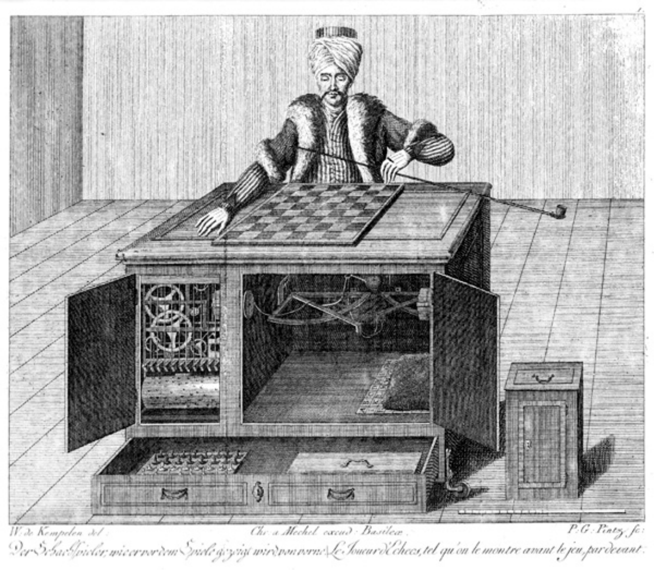
Using Mechanical Turk for Linguistic Experiments

Harry Tily

Massachusetts Institute of Technology

What is Mechanical Turk?


W. de Kempelen del:
Chr. a Mechel excud: Basileæ.
P. G. Pintz fc:
Der Schachſpieler, wie er vor dem Spiele gezeigt wird, von vorne. Le Joueur d'Echecs, tel qu'on le montre avant le jeu, par devant.

What is Mechanical Turk?

A *"labour marketplace"* for tasks that can be done at a computer

Intended to replace artificial intelligence for tasks that can not yet be done well by computers; therefore

- fast
- cheap
- can require no special skills beyond basic human intelligence

What is the appropriate category for this product?

Playstation 2 PS2 Replacement Laser Cable Free Shipping

What is the appropriate category for Playstation 2 PS2 Replacement Laser Cable Free Shipping ?

- PSP Consoles
- GameCube Consoles
- PlayStation 3 Accessories
- PSP Accessories
- PlayStation 2 Accessories
- None of the Above

Who uses Mechanical Turk?

Workers by country:

- United States: 46.80%
- India: 34.00%
- Other: 19.20%

	USA	India
Female	65	30
Male	35	70 (%)

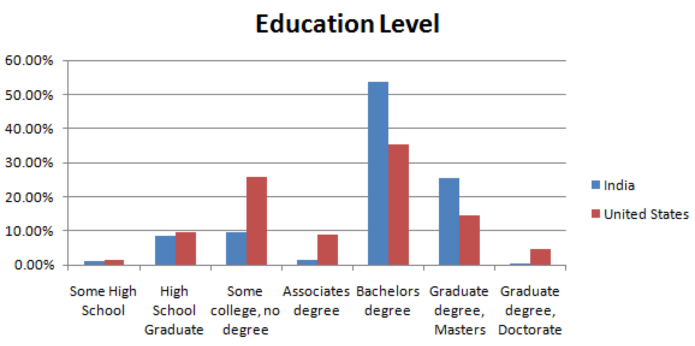
From Ipeirotis (2010)

	USA
Female	58
Male	42 (%)

From Gibson, Piantadosi & Fedorenko (to appear)

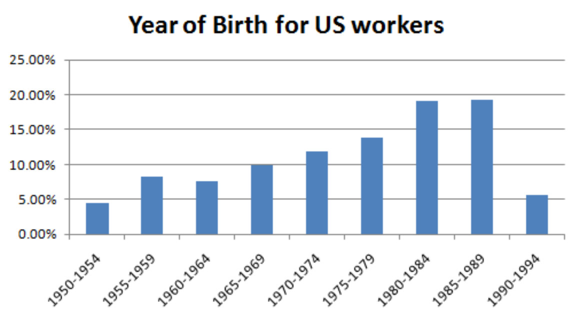
Who uses Mechanical Turk?

