Creating Training Material for Health Informatics: Toward a Science of Annotation

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

Toward a Science of Annotation

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

Acknowledgments

- For **OntoNotes** materials, and for exploring annotation, thanks to
  - 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, thanks to the **IAMTC** team:
  - 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 Miller (MITRE)
- For **additional experience**, thanks to colleagues and students at ISI and elsewhere:
  - ISI: Gully Burns, Zornitsa Kozareva, Andrew Philpot, Stephen Tratz
  - U of Pittsburgh: David Halpern, Peggy Hom, Stuart Shulman
  - Others: Julia Lavid (Complutense U, Madrid)
- For **funding**, thanks to DARPA, the NSF, and IBM

Overview

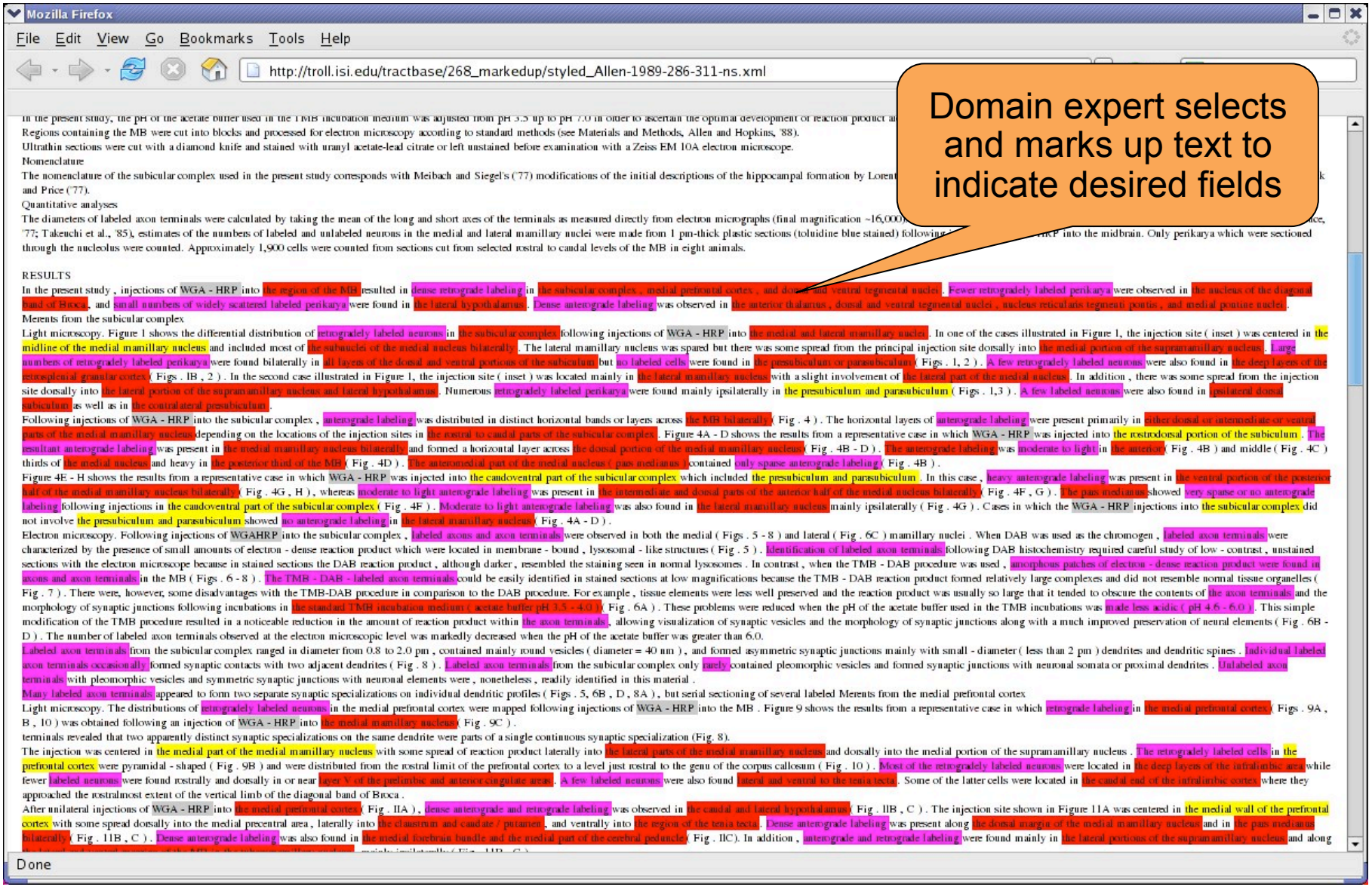
- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

Annotation

- **Definition**: Annotation is the process of adding new information to raw data by humans (annotators). Usually, the information is added in small individual decisions, in many places throughout the data. The addition process usually requires some sort of mental decision that depends both on the raw data and on some theory or knowledge that the annotator has internalized earlier.

- **Typical annotation steps**:
  - Decide which fragment of the data to annotate
  - Add to that fragment a specific bit of information, usually chosen from a fixed set of options

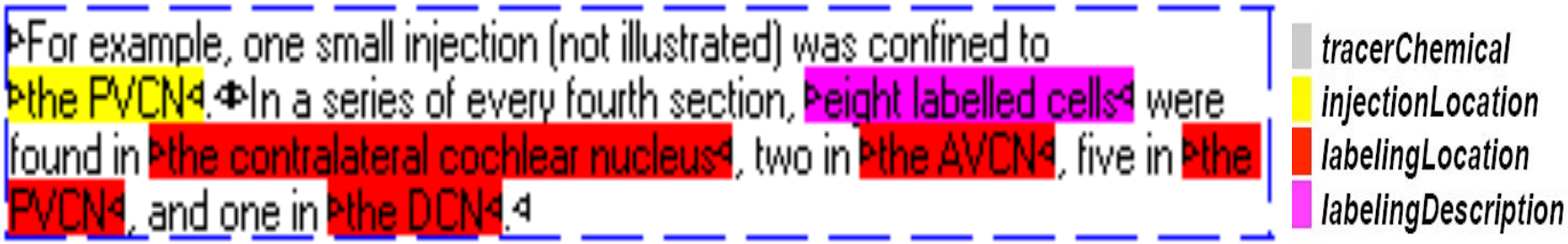
Example: Biomed annotation (Burns, Feng, Hovy 06; 07)



Example output: Single record

1: Identify the information fields in the text

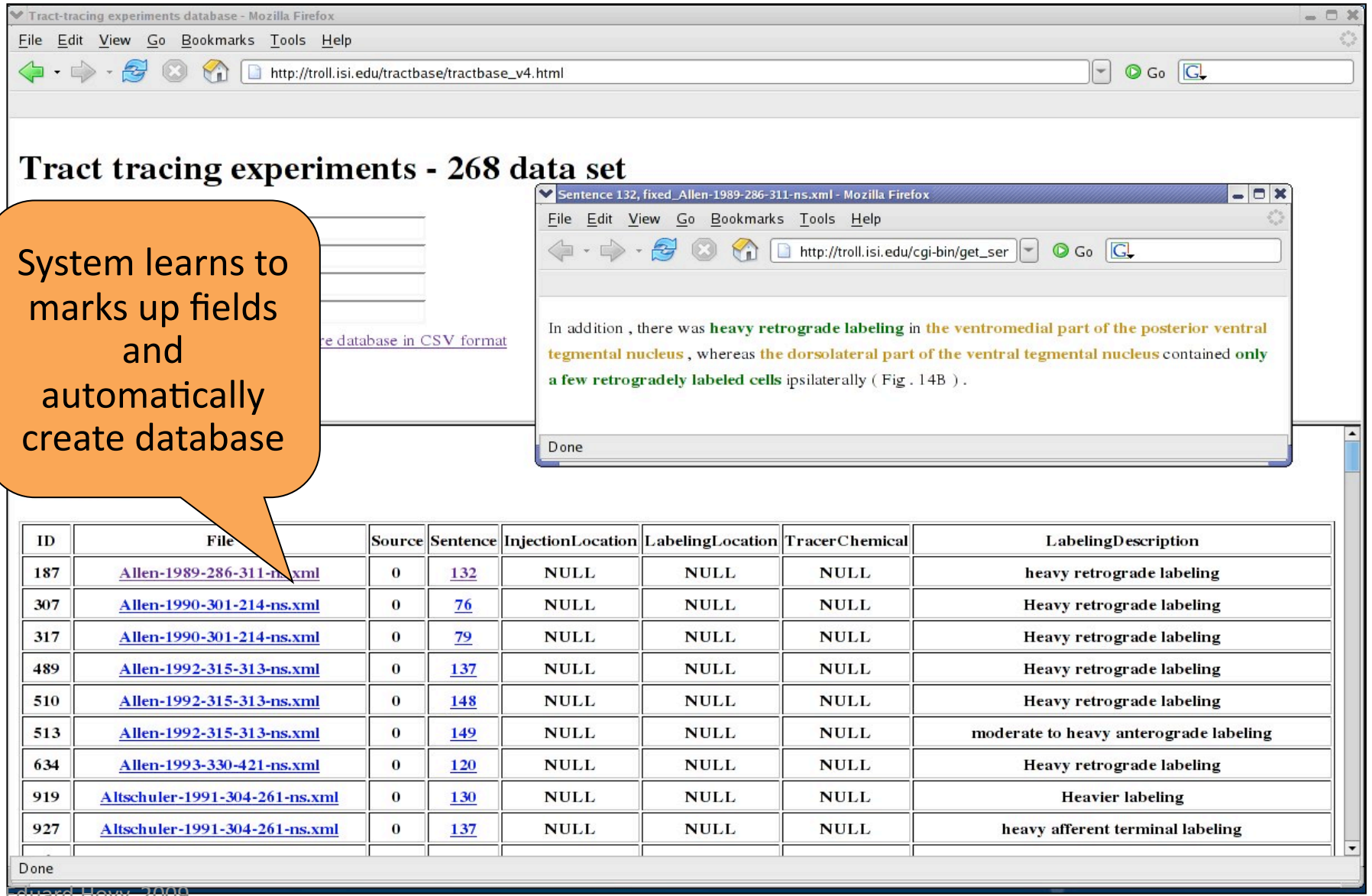
2: Assemble fields into separate experiments:



3: Then output each experiment as a data record:

record #	tracerChemical	injectionLocation	labelingLocation	labelingDescription
	NULL	the PVCN	the contralateral cochlear nucleus the AVCN the PVCN the DCN	eight labelled cells

