Toward a 'Science' of Annotation

Eduard Hovy

Information Sciences Institute
University of Southern California
hovy@isi.edu

USC INFORMATION SCIENCES INSTITUTE

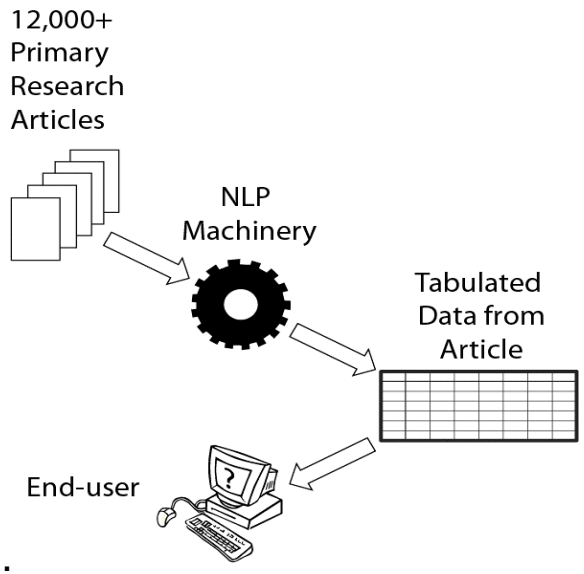
NLP Group at ISI

- **Goals**:
  - **Practical**: Build technology to help people communicate more naturally with computers, and better with one another
  - **Theoretical**: Understand foundations of language and semantics
- **Research methodology**:
  - Integrate **human theorizing** (model building) with **statistical processing** (machine learning) — find optimal balance
  - Evaluate frequently
- **Research group**:
  - One of largest university-based NLP groups in North America (in operation since late 1970s) — approx. 35 people
  - Funding from DARPA, NSF, ARDA, DoD, etc. (= $5M/year)
  - Active interaction with Computational Linguistics worldwide (Best Paper awards; leading roles in prof. societies; service on roadmap committees; etc.)
  - Significant focus on education (2–5 PhD students/year; 2 semester courses per year in CS; MS in CL program joint with Linguistics)

USC INFORMATION SCIENCES INSTITUTE

Information Extraction

- **Information Extraction task**: Identify fragments in texts that express important/useful info; extract; store in database (= IE is a kind of annotation)



- **Why do this?**
  - Create a single coherent database of just the info you care about
  - Cover a large number of sources — many more than you can possibly get by hand
  - Useful for study, reference, and teaching
  - Useful for data mining: finding trends/patterns over time
- **Who uses this today?**
  - Military / Intelligence community
  - Business community
  - Government
  - Biomedical, PoliSci, and other research communities

USC INFORMATION SCIENCES INSTITUTE

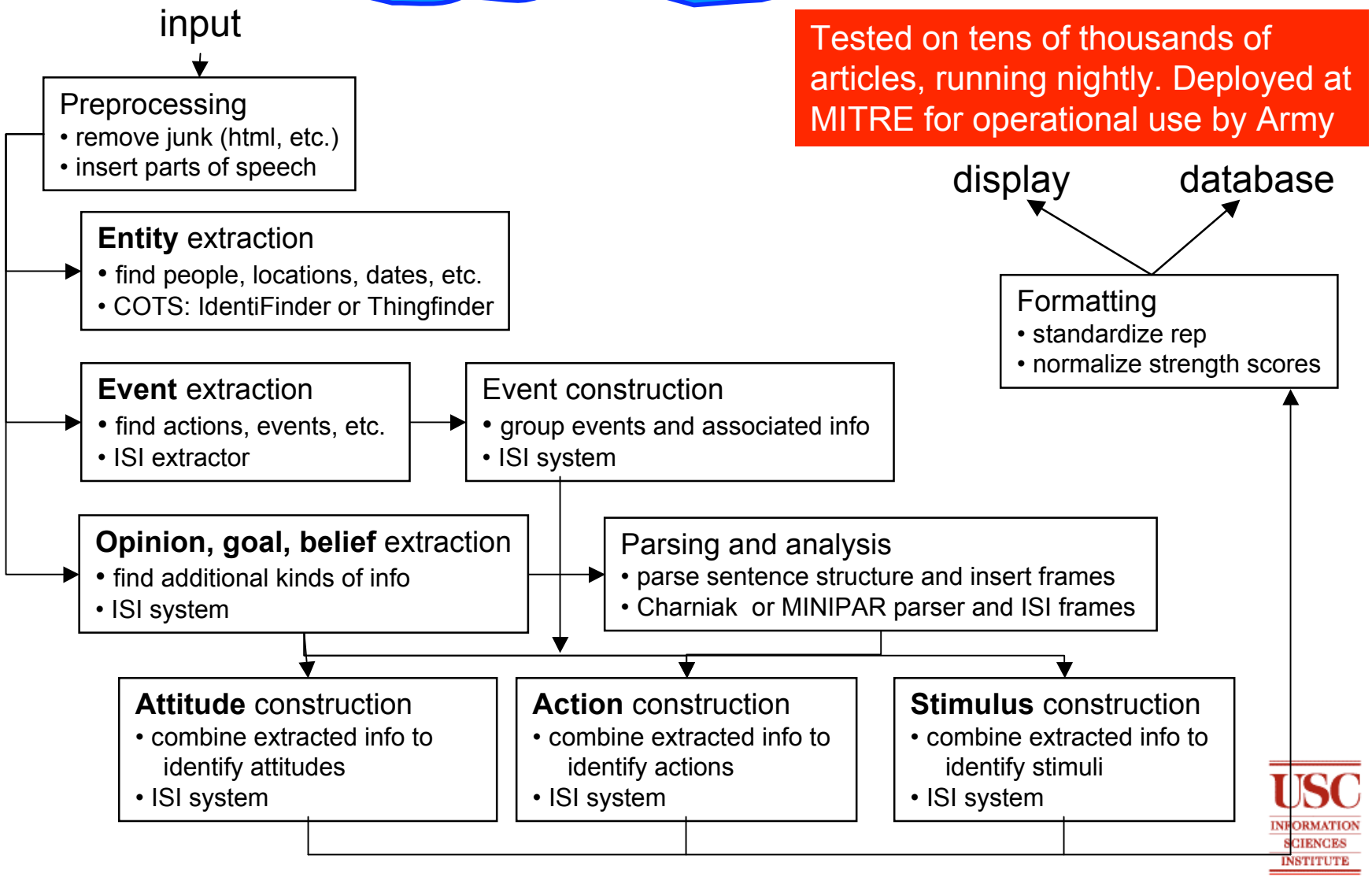
Example: IE of complex notions

- Context: build psychological models of people in focus areas — help Army avoid mistakes in action

- Automatically identify and extract people's *Attitudes, Stimuli,* and *Actions*
- Problem: what are these concepts?
  - Domain experts disagree
  - For Attitude, Opinions are somehow important, but are not everything — also relevant are Goals and Beliefs
- Approach:
  - First extract all simpler pieces (entities, events, goals, beliefs, opinions, etc.)
  - Analyze internal structure of each piece
  - Then try to combine somehow, to identify useful sentences

USC INFORMATION SCIENCES INSTITUTE

Example: Data flow & modules



USC INFORMATION SCIENCES INSTITUTE

Example: Cascading complexity

- **Type**: Attitudes
- **Technology**: Combination of basic factors
  - Goals+Opinions+other things → Attitudes
  - Attitude includes 12 classes:
    - MOTIVATION
    - SUPERLATIVE
    - BELIEF
    - GOAL
    - OPINION
    - RELIGION
    - EXTREMUM
    - REPORT
    - GPE
    - DATE
    - TIME
    - LOCATION

Agree / disagree			# sents at level of agreemnt
Yes(%)	No(%)	System	
100%	0%	high	13
87.50%	12.50%	high	5
0%	100%	none	4
80	20	high	1
75	25	high	1
20	80	medium	1
25	75	medium	1
40	60	medium	1
60	40	high	1
60	40	medium	1

- Each factor has an indicator engine
- Each engine returns a fragment of text (plus usually a score)
- To find the 'attitude strength' of a sentence, we combine the various scores using their *relative strengths*
- This we do by correlating human (SME) judgments

  - Combination function: X = $\sum_i \alpha_i . f_i$ — work in progress to determine optimal coefficients α_l
  - Annotation tests show much higher agreement among domain specialists than for basic factors alone

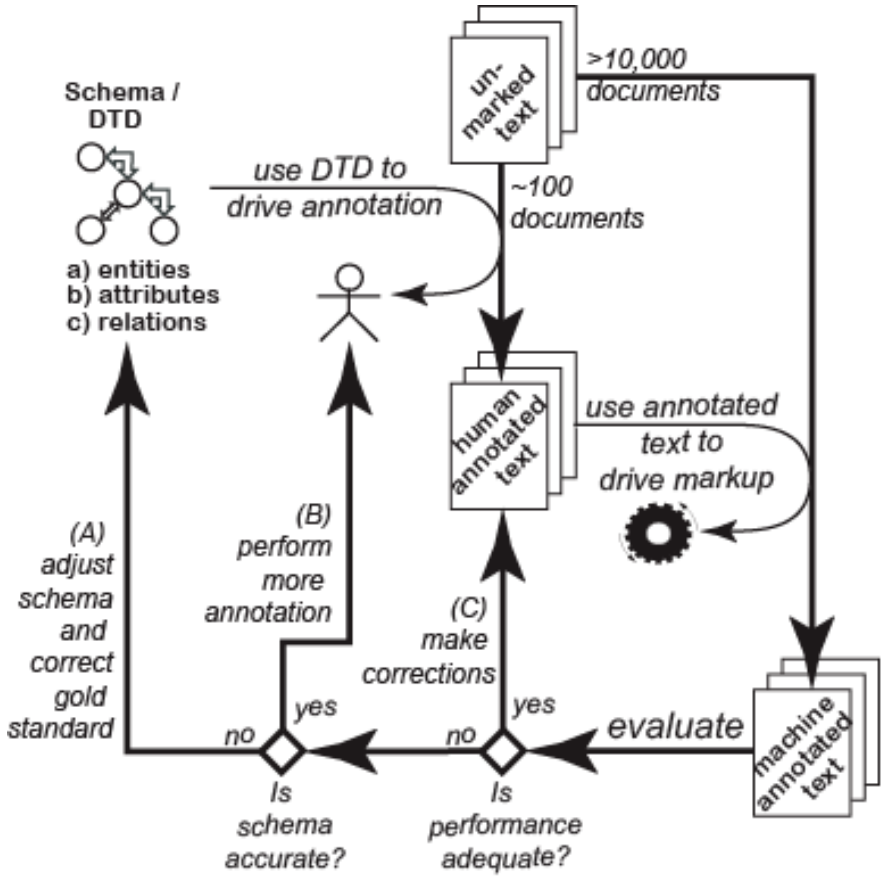
USC INFORMATION SCIENCES INSTITUTE

Annotation for IE

As the items to extract become more complex, IE definition phase becomes harder: move from pre-specified (hard-coded) rules to automated learning…

