Amazon Mechanical Turk and the generation of new linguistic data

Robert Munro

Stanford Linguistics

For the Seminar on Psycholinguistics

April 27 2009

Who is using AMT for linguistic data?

- Mostly tech companies, labeling large corpora to optimize search/classification:
  - Document 'topics'

    "what is the topic of this web-page/thread?"
  - Textual entailment

    "are these two descriptions about the same product?"
  - Tagging

    "is this phrase a person, organization or location?"
  - And many more
- Or NLP researchers, generating new corpora
  - Perhaps a half-dozen (and growing) in the Stanford NLP group

Who are the Turkers?

- Predominantly from the US (stats by Dolores labs):
  - United States 77.1%
  - India 5.3%
  - Philippines 2.8%
  - Canada 2.8%
  - UK 1.9%
- Education (stats by Panos Ipeirotis, NYU)
  - >50% with a Bachelors or higher (self-reported!)
- Languages spoken?
  - Unknown – Urdu is the least widely spoken language that I know of someone using

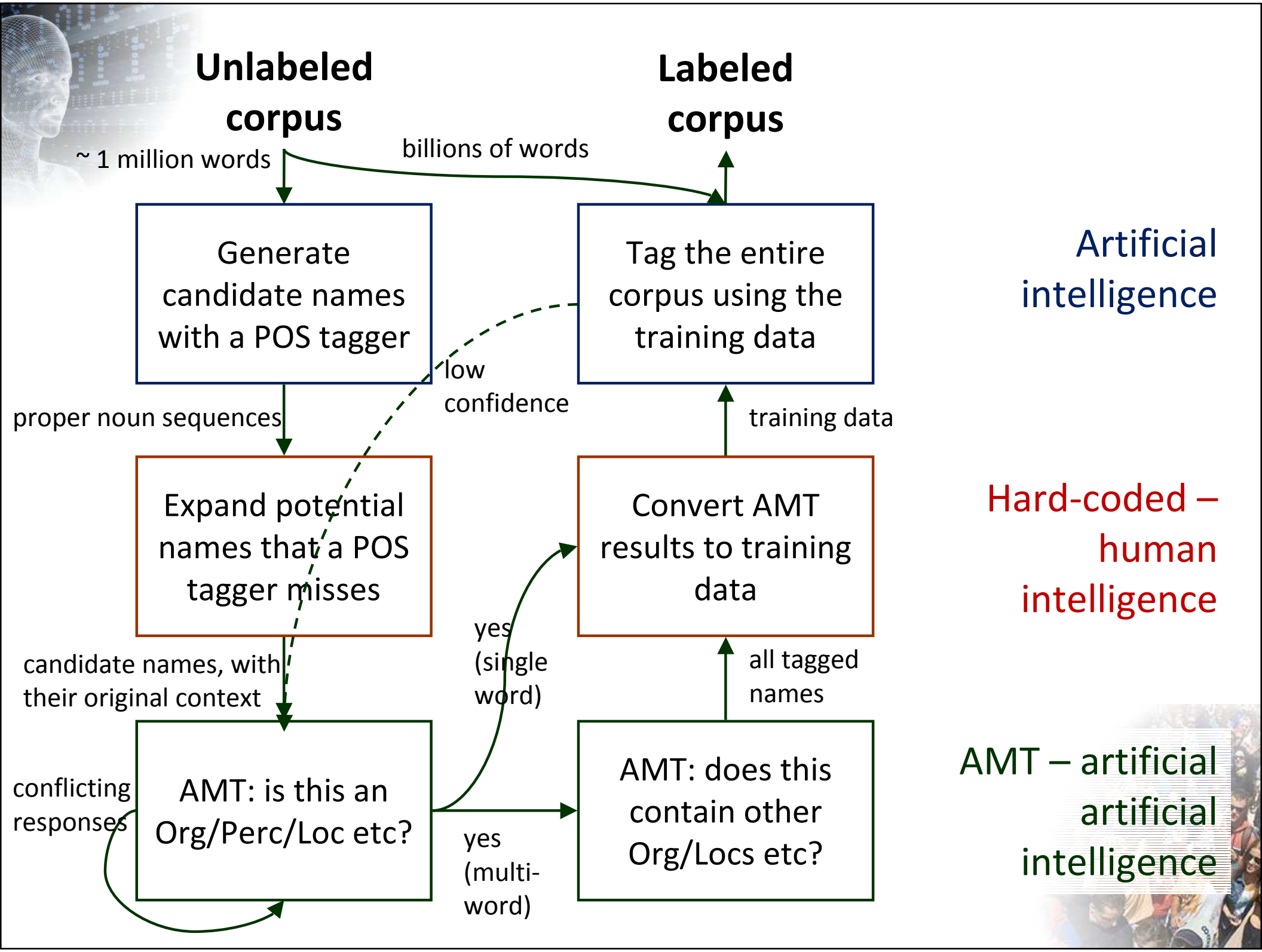
I know what you did last summer

- I built a named entity corpus and tagger at Powerset (now part of Microsoft), using AMT
  - With credit to Pascal Dennis & Valerie Nygaard.
  - Powerset employs many linguists and developers (in harmony) working on natural language search technologies.
  - This task: tagging *all* Wikipedia articles with instances of People, Organizations, Locations and Wars:

    "The Stanford Cardinals battled the University of California, Berkeley on Thursday"

    →

    "The <ORG><ORG>Stanford</ORG> Cardinals</ORG> battled the <ORG><ORG>University of <LOC>California</LOC></ORG>, <LOC>Berkeley</LOC></ORG> on Thursday":


Unlabeled corpus
Labeled corpus
~ 1 million words
