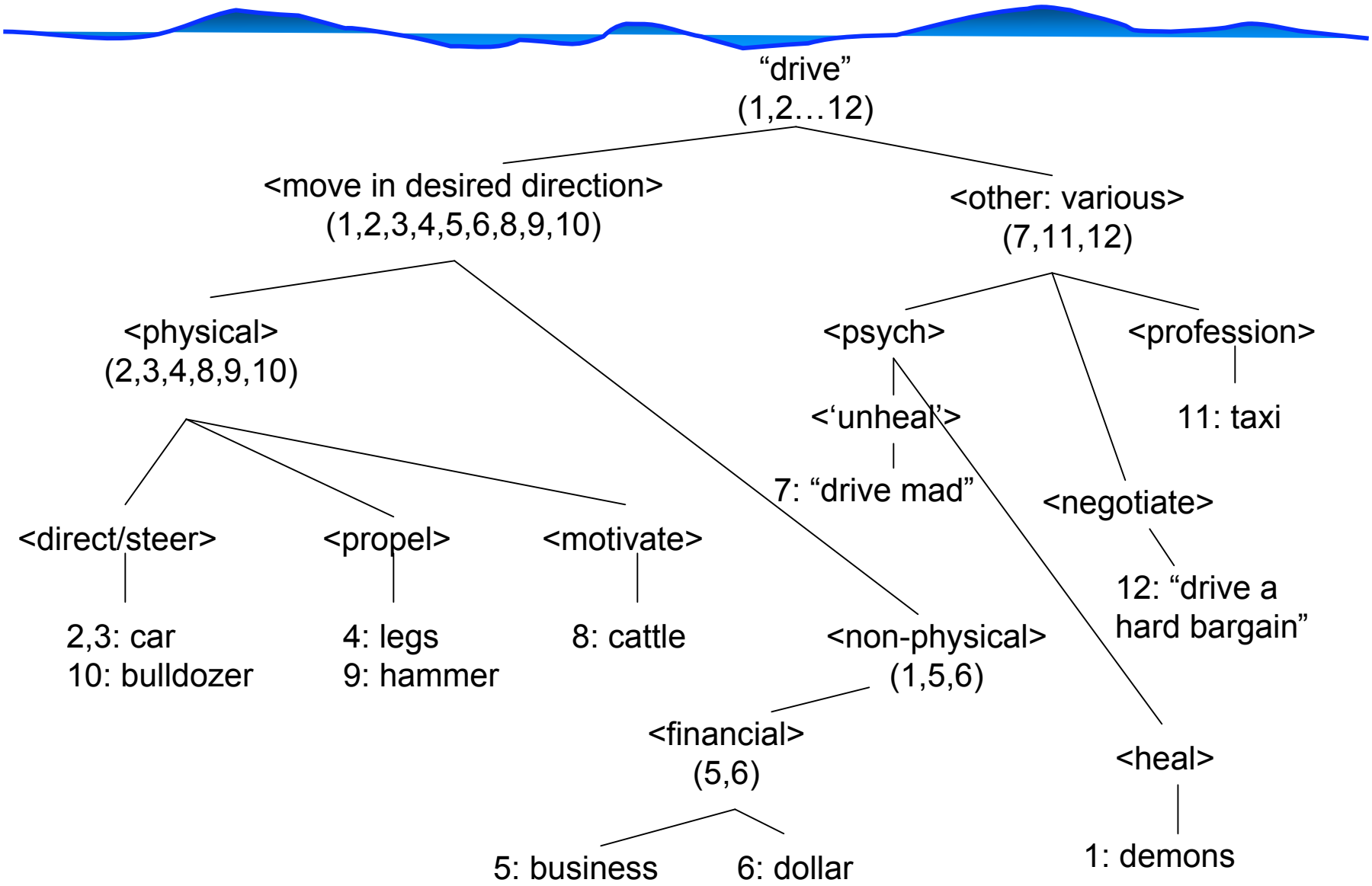
Creating senses: Graduated refinement

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

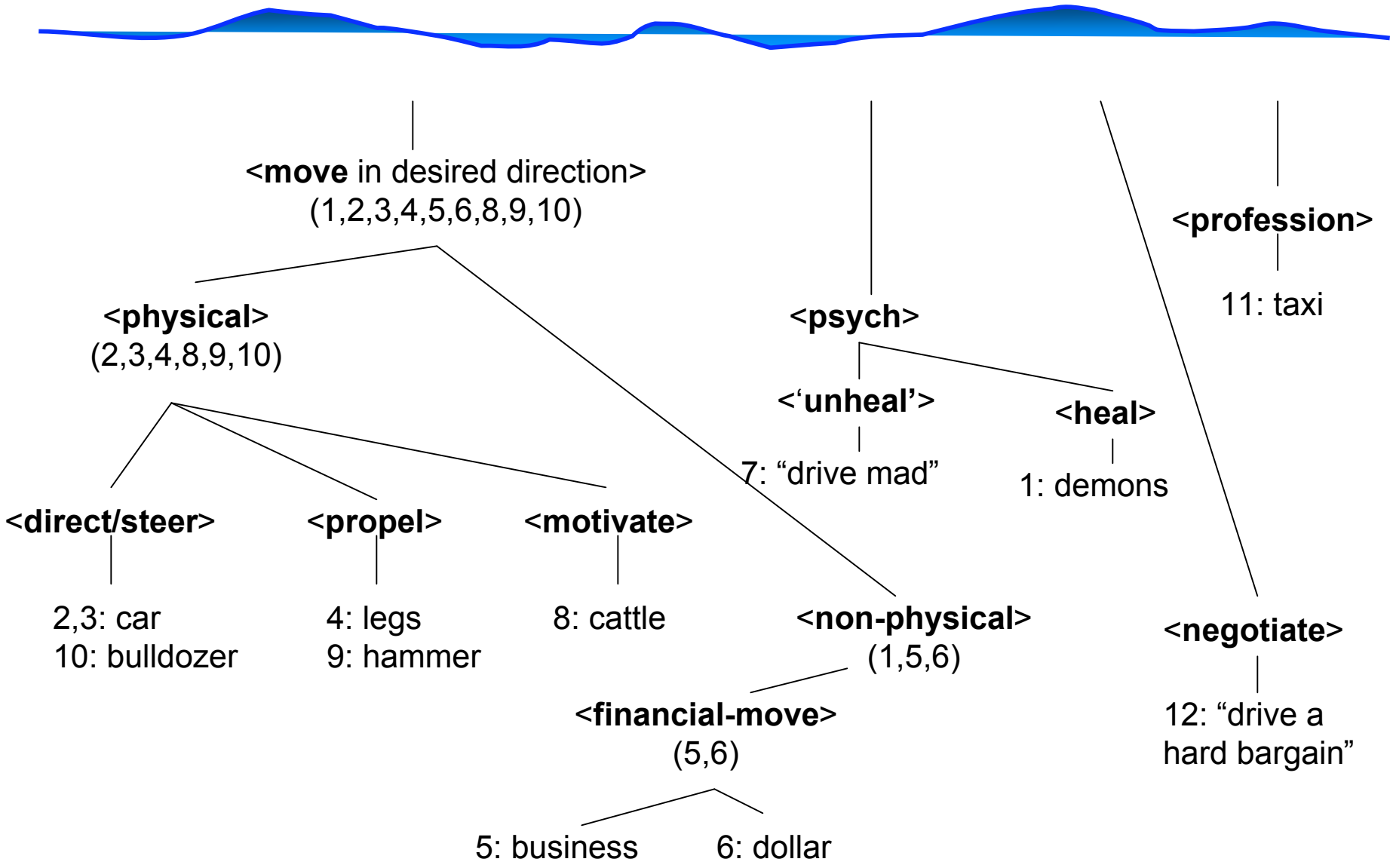
An exercise: “drive”



Deeper semantic “drive”



Ontologizing “drive”



Annotate according to your senses

Choices: *1. direct/steer 2. propel 3. motivate 4. financial 5. unheal 6. heal 7. negotiate 8. profession*

Annotate these sentences:

- *Drive* is a short-lived, Emmy Award-nominated television series created by Tim Minear and Ben Queen.
- LaCie is a leading manufacturer of external storage devices including our award winning selection of hard drives.
- Top 10 Scenic Drives: These roads do more than get you there.
- Test drive the 2007 Microsoft Office system programs today!
- Advances in technology have made it very easy for people to drive a clean vehicle.
- Business travel demand will outpace capacity in 2008 and drive rate increases across air, hotel, car rental and meetings.
- In a good golf game, use your club like a catapult to drive your ball straight.
- What we call a life force, drive, urge, compulsion, or impetus, is intimately conjoined with its opposition, that is, its negation, termination, or lack.
- Variables such as the nominal interest rate that drive exchange rate volatility can fluctuate daily.
- Take care that the attack does not drive his defending arm into his opponent's body or head.

- **Your choices**:

Annotate according to your senses

The reduced choices:

1. Move-in-direction 2. Psych 3. Negotiate 4. Profession

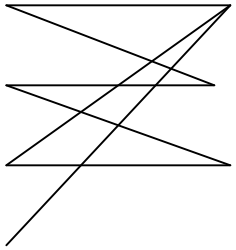
Annotate these sentences:

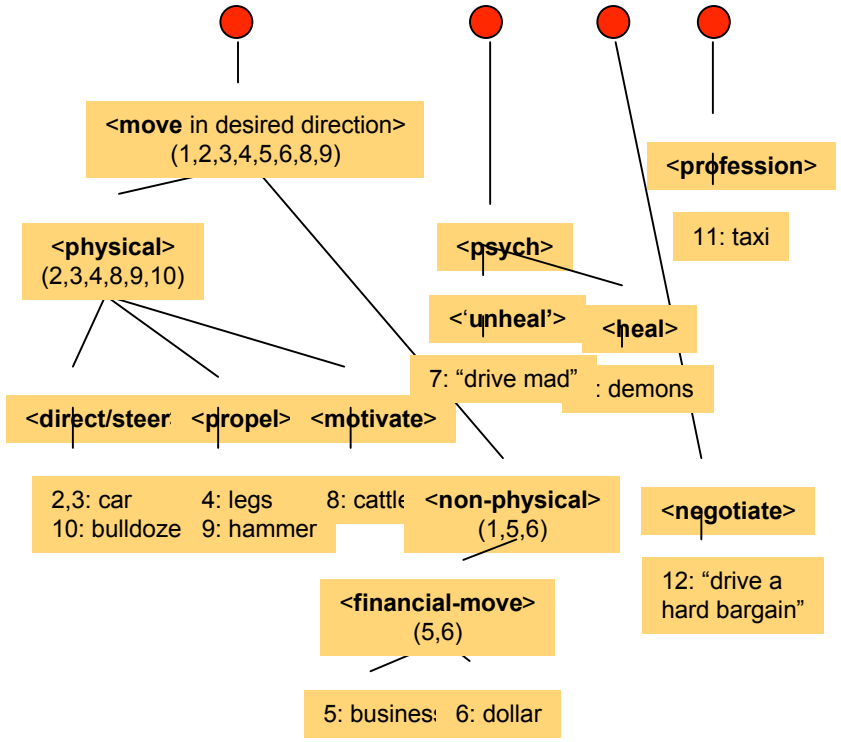
- *Drive* is a short-lived, Emmy Award-nominated television series created by Tim Minear and Ben Queen.
- LaCie is a leading manufacturer of external storage devices including our award winning selection of hard drives.
- Top 10 Scenic Drives: These roads do more than get you there.
- Test drive the 2007 Microsoft Office system programs today!
- Advances in technology have made it very easy for people to drive a clean vehicle.
- Business travel demand will outpace capacity in 2008 and drive rate increases across air, hotel, car rental and meetings.
- In a good golf game, use your club like a catapult to drive your ball straight.
- What we call a life force, drive, urge, compulsion, or impetus, is intimately conjoined with its opposition, that is, its negation, termination, or lack.
- Variables such as the nominal interest rate that drive exchange rate volatility can fluctuate daily.
- Take care that the attack does not drive his defending arm into his opponent's body or head.