…and this requires annotation…

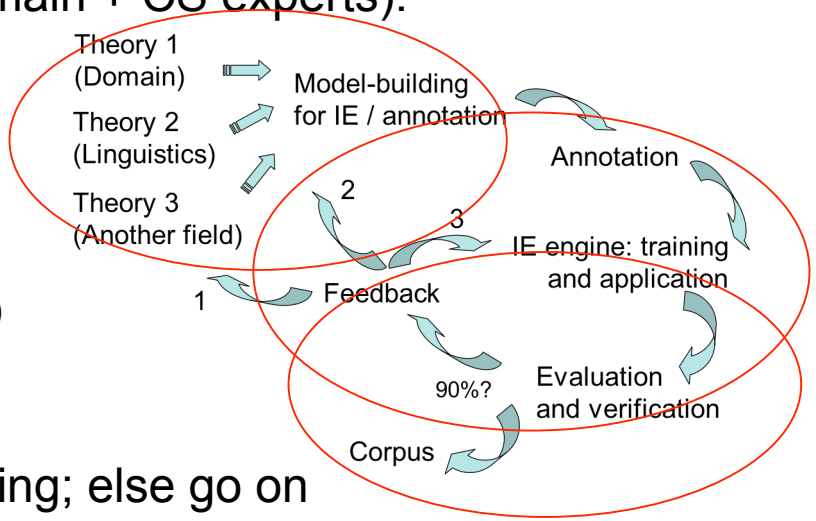
- What is the role of annotation?
- How to define the IE, and how to determine IE acceptability?
- When IE is not acceptable, what can one do?
  - Improve IE algorithms
  - Fix or extend training data
  - Redefine extraction model
  - … etc. …



USC INFORMATION SCIENCES INSTITUTE

Generic IE methodology

- **Design phase** (domain experts):
  - Decide on the information types desired — based on domain theory and model
  - Obtain and prepare document corpus (domain + CS experts)
  - Annotate test documents to determine task feasibility for humans — often this requires further theoretical (model) adjustments or growth
- **Learning phase**: building the system (domain + CS experts):
  - Domain expert: annotate documents
  - CS person: create IE model (identify likely indicator cues; build cue recognizer functions, using words/phrase patterns/ ling. info… as features)
  - CS: Deploy IE learning algorithms (CRF…)
  - Both: evaluate performance on unseen data and assign reliability scores



- **Is the result adequate?** If not, fix something; else go on
- **Application phase**: running the system (domain experts):
  - Obtain more documents
  - Run system (on input, apply all cues for the desired type; collect and merge results using merging functions; save output in database)
  - Reconcile inconsistencies and enjoy the results (or extend your theory!)

USC INFORMATION SCIENCES INSTITUTE

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

USC INFORMATION SCIENCES INSTITUTE

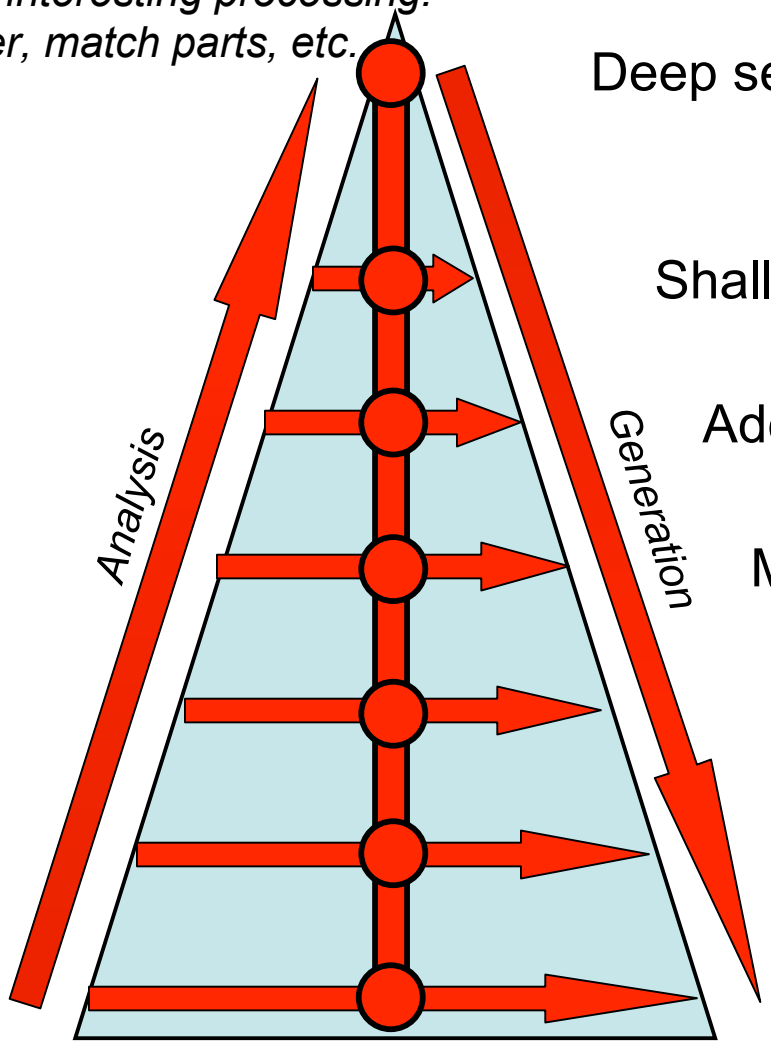
In NLP: Are we entering a new era of corpus building?

- 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’?

USC INFORMATION SCIENCES INSTITUTE

NLP at increasing depths

Do interesting processing: filter, match parts, etc.



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)

- **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

USC INFORMATION SCIENCES INSTITUTE

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**!

- Example: Can this be done for semantics???

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

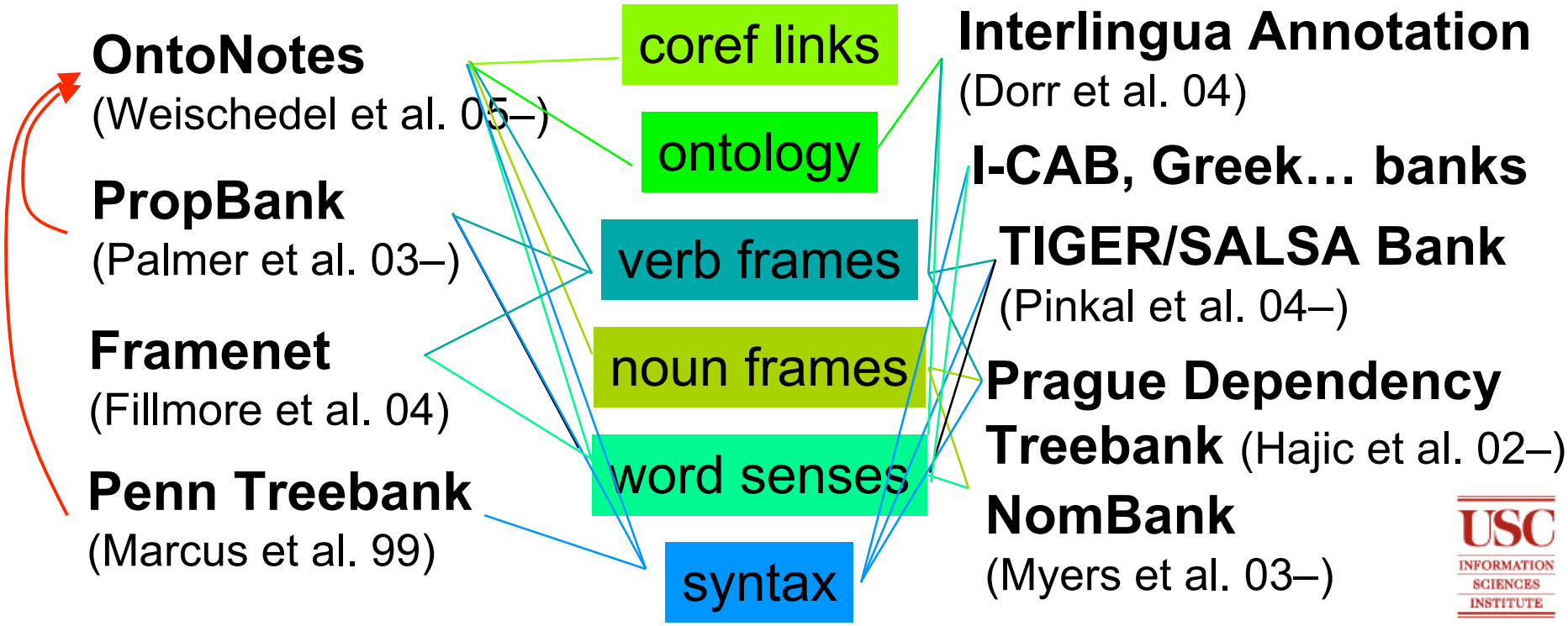
- Increased need for corpora and for annotation raises new questions:
  - **What kinds/aspects of 'domain semantics' to annotate?** …it's hardly an uncontroversial notion…
  - Which corpora? How much?
  - Which computational tools to apply once annotation is 'complete'? When *is* it complete?
  - How 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'?