billions of words
Generate candidate names with a POS tagger
Tag the entire corpus using the training data
low confidence
proper noun sequences
training data
Expand potential names that a POS tagger misses
Convert AMT results to training data
yes (single word)
candidate names, with their original context
all tagged names
conflicting responses
AMT: is this an Org/Perc/Loc etc?
AMT: does this contain other Org/Locs etc?
yes (multi-word)
Artificial intelligence
Hard-coded – human intelligence
AMT – artificial artificial intelligence

Results

- Accuracy:
  - ~90% (with high recall = low type II errors)
  - Compared to 95% inter-annotator agreement from expert annotators using dedicated annotation software
- Errors? - mostly the cultural bias of the Turkers
  - Unusual location names

    “He lived at <LOC>Stow on the Wold</LOC>”
  - Non-anglicized names

    “He went to <PER>Xue</PER> with the problem”
  - Unusual organization names, esp foreign sports teams

    “He played at <ORG>Deportivo Wanka</ORG>”

Comparing Turkers to expert annotators

- The following slides/results are from:
  - Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast - But is it Good? Evaluating Nonexpert Annotations for Natural Language Tasks. *Proceedings of EMNLP*
- A comparison of AMT and expert annotations across corpora from a number of NLP subfields

Tasks

- Affect recognition
  - Strapparava and Mihalcea (2007)

fear("Tropical storm forms in Atlantic") > *fear*("Goal delight for Sheva")

- Word Similarity
  - Miller and Charles (1991)

sim(boy, lad) > *sim*(rooster, noon)

- Textual Entailment
  - Dagan et al. (2006)

if "Microsoft was established in Italy in 1985", then "Microsoft was established in 1985" ?