- **Your choices**:

From lexemes to concepts

Lexical space	Sense space	Concept space
– Words	– Word senses	– Concepts
– Monolingual	– Multilingual	– Interlingual (?)
– “drive”	– Drive1	
– “steer”	– Drive2	
– “fahren”	– Drive3	
– “rijden”	– …	
– …		





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

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Q3: The interface

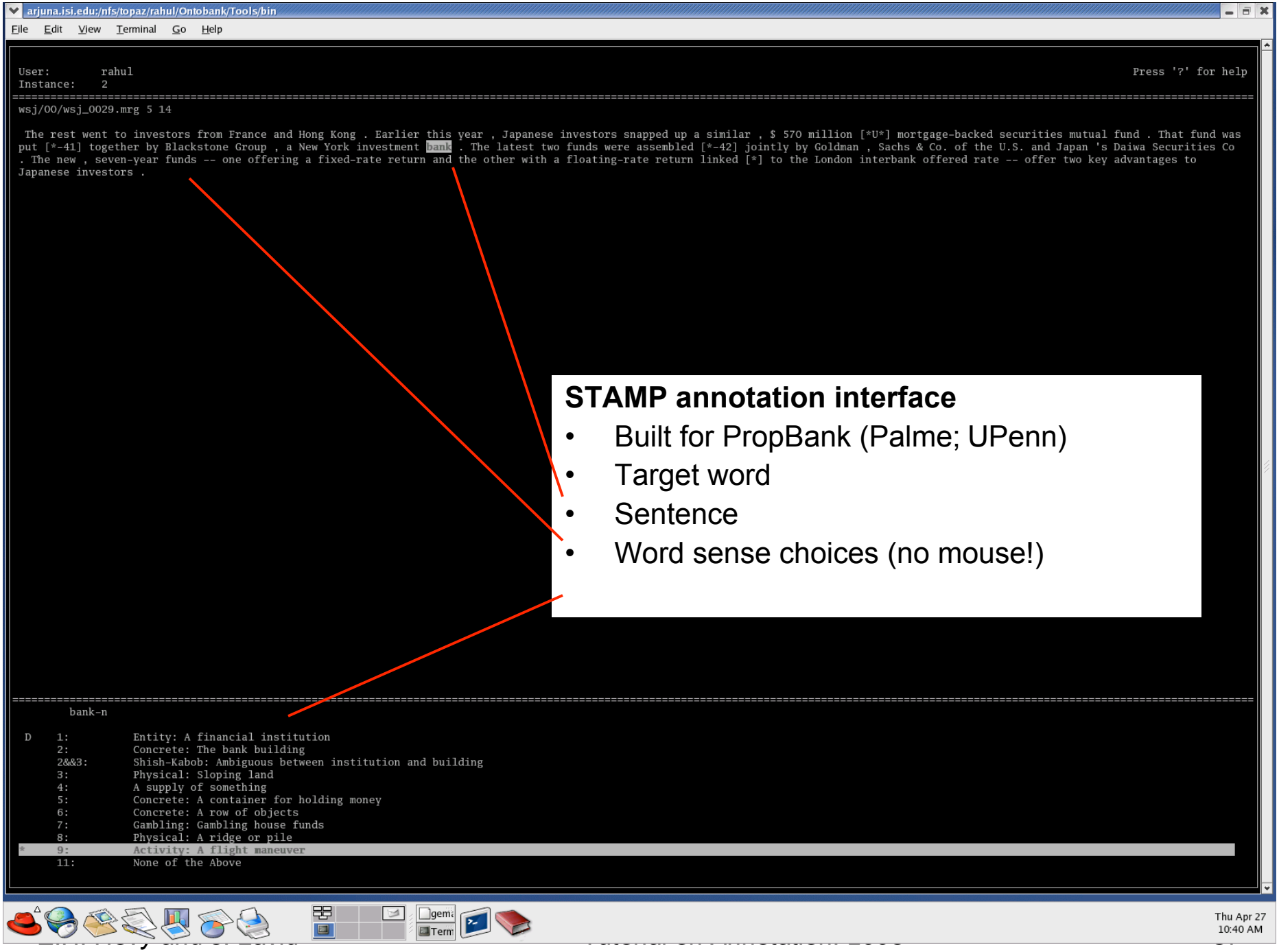
- How to design adequate interfaces?
  - Maximize speed!
    - Create very simple tasks—but how simple? Boredom factor, but simple task means less to annotate before you have enough
    - Don't use the mouse
    - Customize the interface for each annotation project?
  - Don't bias annotators (avoid priming!)
    - Beware of order of choice options
    - Beware of presentation of choices
    - Is it ok to present together a whole series of choices with expected identical annotation? — annotate *en bloc*?
  - Check agreements and hard cases in-line?
    - Do you show the annotator how 'well' he/she is doing? Why not?
- Experts: Psych experimenters; Gallup Poll question creators
- Experts: interface design specialists

Q3: Types of annotation interfaces

- **Select**: choose one of *N* fixed categories
  - Avoid more than 10 or so choices (7 ± 2 rule)
  - Avoid menus because of mousework
  - If possible, randomize choice sequence across sessions
- **Delimit**: delimit a region inside a larger context
  - Often, problems with exact start/end of region (e.g., exact NP) — but preprocessing and pre-delimiting chunks introduces bias
  - Evaluation of partial overlaps is harder
- **Delimit and select**: combine the above
  - Evaluation is harder: need two semi-independent scores
- **Enter**: instead of *select*, enter own commentary
  - Evaluation is very hard

Q3. Available interfaces

- Interfaces/annotation tools:
  - ATLAS.TI: annotation toolkit (www.atlasti.com/)
  - Ad hoc annotation interfaces and tools from the NLP community
  - QDAP annotation center at U of Pittsburgh (www.qdap.pitt.edu)
- Annotation standards:
  - Various XML and other notations
  - Standard backoff and other alternatives
  - Romary and Ide (2007): ISO annotation notation standards committee (ISO TC37 SC4 WG1)
    - Criteria: Expressive adequacy, media independence, semantic adequacy, incrementality for new info in layers, separability of layers, uniformity of style, openness to theories, extensibility to new ideas, human readability, computational processability, internal consistency



STAMP annotation interface

- Built for PropBank (Palme; UPenn)
- Target word
- Sentence
- Word sense choices (no mouse!)

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Monitoring progress and validating the result
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Q4: Annotators

- How to choose annotators?
  - Annotator backgrounds — should they be experts, or precisely not?
  - Biases, preferences, etc.
  - Experts: Psych experimenters

- Who should train the annotators? Who is the most impartial?
  - Domain expert/theorist?
  - Interface builder?
  - Builder of learning system?

- When to train?
  - Need training session(s) before starting
  - Extremely helpful to continue weekly general discussions:
    - Identify and address hard problems
    - Expand the annotation Handbook
  - BUT need to go back (re-annotate) to ensure that there's no 'annotation drift'

How much to train annotators?

- **Undertrain**: Instructions are too vague or insufficient. Result: annotators create their own 'patterns of thought' and diverge from the gold standard, each in their own particular way (Bayerl 2006)
  - How to determine?: Use Odds Ratio to measure pairwise distinguishability of categories
  - Then collapse indistinguishable categories, recompute scores, and (?) reformulate theory — is this ok?
  - Basic choice: EITHER 'fit' the annotation to the annotators — is this ok? OR train annotators more — is this ok?

- **Overtrain**: Instructions are so exhaustive that there is no room for thought or interpretation (annotators follow a 'table lookup' procedure)
  - How to determine: is task simply easy, or are annotators overtrained?
  - What's really wrong with overtraining? No predictive power…

Agreement analysis

Sometimes, one annotator is bad

Sometimes, the senses are bad

Sometimes, the word is just hard

				Annotators		vs. Adjudicator				
noun	**total annotated**	**number adjudicated**	**%adj**	**A1-A2 agr**	**A1-A2 agr%**	**A1-Adj agr%**	**A2-Adj agr%**	**Col G+H**	**What to do**	
term	349	64	18.3	285	81.7	87.5	10.9	98.4	A2 bad	A2=ticrea
amount	310	78	25.2	232	74.8	91.0	8.9	99.9	A2 bad	A2=ticrea
return	281	52	18.5	229	81.5	13.4	84.6	98.0		
payment	270	73	27.0	197	73.0	49.3	50.7	100.0	split	
control	262	161	61.5	102	38.9	26.1	71.4	97.5		
activity	245	140	57.1	108	44.1	10.7	91.4	102.1	A1 bad	A1=mccorley
building	231	38	16.5	193	83.5	36.8	63.2	100.0		
average	220	16	7.3	191	86.8	100.0	0.0	100.0	A2 bad	A2=sklaver
place	205	137	66.8	68	33.2	65.7	26.3	92.0		
support	198	27	13.6	171	86.4	25.9	74.1	100.0		
department	145	0	0.0	145	100.0			0.0		
marketing	167	85	50.9	83	49.7	60.0	40.0	100.0	split	
game	163	60	36.8	125	76.7	86.7	60.0	146.7		
import	157	104	66.2	59	37.6	76.0	29.8	105.8		
competition	152	97	63.8	5	3.3	42.2	57.7	99.9	split	
situation	143	49	34.3	76	53.1	65.3	42.9	108.2		
material	129	30	23.3	99	76.7	10.0	90.0	100.0	A1 bad	A1=tsukerman
form	131	31	23.7	100	76.3	58.1	38.7	96.8	split	
trend	113	28	24.8	86	76.1	17.9	85.7	103.6		
protection	111	41	36.9	70	63.1	22.0	78.0	100.0		
date	102	84	82.4	18	17.6	23.8	72.6	96.4		
requirement	95	86	90.5	9	9.5	95.4	3.5	98.9	A2 bad	A2=mccorley
saving	89	59	66.3	29	32.6	96.6	3.4	100.0	A2 bad	A2=mccorley
structure	87	19	21.8	68	78.2	100.0	0.0	100.0	A2 bad	A2=mccorley
recovery	75	17	22.7	58	77.3	76.5	23.5	100.0		
traffic	57	16	28.1	42	73.7	81.2	6.2	87.4	A2 bad	A2=mccorley
challenge	54	26	48.1	34	63.0	73.0	50.0	123.0		
location	54	17	31.5	37	68.5	88.2	11.8	100.0		
merchant	51	34	66.7	17	33.3	0.0	100.0	100.0	A1 bad	A1=tsukerman
beginning	50	25	50.0	26	52.0	60.0	44.0	104.0	split	

Annotation rates: English

English		#types = 9190								
	avg	**at 3/15**	**3/15 - 4/15**	**4/15 - 5/15**	**5/15 - 6/28**	**6/28 - 8/15**	**8/15 - 9/25**	**9/25 - 12/10**	**12/10 - 2/10**	**2/15 - 3/20**
sensed		136	145	249	315	370	500	630	731	754
			9	104	66	55	130	130	101	23
hours sensing										
d-annot types		138	149	217	272	359	415	465	540	570
(words)			11	68	55	87	56	50	75	30
d-annot types		17.5	18.9	24.3	31.3	43.3	44.7	46.4	47.6	48.6
(% of corpus)			1.4	5.4	7	12	1.4	1.7	1.2	1
hours annotating		353.9	115.1	69.7	106.4	197	56.8	111.2	165.7	352.9
		includes training								includes training
rate sensing (words/hr)										
rate sensing (hrs/word)										
rate d-annot types (words/hr)	0.56		0.10	0.98	0.52	0.44	0.99	0.45	0.45	
rate d-annot types (hrs/word)	3.02		10.46	1.03	1.93	2.26	1.01	2.22	2.21	
rate d-annot types (%corpus /hr)	0.04		0.01	0.08	0.07	0.06	0.02	0.02	0.01	
rate dannot types (hrs/%corpus)	52.97		82.21	12.91	15.20	16.42	40.57	65.41	138.08	

Rate varies widely: due to re-sensing?

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Q5: Annotation procedure

- How to manage the annotation process?
  - When annotating multiple variables, annotate each variable separately, across whole corpus — speedup and local expertise … but lose context
  - The problem of 'annotation drift': shuffling and redoing items
  - Annotator attention and tiredness; rotating annotators
  - Complex management framework, interfaces, etc.
- **Reconciliation**
  - Allow annotators to discuss problematic cases, then continue — can greatly improve agreement but at the cost of drift / overtraining
- Backing off: In cases of disagreement, what do you do?
  - (1) make option granularity coarser; (2) allow multiple options; (3) increase context supporting annotation; (4) annotate only major / easy cases
- Experts: …?

- **Adjudication** after annotation, for the remaining hard cases
  - Have an expert (or more annotators) decide in cases of residual disagreement — but how much disagreement can be tolerated before just redoing the annotation?

Q5: Annotation procedure heuristics

- Overall approach — Shulman's rule: do the easy annotations first, so you've seen the data when you get to the harder cases
- The '85% clear cases' rule (Wiebe):
  - Ask the annotators also to mark their level of certainty
  - There should be a lot of agreement at high certainty — the clear cases
- Hypothesis (Rosé): for up to 50% incorrect instances, it pays to show the annotator possibly buggy annotations and have them correct them (compared to having them annotate anew)

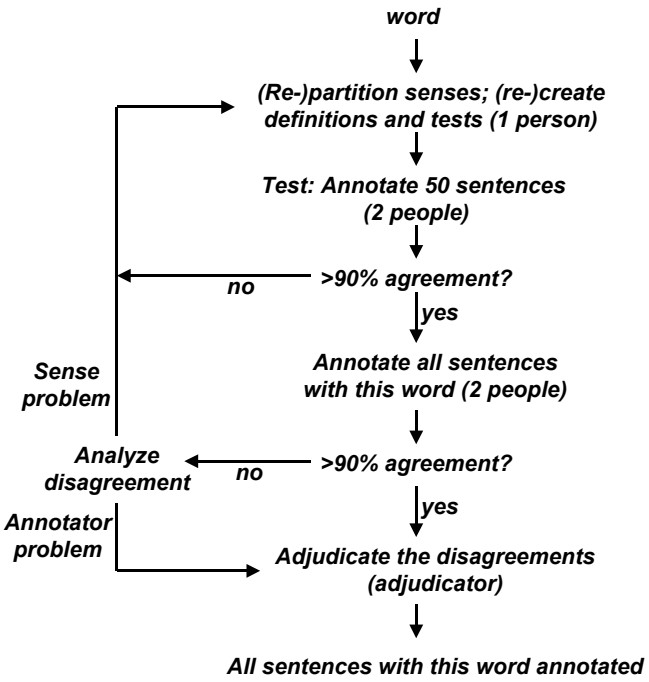
- **Active learning**: In-line process to dynamically find problematic cases for immediate tagging (more rapidly get to the 'end point'), and/or to pre-annotate (help the annotator under the Rosé hypothesis)
  - Benefit: speedup; danger: misleading annotators

OntoNotes annotation procedure

- **Sense creation** process goes by word:
  - Expert creates meaning options (shallow semantic senses) for verbs, nouns, [adjs, advs] … follows PropBank process (Palmer et al.)
  - Expert creates definitions, examples, differentiating features
  - (Ontology insertion: At same time, expert groups equivalent senses from different words and organizes/refines Omega ontology content and structure … process being developed at ISI)
- **Sense annotation** process goes by word, across docs:
  - Process developed in PropBank
  - Annotators manually…
    - See each sentence in corpus containing the current word (noun, verb, [adjective, adverb]) to annotate
    - Select appropriate senses (= ontology concepts) for each one
    - Connect frame structure (for each verb and relational noun)
- **Coref annotation** process goes by doc:
  - Annotators connect co-references within each doc