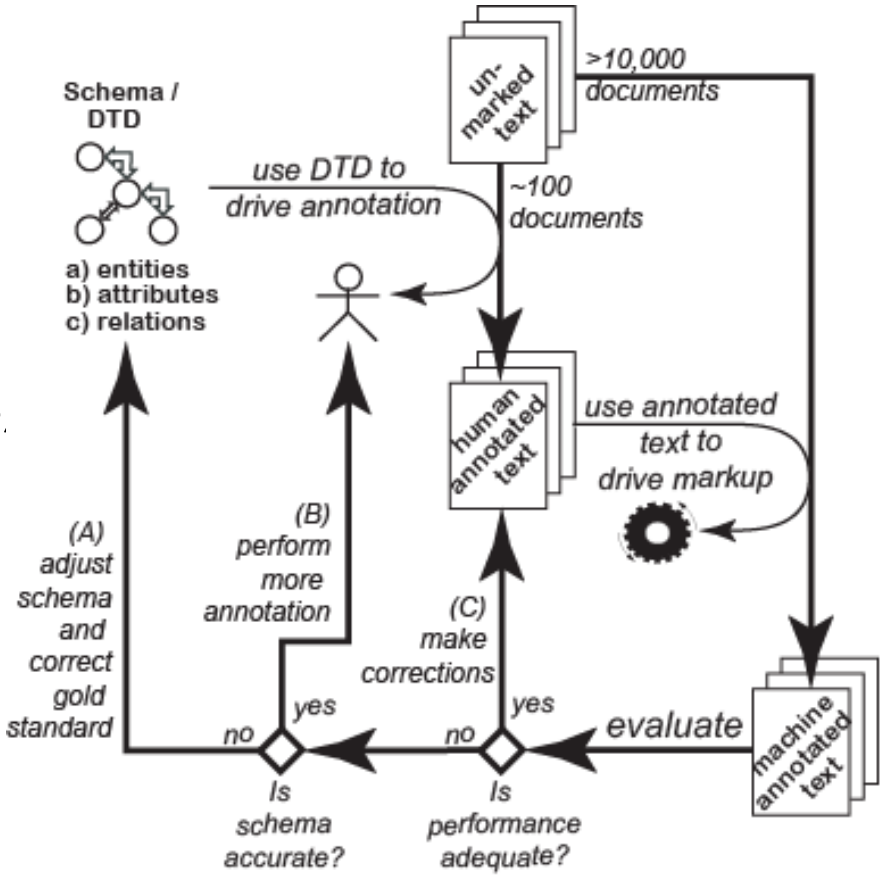
Database: Final result



ID	File	Source	Sentence	InjectionLocation	LabelingLocation	TracerChemical	LabelingDescription
187	Allen-1989-286-311-ns.xml	0	132	NULL	NULL	NULL	heavy retrograde labeling
307	Allen-1990-301-214-ns.xml	0	76	NULL	NULL	NULL	Heavy retrograde labeling
317	Allen-1990-301-214-ns.xml	0	79	NULL	NULL	NULL	Heavy retrograde labeling
489	Allen-1992-315-313-ns.xml	0	137	NULL	NULL	NULL	Heavy retrograde labeling
510	Allen-1992-315-313-ns.xml	0	148	NULL	NULL	NULL	Heavy retrograde labeling
513	Allen-1992-315-313-ns.xml	0	149	NULL	NULL	NULL	moderate to heavy anterograde labeling
634	Allen-1993-330-421-ns.xml	0	120	NULL	NULL	NULL	Heavy retrograde labeling
919	Altschuler-1991-304-261-ns.xml	0	130	NULL	NULL	NULL	Heavier labeling
927	Altschuler-1991-304-261-ns.xml	0	137	NULL	NULL	NULL	heavy afferent terminal labeling

Example annotation pipeline

- 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…



So, why annotate?

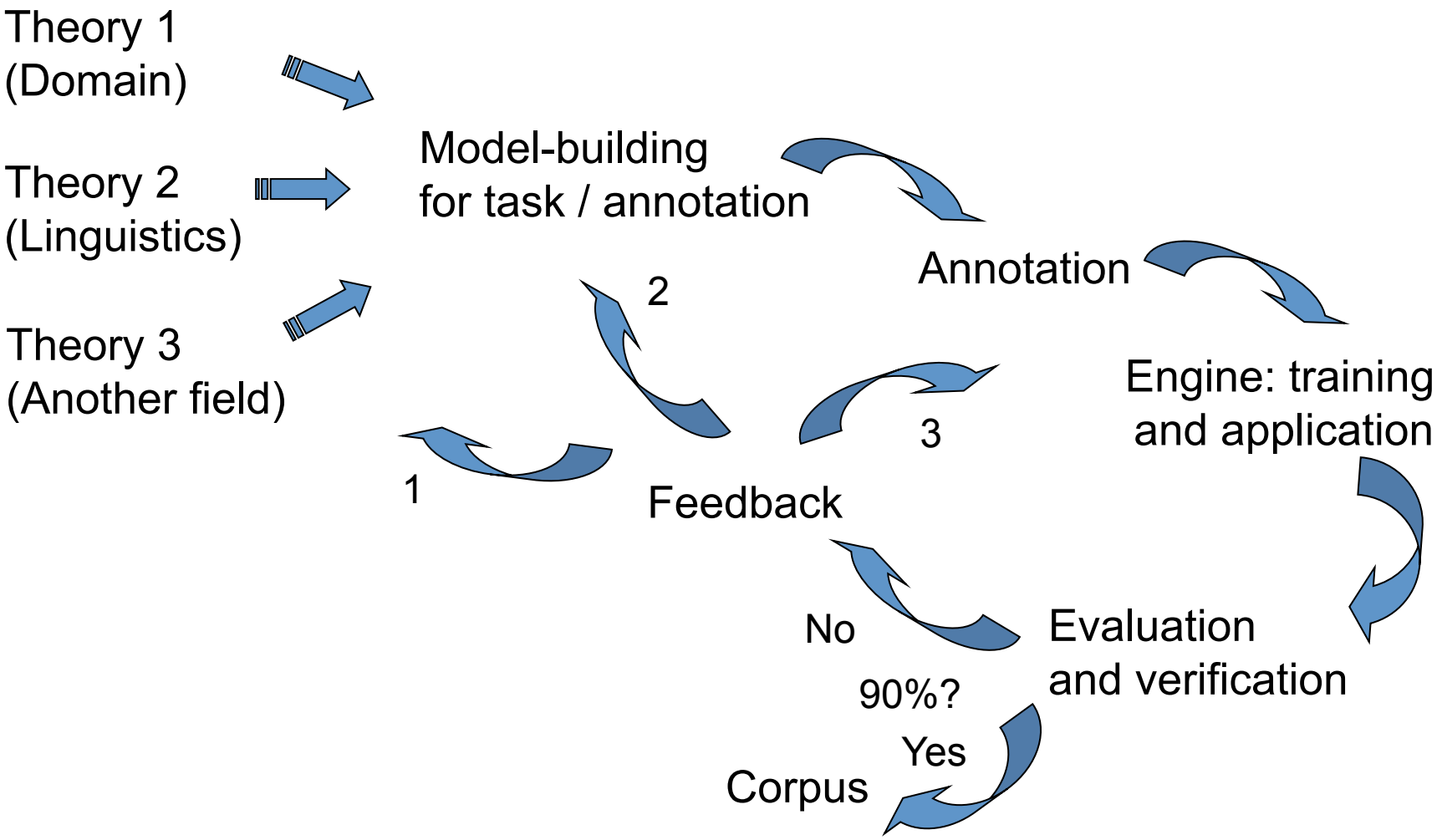
Some reasons for annotation

- **NLP**: Build data to enable machine learning of some application
  - Methodology: Transform pure input text into interpreted/extracted/marked-up input text
    - Have several humans manually annotate texts with info
    - Compare their performance
    - Train a learning algorithm to do the job using the annotated data
    - Also, use the annotated data as standard for system evaluations

- **Biomedicine**: Identify and extract information relevant for future work
  - Typical sources: research articles in journals, other publications
  - Typical details sought: experimental conditions and methods, generalized findings
  - Useful for: Surveying previous work, searching for trends over time, avoiding duplication of experiments

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

The generic annotation pipeline



…The **computational linguistics** reason

- Create (gold standard) text corpora with the kinds of additional information desired:
  - Parts of speech: "Joe$_{NOUN}$ liked$_{VERB}$ the book$_{NOUN}$"
  - Semantic meaning: "Joe$_{PERSON}$ liked the book$_{OBJECT}$"
  - Opinions: "Joe$_{HOLDER}$ liked$_{POSITIVE}$ the book$_{TOPIC}$"
  - etc.
- Teach machine learning algorithms to do this automatically, using the gold standard text
- Apply these algorithms to new text in different NLP application systems: language understanding, QA, machine translation, etc.

…The **ontology building** reason

- Identify examples of the concepts inside the (domain) corpus
- Place these concepts into the ontology at the correct position(s)

…The **semantic web** reason

- Add appropriate semantic information into web pages to create a gold standard corpus
- Train machine learning algorithms to do the same, using the corpus
- Apply the algorithms to newly created web pages

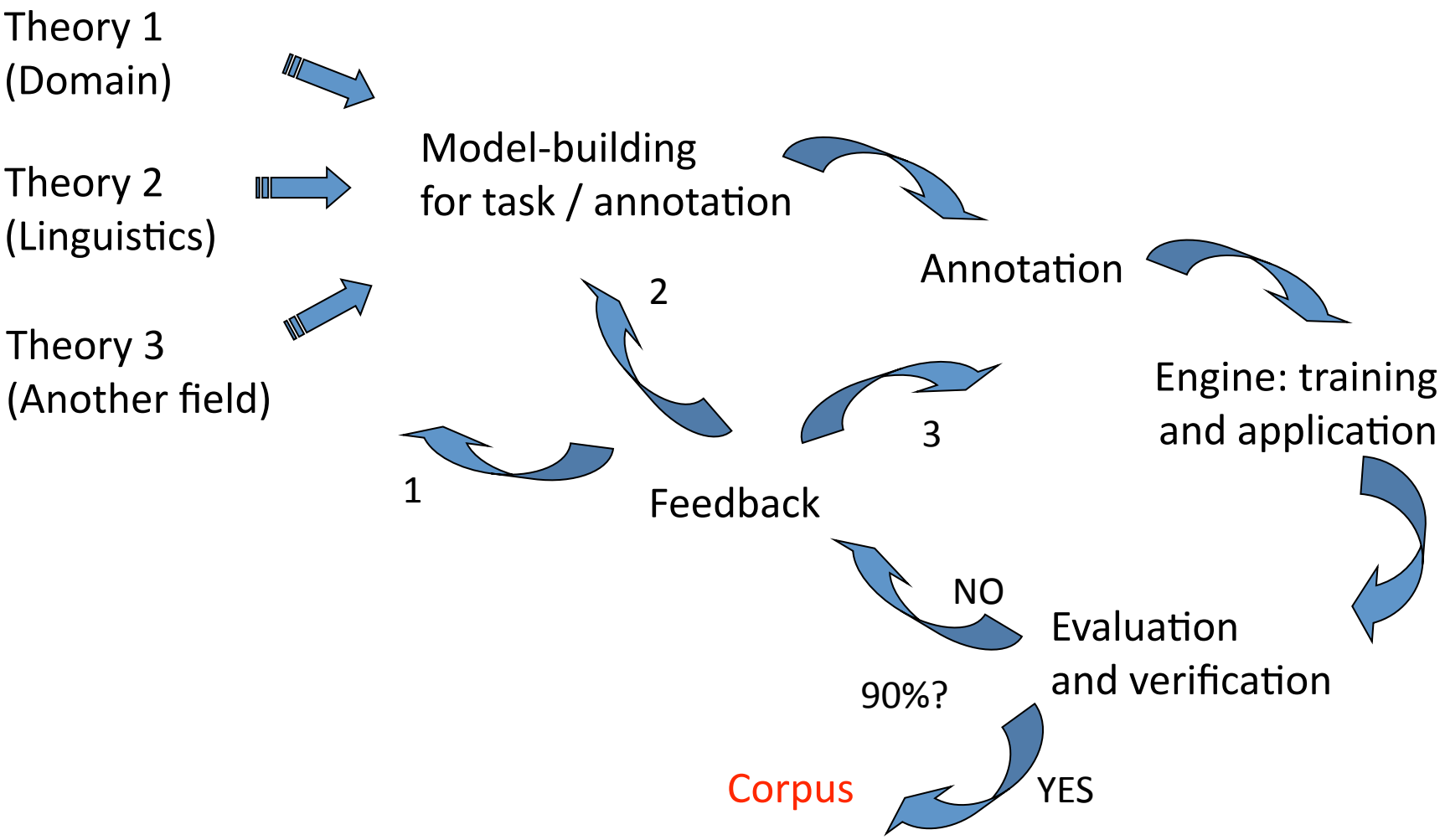
…The **linguistics** reason

- To develop a theory about some aspect of language (syntax, semantics, pragmatics, discourse, etc.), you have to identify and study the phenomenon
- So:
  - You locate examples of it inside text
  - You analyze the variations, defining types
  - You create a general theoretical 'infrastructure' to describe/explain how the types work together

These reasons combine in **semantics**

- Many people believe that computers must transform text into its 'meaning' — semantics
- For this, we have to **build a 'semantic lexicon'** (when taxonomized/structured, this is usually called an ontology)
- So we need a theory of semantics, at least for concepts
- Annotation in this context:
  1. Define/identify/choose textual signal of meaning/concept — e.g., a word or phrase
  2. Select its meaning
  3. Collect all the instances of the same meaning, study them, and deduce their semantic representation
  4. Organize all the meanings/concepts into the ontology

The generic annotation pipeline



Annotation as a kind of methodology

- **Traditional goal**: Create high-accuracy computer applications
  - Old method: Build rules for computer programs
  - New method:
    1. Have humans manually insert/add information into a (text) corpus
    2. Train computers on the corpus to do the same job

- **New goal**: Use annotation as a way to test aspects of a (ontological) theory empirically
  - For this, though, need to systematize the process — 'annotation science'

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'?