Talk overview

1. Introduction: A new role for annotation?
2. Example: Semantic annotation in OntoNotes
3. Toward a science of annotation: 7 questions
4. Conclusion

Semantic annotation projects

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

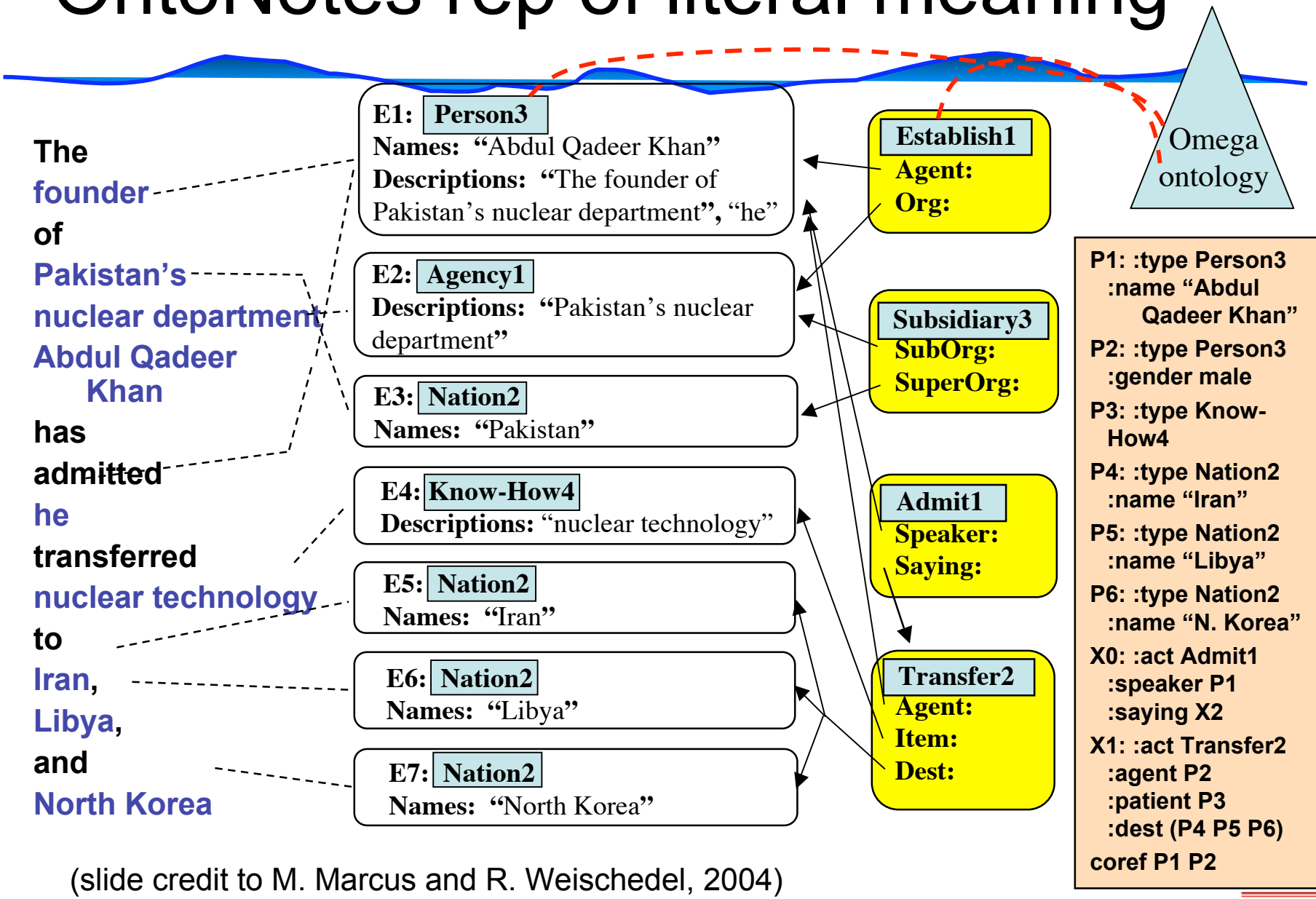


USC INFORMATION SCIENCES INSTITUTE

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 rep of literal meaning



(slide credit to M. Marcus and R. Weischedel, 2004)

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

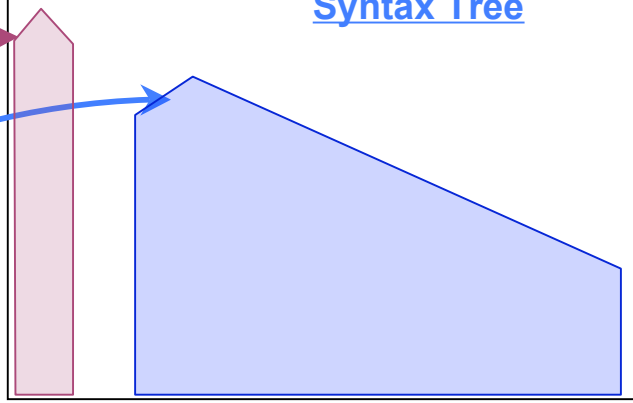
predicate : say
pb sense : 01
on sense : 1

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

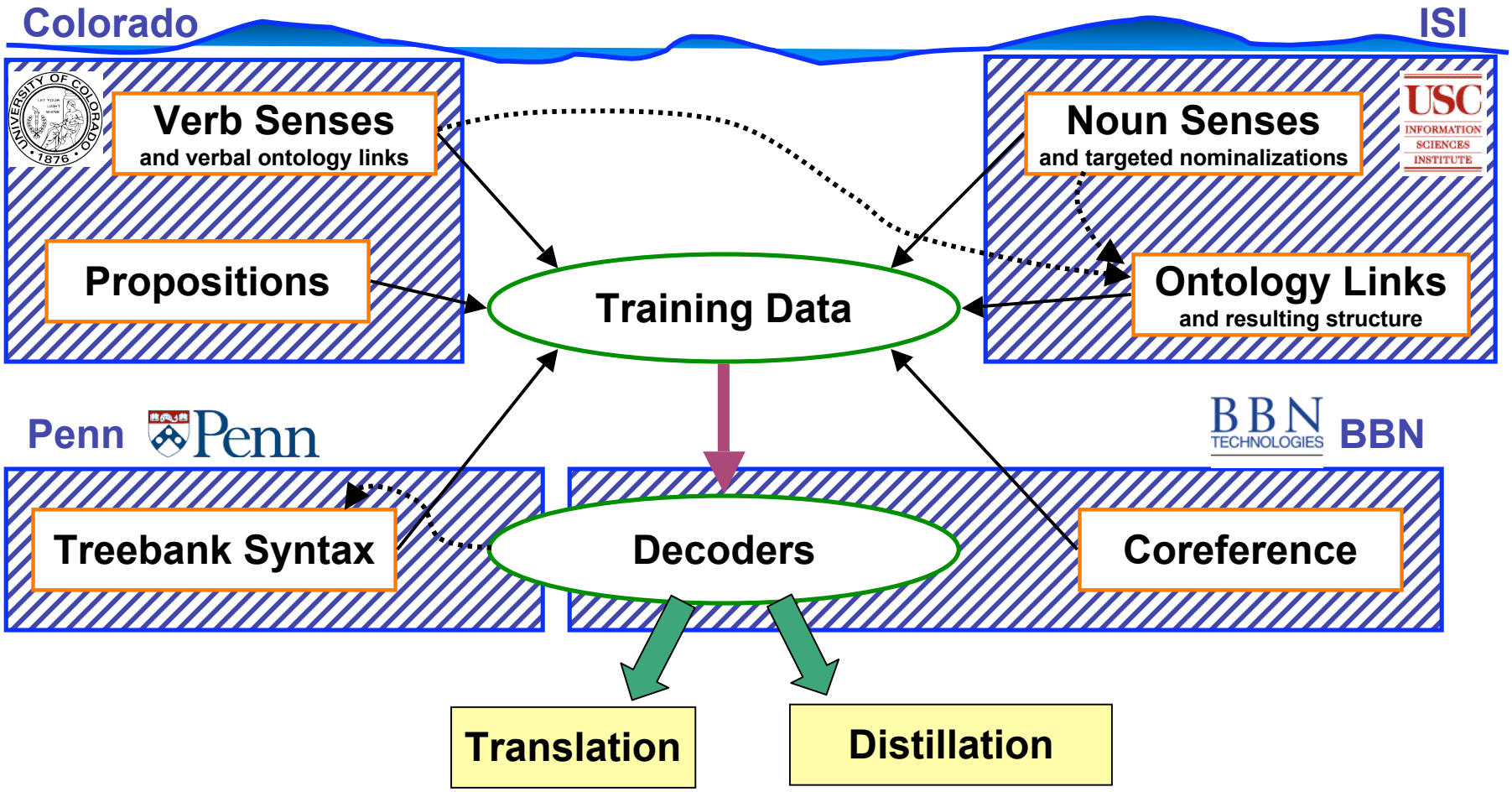
Say.A.1.1.1: DEF "…" EXS "…" FEATS …
POOL [State.A.1.2 Declare.A.1.4…]

Say.A.1.1.2: DEF "…" EXS "…" POOL […]

USC INFORMATION SCIENCES INSTITUTE

Project structure & parts

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



- Syntactic structure
- Predicate/argument structure
- Disambiguated nouns and verbs