Sense annotation procedure

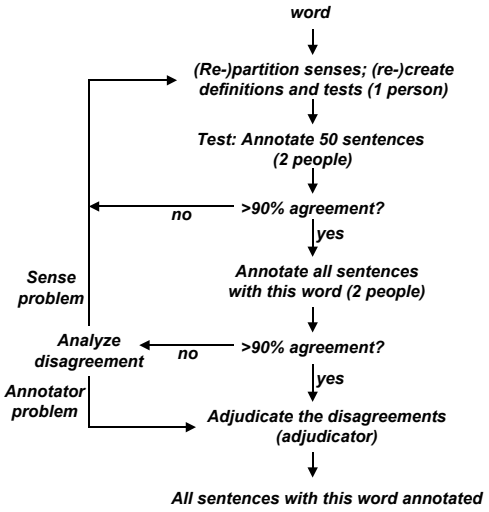
- Sense creator first creates senses for a word
- Loop 1:
  - Manager selects next nouns from sensed list and assigns annotators
  - Programmer randomly selects 50 sentences and creates initial Task File
  - Annotators (at least 2) do the first 50
  - Manager checks their performance:
    - 90%+ agreement + few or no *NoneOfAbove* — send on to Loop 2
    - Else — Adjudicator and Manager identify reasons, send back to Sense creator to fix senses and defs



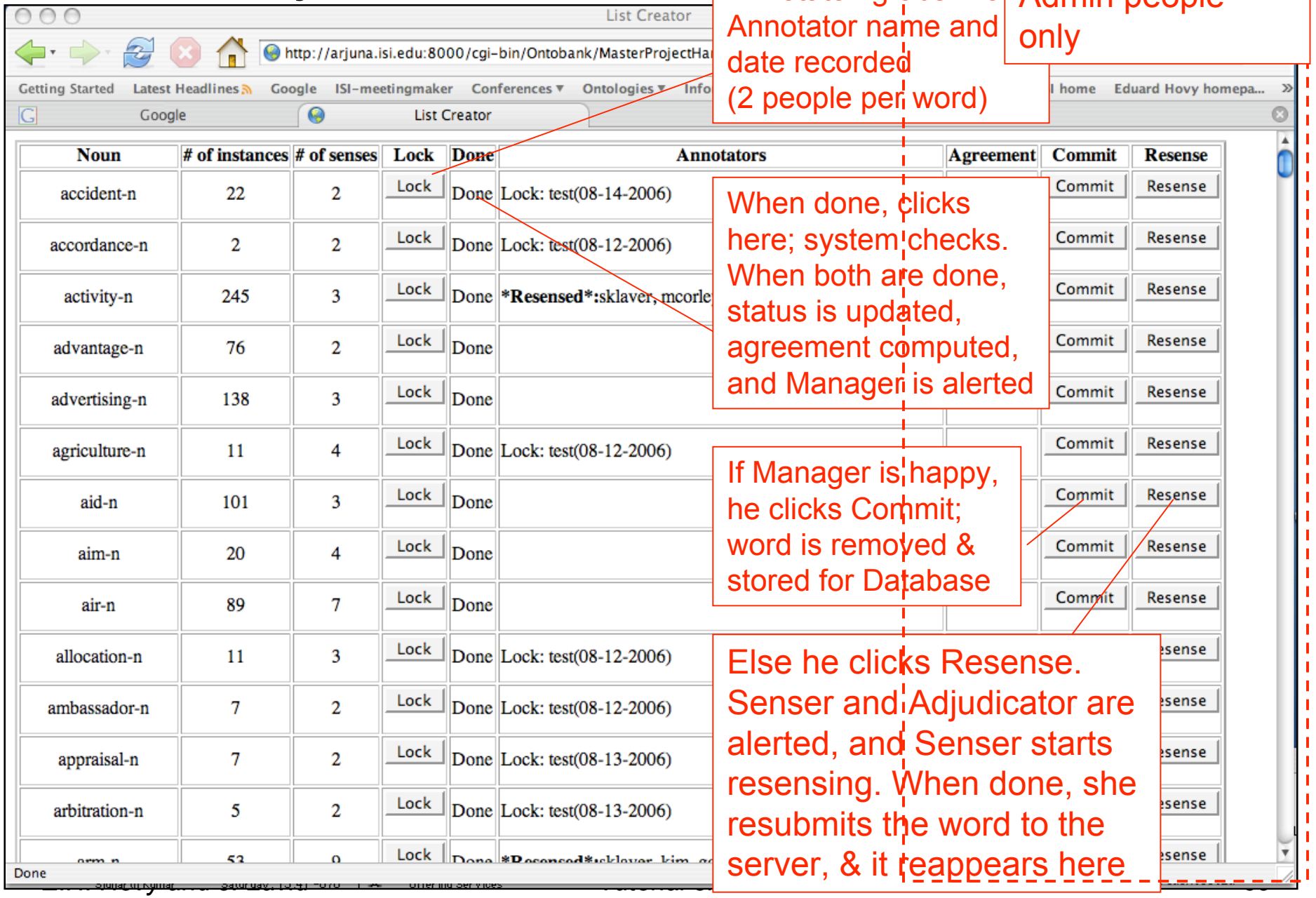
- Loop 2:
  - Annotators (at least 2) annotate all the remaining sentences
  - Manager checks their performance:
    - 90%+ agreement + few or no *NoneOfAbove* — send to Adjudicator to fix the rest
    - Else — Adjudicator annotates differences
    - If Adj agrees with one Annotator 90%+, then ignore other Annotator's work (assume a bad day for the other); else Adj agrees with both about equally often, then assume bad senses and send the problematic ones back to Sense creator

Annotation framework

- Data management:
  - Defined a data flow pathway that minimizes amount of human involvement, and produces status summary files (avg speed, avg agreement with others, # words done, total time, etc.)
  - Several interacting modules:
    - STAMP (built at UPenn, Palmer et al.): annotation
    - Server (ISI): store everything, with backup, versioning, etc.
    - Sense Creation interface (ISI): define senses
    - Sense Pooling interface (ISI): group together senses into ontology
    - Master Project Handler (ISI): annotators reserve word to annotate
    - Annotation Status interface (ISI): up-to-the-minute status
    - Statistics bookkeeper (ISI): individual annotator work



Master Project Handler



Annotator 'grabs' wo
Annotator name and
date recorded
(2 people per word)

This part visible to
Admin people
only

When done, clicks
here; system checks.
When both are done,
status is updated,
agreement computed,
and Manager is alerted

If Manager is happy,
he clicks Commit;
word is removed &
stored for Database

Else he clicks Resense.
Senser and Adjudicator are
alerted, and Senser starts
resensing. When done, she
resubmits the word to the
server, & it reappears here

Status page

Dynamically updated

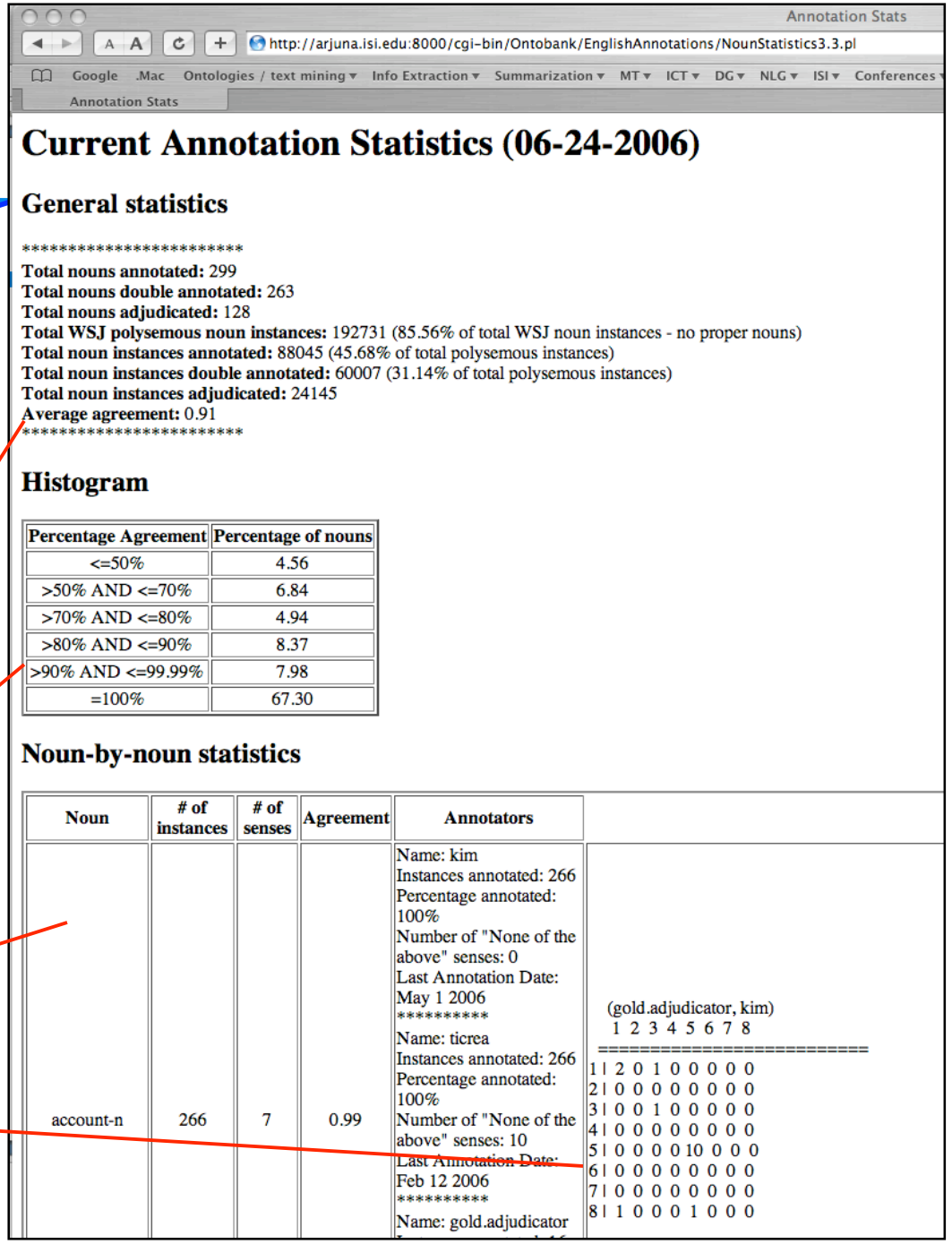
http://arjuna.isi.edu:8000/Ontobank/AnnotationStats.html

Current status: # nouns annotated, # adjudicated; agreement levels, etc.

Agreement histogram

Individual noun stats: annotators, agreement, # sentences, # senses

Confusion matrix for results



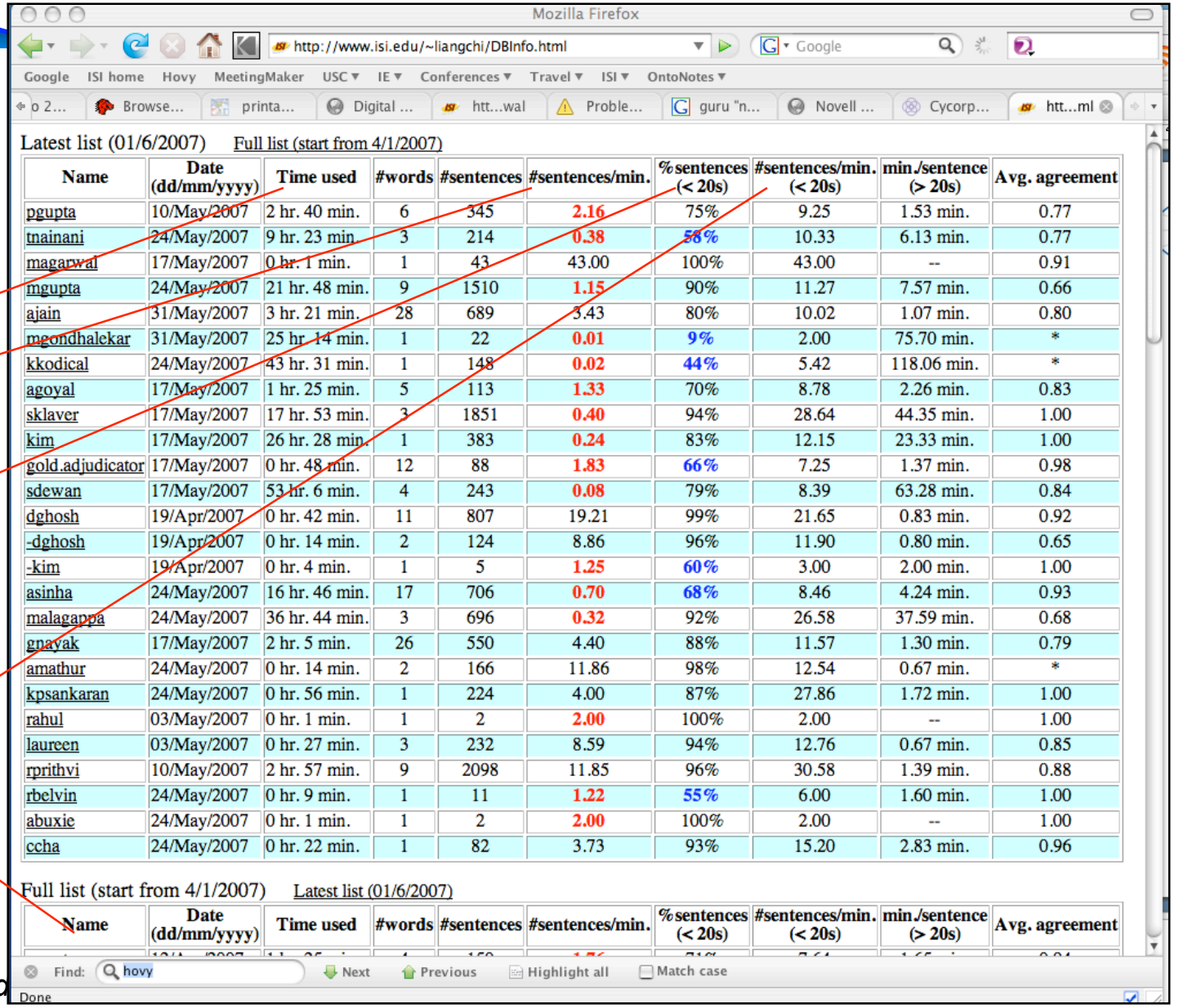
E.H. Hovy and J. Lavid

Annotator work record

Most recent week, each person:

- Total time
- Avg rate
- % of time working at acceptable rate (3/min)
- # sentences at acceptable rate

Full history of each person, weekly



Name	Date (dd/mm/yyyy)	Time used	#words	#sentences	#sentences/min.	%sentences (< 20s)	#sentences/min. (< 20s)	min./sentence (> 20s)	Avg. agreement
pgupta	10/May/2007	2 hr. 40 min.	6	345	2.16	75%	9.25	1.53 min.	0.77
tnainani	24/May/2007	9 hr. 23 min.	3	214	0.38	58%	10.33	6.13 min.	0.77
magarwal	17/May/2007	0 hr. 1 min.	1	43	43.00	100%	43.00	--	0.91
mgupta	24/May/2007	21 hr. 48 min.	9	1510	1.15	90%	11.27	7.57 min.	0.66
ajain	31/May/2007	3 hr. 21 min.	28	689	3.43	80%	10.02	1.07 min.	0.80
mgondhalekar	31/May/2007	25 hr. 14 min.	1	22	0.01	9%	2.00	75.70 min.	*
kkodical	24/May/2007	43 hr. 31 min.	1	148	0.02	44%	5.42	118.06 min.	*
agoyal	17/May/2007	1 hr. 25 min.	5	113	1.33	70%	8.78	2.26 min.	0.83
sklaver	17/May/2007	17 hr. 53 min.	3	1851	0.40	94%	28.64	44.35 min.	1.00
kim	17/May/2007	26 hr. 28 min.	1	383	0.24	83%	12.15	23.33 min.	1.00
gold.adjudicator	17/May/2007	0 hr. 48 min.	12	88	1.83	66%	7.25	1.37 min.	0.98
sdewan	17/May/2007	53 hr. 6 min.	4	243	0.08	79%	8.39	63.28 min.	0.84
dghosh	19/Apr/2007	0 hr. 42 min.	11	807	19.21	99%	21.65	0.83 min.	0.92
-dghosh	19/Apr/2007	0 hr. 14 min.	2	124	8.86	96%	11.90	0.80 min.	0.65
-kim	19/Apr/2007	0 hr. 4 min.	1	5	1.25	60%	3.00	2.00 min.	1.00
asinha	24/May/2007	16 hr. 46 min.	17	706	0.70	68%	8.46	4.24 min.	0.93
malagappa	24/May/2007	36 hr. 44 min.	3	696	0.32	92%	26.58	37.59 min.	0.68
gnayak	17/May/2007	2 hr. 5 min.	26	550	4.40	88%	11.57	1.30 min.	0.79
amathur	24/May/2007	0 hr. 14 min.	2	166	11.86	98%	12.54	0.67 min.	*
kpsankaran	24/May/2007	0 hr. 56 min.	1	224	4.00	87%	27.86	1.72 min.	1.00
rahul	03/May/2007	0 hr. 1 min.	1	2	2.00	100%	2.00	--	1.00
laureen	03/May/2007	0 hr. 27 min.	3	232	8.59	94%	12.76	0.67 min.	0.85
rprithvi	10/May/2007	2 hr. 57 min.	9	2098	11.85	96%	30.58	1.39 min.	0.88
rbelvin	24/May/2007	0 hr. 9 min.	1	11	1.22	55%	6.00	1.60 min.	1.00
abuxie	24/May/2007	0 hr. 1 min.	1	2	2.00	100%	2.00	--	1.00
ccha	24/May/2007	0 hr. 22 min.	1	82	3.73	93%	15.20	2.83 min.	0.96

Name	Date (dd/mm/yyyy)	Time used	#words	#sentences	#sentences/min.	%sentences (< 20s)	#sentences/min. (< 20s)	min./sentence (> 20s)	Avg. agreement

E.H. Hovy and J. Lavid

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Q6.1: Validating annotations

- Evaluating individual pieces of information:
  - What to evaluate:
    - Individual agreement scores between creators
    - Overall agreement averages?
  - What measure(s) to use:
    - Simple agreement is biased by chance agreement — however, this may be fine, if all you care about is a system that mirrors human behavior
    - Kappa is better for testing inter-annotator agreement. But it is not sufficient — cannot handle multiple correct choices, and works only pairwise
    - Krippendorff's alpha, Kappa variations…; see (Bortz 2005; 6th ed; in German)
  - Tolerances:
    - When is the agreement no longer good enough? — why the 90% rule? (Marcus's rule: if humans get *N*%, systems will achieve (*N-10*)%)
  - The problem of asymmetrical/unbalanced corpora
    - When you get high agreement but low Kappa — does it matter? An unbalanced corpus makes choice easy but Kappa low. Are you primarily interested in annotation qua annotation, or in doing the task?
- Experts: Psych experimenters and Corpus Analysis statisticians

Agreement counts: Kappa

- Simple agreement:
  - *A = number choices agreed / total number*
- But what about random agreement? Fix using Cohen's Kappa:
  - *E = expected number of choices agreed / total number*
  - *Kappa = (A - E) / (1 - E)*

'normalize' by removing random agreement (100% - E)

- Example:
  - Assume 100 examples, 50 labeled *A*, and 50 *B*: E_{random} = 0.5
  - Then a random annotator would score 50%: A_{random} = 0.5
  - $Kappa_{random}$ = (0.5 - 0.5) / (1 - 0.5) = 0
  - And an annotator with 70% agreement?: A_{70} = 0.7
  - $Kappa_{70}$ = (0.7 - 0.5) / (1 - 0.5) = 0.2 / 0.5 = 0.4
  - This is much lower than 0.7, but reflects the nonrandom agreement
- Shortcomings of Kappa:
  - Works only to compare 2 annotators (else use *Fleiss's Kappa*)
  - Doesn't apply when multiple correct choices possible
  - Penalizes when choice distribution is skewed — but if that's the nature of the data, then why penalize?

Q6.2: Validating someone's corpus

- But also, evaluate aspects of 'metadata':
  - **Theory and model**:
    - What is the underlying/foundational theory?
    - Is there a model of the theory for the annotation? What is it?
    - How well does the corpus reflect the model? And the theory? Where were simplifications made? Why? How?
  - **Creation**:
    - What was the procedure of creation? How was it tested and debugged?
    - Who created the corpus? How many people? What training did they have, and require? How were they trained?
    - Overall agreement scores between creators
    - Reconciliation/adjudication/purification procedure and experts
  - **Result**:
    - Is the result enough? What does 'enough' mean? (Sufficiency: when the machine learning system shows no increase in accuracy despite more training data)
    - Is the result consistent (enough)?
    - Is it correct? (can be correct in various ways!)
    - How was it used?

Dealing with imbalance

- After a certain amount of annotation, you will almost certainly find ‘imbalance’
- Certain choices underrepresented in the corpus
- Why?
  - Limited/biased corpus selection
  - Biased choice creation
  - Poor annotation

- How can you redress the balance?
- *Should* you?

OntoNotes noun statistics after Yr 1

- For 465 most frequent nouns in WSJ annotated:
  - total senses = about 2080
  - average number of senses per word = 4.47
  - 60.8% of nouns have 2–4 senses
    - "head" has largest number of senses: 32 senses
  - 78.9%of the *polysemous* nouns in WSJ need *only one* sense (predominant sense) (!)
  - 93.3% instances are covered by topmost 2 senses
  - 497 senses (23.9%) do not occur at all (!)
  - 254 nouns (54.6%) have at least one unseen sense (!)
  - Nouns, sorted by entropy of tags
- So: WSJ part of OntoNotes is an *unbalanced corpus* — we need another as well
  - It is very difficult to use such a skewed corpus for identifying infrequent or unseen senses

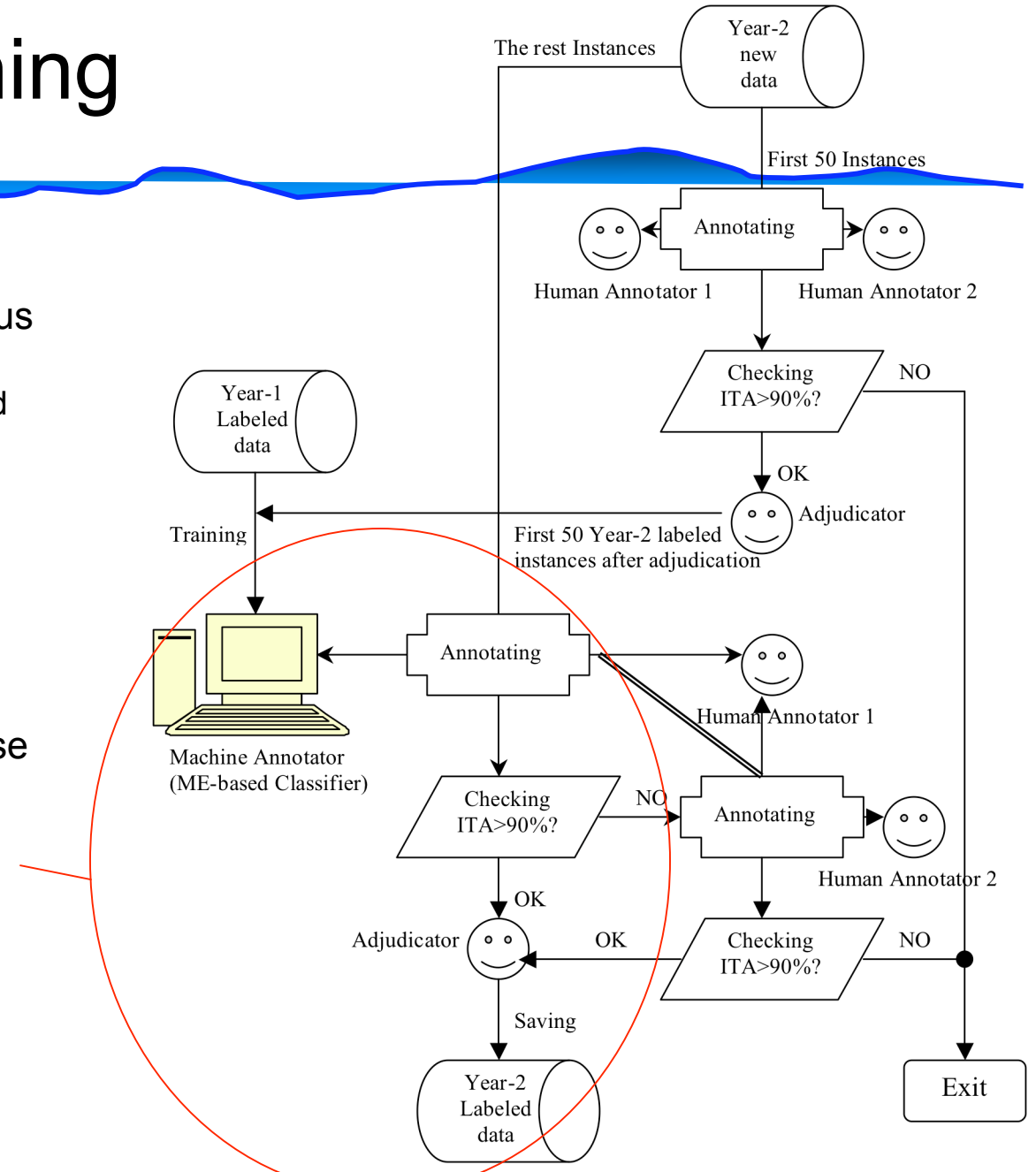
1:**play-n** senses=13 instances=41 agreement=0.490000 entropy:2.003143 distribution:0.239 0.217 0.130 0.109 0.087 0.087 0.065 0.043 0.021739 0.00
2:**control-n** senses=8 instances=262 agreement=0.820000 entropy:1.788795 distribution:0.416 0.168 0.119 0.097 0.065 0.058 0.026 0.023 0.016 0.013
3:**defense-n** senses=8 instances=149 agreement=1.000000 entropy:1.776716 distribution:0.261 0.248 0.164 0.128 0.106 0.058 0.035 0.00 0.00 0.00
4:**bar-n** senses=16 instances=27 agreement=1.000000 entropy:1.767424 distribution:0.310 0.241 0.138 0.103 0.069 0.069 0.069 0.00 0.00 0.00
5:**life-n** senses=9 instances=272 agreement=0.620000 entropy:1.763602 distribution:0.333 0.187 0.184 0.075 0.066 0.043 0.026 0.020 0.014 0.006
…
225:**arbitration-n** senses=2 instances=5 agreement=0.800000 entropy:0.500402 distribution:0.80 0.20 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
226:**condition-n** senses=4 instances=171 agreement=0.930000 entropy:0.500055 distribution:0.817 0.178 0.005 0.00 0.00 0.00 0.00 0.00 0.00 0.00
227:**value-n** senses=5 instances=403 agreement=0.000000 entropy:0.497184 distribution:0.849 0.123 0.028 0.00 0.00 0.00 0.00 0.00 0.00 0.00
228:**technology-n** senses=2 instances=203 agreement=0.820000 entropy:0.494763 distribution:0.871 0.072 0.049 0.008 0.00 0.00 0.00 0.00 0.00 0.00
229:**policy-n** senses=2 instances=382 agreement=1.000000 entropy:0.493565 distribution:0.805 0.195 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
…
397:**accordance-n** senses=2 instances=2 agreement=1.000000 entropy:0.000000 distribution:1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
…same all the way down

Adapting active learning for WSD

- **Problem**: Human annotation is expensive and time-consuming
  - Can we use Active Learning to minimize human annotation effort?
- **Imbalance** is a problem:
  - WSJ sense distribution is very skewed — creates large discrepancy between the prior probabilities of the individual senses:
    - For all annotated nouns: about 78.9% of nouns are covered by the first sense, and about 93.3% by the top two senses
    - For only the nouns with high agreement: 86% are covered by top sense; 95.9% by top 2 senses; 98.5 by top 3
    - 497 senses (23.9%) do not occur at all (!)
    - 254 nouns (54.6%) have at least one unseen sense (!)
  - Calculated entropy of sense distributions; sorted into three classes:
    - Extremely imbalanced — almost all instances (97%+) are same sense
    - Highly imbalanced — 85%–97% of instances are dominant sense
    - Somewhat imbalanced — more flat distribution over senses
- **Active learning** is promising way to enrich OntoNotes
  - But need to balance infrequent senses — how?