Some (meta-) issues

- **Degree of automation**: fully manual, automated-assist, semi-automated, fully automated
- **Level and kind of interpretation**: entity identification, semantic interpretation (e.g., entity class), rephrasing/summary, full curation (i.e., reformulation of content in order e.g. to fill a knowledge base)
- **Step where the annotation occurs**: by the source author, an editor, a referee, the publisher; also by a single annotator, by a user community
- **Access to the interface**: special, web plug-in, wiki

Overview

- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

Amazon's Mechanical Turk

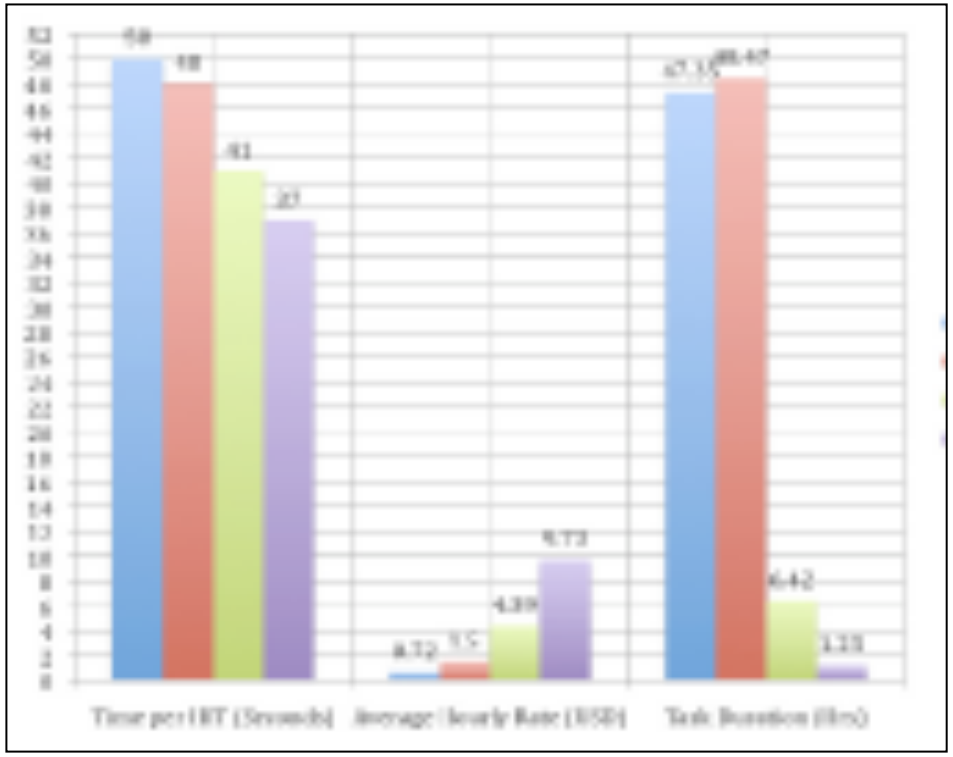
- Service offered by Amazon.com at
  https://www.mturk.com/mturk/welcome
- Researchers post annotation jobs on the MTurk website:
  - Researcher specifies annotation task and data
  - Researcher pays money into Amazon account, using credit card
  - Researcher specifies annotator characteristics and payment (typically, between 1c and 10c per annotation decision)
- People perform the annotations via the internet
  - People sign up
  - May select any job currently on offer
  - May stop at any time: simply bail out and leave
  - Researchers can rate annotators (star ratings, like eBay)
- (Each individual decision in MTurk is called a "hit")

Some questions for MTurk

- How much to pay them?
  - Feng et al.: 5c a hit
  - Graph: completion rate vs amt paid
- How large to make each job?
  - Too large and people flake out
  - Graph: completion rate vs. job length
- How many annotators to hire?
  - Graph: completion ratios vs. ?
- What to do with the poor annotators?
  - When can you discard them without cheating?

Using MTurk (Feng et al. 09)

Payment	1c	2c	5c	10c
Agreement	.807	.858	.935	.903



Time per HIT Avg hourly rate Task duration

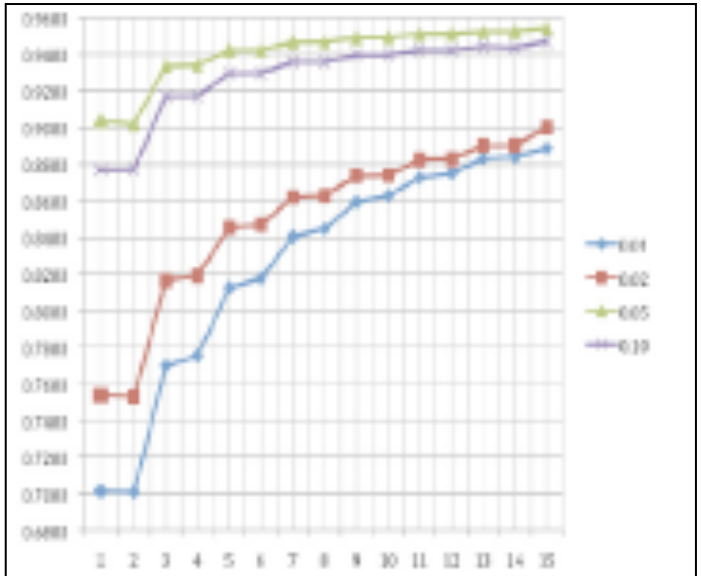


Figure 4. Agreement with experts.

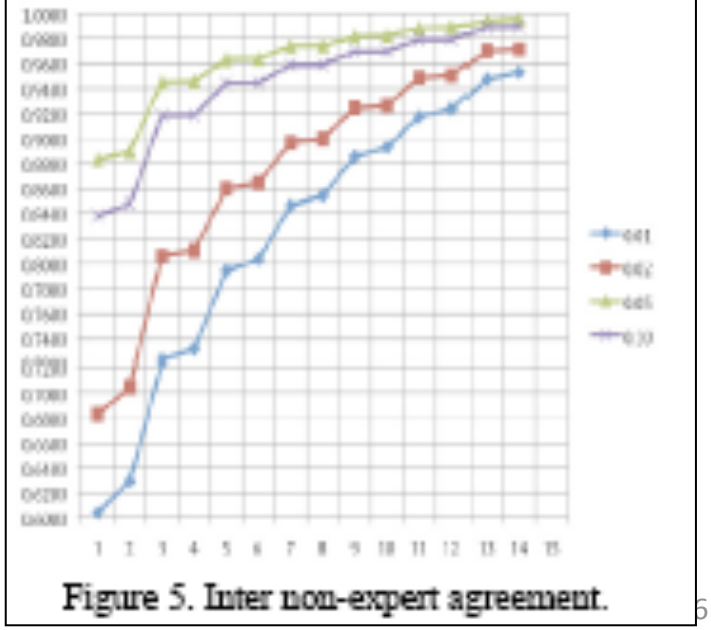


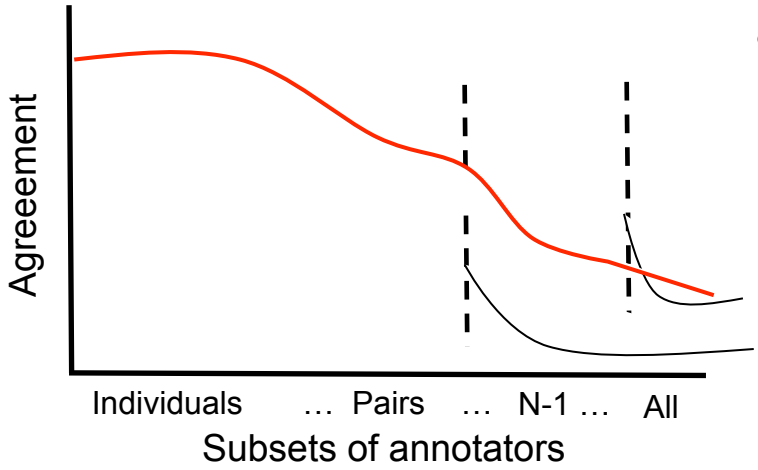
Figure 5. Inter non-expert agreement.

Throwing out annotators

- Every MTurk exercise always includes weirdos whose answers seem random (or automated)
- If you throw them out, what's the limit?

(If you throw out all the bad' ones, you get high agreement—but is this a true reflection of the world? You do annotation precisely to *test* whether agreement is reachable!

- An idea:



- Each time you discard an annotator, agreement goes up
- Discard only while *Δ%Agrmnt > Δ%annotrs*

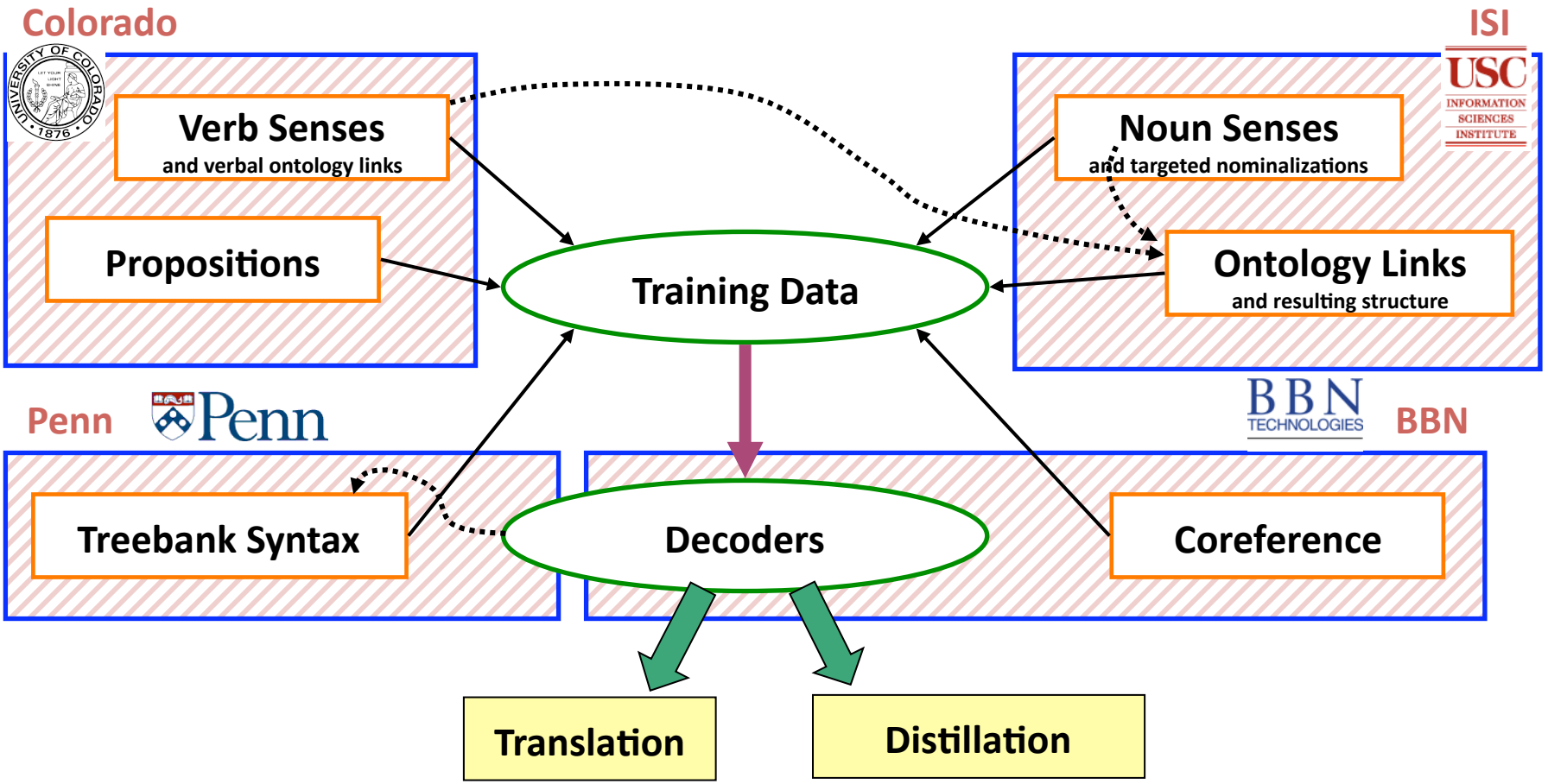
1. Remove worst guy
2. Remove next-worst guy