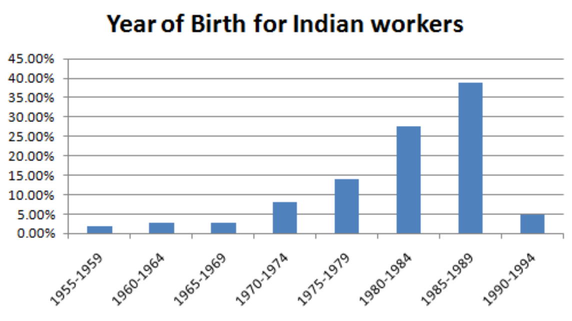
Ipeirotis (2010):



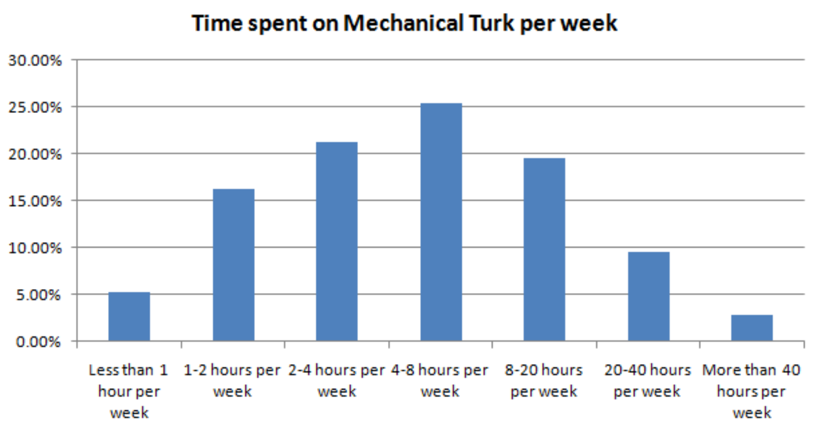
Gibson, Piantadosi & Federenko (to appear), US workers:

- no high school: 2%
- high school: 40%
- college degree: 41%
- graduate degree: 17%


Year of Birth for US workers


Year of Birth for Indian workers

Who uses Mechanical Turk?



(All these plots from Ipeirotis 2010: `http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html`)

What can be done on Mechanical Turk?

Anything you can display in a web browser!

- *Easy:* Questionnaires, surveys, rating tasks
- *Harder:* Interactive tasks, questions with feedback
- *Still Possible:* Reaction time tests, multimedia displays, voice and maybe even eye movement recording

What can be done on Mechanical Turk?

Anything you can display in a web browser!

- *Easy:* Questionnaires, surveys, rating tasks
- *Harder:* Interactive tasks, questions with feedback
- *Still Possible:* Reaction time tests, multimedia displays, voice and maybe even eye movement recording

Turkolizer

A collection of scripts designed to help you set up standard linguistic experiments

Especially: Grammmaticality judgements, rating tasks,

Documented in Gibson, Piantadosi & Fedorenko (to appear)

Recap: Experimental design

What does an experiment do?

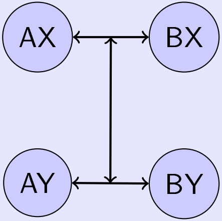
- Elicit behavioral data
  (DV: reaction times, ratings, judgements)
- Allow comparison of data elicited in different *conditions*

Conditions correspond to levels of the quantity of interest
(IV: construction, frequency, etc)

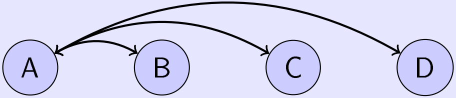
Conditions should be designed to differ *only* in the quantity of interest

Recap: Experimental design

2x2 design



Nx1 design



Compare a baseline (A) with other conditions (here N=4)

Does the difference between A and B vary with X/Y?