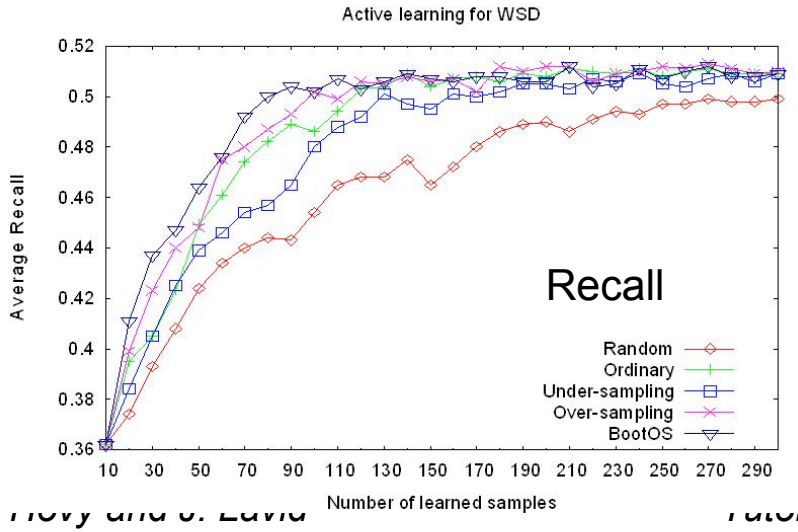
Role of learning

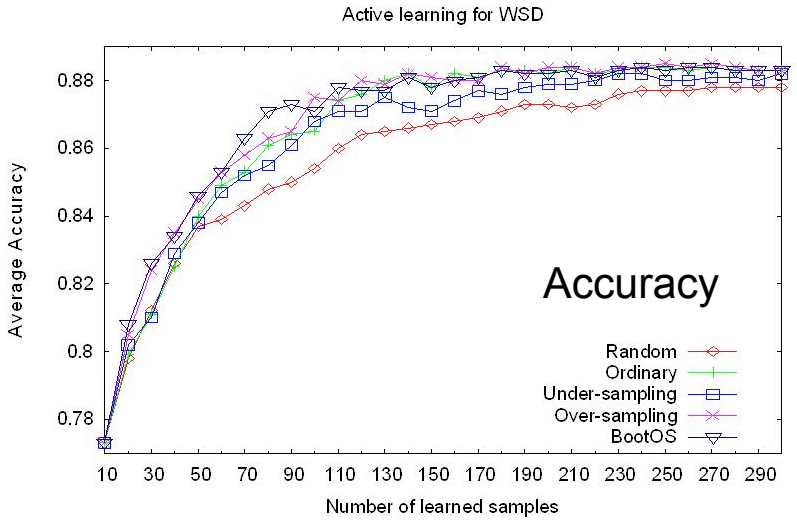
- Approach:
  - Use results of Yr1 corpus to train sense classifier
    - if word is skewed, add pre-annotation of 50 examples from Yr2
  - Apply classifier to Yr2 corpus
  - Compare results with Annotator 1 output
  - If agreement >90%, do not use Annotator 2; else use Annotator 2
- Benefit: may save one annotator on Yr2 text
- Experiment: build sense classifier and try various sampling techniques
  - 5-fold cross-validation
  - 80% training, 20% test



Dealing with imbalance (Zhu and Hovy, EMNLP-07)

- Idea:
  - **Undersampling**: remove majority class instances (up to 0.8x)
  - **Oversampling**: add randomly chosen copies (duplicates) of minority class instances (up to 1.8x)
  - **Bootstrap Oversampling**: like oversampling, but construct new samples using k-NN and similarity functions
- Experiments:
  - Which sampling method?
  - When to stop sampling process?





Active Learning trials

- Experiments on nouns:
  - Setup:
    - 17 nouns; 6 of them fully double-annotated
    - ITA: 13 over 90%; 4 over 80%
  - Expt 1:
    - Trained on Yr I corpus, tested on Yr II
    - Results: 9 over 90%; 3 over 80%; 4 over 70%; 1 over 50% (average 84%)
  - Expt 2:
    - Trained on Yr I + top 50 instances of Yr II corpora; tested on rest of Yr II
    - Results: 10 over 90%; 3 over 80%; 2 over 70%; 2 over 60% (average 87%)
  - Predictiveness: If machine agreement is high with Human1, is it also with Human2?
    - Yes: in only 1 case (of 6) is the H2 agreement significantly lower
- Bottom line: Can save some time—**more than half the frequent nouns can be machine-annotated**, replacing one person

Results

(Zhu and Hovy 07)

Nice outcome: Can save around 50% annotation effort for frequent-enough nouns

Pairwise agreements (2 Humans, 2 Machine systems)

Word	#Instances	H1 - H2	H1 - M1	H1 - M2	M1 - M2	Human 1	Human 2	H2 - M1	H2 - M2	M1 - M2
people-n	1288	0.96	0.83	0.85	0.93	sklaver	kim(51)			0.93
country-n	783	0.90	0.99	0.99	0.99	sklaver	kim(51)			0.99
today-n	684	0.92	0.72	0.73	0.98	kim	sklaver(50)			0.98
development-n	563	0.96	0.95	0.98	0.96	kim	sklaver(50)			0.96
trade-n	431	0.88	0.90	0.91	0.99	kim	sklaver(50)			0.99
company-n	423	0.98	0.99	1.00	1.00	kim	sklaver(50)			1.00
area-n	410	0.82	0.72	0.72	0.93	kim	sklaver(115)			0.93
state-n	368	0.94	0.51	0.61	0.87	kim	sklaver(50)			0.87
number-n	360	1.00	0.79	0.84	0.93	asinha	kim(51)			0.93
economy-n	355	0.98	1.00	1.00	1.00	kim	sklaver(51)			1.00
system-n	313	0.82	0.53	0.64	0.85	kim	magarwal(51)			0.85
group-n	283	1.00	1.00	1.00	1.00	ajain	mgupta	1.00	1.00	1.00
management-n	165	0.87	0.75	0.80	0.90	kpsankaran	mgupta	0.76	0.83	0.90
role-n	138	0.92	0.93	0.94	0.97	kpsankaran	ajain	0.89	0.89	0.97
director-n	132	0.95	0.88	0.93	0.95	kpsankaran	mgupta	0.91	0.95	0.95
death-n	109	0.94	0.95	0.96	0.99	sdewan	ajain	0.93	0.92	0.99
food-n	91	0.97	0.92	0.97	0.96	agoyal	tnainanii	0.96	0.98	0.96
AVG AGREEMENT		**0.94**	**0.84**	**0.87**	**0.95**			**0.91**	**0.93**	**0.95**

trained on on YI only

trained on YrI+50 of YrII

last 6 words fully double-annotated

Predictiveness

M1 = machine trained on YrI only
M2 = machine trained on YrI+50 of YrII

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Q7: Delivery

- It's not just about annotation…

  How do you make sure others use the corpus?

- Technical issues:
  - Licensing
  - Distribution
  - Support/maintenance (over years?)
  - Incorporating new annotations/updates: layering
  - Experts: Data managers

Problems with multiple annotation layers

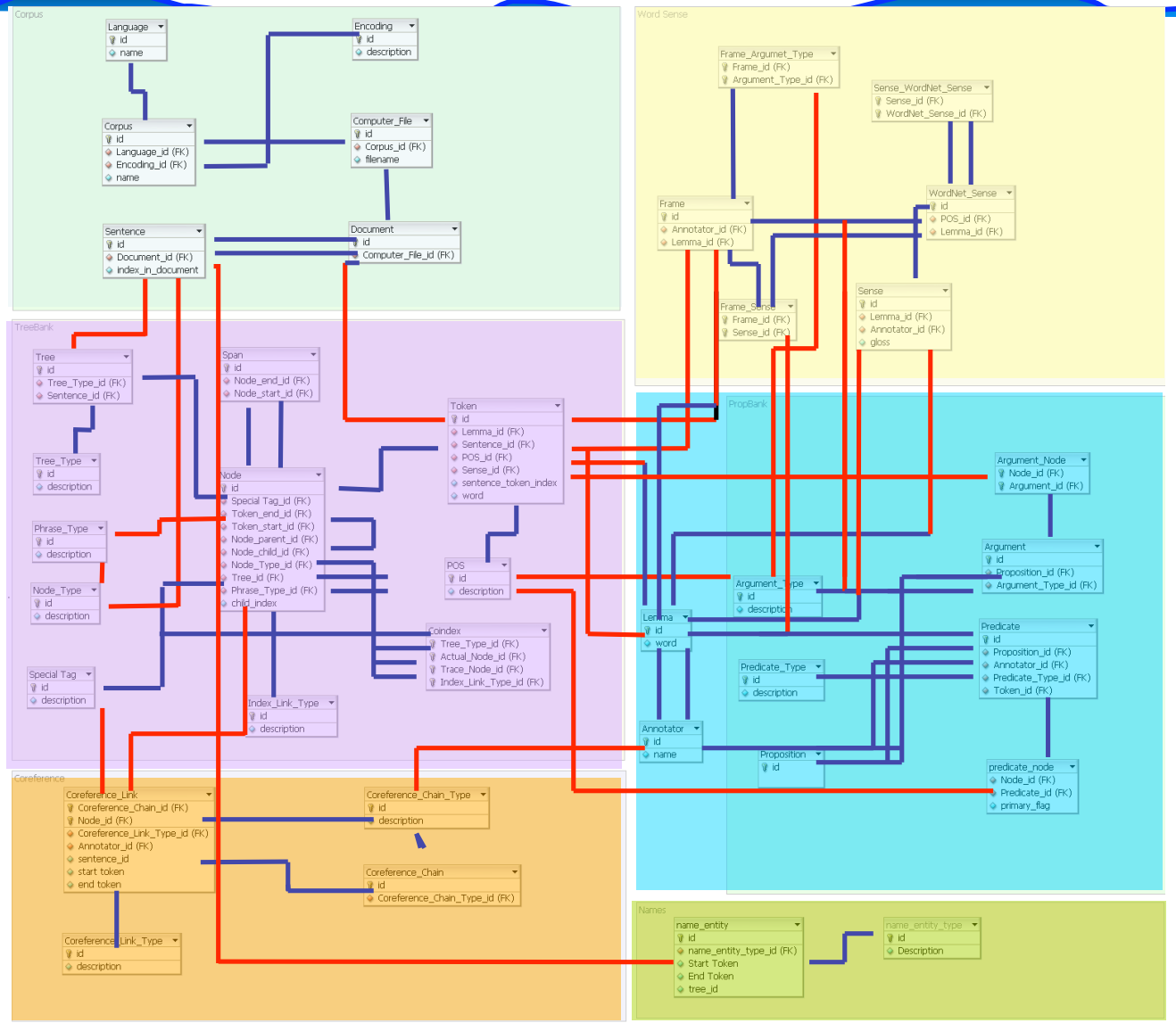
(Slide by Sameer Pradhan, BBN)

- Problems:
  - Not previously available or integrated
    - Most projects address only a single annotation type (layer)
    - And when multiple, 'annotation units' may not align
      - Each layer encoded separately as individual files, requiring supporting documentation for interpretation
  - Not previously completely consistent
    - E.g., mismatches between Treebank and PropBank
  - Not previously user friendly (raw text format…)
- Goal: Provide a bare-bones representation independent of the individual semantics that can
  - Efficiently capture intra- and inter- layer semantics
  - Maintain component independence
  - Provide mechanism for flexible integration
  - Integrate information even at the lowest level of granularity
  - Allow easy cross-layer queries

OntoNotes Solution: **Relational Database + Object Oriented API**

Database: Unified relational rep

(Slide by Sameer Pradhan, BBN)