Example project: OntoNotes

- Partners: BBN (Weischedel), U of Colorado (Palmer), U of Pennsylvania (Marcus), USC/ISI (Hovy)
- Time frame: 2006–2010
- 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 (PropBank)
  - Manually link **anaphoric references**
  - Manually construct supporting **ontology of senses**

Project structure

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



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

- Coreference links
- Ontology
- Decoders

Example task: Word senses in Ontonotes

- Create a corpus of 'semanticized' text by **annotating JUST the semantic sense(s)** of every noun, verb, adjective, and adverb…
- Why?
  - **Build ontologies** to reflect (domain) text
  - Enable **computer programs to learn to assign correct senses** automatically, for better info extraction, machine translation, summarization, question answering, (web) search, etc.
  - Begin to **understand the distribution of principal semantic features** (*animacy, concreteness*, etc.) at large scale

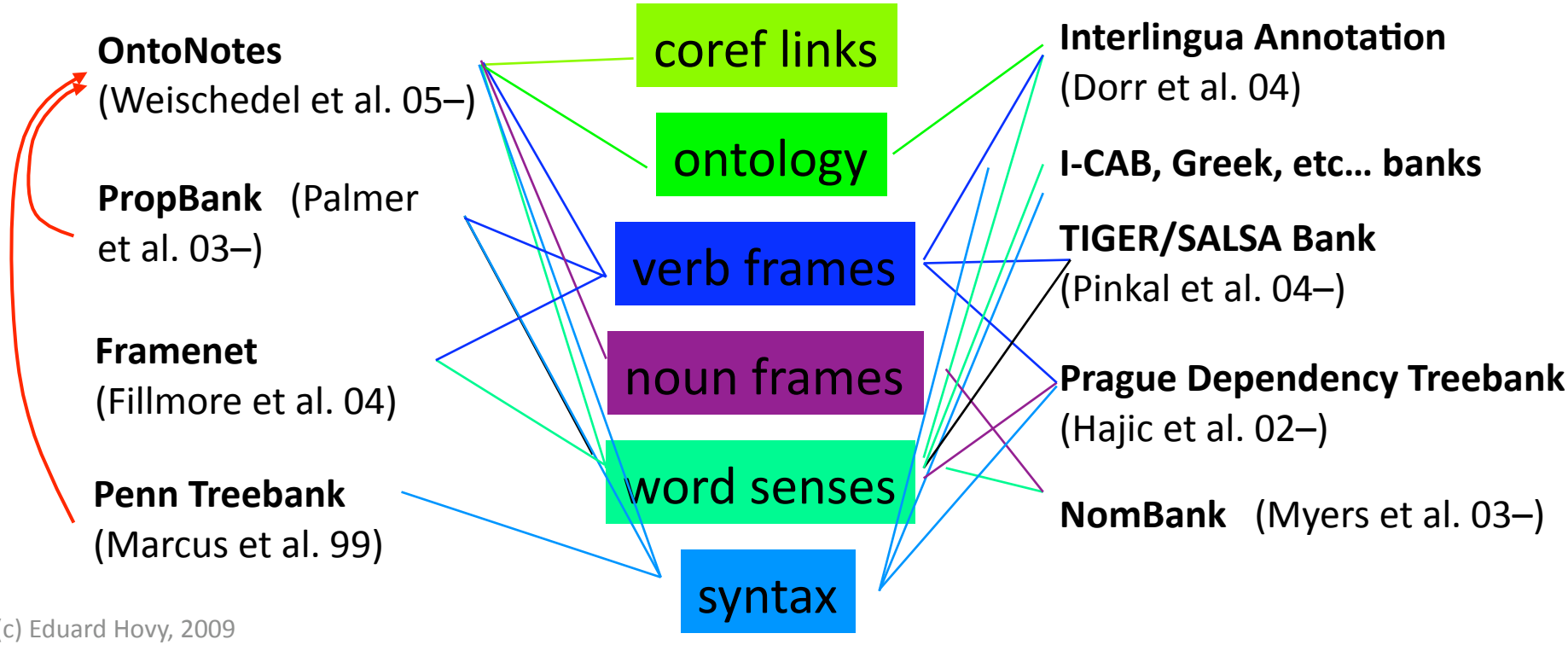
Ensuring trustworthiness/stability

- 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

Recent 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.)
  - Modalities (Dorr 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)
- China:
  - TCT and CEB corpora, Tsinghua U, Beijing (Zhou et al.)

Overview

- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

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. **The annotators**
   - Choosing the annotators — what background? How many?
   - How to avoid overtraining? And undertraining? How to even know?
4. **Annotation procedure**
   - How to design the exact procedure? How to avoid biasing annotators?
   - Reconciliation and adjudication processes among annotators
5. **Interface design**
   - Building the interfaces. How to ensure speed and avoid bias?
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-get 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/

Representativeness, balance, and timeliness

- When is a corpus representative?
  - "stock" in *Wall Street Journal* is never the soup base
  - Def: When what we find for the sample corpus also holds for the general population / textual universe (Manning and Schütze 99)
  - Degrees of representativeness (Leech 2007) needed:
    - High → for making objective measures (Biber 93; 07)
    - Low → examples only, for linguistic judgments (BNC, Brown, etc.)

- How to balance genre, era, domain, etc.?
  - Decision depends on (expected) usage of corpus (Kilgarriff and Grefenstette CL 2003)
  - Does balance equal proportionality? But proportionality of what?
    - Variation of genre (= news, blogs, literature, etc.)
    - Variation of register (= formal, informal, etc.) (Biber 93; 07)
    - Not text production, but text reception (= number of hearers/readers) (Czech Natn'l Corpus)
    - Variation of era (= historical, modern, etc.)

Social, political, funding issues

- How do you ensure agreement / complementarity with other corpora? Should you bother?
- How do you choose which phenomena to annotate? Need high payoff…
- How do you convince funders to invest in the effort?

Suggested methodology

- Pick a specific day
- Take a 'snapshot' of the text world: count the sizes of that day for all the various genres and domains of interest, from various collections
  - Library of Congress
  - A large publisher
  - A large portion of the web
  - A group of newspapers

  This provides the desired distribution
- Then try to obtain (license) the correct proportions of each of the genres and domains
- Assemble as much text as possible into the corpus
- Show statistics of desired and actual distributions

Setting up OntoNotes: Word statistics

Number of word tokens/types in 1000-word corpus

(95% confidence intervals on 85213 trials)

Nouns: approx. 50% of tokens

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

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

Polysemy of verbs and nouns

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%)

Coverage in WSJ and Brown Corpus of most frequent *N* polysemous-2 nouns

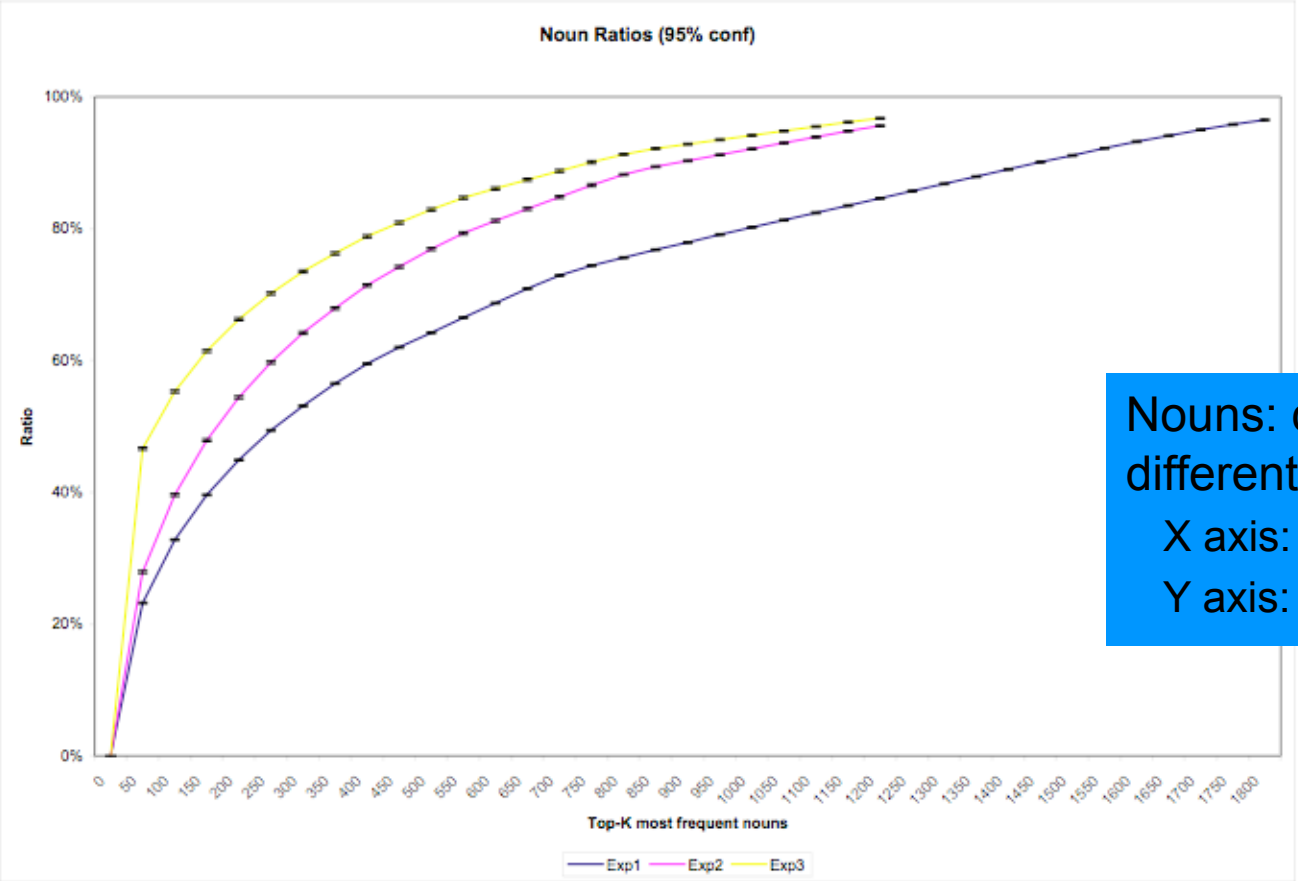
Nouns	Tokens (total 205442)	
100	76420	37%
500	140453	68%
1000	167715	82%
1500	181412	88%
2000	189641	92%

OntoNotes decisions

- Year 1: started with what was available
  - Penn Treebank, already present, allowed immediate proposition and sense annotation (needed this for verb structure annotation)
  - Problem: *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: How many nouns to annotate?