- Coreference links
- Ontology
- Decoders

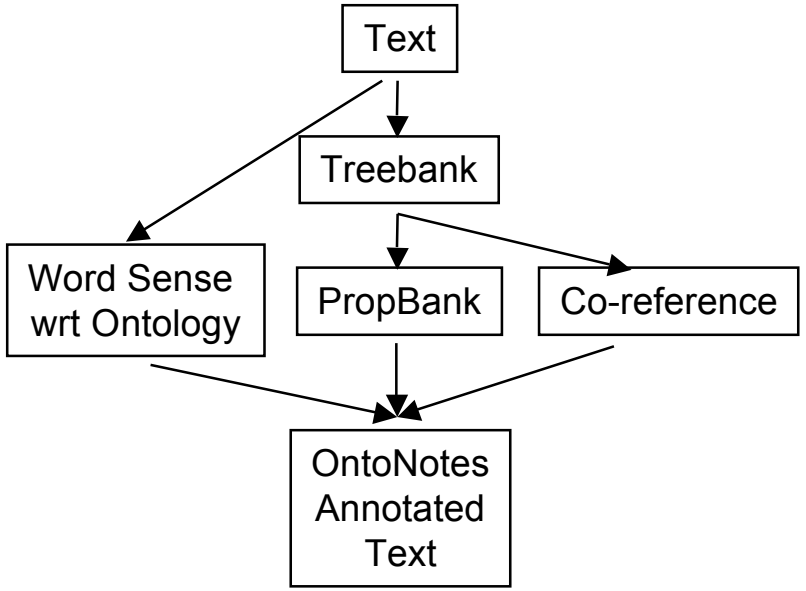
OntoNotes 1

- OntoNotes colleagues:
  - Sentence structure: U of Pennsylvania: Mitch Marcus et al.
  - Verb meanings: U of Colorado: Martha Palmer, Ann Houston, et al., plus about 20 annotators
  - Noun meanings: ISI: Robin Belvin, Ann Houston, Bonnie Glover Stalls, Rahul Bhagat, Mani Alagappan, Andrew Philpot, Jingbo Zhu, plus about 35 annotators
  - Coreference links: BBN: Ralph Weischedel, Lance Ramshaw, Sameer Pradhan, et al., plus 5 annotators

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

OntoNotes 2

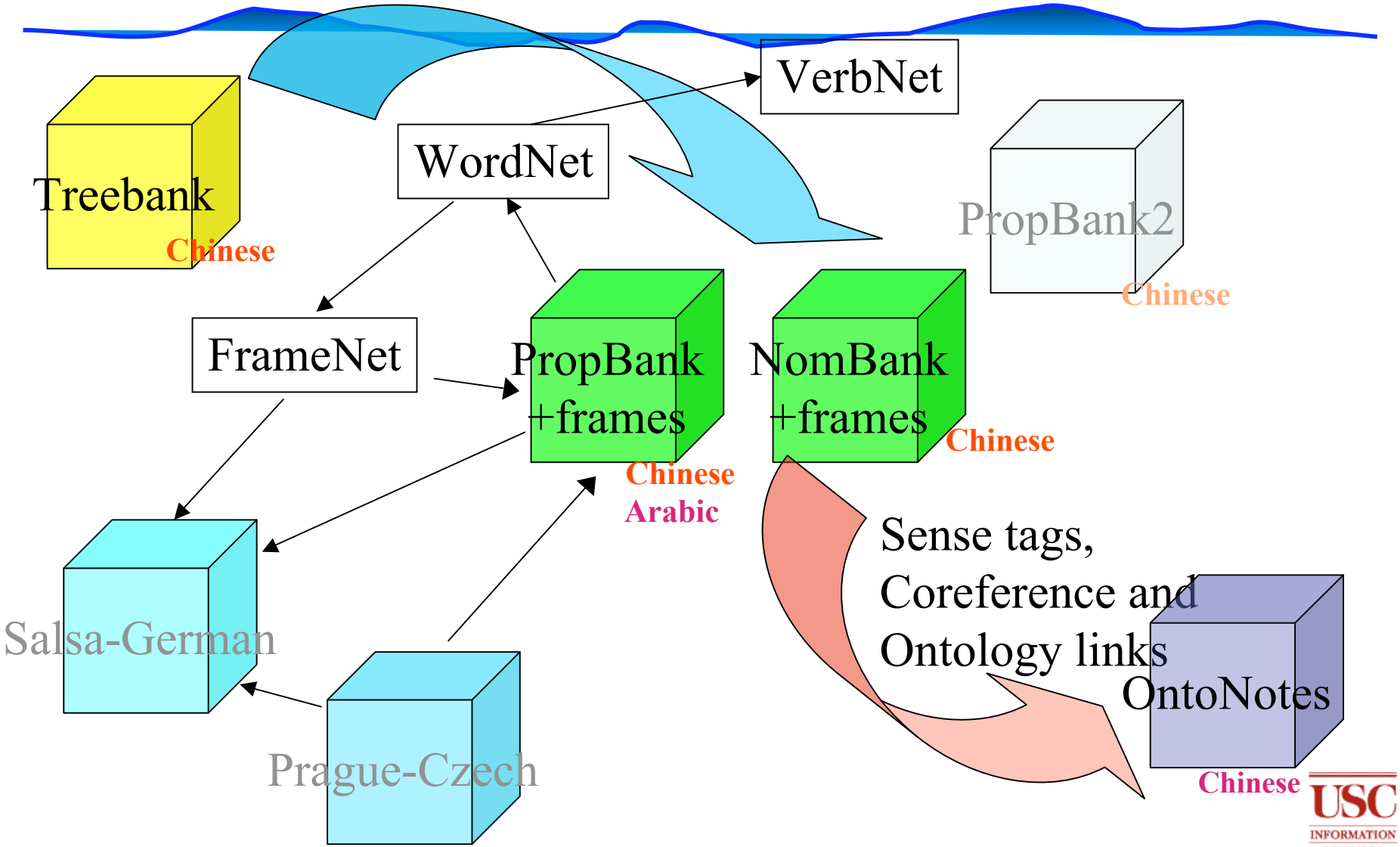
- History:
  - PropBank (2002–): verb annotation procedure developed
  - OntoNotes Feasibility Study (2004): Test corpus built, with coref annotation
  - Project started October 2005 (English); Chinese added 2006; Arabic in 2007
  - Possible to continue until 2009, funding permitting
- Potential for the near future: semantics 'bank'
  - May energize lots of research on semantic analysis, reps, etc.
  - May enable semantics-based IR, QA, MT, etc.



USC INFORMATION SCIENCES INSTITUTE

OntoNotes antecedents

(Palmer et al.)



USC INFORMATION SCIENCES INSTITUTE

Even so: Many words untouched!

Results of automated annotation using 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.

USC INFORMATION SCIENCES INSTITUTE

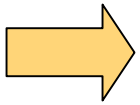
Three major subtasks

- 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

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

USC INFORMATION SCIENCES INSTITUTE

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

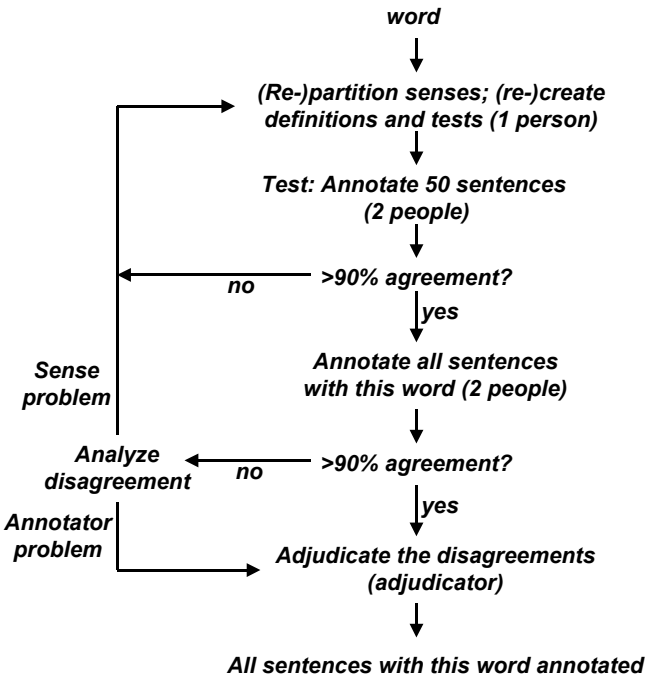
USC INFORMATION SCIENCES INSTITUTE

Ensuring trustworthiness/stability

- Problematic issues:
  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 (from PropBank): "**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

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

Pre-project test: Can it be done?

- Annotation process and tools developed and tested in PropBank (Palmer et al.; U Colorado)
- Typical results (10 words of each type, 100 sentences each):

	Round1 → Round2 → Round 3		
	tagger agreement	**# senses**	**time (min/100 tokens)**
verbs	**.76 → .86 → .91**	**4.5 → 5.2 → 3.8**	**30 → 25 → 25**
nouns	**.71 → .85 → .95**	**7.3 → 5.1 → 3.3**	**28 → 20 → 15**
adjs	**.87 → – → .90**	**2.8 → – → 5.5**	**24 → – → 18**

(by comparison: agreement using WordNet senses is 70%)

USC INFORMATION SCIENCES INSTITUTE

Before we start: Word statistics

Number of word tokens/types in 1000-word corpus
(95% confidence intervals on 85213 trials)

1000-word corpus	tokens	types
verbs	125.3	87.3
nouns	446.6	288.7
adjectives	103.2	80.6

Nouns: 57.2% of tokens
Monosemous nouns
(but not names etc.):
14.6% of tokens
= 25.6% of nouns