Corpus

Trees

Coreference

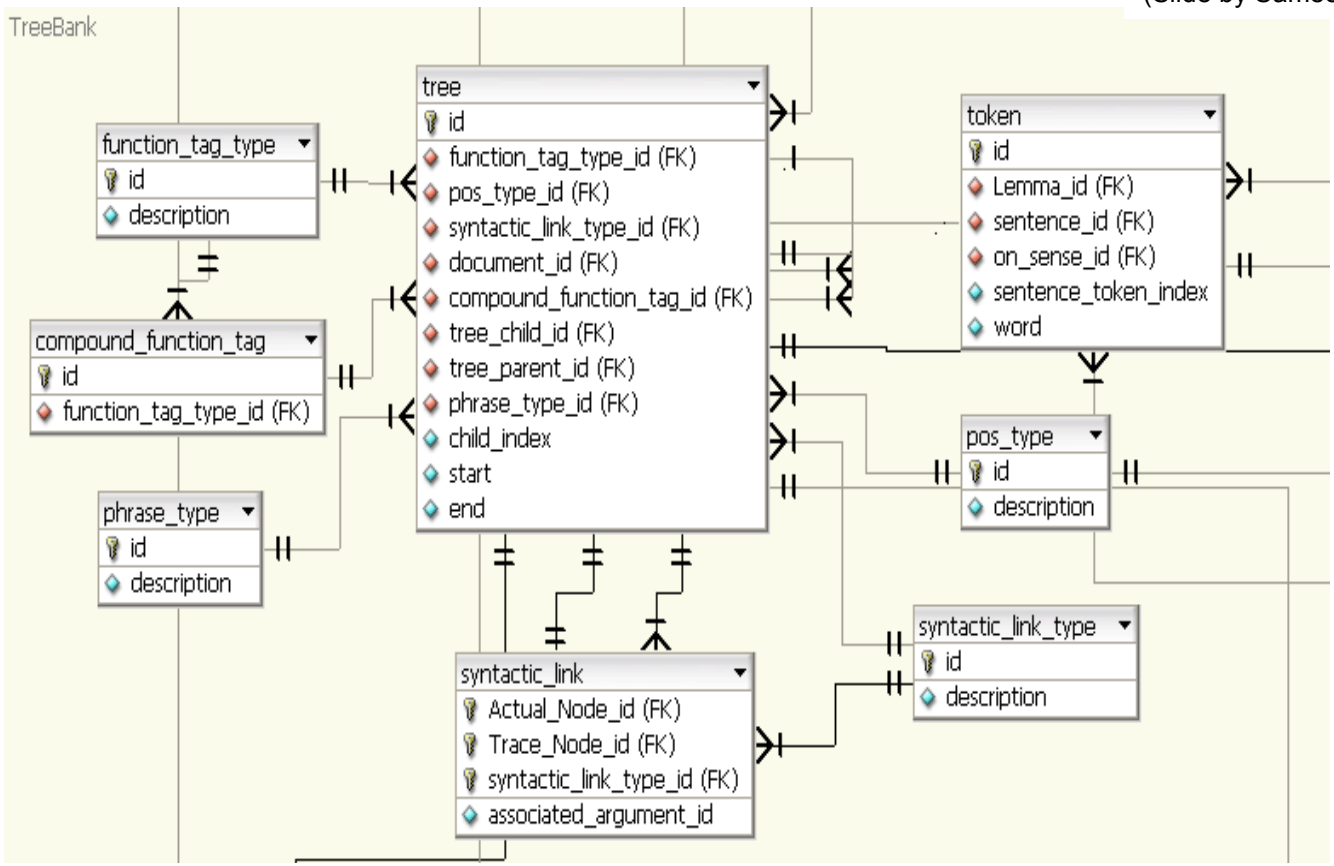
Senses

Propositions

Names

Example: DB representation of syntax

(Slide by Sameer Pradhan, BBN)



- Treebank tokens (stored in the Token table) provide the common base
- The Tree table stores the recursive tree nodes, each with its span
- Subsidiary tables define the sets of function tags, phase types, etc.

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Exercise 2: Ontologizing

- For each noun, create senses
  - Manual procedure (1 person)
- Verify senses
  - Corpus annotation (2 people + adjudicator)

Sense creation & annotation

- Group synonymous senses into sense pools
  - Manual procedure (1 person for nouns, 1 for verbs)
- Verify pool contents
  - Google + manual procedure
- Attach pool into ontology Upper Model
  - Manual attachment (3 people for nouns, 2 for verbs)
- Verify attachment agreement

Why ontologize?

Word senses alone are ok, but you can do more

1. Synonymous word senses grouped together
   (*tightest grouping, by synonymy*)
   - MT and Distillation: use word replacements

2. Sense groups ('pools') taxonomized to allow inheritance
   (*looser grouping: semantic relatedness*)
   - Distillation: use for compaction of sentences

3. Pertinent features added to sense pools
   - MT: use to translate the Chinese '*de*'?—choose approp prep/etc.
   - Distillation: use for output generation—choose approp answer form

Why an ontology?

- Current HLT systems depend on impoverished text models:
  - Bags of words, ngram word sequences, syntactic structure
- OntoNotes provides a (very slightly) deeper and more semantic (meaning-based) representation that:
  - Resolves meaning ambiguity of words in terms of senses
  - Connects the word senses to an ontology of symbols
  - The ontology symbols are organized in semantic clusters
  - The symbols also contain features
- **Why not just senses?**
- For more effective HLT systems, it may be useful to exploit the symbols' organization and features
  - Applications (Information Extraction, Question Answering, Summarization…) and tasks (entailment, semantic analysis for learning by reading, etc.) all use inference
  - Ontology may support limited inference for **term expansion, term substitution, term matching, structure matching**, etc.

Noun and verb sense creation

- Performed by Ann Houston in Boston
- Sense groups created:
  - 4 to 6 nouns sense-created per day
  - Max: “head”, with 15 senses
  - Verb procedure creates senses by grouping WordNet senses (PropBank)
  - Noun procedure taxonomizes senses into trees, with differentiae at each level, for insertion into ontology
  - For each sense, add features
- Group senses into semantic ‘concepts’
- Sense groups manually inserted under Omega Upper Model

price
Sense 1: +abstract +quantity +monetary_value
Sense 2: +physical +activity +complex (not single event or action) +effort

Grouped with sense of “sacrifice”

Grouped with senses of “value”, “cost”

examples and tests
WN groups
differentiae

<inventory lemma="**price-n**">
<**sense n="1"** type="" name=**"cost or monetary value of goods or services"** group="1">
<diff> +quantity +monetary_value </diff>
<comment> PRICE of NP -> NP's[+good/+service] PRICE[+exchange_value] </comment>
<examples>
The price of gasoline has soared lately.
I don't know the prices of these two fur coats.
The museum would not sell its Dutch Masters collection for any price.
The cattle thief has a price on his head in Maine.
They say that every politician has a price.
</examples>
<mappings> <wn version="2.1">1,2,4,5,6</wn> <omega> </omega> </mappings>
</sense>

<**sense n="2"** type="" name=**"sacrifice required to achieve something"** group="1">
<diff> +activity +complex +effort </diff>
<comment> PRICE{+effort] PREP(of/for)/SCOMP NP[+goal/+result] </comment>
<examples>
John has paid a high price for his risky life style.

Sense Pool P2968: *tank*

Features (from individual noun senses and for pool overall)

Senses from individual nouns (some from sense creator, some from WordNet or MIKRO)

E.H. Hovy and J. Lavid

Noun pool P2968

Definition(s):

1. Jail - holding cell, detention center, place of incarceration
2. a dentention center, locked cell where a prisoner is kept.
3. bullpen.a.n.2

Local features(s):

[+building] [+incarceration] [+room] [+locked] [+center] [+detention]

Pool sense(s):

bullpen.a.n.2
- a detention center for prisoners or refugees
- BULLPEN[+center][+detention][+prisoner/+refugee]

bullpen.o.n.1 (≡ |PRISON-BUILDING|)
- a place of confinement for those who are convicted by or are awaiting trial

hold.a.n.7
- a jail cell
- HOLD[+entity][+artifact][+building][+incarceration][+room][+locked]

hold.o.n.6 (≡ |hold<cell|)
- a cell in a jail or prison

jail.a.n.1
- place of confinement, prison
- JAIL[+entity][+artifact][+building][+incarceration]

jail.o.n.1 (≡ |PRISON-BUILDING|)
- a place of confinement for those who are convicted by or are awaiting trial

jail.o.n.3 (≡ |jail|)
- a correctional institution used to detain persons who are in the lawful custody of the gove
  a sentence)

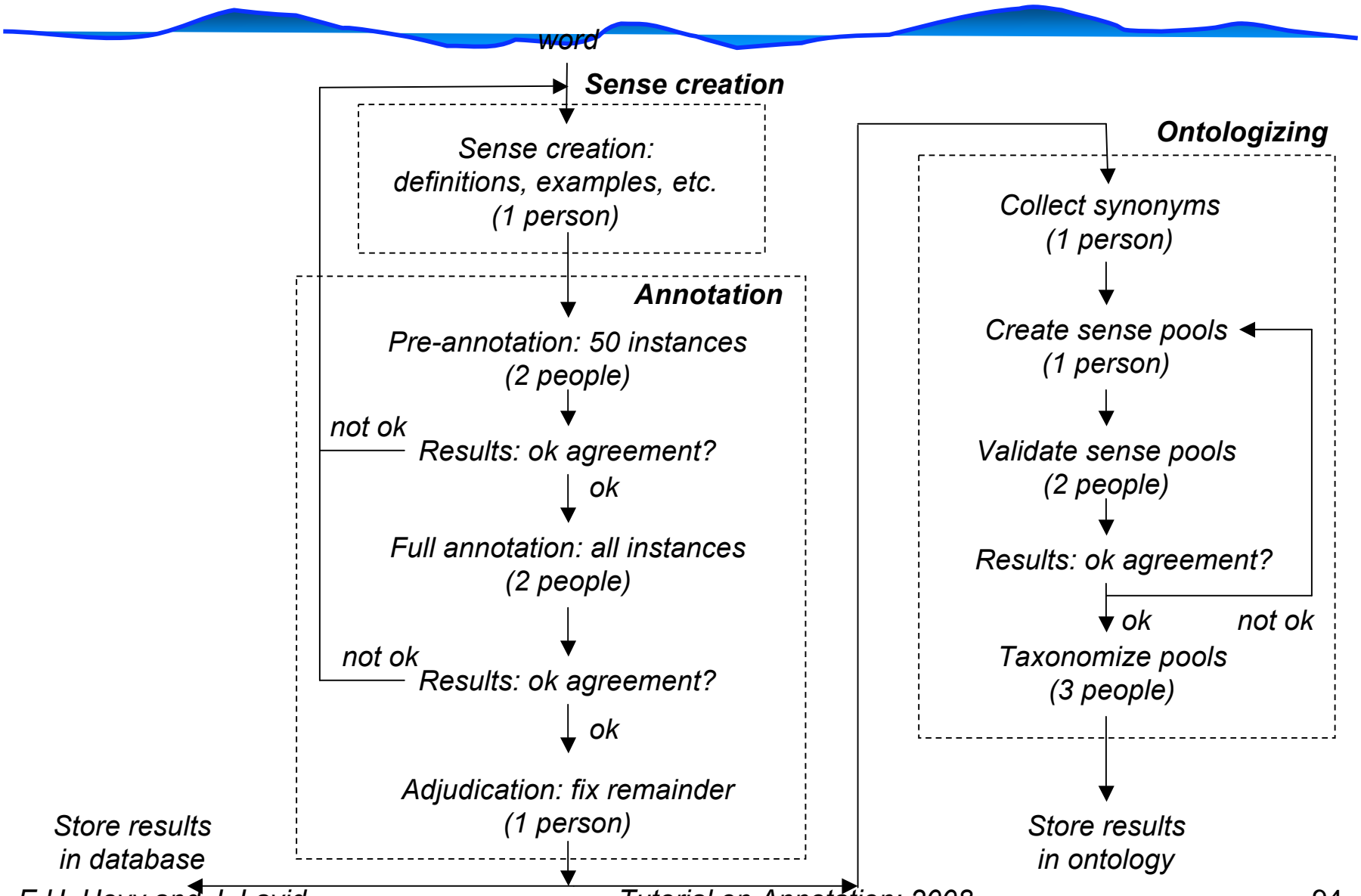
tank.a.n.4
- a detention center for prisoners or drunks
- TANK[+center][+detention][+prisoner/+drunkard]

tank.o.n.2 (≡ |PRISON-BUILDING|)
- a place of confinement for those who are convicted by or are awaiting trial

tank.o.n.6 (≡ |tank<cell|)

Done

Complete procedure



Verifying pools: Normalization & cutoff

(Yu and Hovy 08)

Raw substitution frequency				C(W→Bridge) = x		
bridge		causeway	overpass	viaduct	tunnel	error
bridge over the	1,250,000	9,890	19,600	29,000	33,300	162,000
Raw term frequency				C(W) = y		
bridge		causeway	overpass	viaduct	tunnel	error
	222,000,000	7,960,000	3,190,000	4,400,000	66,200,000	506,000,000
Normalization				NC(W→Bridge) = x/y		
bridge		causeway	overpass	viaduct	tunnel	error
bridge over the	0.006	0.001	0.006	0.007	0.0005	0.0003
Ratio with "bridge" after normalization				NC(W→Bridge) / NC(Bridge)		
Bridge		causeway	overpass	viaduct	tunnel	error
bridge over the	0.006	0.001	0.006	0.007	0.0005	0.0003
Ratio		**0.17**	**1**	**1**	**0.08**	0.05

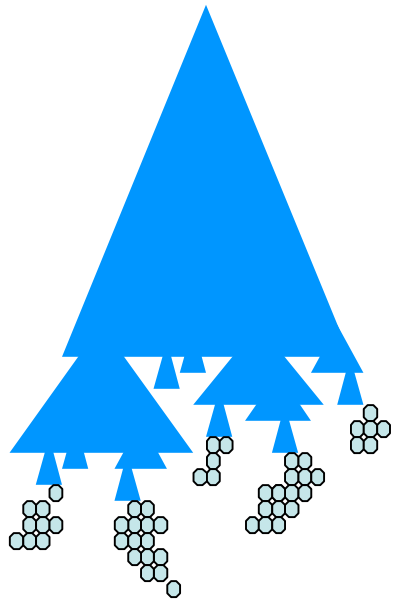
bridge	causeway	overpass	viaduct	tunnel
2-gram	0.373	0.559	0.644	0.548
3-gram	0.176	0.383	0.346	0.397
4-gram	0.176	0.288	0.157	0.282
5-gram	0.000	0.000	0.000	0.000
causeway	bridge	overpass	viaduct	tunnel
2-gram	0.713	0.656	0.484	0.628
3-gram	0.464	0.375	0.250	0.346
4-gram	0.713	0.440	0.186	0.469
5-gram	0.000	0.000	0.000	0.000
overpass	bridge	causeway	viaduct	tunnel
2-gram	0.791	0.648	0.779	0.526
3-gram	0.799	0.575	0.424	0.430
4-gram	0.779	0.276	0.135	0.300
5-gram	0.000	0.000	0.000	0.000
viaduct	bridge	causeway	overpass	tunnel
2-gram	0.549	0.399	0.551	0.494
3-gram	0.630	0.367	0.373	0.586
4-gram	0.585	0.321	0.397	0.556
5-gram	0.000	0.000	0.000	0.000
tunnel	bridge	causeway	overpass	viaduct
2-gram	0.586	0.480	0.625	0.530
3-gram	0.537	0.444	0.536	0.701
4-gram	0.353	0.153	0.166	0.061
5-gram	0.000	0.000	0.000	0.000

	bridge	causeway	overpass	viaduct	tunnel
bridge		0.176	0.288	0.157	0.282
causeway	0.713		0.440	0.186	0.469
overpass	0.779	0.276		0.135	0.300
viaduct	0.585	0.321	0.397		0.556
tunnel	0.353	0.153	0.166	0.061	