Effect of corpus/genre on coverage:



Nouns: coverage for 3 different corpora

- X axis: most-freq *N* nouns
- Y axis: coverage of nouns

Overview

- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

Q2: 'Instantiating' the theory

- Most complex question: What **phenomena** to annotate, with which **options**?

- Goal: for practical task (e.g., Info Extraction), for theory building (linguistics), or both?
  - The 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: What's the right level of 'granularity'?

How ‘deeply’ to instantiate the theory?

- What representation/annotation classes to use?
  - Design rep scheme / formalism very carefully — simple and transparent
  - ? Depends on theory — but also (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

Hints for instantiating the theory

- Issues:
  - Before building the theory, you don't know **how many categories (types) really appear** in the data
  - When annotating, you don't know **how easy it will be for the annotators to identify all the categories** your theory specifies
- Likely problems:
  - **Categories not exhaustive** over phenomena in the data
  - **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
  - Create and constantly update the **Annotator Handbook**
    - (Penn Treebank Codebook: 300 pages!)
  - Modify your categories as needed—is the problem with the annotators or the theory? Make sure the annotators are not inadequate…
  - Measure the annotator agreement and disagreement (see below)

Measuring aspects of (dis)agreement

(Lipsitz et al., 1991; Teachman, 1989)

- **Precision** (correctness)
  - P_i = *#correct / N*
  - *Measures correctness of annotators: conformance to gold standard*
  - Corresponds to 'easiness' of category and choice
- **Entropy** (ambiguity, regardless of correctness)
  - $E_i = -\sum_i P_i \,.\, ln\, P_i$
  - *Measures dispersion of annotator choices (the higher the entropy, the more dispersed: 0 = unambiguous)*
  - Indicates clarity of definitions
  - Example: 5 annotators, 5 categories:

	C1	C2	C3	C4	C5	E_i
Ex 1	1	1	1	1	1	1.61
Ex 2	5	0	0	0	0	0
Ex 3	0	3	2	0	0	0.67
…	…	…	…	…	…	…

Can normalize:

$NH_i = 1 - E_i / E_{i\text{-}max}$

$E_{i\text{-}max}$ = *max* E_i *for each example*

(NH $\rightarrow$ 1 means less ambiguity)

(Bayerl, 2008)

Distinguishability of classes

- **Odds Ratio** — for two categories
  - $OR_{xy} = \frac{f_{xx} f_{yy}}{f_{xy} f_{yx}}$ *(where f_{xy} means annotator 1 chooses x and annotator 2 chooses y)*
  - *Measures how much the annotators confuse two categories: collapsability of the two (= 'neutering' the theory)*
  - High values mean good distinguishability; small means indistinguishable

- **Collapse method**
  - Create two classes (class *i* and all the others, collapsed) and compute agreement
  - Repeat over all classes
  - See (Teufel et al. 2006)

Q2: Theory and model

- First, you obtain the theory and annotate
- But sometimes the theory is controversial, or you simply cannot obtain stability (using the previous measures)
- All is not lost! You can 'neuter' the theory and still be able to annotate, using a more neutral set of classes/types
  - 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 cruder / less detailed ones

- **When to neuter?** — you must decide acceptability levels for the measures
- **How much to neuter?** — do you aim to achieve high agreement levels? Or balanced class representativeness for all categories?
- What does this say about the theory, however?

OntoNotes acceptability threshold: 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

Overview

- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

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

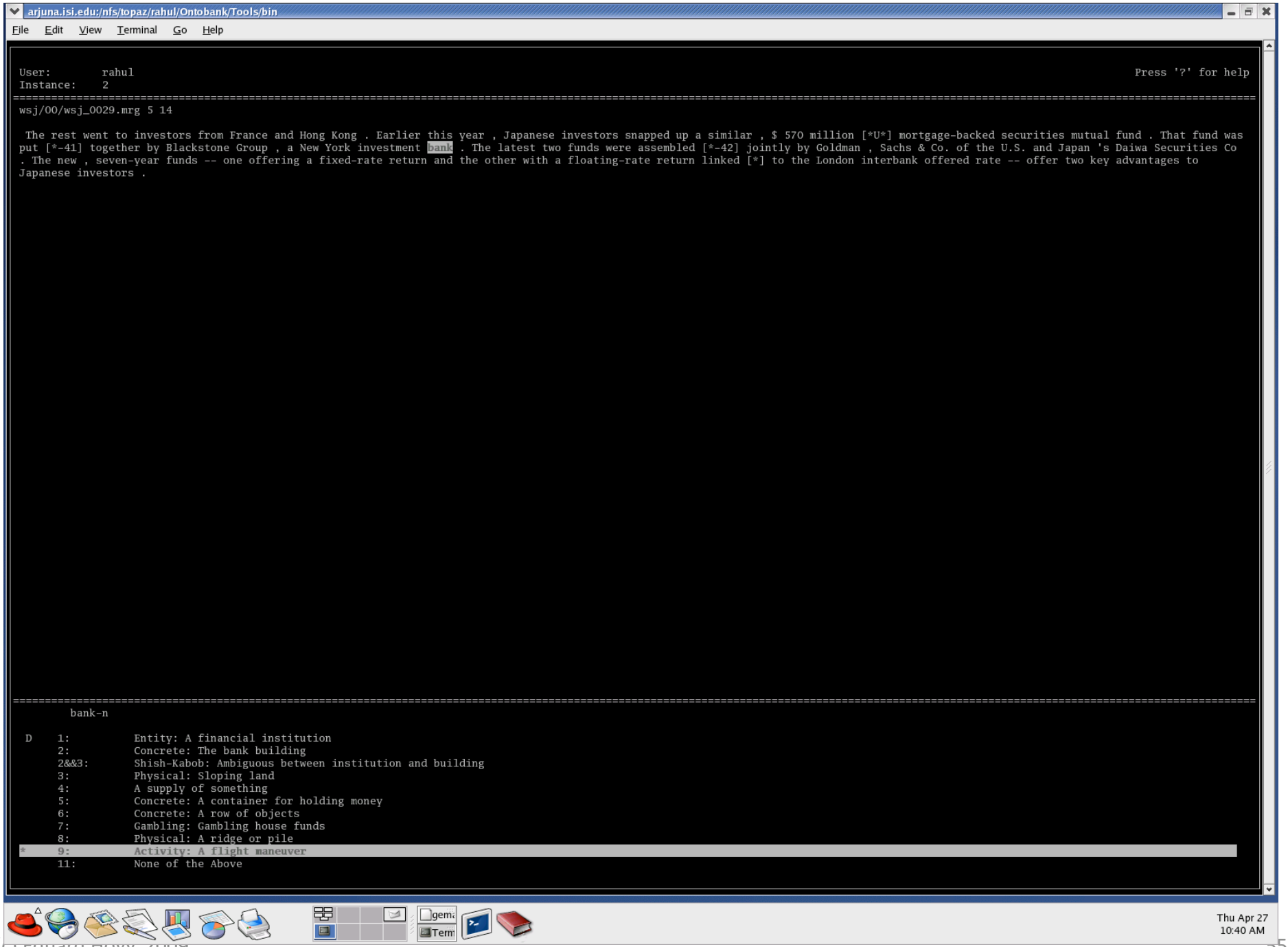
- **Delimit**: delimit a region inside a larger context
  - Often, problems with exact start/end of region (e.g., exact NP) — but preprocessing and pre-delimiting chunks introduces bias
  - Evaluation of partial overlaps is harder

- **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

Some 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


arjuna.isi.edu:/nfs/topaz/rahul/Ontobank/Tools/bin
File Edit View Terminal Go Help
User: rahul
Instance: 2
Press '?' for help
wsj/00/wsj_0029.mrg 5 14
The rest went to investors from France and Hong Kong . Earlier this year , Japanese investors snapped up a similar , $ 570 million [*U*] mortgage-backed securities mutual fund . That fund was put [*-41] together by Blackstone Group , a New York investment bank . The latest two funds were assembled [*-42] jointly by Goldman , Sachs & Co. of the U.S. and Japan 's Daiwa Securities Co . The new , seven-year funds -- one offering a fixed-rate return and the other with a floating-rate return linked [*] to the London interbank offered rate -- offer two key advantages to Japanese investors .
bank-n
D 1: Entity: A financial institution
2: Concrete: The bank building
2&&3: Shish-Kabob: Ambiguous between institution and building
3: Physical: Sloping land
4: A supply of something
5: Concrete: A container for holding money
6: Concrete: A row of objects
7: Gambling: Gambling house funds
8: Physical: A ridge or pile
* 9: Activity: A flight maneuver
11: None of the Above
Thu Apr 27
10:40 AM

Overview

- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

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…

Ontonotes agreement analysis

Sometimes, one annotator is bad

Sometimes, the choices 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	

Overview

- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

Q5: Annotation procedure

- How to manage the annotation procedure?
  - 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** *during annotation*
  - 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?

The Adjudicator

- Function: Deal with inter-annotator discrepancies
  - Either: produce final decision
  - Or: send choice back to theoretician for redefinition (and then, reannotation)
- Various possible modes:
  - **Adjudicator as extra (super) annotator**:
    - Is presented with all instances of disagreement
    - Does not see annotator choices; just makes own choice(s)
    - After that, usually considers annotator decisions
  - **Adjudicator as judge**:
    - Is presented with all instances of disagreement, together with annotator choices
    - Makes final ruling
- In OntoNotes noun annotation: Adjudicator as judge
  - "I find that seeing their choices helps me to understand how the individual annotators are thinking about the senses. It helps me to determine if there is a problem with the sense or if the annotator may have misunderstood a particular sense. Sometimes I notice if an annotator tends to stick with a particular sense if the context is not clear or is not looking at enough of the larger context before make the a choice."
  - "There are many times when neither annotator choice is wrong, but I still find one to be better than the other (I'm guessing somewhere between 20% – 30%). It is more rare that I feel that the instance is so ambiguous that either choice is equally good…"
  - Adjudicator disagreed with *both* annotators very seldom

Q5: Annotation procedure heuristics

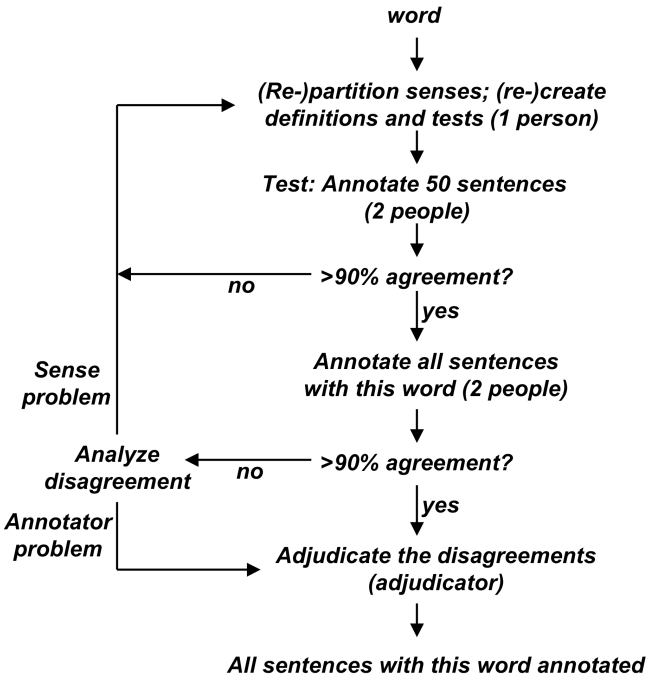
- Overall approach — Shulman's rule: do the easy annotations first, so you've seen the data when you get to the harder cases
- The '85% clear cases' rule (Wiebe):
  - Ask the annotators also to mark their level of certainty
  - There should be a lot of agreement at high certainty — the clear cases
- 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)
  - **(Here can use active learning**: 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