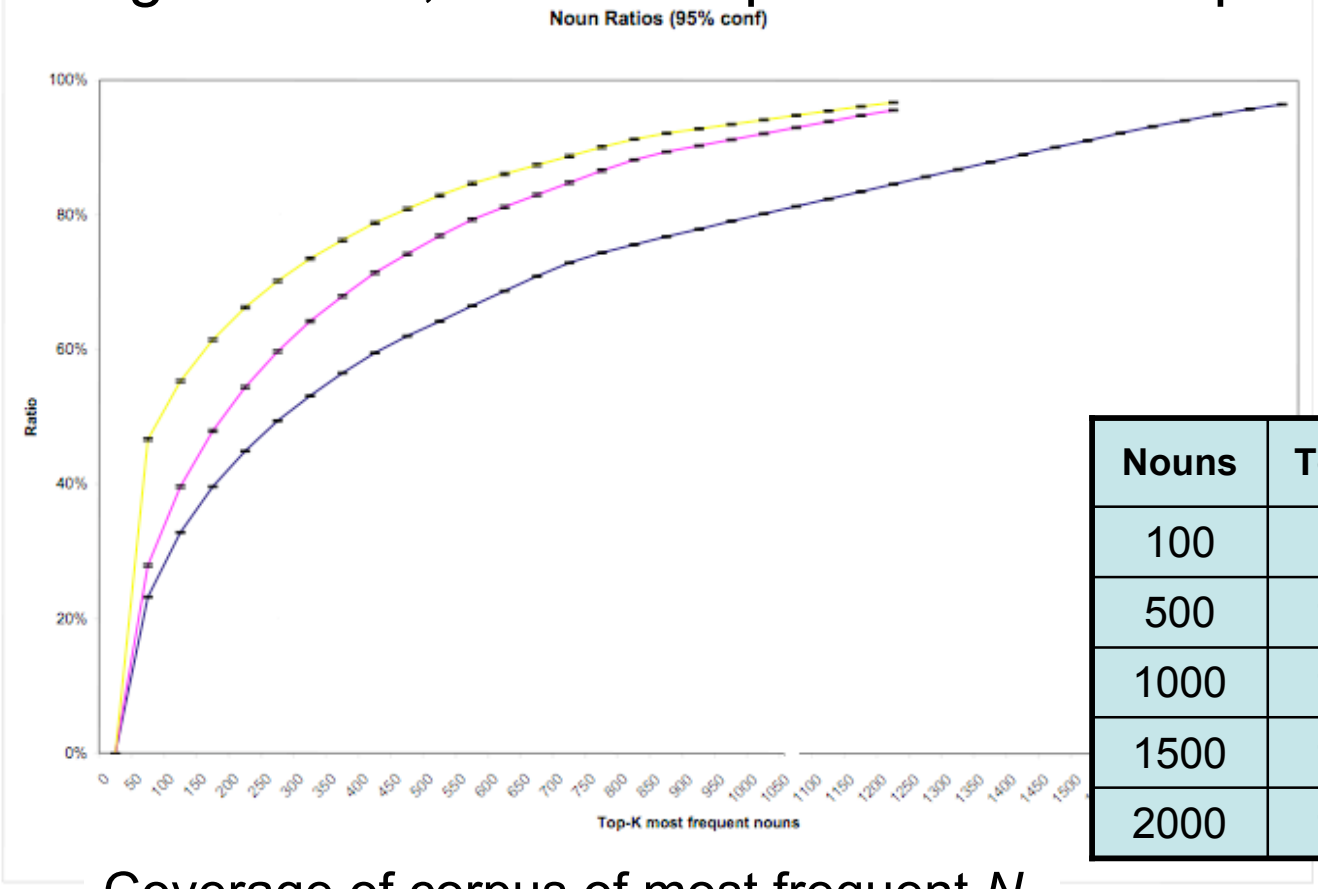
Polysemy

250K WSJ	verbs		nouns	
total	2341		5421	
1 WN sense	428	(18%)	1751	(32%)
2 or 3 senses	966	(41%)	2159	(40%)
4+ senses	947	(40%)	1511	(28%)

USC INFORMATION SCIENCES INSTITUTE

Before we start:
Noun coverage, various corpora

Coverage in WSJ, Brown corpora of most frequent *N* nouns



Nouns	Tokens (total 205442)	
100	76420	37%
500	140453	68%
1000	167715	82%
1500	181412	88%
2000	189641	92%

Coverage of corpus of most frequent *N* polysemous-2 nouns (WSJ+Brown)

USC INFORMATION SCIENCES INSTITUTE

Compound noun groups

- Problem: N-N compounds ("*kitchen knife*", "*party animal*", etc.)
  - Do not want to annotate each noun independently (a *party animal* is neither a *party* nor an *animal*)
- Solution: automatically find multiple-noun tuples (pairs, triples, etc.) with high co-occurrence
  - Pantel at ISI used pointwise mutual information algorithm to identify high-reliability tuples (up to 4-grams)
  - Found 35,700 tuples
- Linking into Omega:
  - Automatically generated Omega superconcepts
  - Quasi-random check of 40 pairs showed about 72.5% accuracy
  - 1951 of the tuples cannot be attached into Omega because the head noun does not exist (e.g. proper nouns)
  - File at http://www.isi.edu/~pantel/wninte.txt

wsj/00/wsj_0003.mrg	asbestos fiber
wsj/00/wsj_0003.mrg	protection agency
wsj/00/wsj_0007.mrg	engineering industry
wsj/00/wsj_0008.mrg	government debt
wsj/00/wsj_0008.mrg	borrowing authority
wsj/00/wsj_0009.mrg	marketing arm

wsj/00/wsj_0009.mrg	auto maker
wsj/00/wsj_0009.mrg	marketing manager
wsj/00/wsj_0009.mrg	marketing executive
wsj/00/wsj_0010.mrg	boca raton
wsj/00/wsj_0099.mrg	air force contract
wsj/00/wsj_0099.mrg	intelligence data

USC INFORMATION SCIENCES INSTITUTE

Nouns to be handled

Monosemous	
1253	trading
1117	investor
867	firm
585	tax
581	trader
567	chairman
566	income
462	asset
420	spokesman
338	customer
336	transaction
335	employee
324	shareholder
309	consumer
292	ad
…	…
1	academe
1	abstention
1	absorber

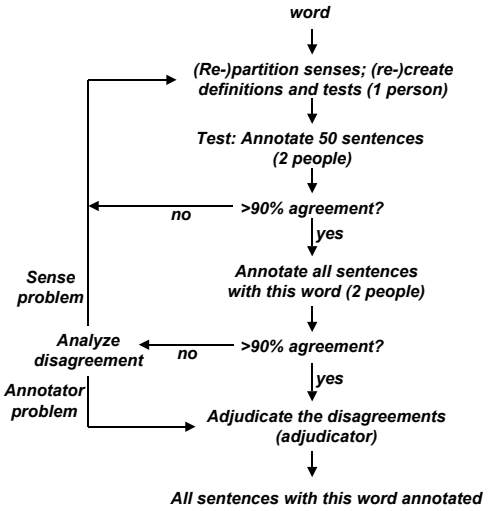
Polysemous	
4178	year
3095	market
1796	sale
1467	month
1308	business
1253	trading
1211	rate
1141	time
1140	president
1117	investor
1053	day
1025	government
1012	quarter
974	bank
944	group
…	…
1	industrialization
1	globalization
1	diving

Polysemous unsensed	
4076	company
3196	share
2393	stock
1874	price
1149	bond
1122	week
998	analyst
996	cent
919	interest
867	firm
774	product
759	earnings
749	industry
696	executive
667	money
…	…
1	aberration
1	abandonment
1	abacus

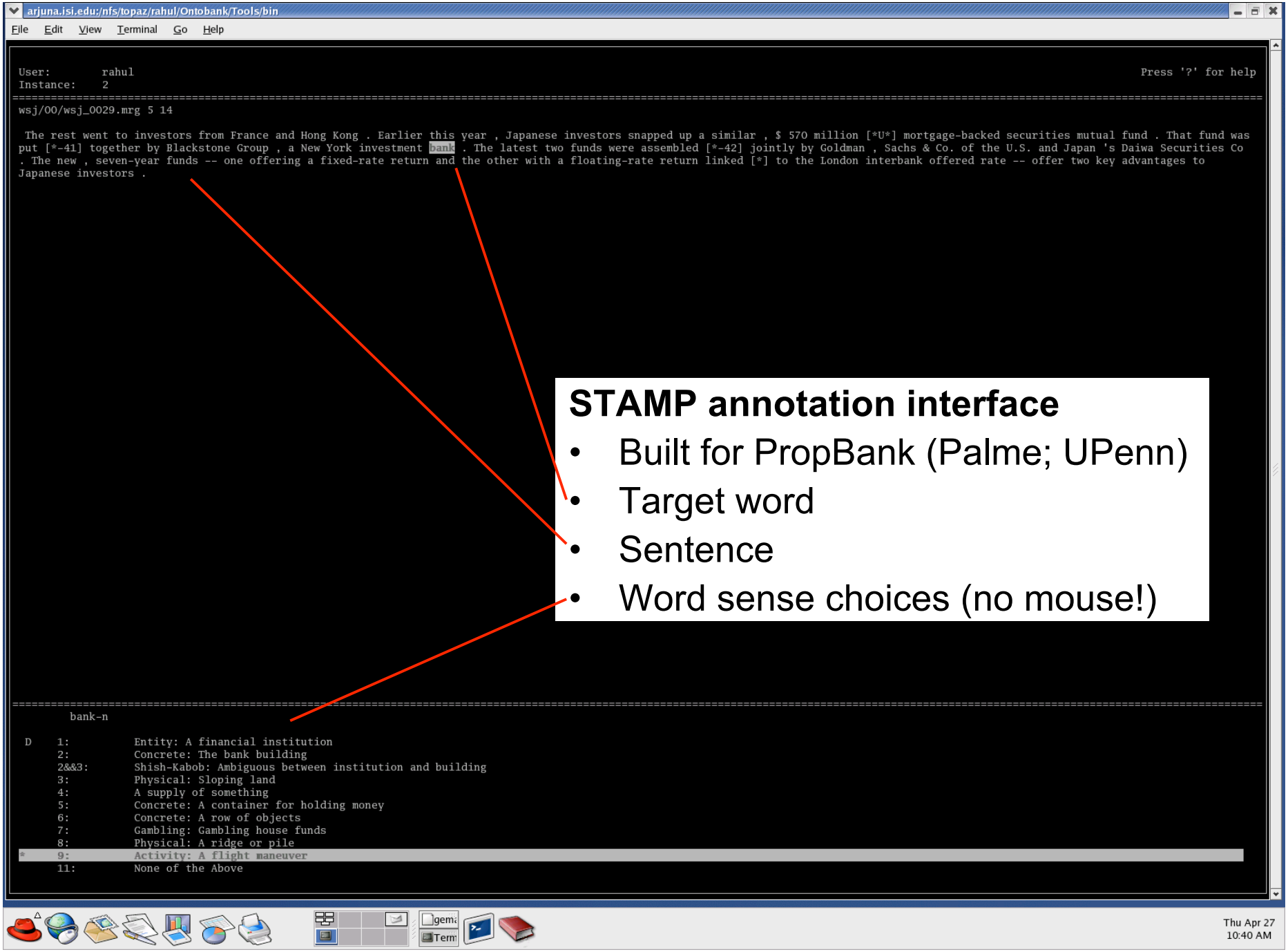
USC INFORMATION SCIENCES INSTITUTE

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, version
    - 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



