6th Annual Meeting, Swedish e-Science Research Centre

Lexical innovation, word games, and the Internet

Janet B. Pierrehumbert

Oxford e-Research Centre, University of Oxford
Department of Linguistics, Stockholm University
Department of Linguistics, Northwestern University

New Zealand Institute of Language Brain and Behaviour, University of Canterbury

WORDOVATORS

Home | About | People | Publications | Jobs | Games

Welcome to Wordovators

How People Make New Words

Play Games

The Wordovators team. Front row (from left): Christoph Bartneck (NZILBB), Janet B. Pierrehumbert (Northwestern), Jen Hay (NZILBB), Stephanie Stokes (NZILBB). Back row (from left): Chun-Liang Chan (Northwestern), Kayo Takasugi (Visual Voice), Aristotle, Kiwi PaPeRo. Photo credit: University of Canterbury

Announcements

- Postdoctoral Fellow: Northwestern University, Evanston IL

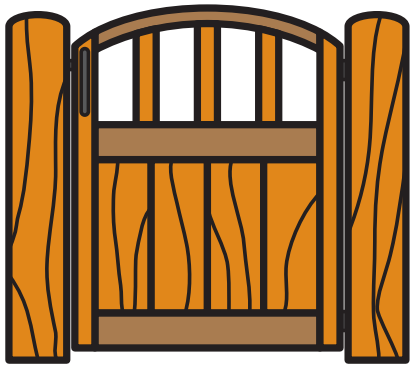
News

- Northwestern press release
- Wordovators featured in Northwestern Annual Research Report 2012
- Kick Off meeting

Sponsors

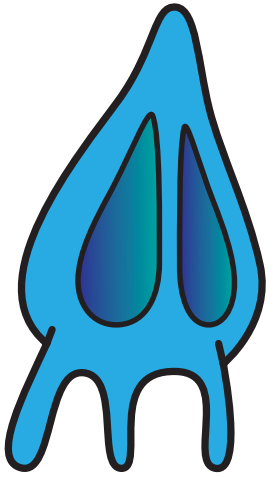
JOHN TEMPLETON FOUNDATION

fox
azeria
kettu
zorro



gate
atea
portti
puerta

mushroom
onddo
sieni
seta



?

How many words have you learned?

- Total vocabulary size of educated English-speaking adults: about 100,000.
- That's including complex words like *hotdog*, *blindfold* and *systems programmer*.
- Languages like Finnish have many more words:

elämä	n	tapa	muutoks	i	lla
life	of	style	change	-s	with

- Rare and novel words are a major challenge for speech and language technology.

Zipf's Law (rank-frequency distribution of words $\approx$ a power law).

A few words are frequent. Many words are rare. The lower the frequency, the more words there are. Let's think about continuing these lists:

1/1000	1/10000	1/100000	1/1000000	1/10000000
should	right	delicious	graduate	swampland
than	move	weird	goldfish	thunk
only	hard	understanding	encyclopedia	escapologist
people	sat	light	carnation	zirconium
also	easily	duck	thrifty	sitka
me	summer	propaganda	transcendence	trangia

(Frequency data from British National Corpus.)

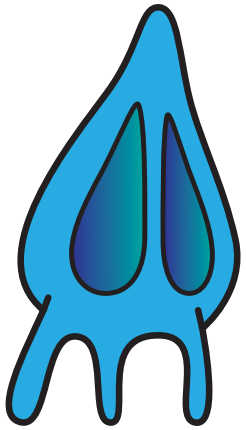
Loss of words.

- What if we just used words at random? A rare word could have a run of bad luck; if nobody uses it, nobody learns it.
- Is that what happened to *snivvy*?
- Also, people change what they talk about.
- How much *stining* do you and your friends do?

The motivation for the Wordovators project

- People know a huge number of words.
- Words are continually lost.
- But people continue to know a huge number of words. People who speak the same language know lots of the same words.
- So new words must be created all the time.
- Lexical diversity seems like biodiversity. New word = new species. Success of new word = success of a new species.

New words must have a sound structure that is allowable in the language



What's this?

Some English candidates	Some Welsh candidates
cratict froure grocid reptagin	hrondd oethyn hwynol acynni
Source: Shannon (1948) 3-letter sequence model	Source: A. Schumacher 2-letter sequence model trained on the CEG Welsh corpus.

But this is not the whole story

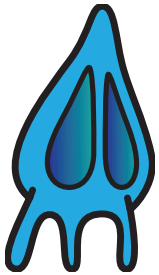
- Made up: Dasani, Swiffer (by Lexicon Branding); Francelle, Zeshawn (African-American).
- Existed in other languages, modified to fit English: aroha, cassata, kumara, jihad.
- MORPHOLOGY (broadly construed): linguicide, manvacation, turkeypalooza, brekkie, nerdify, Alka-Seltzerize.