Ontonotes 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



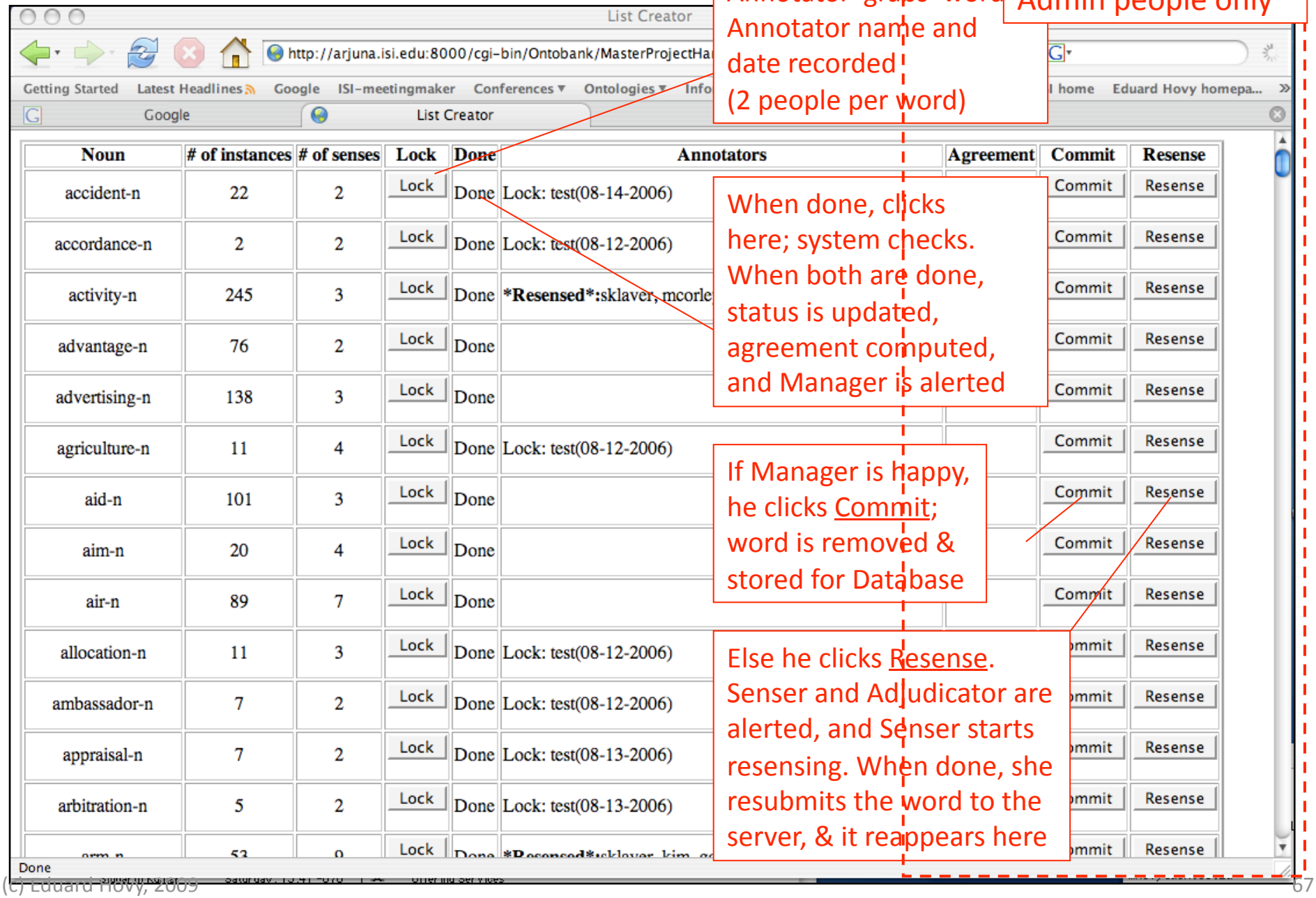
- Loop 2:
  - Annotators (at least 2) annotate all the remaining sentences
  - Manager checks their performance:
    - 90%+ agreement + few or no *NoneOfAbove* — send to Adjudicator to fix the rest
    - Else — Adjudicator annotates differences
    - If Adj agrees with one Annotator 90%+, then ignore 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

Ontonotes 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.)

- Need several interfaces and systems:
  - 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 a word to annotate
  - Annotation Status interface (ISI): up-to-the-minute status
  - Statistics bookkeeper (to be built): individual annotator stats

Master Project Handler



Annotator 'grabs' word
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

Ontonotes noun status page

Dynamically updated

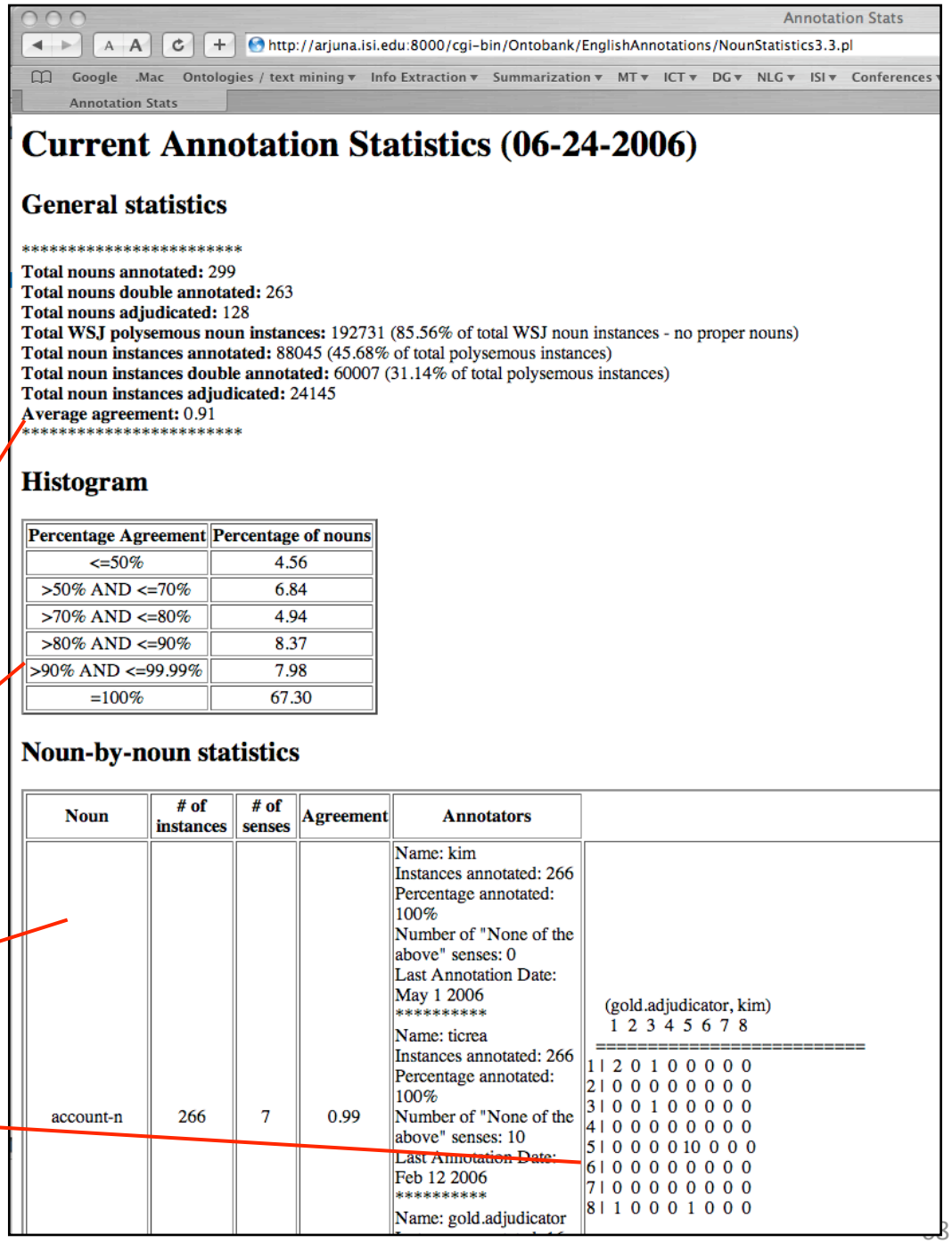
http://arjuna.isi.edu:8000/Ontobank/AnnotationStats.html

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

Agreement histogram

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

Confusion matrix for results

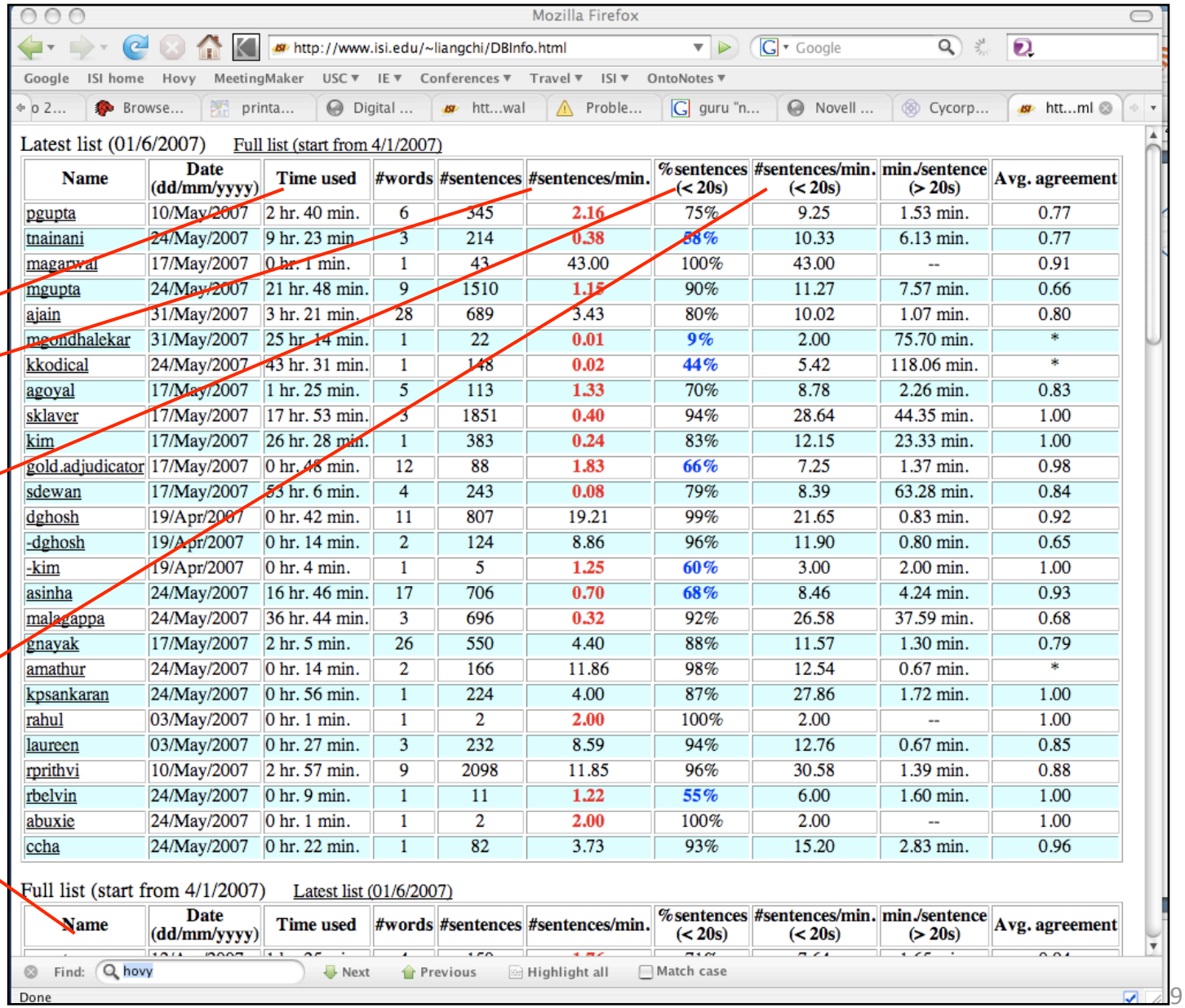


Ontonotes annotator work record

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

Ontonotes 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?