Example design

Research question: Are passive relative clauses favoured when they modify an animate noun phrase?
Hypothesis: Inanimate NPs make passive subject-extracted relative clauses sound more natural.

- The politician that was described by the journalist appeared in the news.
- The accident that was described by the journalist appeared in the news.

2*1 design : animate vs inanimate

Example design

Research question: Are passive relative clauses favoured when they modify an animate noun phrase?
Hypothesis: Inanimate NPs make passive subject-extracted relative clauses sound more natural.

- The politician that was described by the journalist appeared in the news.
- The accident that was described by the journalist appeared in the news.

2*1 design : animate vs inanimate

Example design

Possible confound: any difference is not due to the passive constuction, but differences between the NPs

politician vs **accident**

New hypothesis: Inanimate NPs make passive relative clauses sound more natural than they do active relative clauses.

- The politician that was described by the journalist....
- The accident that was described by the journalist....
- The politician that the journalist described....
- The accident that the journalist described....

2*2 design : animate vs inanimate, passive vs active

New hypothesis: Inanimate NPs make passive relative clauses sound more natural than they do active relative clauses.

- The politician that was described by the journalist....
- The accident that was described by the journalist....
- The politician that the journalist described....
- The accident that the journalist described....

2*2 design : animate vs inanimate, passive vs active

Lists

Most common case: **within subject** manipulations

Who sees what?

	Animate		Inanimate	
item	Active	Passive	Active	Passive
1				
2				
3				
4				
5				
6				
7				
⋮				

Lists

Most common case: **within subject** manipulations

Who sees what?

	Animate		Inanimate	
item	Active	Passive	Active	Passive
1	1			
2		1		
3			1	
4				1
5	1			
6		1		
7			1	
⋮				

Lists

Most common case: **within subject** manipulations

Who sees what?

	Animate		Inanimate	
item	Active	Passive	Active	Passive
1	1	2		
2		1	2	
3			1	2
4	2			1
5	1	2		
6		1	2	
7			1	2
⋮				

Lists

Most common case: **within subject** manipulations

Who sees what?

	Animate		Inanimate	
item	Active	Passive	Active	Passive
1	1	2	3	
2		1	2	3
3	3		1	2
4	2	3		1
5	1	2	3	
6		1	2	3
7	3		1	2
⋮				

Lists

Most common case: **within subject** manipulations

Who sees what?

	Animate		Inanimate	
item	Active	Passive	Active	Passive
1	1	2	3	4
2	4	1	2	3
3	3	4	1	2
4	2	3	4	1
5	1	2	3	4
6	4	1	2	3
7	3	4	1	2
⋮				

Creating a HIT template on Turk

- the contents of your HIT is given by an HTML template
- the template contains *variables* which look like this: `${variable1}`
- a different subset of your items will be substituted in for those variables depending on the list

Back to Turkolizer

Step 1: Prepare an "item file" of your materials

```
# passanim 1 anim_pas
The politician that was described by the journalist appeared in the new
? Did the journalist appear in the news? No

# passanim 1 inan_pas
The accident that was described by the journalist appeared in the news.
? Did the journalist appear in the news? No

# passanim 1 anim_act
The politician that the journalist described appeared in the news.
? Did the journalist appear in the news? No

# passanim 1 inan_act
The accident that the journalist described appeared in the news.
? Did the journalist appear in the news? No
```

Turkolizer lists

Step 2: Run `turkolizer.py` to generate the lists

```
$ python turkolizer.py
hal@kitsune:~/work/turkolizer$ python turkolizer.py
Please enter the name of the text file: itemfile.txt
Please enter the desired number of lists: 4
Please enter the desired number of in-between trials: 1
Please enter the desired number of fillers in the beginning of each
list: 1

Processing the text file...
```

Starting your experiment

Go to the **Publish** tab and choose the template you prepared.