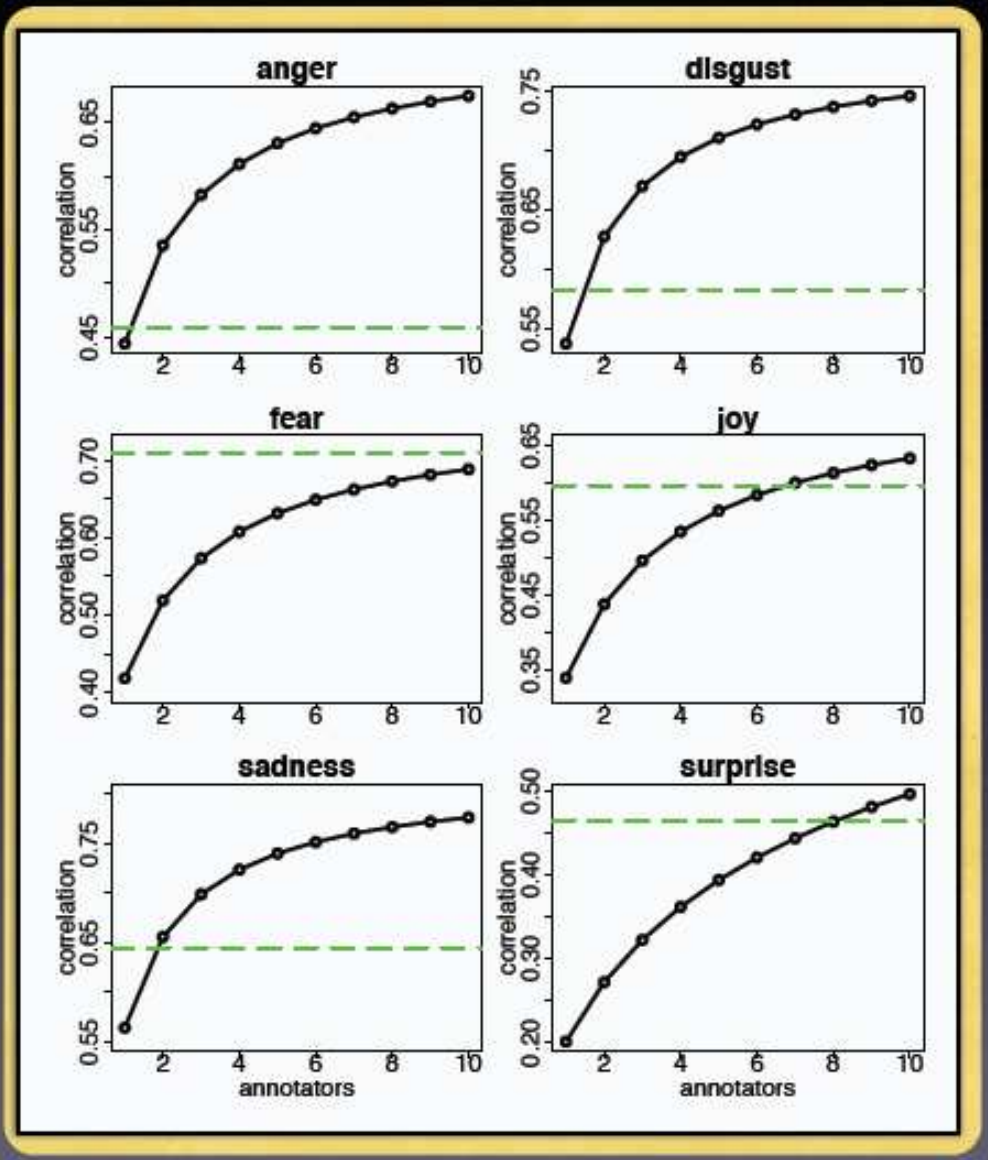
- WSD
  - Pradhan et al. (2007)

"a bass on the line" vs. "a funky bass line"

- Temporal Annotation
  - Pustejovsky et al. (2003)

ran happens before *fell* in:
"The horse ran past the barn fell."