Overview

- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

Adapting active learning for WSD

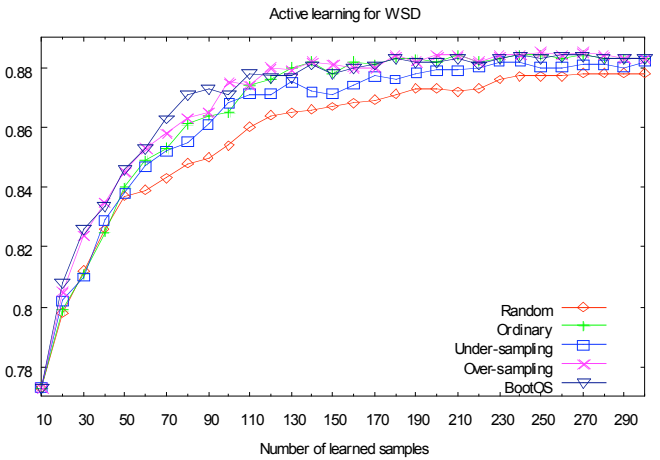
- **Problem**: Human annotation is expensive and time-consuming
  - Can we use Active Learning to minimize human annotation effort?
- **Imbalance** is a problem:
  - WSJ sense distribution is very skewed — creates large discrepancy between the prior probabilities of the individual senses:
    - For all 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?

Active learning for auto. WSD

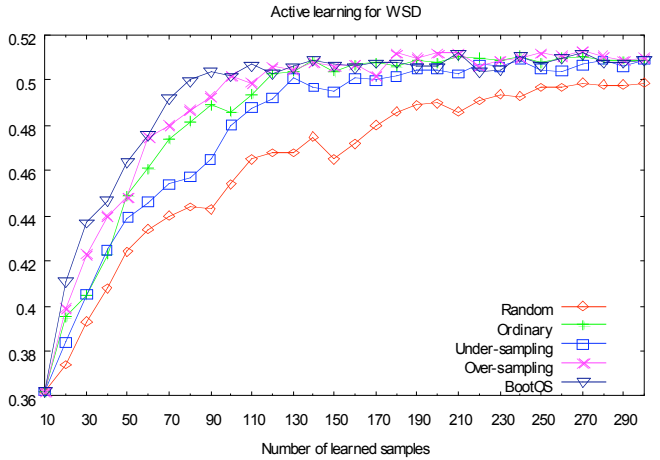
- Sabbatical visit: Prof. Jingbo Zhu (Northeastern U, China):
  - Here for one year, since October
- Stage 1: collect statistics (English only)
  - For all the annotated nouns: about 78.9% of nouns are covered by the first sense, and about 93.3% by the top two senses
  - For only the nouns with high agreement: 86% are covered by top sense; 95.9% by top 2 senses; 98.5 by top 3
  - Calculated entropy of sense distributions; sorted; found 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
- Stage 2: build active learner
  - Focus on *highly imbalanced* cases (cannot improve on *extreme* case, and *somewhat* case is ok for typical WSD learning)
  - Two uses:
    - See whether we can replace one human annotator by machine, after initial annotation allows WSD system to gear up
    - Enable us to cull lower-freq senses from other corpora automatically, for annotation

Results of classifier training

- 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 find new samples from list using k-NN and similarity functions



Macro-average	Ordinary	Under-sampling	Over-sampling	BootOS
Accuracy (%)	89.03	72.42	89.30	<u>89.40</u>
Precision (%)	72.40	60.33	<u>73.94</u>	73.18
Recall(%)	64.51	<u>71.18</u>	66.37	70.94
F1(%)	67.87	64.78	69.61	<u>71.73</u>



Overview

- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

What to measure?

- Fundamental assumption: The work is trustworthy when independent annotators agree
- But what to measure for 'agreement'?

Q6.1 Measuring **individual agreements**

- Pairwise agreements and averages

Q6.2 Measuring **overall group behavior**

- Group averages and trends

Q6.3 Measuring **characteristics of corpus**

- Skewedness, internal homogeniety, etc.

6.1: Measuring individual agreements

- Evaluating individual pieces of information:
  - What to evaluate:
    - **Individual** agreement scores between creators
    - **Overall** agreement averages and trends
  - 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 (almost all decisions have one value) 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 scoring: Kappa

- Simple agreement:

  A = number choices agreed / total number of choices

- But what about random agreement?
  - Annotators might agree by chance!
  - So 'normalize': compute expected (chance) agreement

    E = expected number of choices agreed / total number

- Remove chance, using Cohen's Kappa:

  Kappa = (A - E) / (1 - E)

  Ratio with perfect annotation: (100% - E)

- Example:
  - Assume 100 examples, 50 labeled *A*, and 50 *B*: $E_{random} = 0.5$
  - Then a random annotator would score 50%: $A_{random} = 0.5$
  - But $Kappa_{random} = (0.5 - 0.5) / (1 - 0.5) = 0$
  - And an annotator with 70% agreement?: $A_{70} = 0.7$
  - $Kappa_{70} = (0.7 - 0.5) / (1 - 0.5) = 0.2 / 0.5 = 0.4$
  - 0.4 is much lower than 0.7, and reflects only the nonrandom agreement

Problems with Kappa

- Problems:
  - Works only for comparing 2 annotators
  - 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?
- Some solutions:
  - For more than 2 annotators use *Fleiss's Kappa*
  - Choosing more than one label at a time
    - Extension by Rosenberg and Binkowski (2004)
    - But may return low Kappa even if agreement on two labels (Devillers et al. 2006)
  - For skewed distributions, perhaps just use agreement

Extending Kappa

- Choosing more than one label at a time
  - Extension by Rosenberg and Binkowski (2004)
  - But may return low Kappa even if agreement on two labels (Devillers et al. 2006)

- Krippendorff's (2007) alpha:
  - Observed disagreement D_o
  - Expected disagreement D_e
  - Perfect agreement: $D_o = 0$ and $\alpha = 1$
  - Chance agreement: $D_o = D_e$ and $\alpha = 0$
  - Advantages:
    - Any number of observers, not just two
    - Any number of categories, scale values, or measures
    - Any metric or level of measurement (nominal, ordinal, interval, ratio …)
    - Incomplete or missing data
    - Large and small sample sizes alike, no minimum cutoff

$$\alpha = 1 - \frac{D_o}{D_e}$$

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck\ \text{metric}}\, \delta^2_{ck}$$

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_{k\ \text{metric}}\, \delta^2_{ck}$$

(Reidsma & Carletta, CompLing 2007)

Problems with simple uses of Kappa

- Important: Effect of pattern in annotator disagreement:
  - Simulated disagreements (random or not), trained neural network classifier
  - No pattern: disagreement is random, and machine learning algorithms ignore it
  - Yes pattern: disagreement introduces additional 'trends', and machine learners' results give artificially high Kappa scores

X-axis: Kappa of data
Y-axis: true accuracy
Blue: compared to real data
Red: or to disagreement data
(on 4 data correlation levels)

- What to do?
  - Calculate per-class reliability score (Krippendorff 2004)
  - Study annotators' choices on disagreed items (Wiebe et al. 1999)
  - Try to find what caused disagreements — change schema, get new annotators, etc. (Bayerl and Paul 2007)

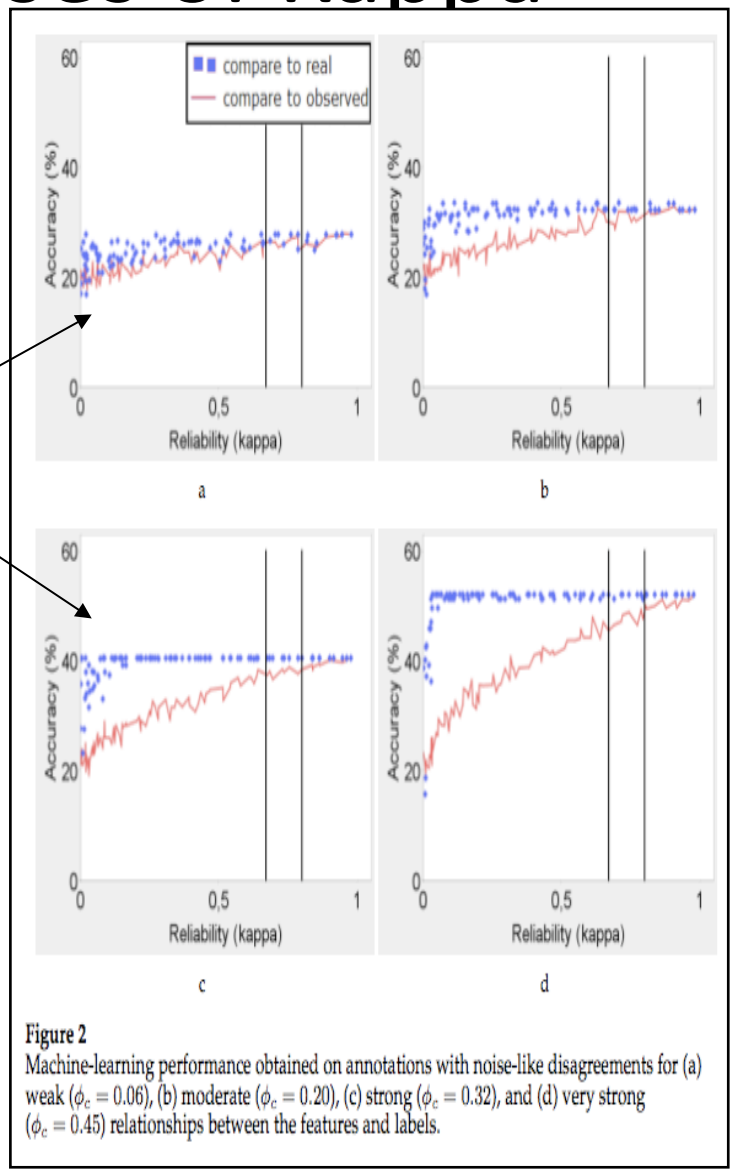
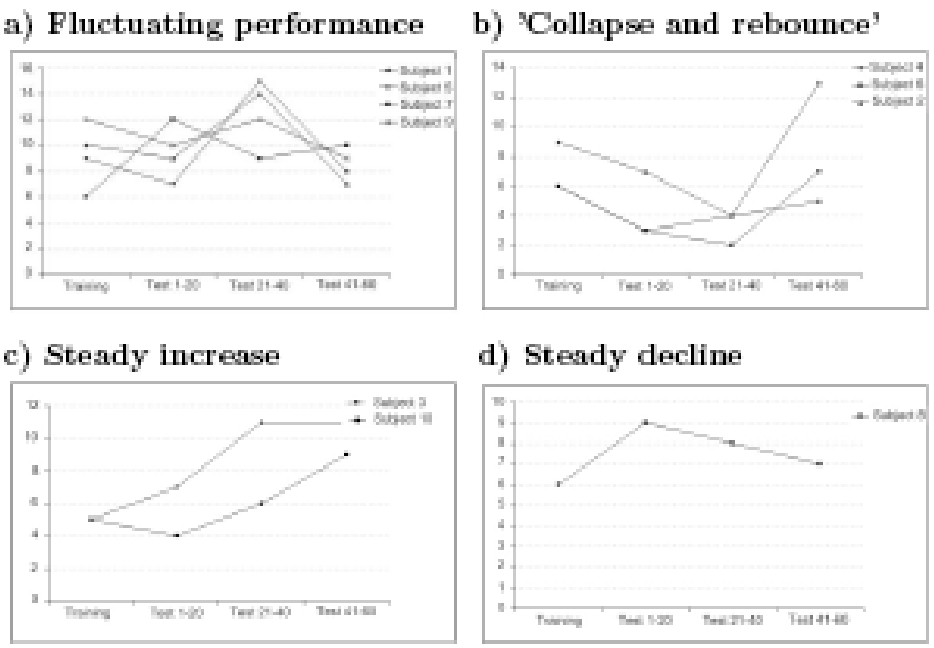


Figure 2
Machine-learning performance obtained on annotations with noise-like disagreements for (a) weak ($\phi_c = 0.06$), (b) moderate ($\phi_c = 0.20$), (c) strong ($\phi_c = 0.32$), and (d) very strong ($\phi_c = 0.45$) relationships between the features and labels.