Click on the **upload** button and select the `.csv` file that Turkolizer created.

That's it!

Interactive and multimedia experiments

Again: Anything you can display in a web browser, you can include in a Turk HIT!

- sound
- video
- interactive “games”

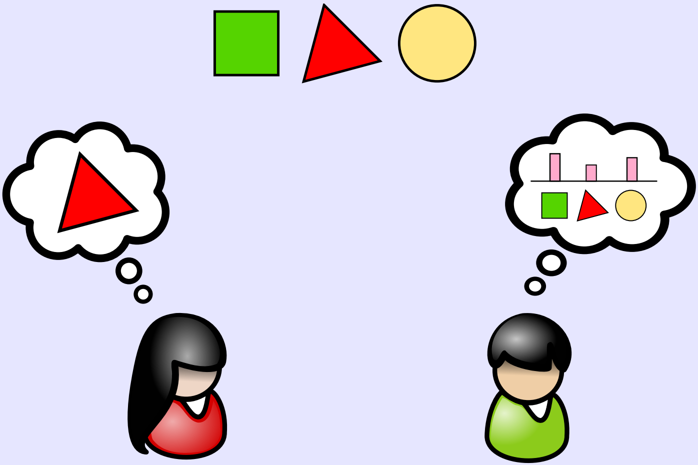
Case study 1: Referring expression choice

Do speakers/writers choice referring expression types which are appropriate to the comprehender's level of uncertainty?

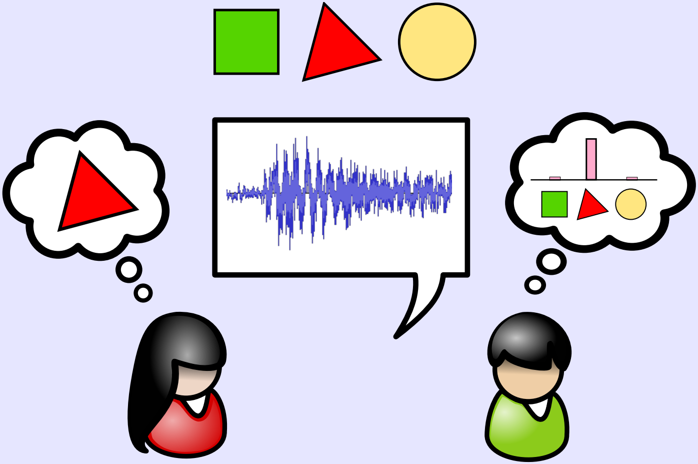
Hypothesis: Referring expression type can be predicted from comprehender uncertainty about an upcoming reference. Longer and more detailed expressions will be used when comprehenders are more uncertain.

(This research described in Tily & Piantadosi 2009)

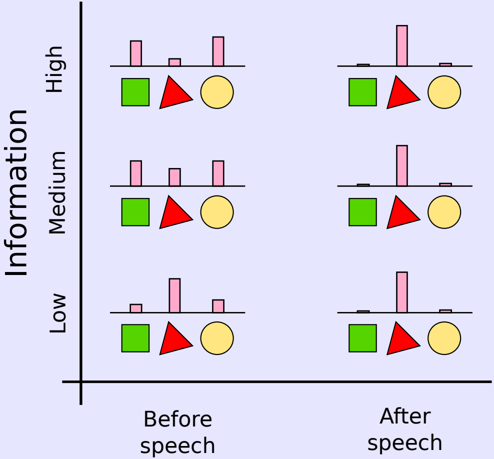
Communication as belief update



Communication as belief update



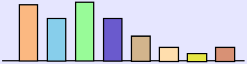
Information in belief update



An actual text

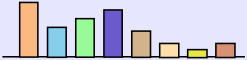
Bob Stone stewed over a letter from his manager putting him on probation for insubordination.