← Do you recognize the source of this neologism?

Morphology: Theory of how words are made from meaningful parts

- You store in your mind simple forms *(cat)*, complex forms *(kindness)*, compounds *(catfish)*.
- Structure is gradient. It depends on:
  - The sequences of sounds in the word. *Fifth Third Bank* can definitely be divided because of the the ”ftht h” sequence.
  - Semantic transparency. *hotdog* can sort of be divided because it’s hot, but it’s not really a dog.
  - Word frequency. Frequent words are easily seen inside less frequent words. *worlds* can definitely be divided because *world* is 58 times as frequent as *worlds*. *stairs* is less dividable because it is 10 times as frequent as *stair*.
  - Subparts of stored words can be recombined to make new words. The more often a subpart occurs, the more likely to be recombined.

Convergence on new words: The Naming Game



What's this?

1: I invent a word and say it.

Me	You
bluekin	

2: You add the word to your lexicon.

A: You had no word before.

Me	You
bluekin	bluekin

B: You had a word. Now you have two words:

Me	You
bluekin	poddic bluekin

3: Now you talk to someone else. You make a random choice from your words.

You	Him
poddic bluekin	seedle

A: Your word is new to him. He adds it.

You	Him
poddic bluekin	poddic seedle

B: He knows your word. Both of you eliminate the other words.

You	Him
poddic ~~bluekin~~	poddic ~~seedle~~

You two have agreed! It's a poddic!
(But maybe I still think it's a bluekin.)

4: Now everyone talks to another person.

Simulation results (Steels, Baronchelli, many others)

- Eventually, people agree on a name for the thing.
- Once they've agreed, the name doesn't change.
- Highly idealized. Some issues:
  - How fast is "eventually"? Have real human languages had enough time to converge or not?
  - Can't people have multiple names for the same thing, used in different contexts?

A case of incomplete convergence

- Prof Kerry Emanuel (MIT) versus Dr. Rasmus Benestad (Norwegian Met Office)
- Members of the same scientific community.
- Both posting on-line about extreme weather.
- Specifically, about whether there is a link between climate change and extreme weather.

Kerry Emmanuel

Among the more consequential effects of global climate change is a possible change in tropical cyclone activity. We are most concerned with three aspects of hurricane activity: their frequency, their intensity, and their geographical distribution. Any change in the frequency with which hurricanes strike populated land is of obvious concern ...

Some unusual words: cyclogenesis, oft-stated, refusards, paleotempestology.

Rasmus Benestad

The Atlantic hurricane season will soon be upon us again , and no doubt many people will recall last years devastating Hurricanes that swept across Florida. There was a great deal of press about these storms, as 3 major hurricanes and 5 tropical storms made landfall in the US ...

Some unusual words: SSTs, r-script, iid-rule, pole-equator, dewpoint.

Who is Dr. X?

Who would think that Internet, ideas, disease, money, birds, and climate literacy have anything in common? Recent progress on complex systems and network theory suggests that they all can be described in terms of a Levy flight....

Some unusual words: fsm-climate-protagonists, SPPI, saturn-tide, himalayanglaciermeltrategate.

Comparing word-formation strategies

Novel forms per thousand words of text.

	Emanuel	**Benestad**	**Dr X.**
compounds	4.2	8.0	8.0
acronyms	0.4	2.1	2.1

... Dr. X is Rasmus Benestad.

What about words in general? A study of language on-line

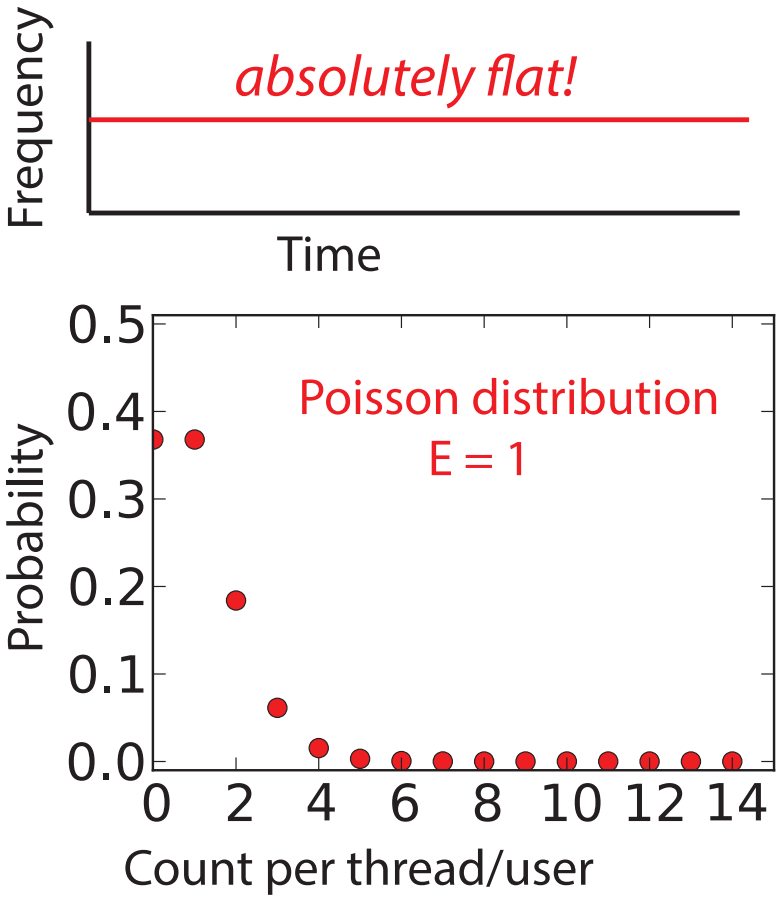
Altmann, Pierrehumbert & Motter (2011) "Niche as a determinant of word fate"