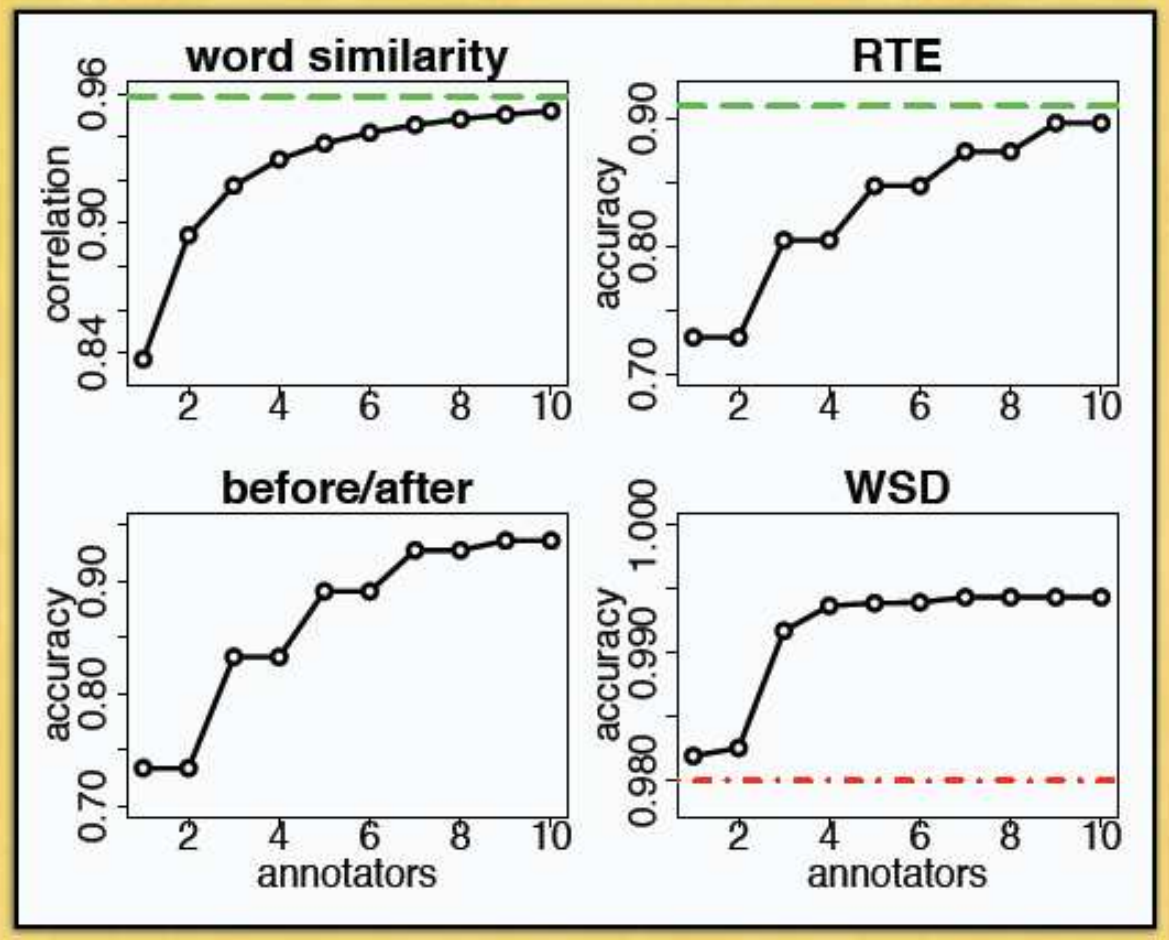
Interannotator Agreement



Emotion	Expert ITA	Aggregate Nonexpert ITA
Anger	0.459	0.675
Disgust	0.583	0.746
Fear	0.711	0.689
Joy	0.596	0.632
Sadness	0.645	0.776
Surprise	0.464	0.496
Valence	0.759	0.844
All	0.603	0.694