Mr. Stone thought...



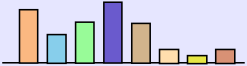
An actual text

Bob Stone stewed over a letter from his manager putting him on probation for insubordination.
Mr. Stone thought the discipline was unfair;...



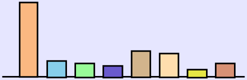
An actual text

Bob Stone stewed over a letter from his manager putting him on probation for insubordination. Mr. Stone thought the discipline was unfair; he believed that...



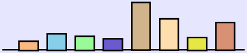
An actual text

Bob Stone stewed over a letter from his manager putting him on probation for insubordination.
Mr. Stone thought the discipline was unfair; he believed that his manager wanted to get rid of...



An actual text

Bob Stone stewed over a letter from his manager putting him on probation for insubordination.
Mr. Stone thought the discipline was unfair; he believed that his manager wanted to get rid of him for...



An actual text

Bob Stone stewed over a letter from his manager putting him on probation for insubordination. Mr. Stone thought the discipline was unfair; he believed that his manager wanted to get rid of him for personal reasons.

Testing predictability of *meaning*

- exactly the task you just saw
- Mechanical Turk participants see text piece by piece
- they guess *coreference* with previous NPs

- we use 82 texts from Wall St Journal
- truncate after 30th NP if longer, yielding 2211 NPs
- 50 participants see each NP in each text
- estimate per NP *surprisal* as $-log \frac{\#correct}{50}$
- only look at *repeated mentions* (25%)

Testing predictability of *meaning*

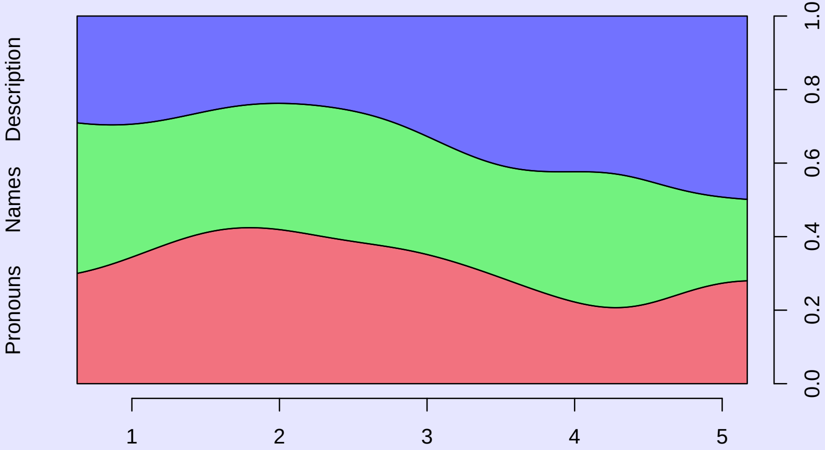
- exactly the task you just saw
- Mechanical Turk participants see text piece by piece
- they guess *coreference* with previous NPs

- we use 82 texts from Wall St Journal
- truncate after 30th NP if longer, yielding 2211 NPs
- 50 participants see each NP in each text
- estimate per NP *surprisal* as $-log \frac{\#correct}{50}$
- only look at *repeated mentions* (25%)