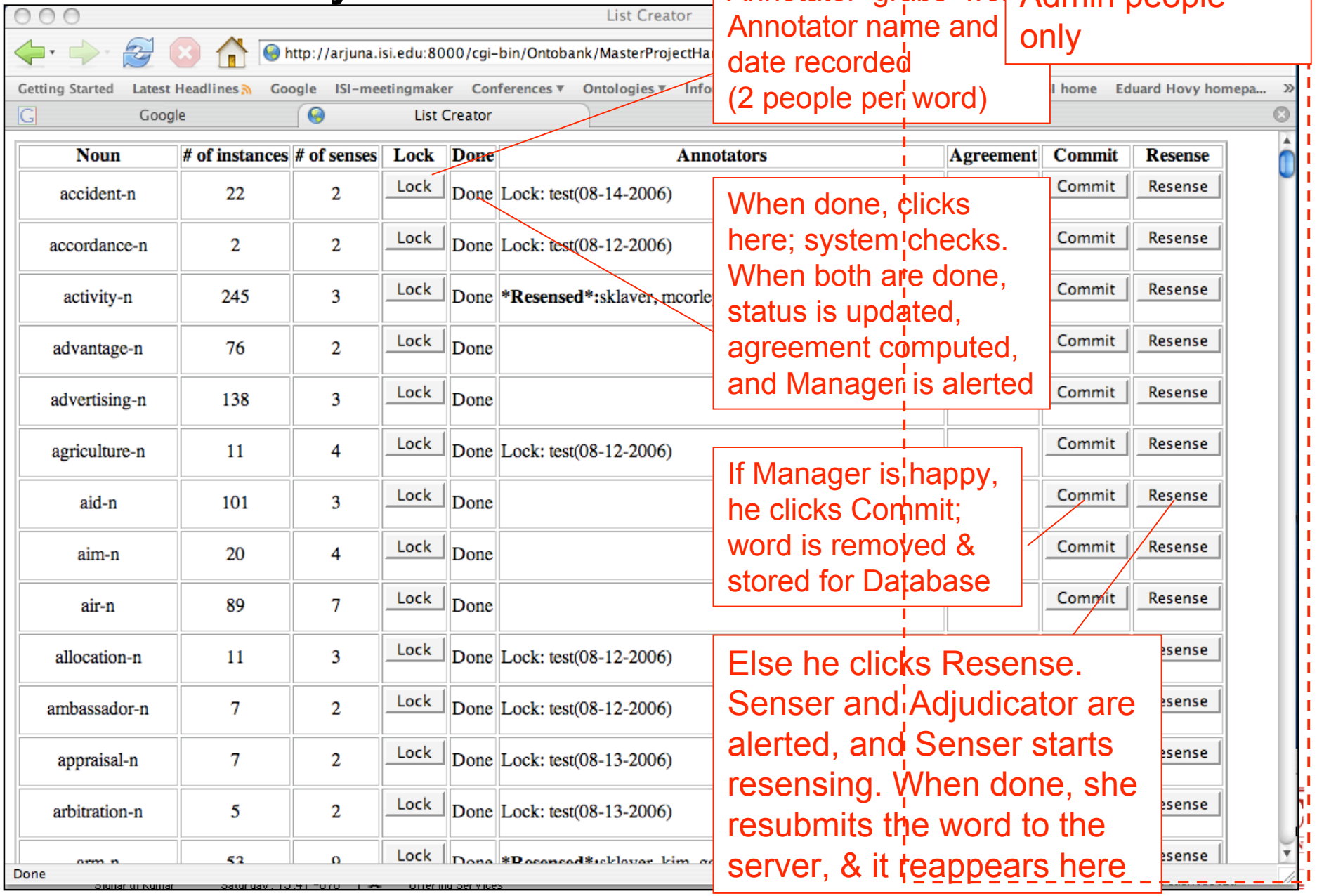
USC INFORMATION SCIENCES INSTITUTE



STAMP annotation interface

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

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

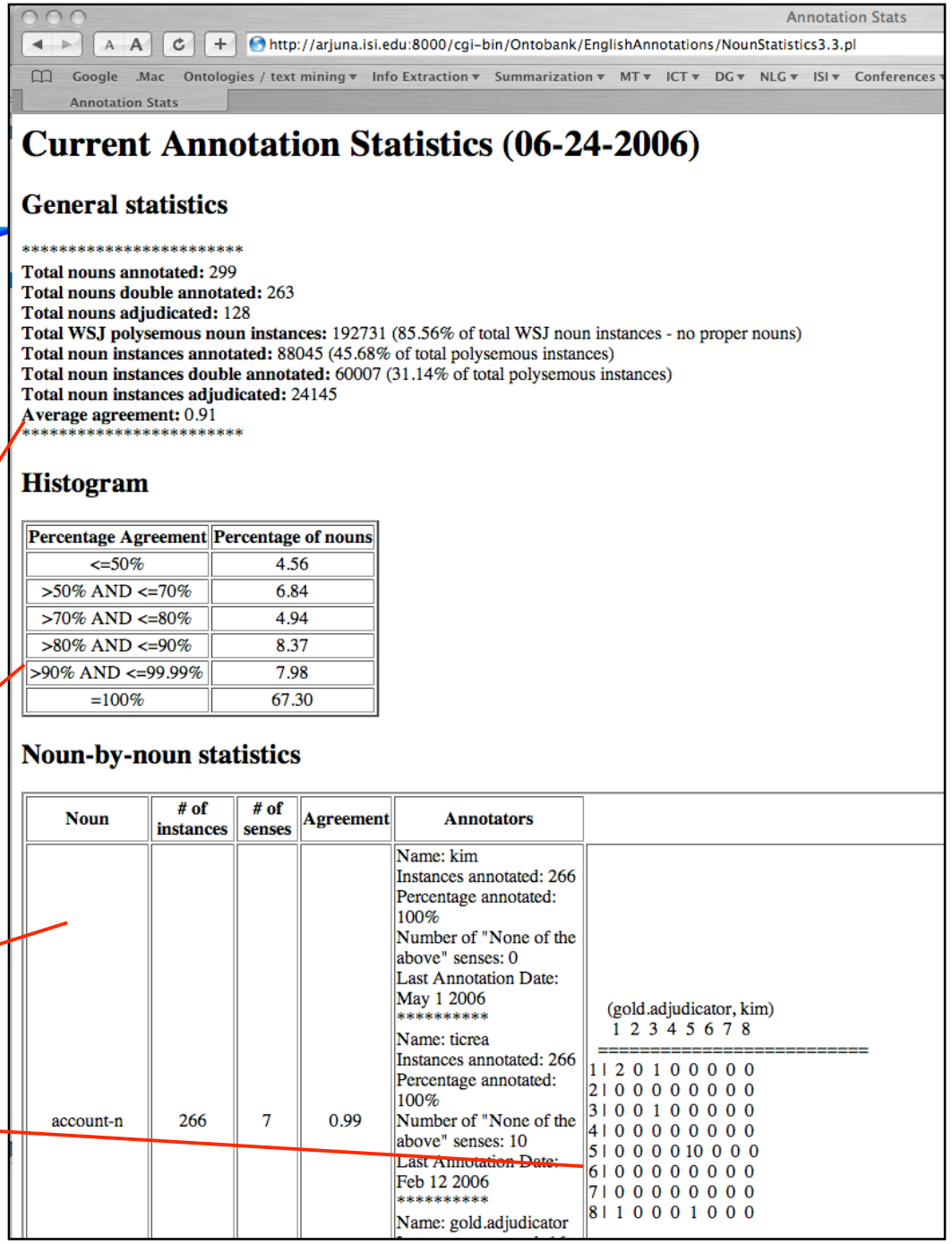
http://arjuna.isi.edu:8000/Ontobank/AnnotationStats.html

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

Agreement histogram

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

Confusion matrix for results



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?