{bridge, causeway}	0.445
{bridge, overpass}	0.534
{bridge, viaduct}	0.371
{bridge, tunnel}	0.318
{causeway, overpass}	0.358
{causeway, viaduct}	0.254
{causeway, tunnel}	0.311
{overpass, viaduct}	0.266
{overpass, tunnel}	0.233
{viaduct, tunnel}	0.309
{bridge, causeway, overpass}	0.445
{bridge, causeway, viaduct}	0.356
{bridge, causeway, tunnel}	0.358
{bridge, overpass, viaduct}	0.390
{bridge, overpass, tunnel}	0.361
{bridge, viaduct, tunnel}	0.332
{causeway, overpass, viaduct}	0.293
{causeway, overpass, tunnel}	0.301
{causeway, viaduct, tunnel}	0.291
{overpass, viaduct, tunnel}	0.269
{bridge, causeway, overpass, viaduct}	0.371
{bridge, causeway, overpass, tunnel}	0.366
{bridge, causeway, viaduct, tunnel}	0.334
{bridge, overpass viaduct, tunnel}	0.338
{causeway, overpass, viaduct, tunnel}	0.288
{bridge, causeway, overpass, viaduct, tunnel}	0.340

Omega after OntoNotes

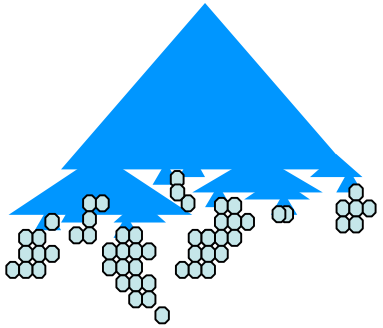
- Old Omega:
  - 120,000 concepts: Middle Model mostly WordNet
  - Upper Model derived through alignments, in earlier work
  - Essentially no formally defined features
  - Fixed hierarchical structure
- Post-OntoNotes Omega:
  - New Upper Model, carefully defined
    - Sense groups manually aligned (and validated) under Upper Model
  - Middle Model: 60,000 concepts?
    - Granularity validated by 90% rule
    - Each concept is a sense group, defined with features
    - No fixed hierarchical structure
  - Instance base:
    - Each concept linked to many example sentences
    - Augment existing instance databases
  - Usage: Used in BBN's GALE Distillation system



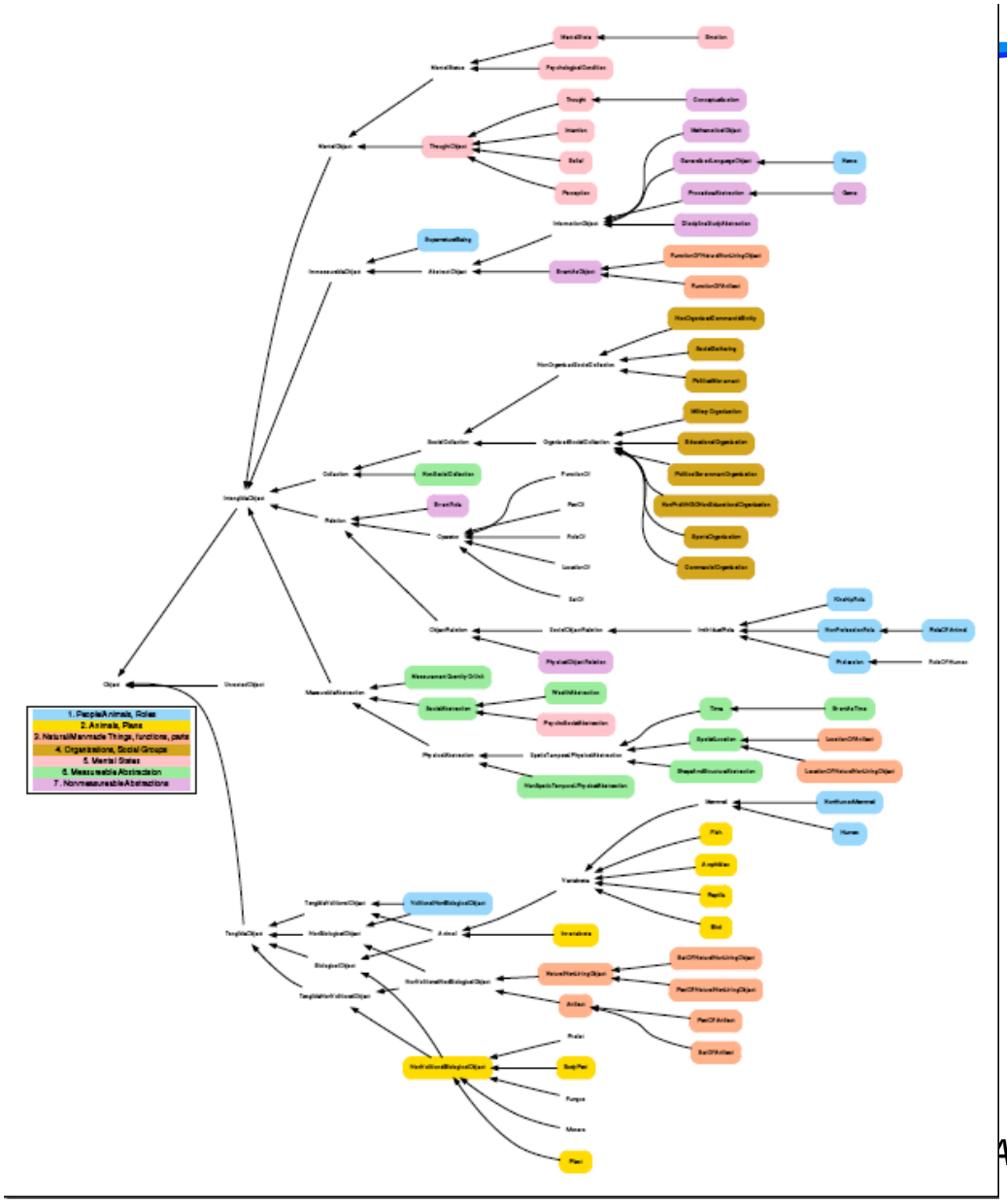
Ontology construction

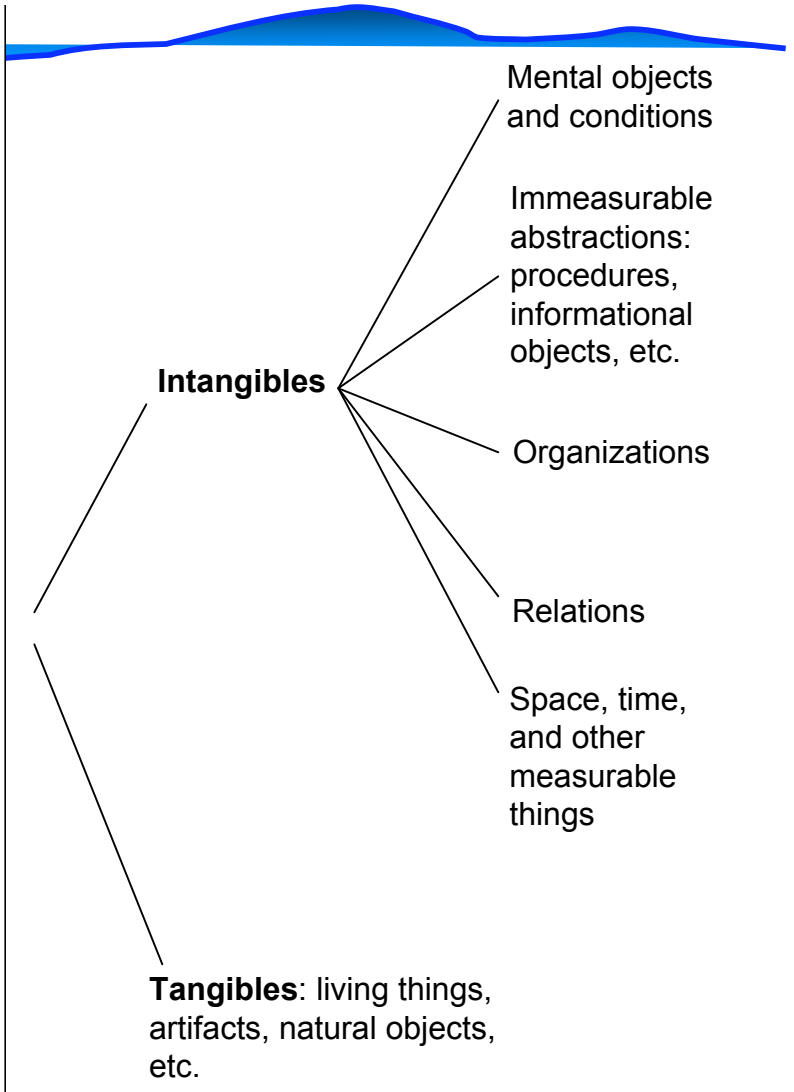
- Goal: Cluster together sense pools by semantic similarity
  - Provide validated initial clustering into major semantic types
  - Enable subsequent more fine-grained subdivision into smaller and deep taxonomies based on specific feature prioritization

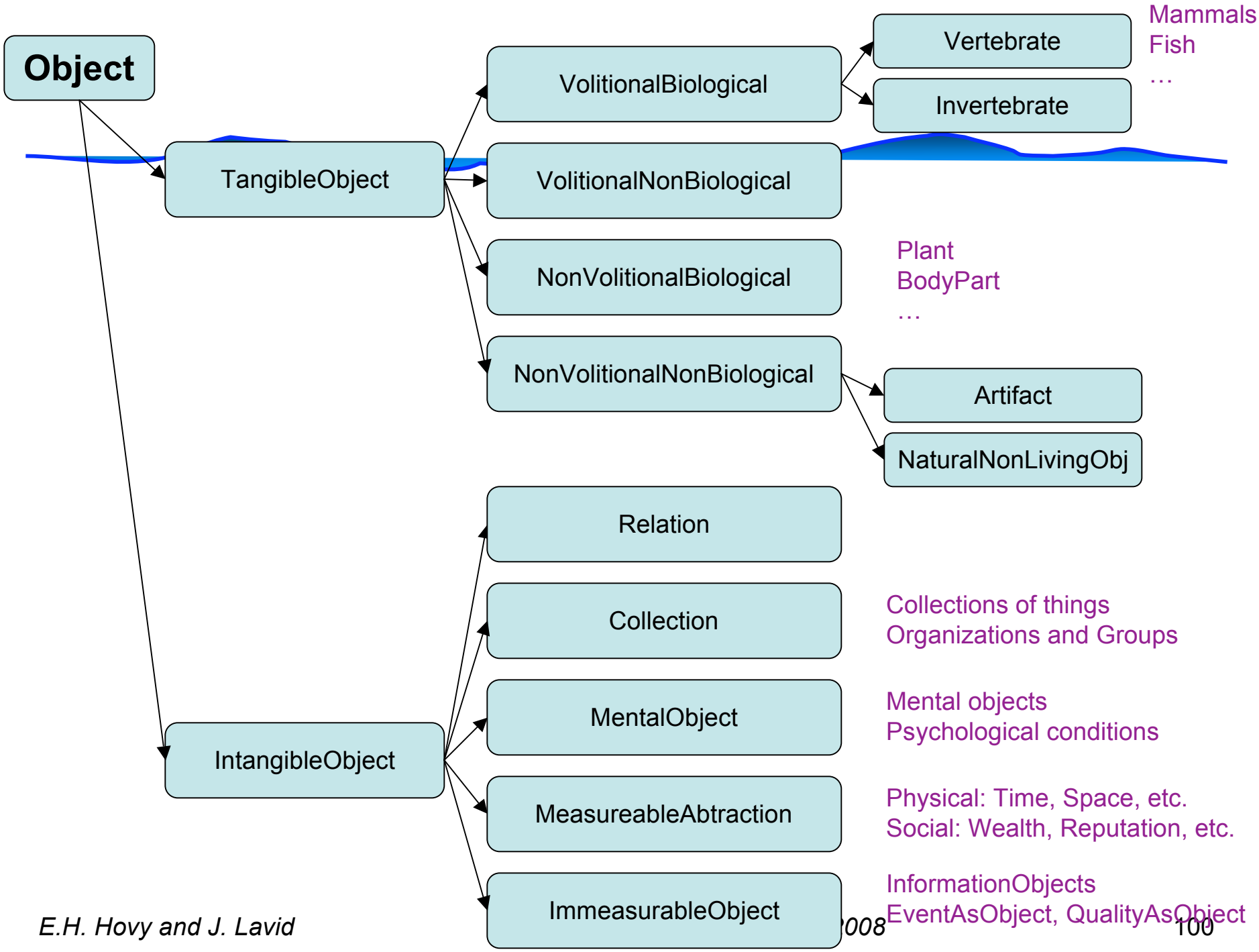
- Overall framework: Omega ontology
  - Upper Model:
    - Very high-level generalizations that partition concepts
    - Approx. 90 nodes (Objects) and 30 nodes (Events)
  - Middle Model:
    - Under construction: Sense pools manually attached to appropriate Upper Model node(s)
    - Linking treated as annotation (3 linkers per pool); apply agreement threshold (just as for annotation)
    - No fixed hierarchical structure: feature order specified by user, which gives hierarchy
  - Progress: about 1000 ontologized to date; still to be finalized