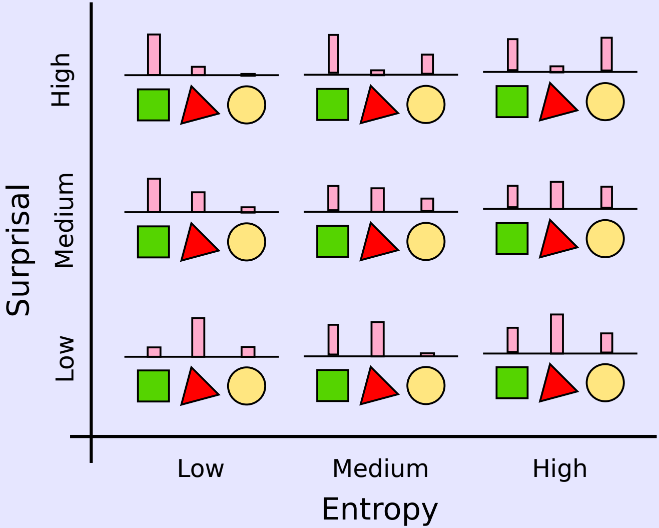
Expression choice

Pronouns	Names	Descriptions
him	Bob Stone	a letter
he	Mr. Stone	his manager
it	the U.S.A.	the discipline
theirs	Kobe Steel Ltd.	hot-dipped galvanized steel products

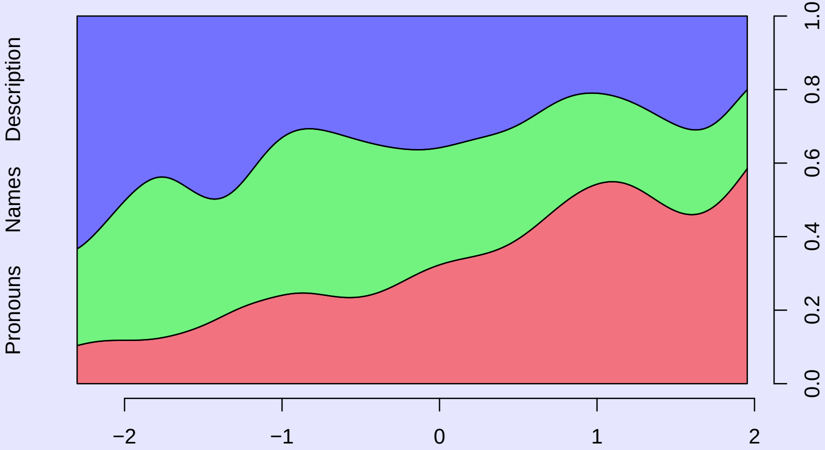
Expression type as a function of surprisal



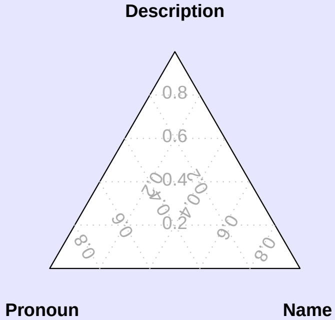
Surprisal vs entropy



Expression type as a function of (residual) entropy

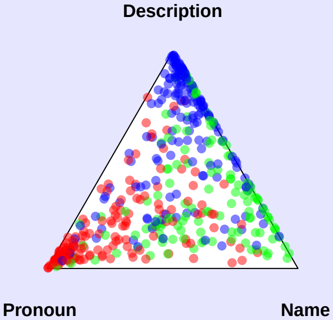


Model



Fitted a (multinomial) regression model to predict expression type as a function of multiple factors (e.g. last mention distance, grammatical function, number of referents in discourse) including **surprisal** and **entropy** over comprehenders' guesses.

Model



Classification accuracy: .68

	Pronouns	Names	Descriptions
Precision:	.72	.59	.69
Recall:	.73	.51	.76

Conclusions

- people tend to use pronouns or names when the referent is not *surprising* (i.e., when comprehenders are good at guessing)
- when the referent is more surprising, people use descriptions more when there are competing referents and pronouns when there are not

This kind of research is only possible with huge samples...
Mechanical Turk allowed us to collect 100,000 judgements