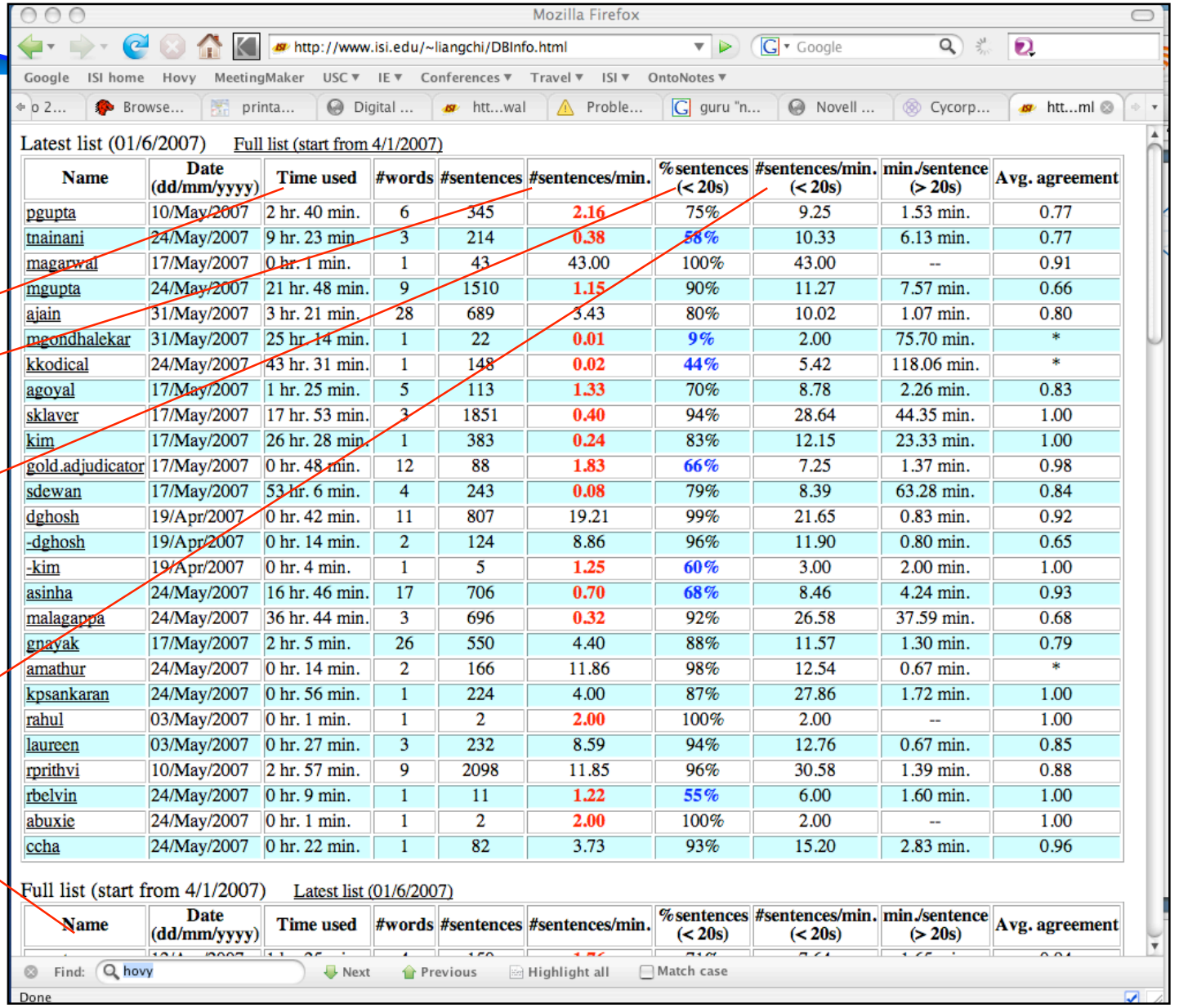
USC INFORMATION SCIENCES INSTITUTE

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

Full list (start from 4/1/2007) Latest list (01/6/2007)

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

English noun annotation stats

- Annotators at ISI:
  - About 9 regular annotators for English, 6 for Chinese
  - All trained on “bank”
  - Weekly telecons for discussion
- Status (October 06):
  - Avg. agreement: 91%
  - Residual disagreements adjudicated by linguist
  - Slow start, but speeding up

	English nouns						
2006	**Date**	**Sensed**	**Sensing target**	**Double-annotated**	**Double-ann coverage**	**Adjudica-ted**	**Final target**
Jan		0	200	0		0	150
Feb		59	250	102		13	200
Mar		136	375	138	15.5%	44	275
Apr		145	450	149	18.9%	44	350
May	5/15	249	525	217	24.3%	51	450
Jun	6/28	315	625	272	31.3%	128	550
Jul			720				650
Aug	8/15	~370	850	359	43.3%	149	800
Sep	9/25	500	980	415	44.7%	214	950
Oct	10/30	568	1100	451	46.1%	221	1100
Nov							
Dec							

USC INFORMATION SCIENCES INSTITUTE

Some preliminary statistics

- 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

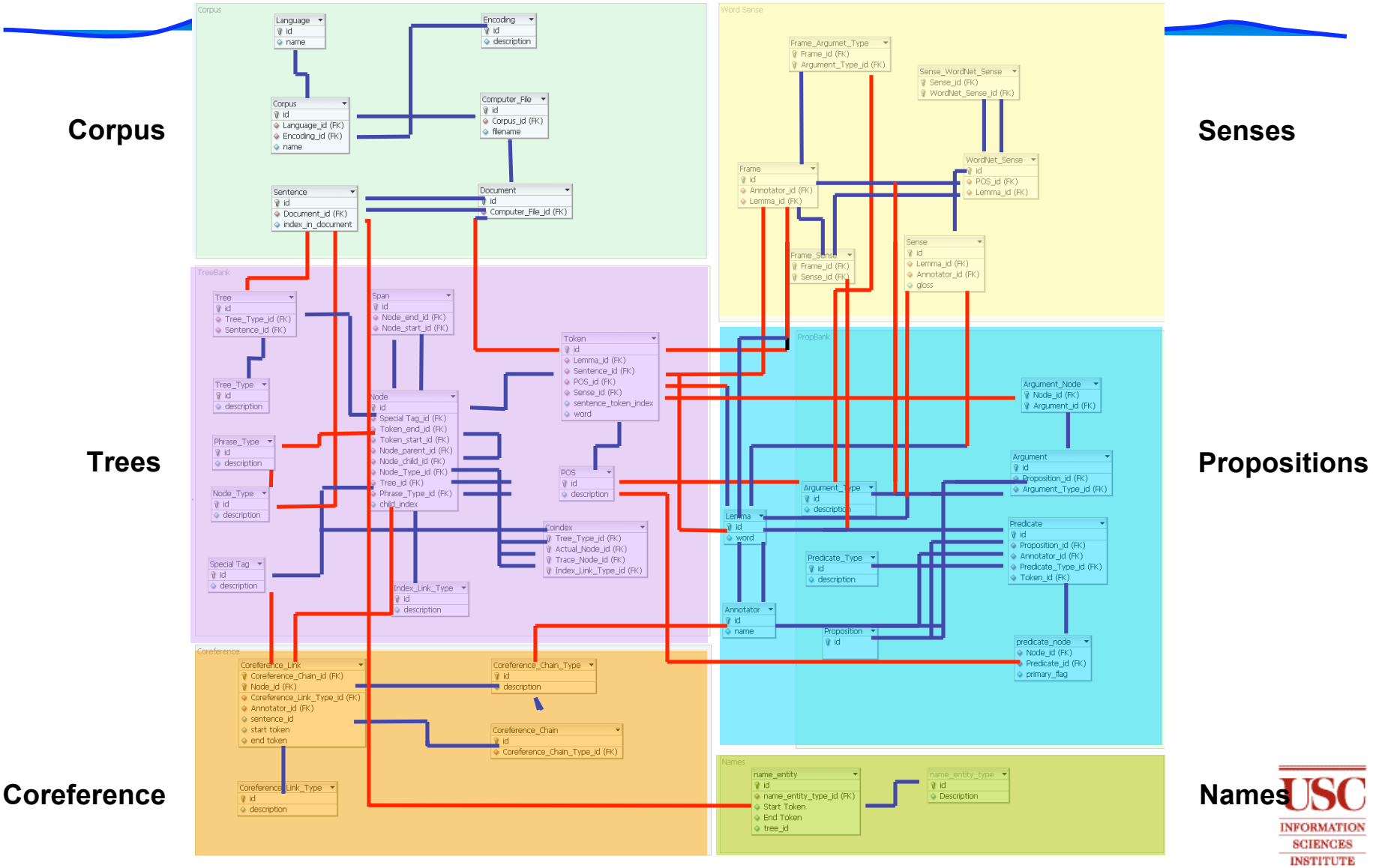
1:**play-n** senses=13 instances=41 agreement=0.490000 entropy:2.003143 distribution:0.239 0.217 0.130 0.109 0.087 0.087 0.065 0.043 0.021739 0.00
2:**control-n** senses=8 instances=262 agreement=0.820000 entropy:1.788795 distribution:0.416 0.168 0.119 0.097 0.065 0.058 0.026 0.023 0.016 0.013
3:**defense-n** 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

Current status (month 23)

- Text annotated:
  - Newswire text:
    - 300K words of Wall Street Journal: 800+ verbs and 500+ nouns; verb arg structure; coref links
  - Primarily broadcast news:
    - Broadcast news: 200K English and 300K Chinese (same number of nouns and verbs; all corefs)
    - Starting with 100K Arabic newswire data
  - Next year:
    - Broadcast conversations (200K words of English, 150K Chinese; all corefs)
  - Later:
    - Weblogs, newsgroups, etc.
- Ontology: Terms now being converted to concepts and taxonomized/inserted into overall structure