Number of nonexpert annotators required to match expert ITA, on average: 4

Interannotator Agreement



Task	Expert ITA	Aggregate Nonexpert ITA
Affect Recognition	0.603	0.694
Word Similarity	0.958	0.952
Textual Entailment	0.91	0.897
Temporal Annotation		0.940
WSD		0.994

Works for improving system performance, too:
Training on nonexpert annotations yields 7% higher accuracy for affect recognition

Cost Summary

Task	Total Labels	Cost in USD	Time in hours	Labels / USD	Labels / Hour
Affect Recognition	7000	$2.00	5.93	3500	1180.4
Word Similarity	300	$0.20	0.17	1500	1724.1
Textual Entailment	8000	$8.00	89.3	1000	89.59
Temporal Annotation	4620	$13.86	39.9	333.3	115.85
WSD	1770	$1.76	8.59	1005.7	206.1
All	21690	$25.82	143.9	840.0	150.7

Comparing Turkers to expert annotators

- Many errors by AMT were close votes, eg 6-4:
  - The closeness could flag these for manual inspection
- Some annotators were better than others:
  - Better results are possible by giving more weight to the opinion of "better" annotators
- The full paper and data sets are available at:
  - http://ai.stanford.edu/~rion/annotations

Beyond labeling – experimental data!

- I'm currently looking at the relationship between sentential and nominal semantic roles

  "The investigation of the police took 3 weeks to complete"

  1) Someone investigated the police and it took 3 weeks

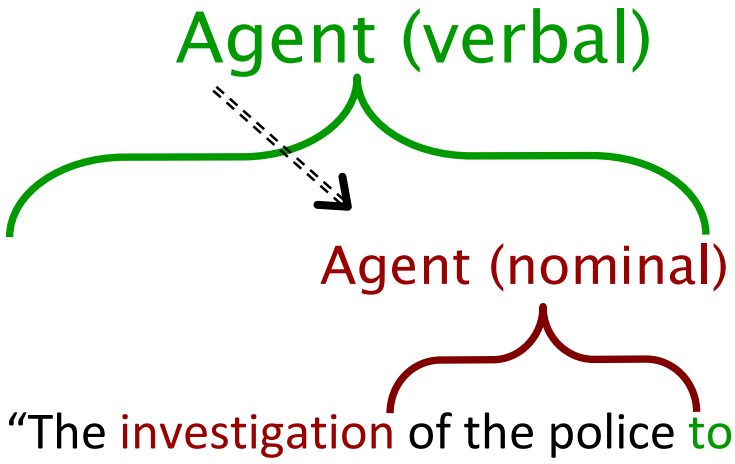
  2) The police investigated someone and it took 3 weeks

- Comparing this to:

  "It took three weeks to complete the investigation of the police"

Beyond labeling – experimental data!

- A tendency for role harmony



"The investigation of the police took 3 weeks to complete"

- 1000s of responses from 100s of people for a few dollars
  - By comparison, I am still waiting for results from the undergraduate subject pool

Overcoming AMT limits

- Ordering effects (cannot randomize questions)
  - you can release the different questions days apart – no need to 'batch' your questions to keep the subject busy for an hour
- Random click-throughs (about 0.5%)
  - Prune outliers
  - Set minimum acceptance
  - Make a random response support the null-hypothesis
- Demographics
  - Ask!
- Eye-tracking and electrodes
  - Not yet…

Why it should change linguistics

- Backup your linguistic intuitions!
  - Test your every whim for $1 in 10 minutes
- Generate/tag your own corpus
  - Goodbye Switchboard?

Tips

- You are competing in a market – if your HITs are fun you will get more responses
  - "Cats vs Dogs"

    Customer ID: XXXXXXXXXXX

    I HAVE READ THE REQUEST FOR THE ANSWER TO WHETER YOU LIKE CATS OR DOGS AND I WOULD LOVE TO RESOND TO IT

  - Online Word Games – free data! (Vickrey et al., 2008)
- For English language judgments, time the release for around 9am Eastern Time
- Add an optional field for feedback