Case study 2: Artificial grammmar

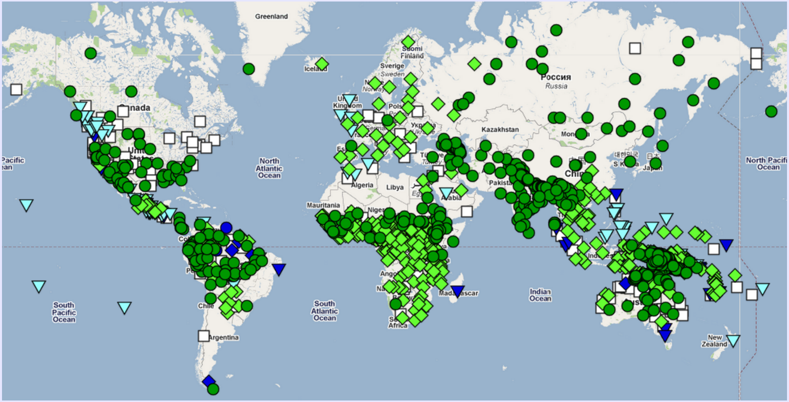
Certain types of language may be harder to learn or use than others

It isn't always possible to get data or speakers for languages which are theorized to be hard to use, because they do not exist!

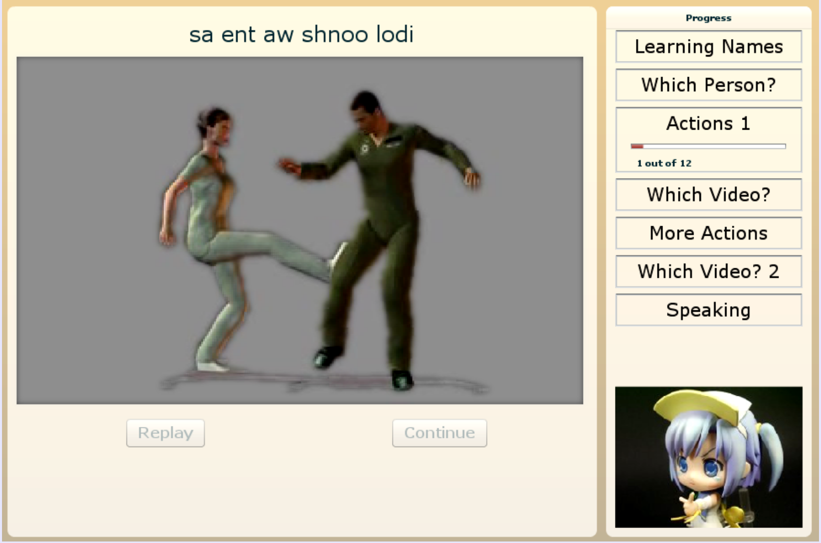
By carefully constructing and testing *artificial* languages, we can determine what properties of language people find easy to learn or use

(see e.g. Hudson Kam & Newport 2005; i.a.)

Basic word order typology



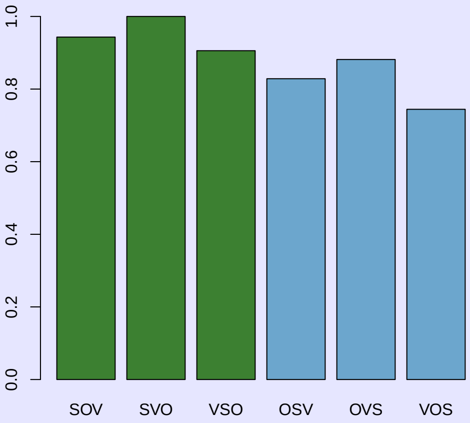
SOV	SVO	VSO	VOS	OVS	OSV	
●	◆	▽	●	◆	▼	
45	42	9	3	1	0	(%)



sa ent aw shnoo lodi

Word order correctness

SO languages are easier to learn than OS



(Tily, Frank & Jaeger submitted)

Conclusions

Mechanical Turk makes simple survey-like experiments **easy**

And it makes many new types of research **possible**

Freely available tools will help you turn a research hypothesis into an experiment that can be run on Turk quickly and easily (Gibson, Piantadosi & Fedorenko to appear)