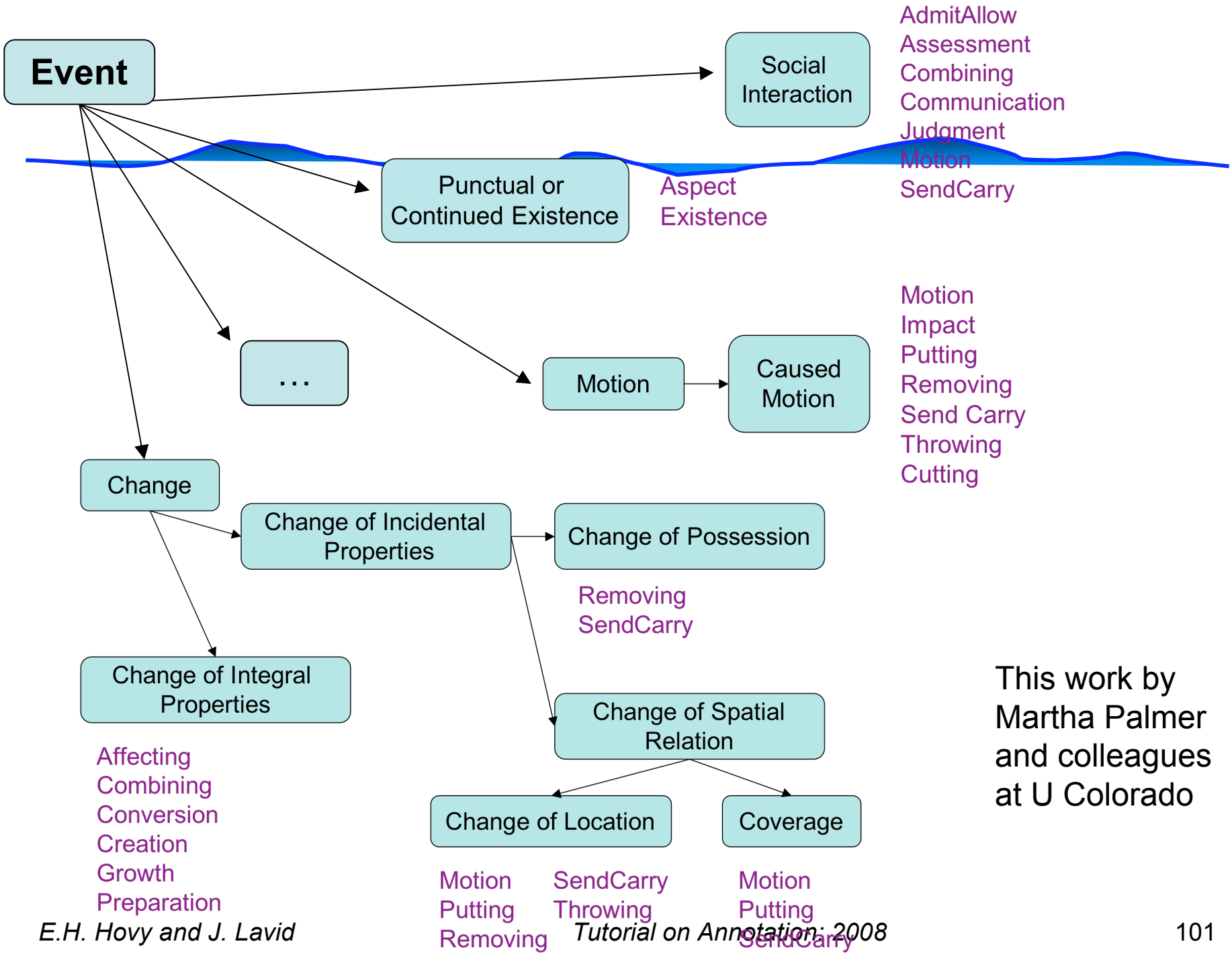


Omega Upper Model for Objects



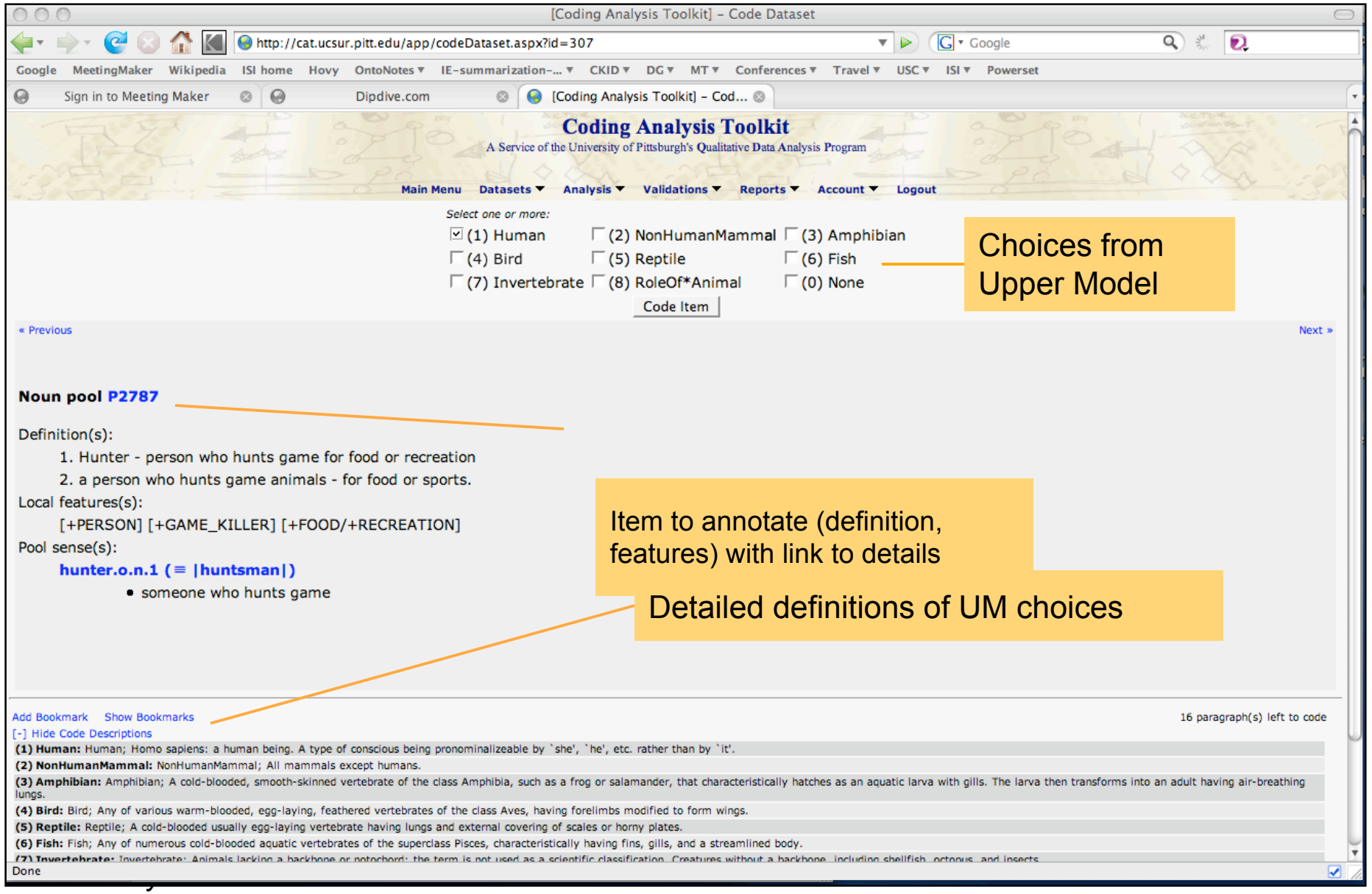



Object
TangibleObject
IntangibleObject
VolitionalBiological
VolitionalNonBiological
NonVolitionalBiological
NonVolitionalNonBiological
Vertebrate
Invertebrate
Mammals
Fish
…
Plant
BodyPart
…
Artifact
NaturalNonLivingObj
Relation
Collection
MentalObject
MeasureableAbtraction
ImmeasurableObject
Collections of things
Organizations and Groups
Mental objects
Psychological conditions
Physical: Time, Space, etc.
Social: Wealth, Reputation, etc.
InformationObjects
EventAsObject, QualityAsObject



This work by Martha Palmer and colleagues at U Colorado

Ontology Attachment Interface



Ontologizing Handbook extract

Sweep 5
Mental states

5.1 Emotion

An emotion, like happiness, peace, anger, etc. Not a psychological condition like schizophrenia. When the concepts is an interpersonal relationship, then it is also a PsychoSocialAbstraction.

5.2 MentalState

This concept represents the mental states of an individual (not a society as a whole) that are neither a (typically clinical) psychological condition (like schizophrenia or bipolar disorder) nor an emotion. (Emotion is a subtype of MentalState.) Typically, MentalStates that are not emotions have longer duration, and can be thought of as habits or traits, such as excitability, placidity, etc., or conditions, such as being focused. Other possible examples are interest (as in: showing interest in), calmness, and the mental state of being glad that a certain politician did not win an election.

Some people feel these are in fact emotions, and should classify such concepts under Emotion instead. So ultimately this may be an empty concept, in which case we will remove it.

5.3 PsychoSocialAbstraction

Abstractions that express human relationships, such as friendship, companionship, etc. These concepts simultaneously have a social and a psychological/emotional component, and tend to be seen from the perspective of an individual, rather than as the sum over a society. Many of them are also linked to Belief, MentalState, or Emotion. They are loosely measurable (though not as precisely quantifiable as PhysicalAbstractions), since one can talk about strong or weak friendship. This concept is a direct child of SocialAbstraction.

5.4 PsychologicalCondition

A (typically clinical) psychological condition, like schizophrenia or bipolar disorder. Not an emotion.

5.5 ThoughtProcess

Concepts denoting a mental process. They may have short duration, such as a thought, an impression, or a perception, or longer duration, such as a deduction, reasoning, or puzzling-out procedure. There is a relationship here with EventAsObject and/or ProceduralAbstraction.

5.6 ImaginaryObject

Imaginary objects, such as unicorns and dragons, the entities in novels and stories (people, places, things, and events), the stuff of dreams, etc. But objects with volition that are claimed to have 'real' (albeit spiritual) existence, such as gods, angels, and ghosts, are classified as SupernaturalBeings.

5.7 NonImaginaryThoughtObject

Non-imaginary mental objects, such as goals, intentions, beliefs, mental images (of real objects), impressions made by someone or some experience, memories, etc. (This is in contrast to imaginary mental objects such as dragons and characters in novels.) But objects with volition that are claimed to have 'real' (albeit spiritual) existence, such as gods, angels, and ghosts, are classified as SupernaturalBeings.

Exercise

Tutorial overview

- Introduction: What is annotation, and why annotate?
- The example project: OntoNotes
- The seven questions of annotation
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
    - Exercise 1: Seeing what we've learned
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Discussion
    - Exercise 2: Practice
- Conclusion

Annotation as a science

- Increased need for corpora and for annotation raises new questions:
  - **What kinds/aspects of 'domain semantics' to annotate?** …it's hardly an uncontroversial notion…
  - Which corpora? How much?
  - Which computational tools to apply once annotation is 'complete'? When *is* it complete?
  - How to manage the whole process?
- Results:
  - A new hunger for annotated corpora
  - A new class of researcher: the Annotation Expert

- Need to systematize annotation process — BUT: How rigorous is Annotation as a 'science'?

Writing a paper in the new style

- How to write a paper about an annotation project (and make sure it will get accepted at LREC, ACL, etc.)?
- Recipe:
  - Problem: phenomena addressed
  - Theory
    - Relevant theories and prior work
    - Our theory and its terms, notation, and formalism
  - The corpus
    - Corpus selection
    - Annotation design, tools, and work
  - Agreements achieved, and speed, size, etc.
  - Conclusion
    - Distribution, use, etc.
    - Future work

Current equiv

problem

past work

training
algorithm

evaluation

distribution

Some current technology and work

- Wide variety of **NLP / machine learning technology** available to learn to mimic human annotations:
  - Simple phrasal patterns (regular expressions)
  - Automated phrasal pattern learning algorithms
  - Markov Models and Conditional Random Fields

- **Kinds of information** typically used for learning **experiments in NLP community**:
  - Parts of speech — solved problem for many languages
  - Named Entities (people, places, organizations, dates, amounts, etc.) — e.g., BBN's IdentiFinder
  - Syntactic structure — somewhat solved for some languages
  - Word senses and argument structure (lexico-semantics)
  - Opinions (both *good/bad* judgments and *true/false* beliefs)
  - Coreference links (pronouns and other anaphora)
  - Discourse structure
  - Various other semantic phenomena — more experimental

In conclusion…

Annotation is **both**:

- A mechanism for providing new training material for machines
- A mechanism for theory formation and validation — in addition to domain specialists, annotation can involve linguists, philosophers of language, etc. in a new paradigm

It’s not only NOT the most boring thing the world…
…it’s actually becoming COOL
(obviously, since we are here now, in this tutorial)

Thank you!

Some readings

- Stability of annotator agreement:
  - Lipsitz, S.R., N.M. Laird, and D.P Harrington. 1991. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* 78(1): 156–160.
- Validation:
  - Bortz, J. 2005. *Statistik für Human- und Sozialwissenschaftler*. Springer Verlag.
  - Cohen's Kappa: Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, pp 37–46.
- OntoNotes:
  - Hovy, E.H., M. Marcus, M. Palmer, S. Pradhan, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% Solution. Short paper. *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006).*
  - Pradhan, S., E.H. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel 2007. OntoNotes: A Unified Relational Semantic Representation. *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC-07).*