Trends in annotation correctness

- If you have a Gold Standard, you can check how annotators perform over time — 'annotator drift'

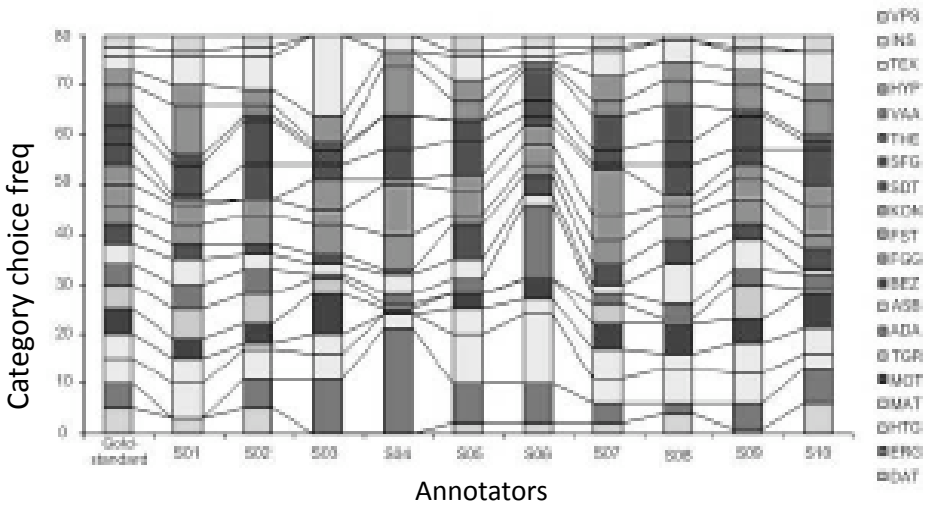
- Example (Bayerl 2008)
  - 10 annotators
  - Averaged precision (correctness) scores over groups of 20 examples annotated
  - a) some people go up and down
  - b) some people slump but then perk up
  - c) some people get steadily better
  - d) some just get tired…
  - Suggestion: Don't let annotators work for too long at a time
- Why the drift?
  - Annotators develop own models
  - Annotators develop 'cues' and use them as short cuts — may be wrong



6.2: Measuring group behavior 1

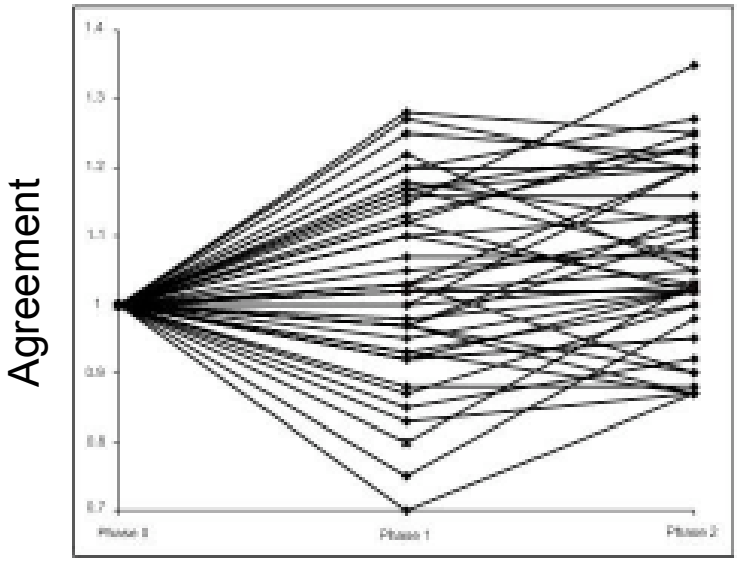
Compare behavior/statistics of annotators as a group

- Distribution of choices, and change of choice distribution over time — 'agreement drift'
  - Example (Bayerl 2008)
    - 10 annotators, 20 categories
    - Annotator S04 uses only 3 categories for over 50% of the examples, and ignores 30% of categories — not good
  - Check for systematic under- or over-use of categories (for this, compare against Gold Standard, or against majority annotations)



Measuring group behavior 2

- Study pairwise (dis)agreement over time
  - Count number of disagreements between each pair of annotators at selected times in project
  - Example (Bayerl 2008), scores normalized:
    - Trend: higher agreement in Phase 2
    - Puzzling: some people 'diverge'
  - Implication: Agreement between people at one time is not necessarily a guarantee for agreement at another



6.3: Measuring characteristics of corpus

1. Is the corpus consistent (enough)?
   - Many corpora are compilations of smaller elements
   - Different subcorpus characteristics may produce imbalances in important respects regarding the theory
   - How to determine this? What to do to fix it?

2. Is the annotated result enough? What does 'enough' mean?
   - (Sufficiency: when the machine learning system shows no increase in accuracy despite more training data)

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?

Imbalance in OntoNotes

- **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 find potentially infrequent senses in order to balance them!

Overview

- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

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 with Ontonotes-type data:
  - 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

Overview

- Introduction: What is annotation, and why annotate?
- Practical examples: Amazon Mech. Turk and OntoNotes
- The seven questions of annotation science
  - Q1: Selecting a corpus
  - Q2: Instantiating the theory
  - Q3: Designing the interface
  - Q4: Selecting and training the annotators
  - Q5: Designing and managing the annotation procedure
  - Accelerating annotation with active learning
  - Q6: Validating results
  - Q7: Delivering and maintaining the product
- Annotation today: The current research world

Questions to ask of an annotated corpus

- **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 the corpus homogeneous (enough)?
  - Is it correct? (can be correct in various ways!)
  - How was it / can it be used?

Some current techniques and areas where annotation may be applied

- 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

Example: SCALE annotation of Modalities

- Coordinated effort in USA (09–), led by Bonnie Dorr, U of Maryland
- Task: Annotate modalities in Urdu text
- Modalities:
  - Def: Modality is an **attitude** on the part of the speaker toward an *action* or *state*. Modality is expressed with bound morphemes or free-standing words or phrases. Modality interacts in complex ways with other grammatical units such as tense and negation.
  - Modality structure: Holder (**H**), Target (**R**), Trigger word/phrase (**M**)
  - 8 modalities in SCALE effort:

    Requirement: does H require R?
    Permissive: does H allow R?
    Success: does H succeed in R?
    Effort: does H try to do R?
    Intention: does H intend R?
    Ability: can H do R?
    Want: does H want R?
    Belief: How strongly does H believe R?
- Associated document lists many semantic phenomena and annotation formats and examples (Dorr 2008)

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

Writing a paper on annotation

- 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 equivalent

problem

past work

training
algorithm

evaluation

distribution

Related work on annotation methodology

(Spiegelman et al., 1953; Bayerl, 2008)

- Material/corpus
  - Type, amount, complexity, familiarity to annotators
- Classification scheme
  - Number and kinds of categories, complexity, familiarity to annotators
- Annotator characteristics
  - Personality, expertise in domain, ability to concentrate, interest in task
- Annotator training procedure
  - Type and amount of training
  - (Experience in domain may better than training!—Bayerl)
- Process
  - Physical situation, length of annotation task, reward system (pay by the amount annotated?—you get speed, not accuracy; pay by agreement?—annotators might unconsciously migrate to default categories, or even cheat)

Some useful readings 1

- **Corpus design**
  - Biber, D. 1993. Representativeness in Corpus Design. *Linguistic and Literary Computing* 8(4): 243–257.
  - Biber, D. and J. Kurjian. 2007. Towards a Taxonomy of Web Registers and Text Types: A Multidimensional Analysis. In M. Hund, N. Nesselhauf and C. Biewer (eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi.
  - Leech, G. 2008. New Resources or Just Better Old Ones? The Holy Grail of representativeness. In M. Hund, N. Nesselhauf and C. Biewer (eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi.
  - Cermak, F and V. Schmiedtová. 2003. The Czech National Corpus: Its Structure and Use. In B. Lewandowska-Tomaszczyk (ed.) *PALC 2001: Practical Applications in Language Corpora*. Frankfurt am Main: Lang, 207–224.

- **Stability of annotator agreement**
  - Bayerl, P.S. 2008. The Human Factor in Manual Annotations: Exploring Annotator Reliability. *Language Resources and Engineering*.
  - 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.
  - Teufel, S., A. Siddharthan, and D. Tidhar. 2006. An Annotation Scheme for Citation Function. *Proceedings of the SIGDIAL Workshop*.

Some useful readings 2

- **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.
  - Kappa agreement studies and extensions:
    - Reidsma, D., and J. Carletta. 2008. Squib in *Computational Linguistics*.
    - Devillers, L., R. Cowie, J.-C. Martin, and E. Douglas-Cowie. 2006. Real Life Emotions in French and English TV Clips. *Proceedings of the 5th LREC*, 1105–1110.
    - Rosenberg, A. and E. Binkowski. 2004. Augmenting the Kappa Statistics to Determine Interannotator Reliability for Multiply Labeled Data Points. *Proceedings of the HLT-NAACL Conference*, 77–80.

- **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).*

Some useful readings 3

- **Formats and standards:**
  - Dorr, B.J., S. Muhammad, and B. Onyshkevych. 2009. From Linguistic Annotations to Knowledge Objects. HLTCOE Technical Report number 1.

- **General collections of annotation papers:**
  - Coling 2008 workshop on human judgments in Computational Linguistics: http://workshops.inf.ed.ac.uk/hjcl/.

Annotation references

- Baumann, S., C. Brinckmann, S. Hansen-Schirra, G-J. Kruijff, I. Kruijff-Korbayov´a, S. Neumann, and E. Teich. 2004. Multi-dimensional annotation of linguistic corpora for investigating information structure. *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*. Boston, MA.
- Calhoun, S., M. Nissim, M. Steedman, and J. Brenier. 2005. A framework for annotating information structure in discourse. *Froceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. Ann Arbor, MI.
- Hovy, E. and J. Lavid. 2007a. Classifying clause-initial phenomena in English: Insights for Microplanning in NLG. *Proceedings of the 7th International Symposium on Natural Language Processing (SNLP).* Pattaya,Thailand.
- Hovy, E. and J. Lavid. 2007b. Tutorial on corpus annotation. Presented at *7th International Symposium on Natural Language Processing (SNLP).* Pattaya,Thailand.
- Miltsakaki, E., R. Prasad, A.Joshi, and B.Webber. 2004. Annotating discourse connectives and their arguments. *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation.* Boston, MA.
- Prasad, R., E. Miltsakaki, A. Joshi, and B. Webber. 2004. Annotation and data mining of the Penn Discourse TreeBank. *Proceedings of the ACL Workshop on Discourse Annotation.* Barcelona, Spain.
- Spiegelman, M., C. Terwilliger, and F. Fearing. 1953. The Reliability of Agreement in Context Analysis. *Journal of Social Psychology* 37:175–187.
- Wiebe, J., T. Wilson, and C. Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39(2/3):164–210.

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

Thank you!

Questions and Discussion