Database: Unified relational rep

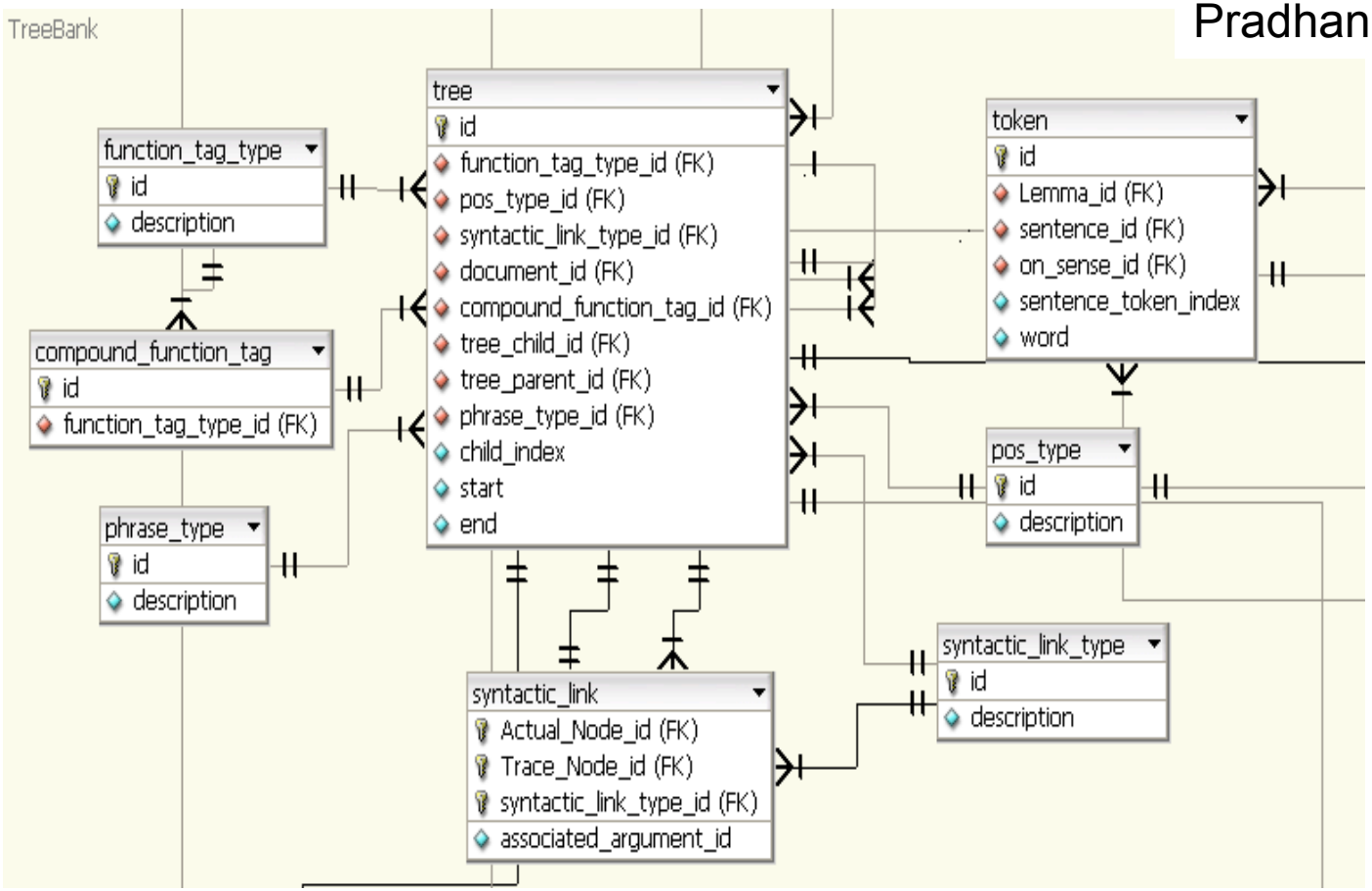
(Slide by Sameer Pradhan, BBN)



USC INFORMATION SCIENCES INSTITUTE

Example: DB representation of syntax

(Slide by Sameer Pradhan, BBN)



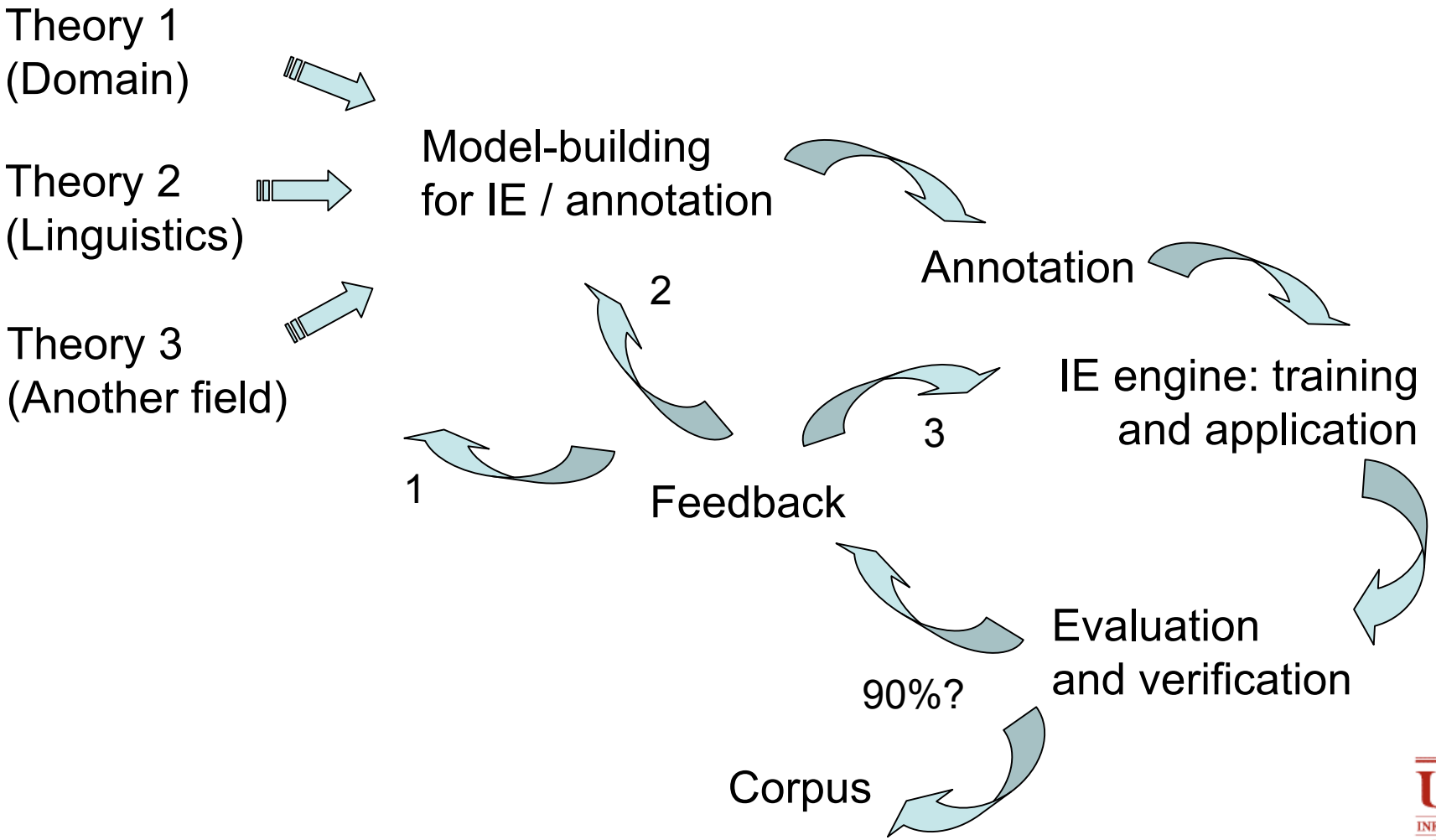
- **Treebank tokens (stored in the Token table) provide the common base**
- **The Tree table stores the recursive tree nodes, each with its span**
- **Subsidiary tables define the sets of function tags, phase types, etc.**

USC INFORMATION SCIENCES INSTITUTE

Talk overview

1. Introduction: A new role for annotation?
2. Example: Semantic annotation in OntoNotes
3. Toward a science of annotation: 7 questions
4. Conclusion

The generic annotation pipeline



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

- Choose carefully—the future will build on your work!
  - (When to re-use something?—Today, we're stuck with WSJ…)
- **Technical issues**: *Balance, representativeness, and timeliness*
  - When is a corpus representative? —"stock" in WSJ is *never* the soup base
    - Methodology of 'principled' corpus construction for representativeness (even BNC process rather ad hoc)
  - How to balance genre, era, domain…See (Kilgarriff and Grefenstette, CL 2003)
    - Effect of (expected) usage of corpus
  - 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?

USC INFORMATION SCIENCES INSTITUTE

Q1. Prep: What's available

- Corpus collections are worth their weight in gold
  - Unencumbered by copyright
  - Available to whole community — standardized results for comparison
- Raw and processed text and speech:
  - Linguistic Data Consortium (LDC), UPenn: www.ldc.upenn.edu/

Q2: Instantiating the theory

- What to annotate? 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
  - 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
- Likely problems:
  - Categories not exhaustive over phenomena
  - Categories badly defined / unclear (intrinsic ambiguity, or relying on bg knowl?)
- Measuring stability — measures of agreement: (Teachman 1989)
  - Precision (correctness) = P_i = *#correct / N*
  - Entropy (ambiguity, regardless of correctness) = $-\sum_i P_i \,.\, ln\, P_i$ (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
  - E.g., from Case Roles (*Agent, Patient, Instrument*) to PropBank's roles (*arg0, arg1, argM*) — user chooses desired role labels and maps PropBank roles to them

- What does this say about the theory, however?

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. Interface: What's available

- Interfaces/annotation tools:
  - ATLAS.TI: annotation toolkit (www.atlasti.com/)
  - Ad hoc annotation interfaces and tools from the NLP community
- 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

Q4: Annotators

- How to choose annotators?
  - Annotator backgrounds — should they be experts, or precisely not?
  - Biases, preferences, etc.
  - Experts: Psych experimenters
- How much to train the 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…
- Who should train the annotators?
  - Is it ok for the interface builder, or the learning system builder? — not: they have an agenda

USC INFORMATION SCIENCES INSTITUTE

Q5.1: 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.
- The Wiebe '85% clear cases' rule
  - Ask the annotators also to mark their certainty
  - There should be a lot of agreement at high certainty — the clear cases
- 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
- Adjudication
  - Have an expert (or more annotators) decide in cases of residual disagreement — but how much disagreement can be tolerated before just redoing the annotation?
- Experts: …?

USC INFORMATION SCIENCES INSTITUTE

Q5.2: Annotation procedure

- Overall approach —Shulman's rule: do the easy annotations first, so you've seen the data when you get to the harder cases
- The Rosé hypothesis: 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

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 05; 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

USC INFORMATION SCIENCES INSTITUTE

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?

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

Talk overview

1. Introduction: A new role for annotation?
2. Example: Semantic annotation in OntoNotes
3. Toward a science of annotation: 7 questions
4. Conclusion

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

USC INFORMATION SCIENCES INSTITUTE

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 workshop)

Thank you!