- Mathematical analysis of dynamics of words in USENET discussion groups comp.os.linux.misc, and rec.music.hiphop.
- Comp.os.linux.misc: 1993-2008. 128,903 users and 140,517 threads.
- Rec.music.hiphop: 1995-2008. 37,779 users and 94,074 threads.
- How are words distributed across users and across threads (topics)?
- Define D^U (Dissemination over Users) and D^T (Dissemination over Threads) as follows:

Baseline: the Naive Bag of Words Model

Each word has a frequency. It is the same for everybody.
Word sequences are made by selecting words at random.
The selection is frequency-weighted.

Predictions for individual words:



Baseline: the amount of bunching expected from this Poisson process.

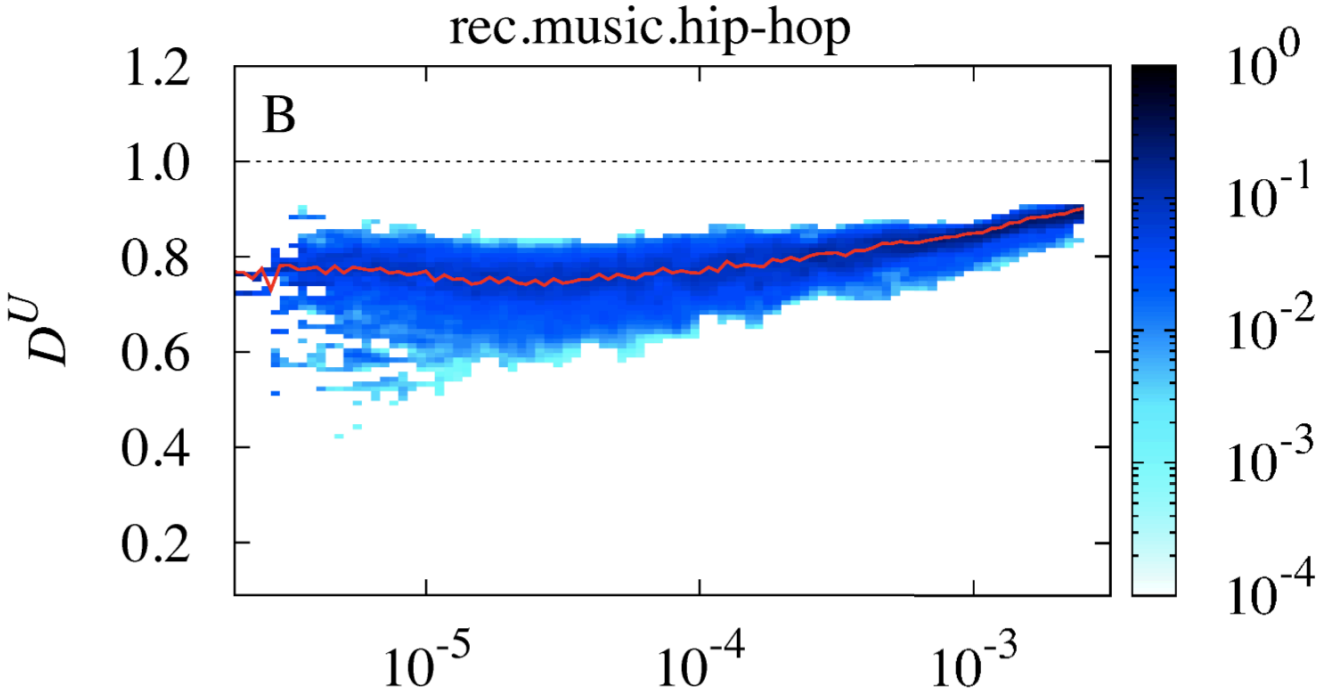
DT: How much is the word bunched up in threads?

DU: How much is the word bunched up in posts by users?

DU = 1: Just as many people use the word as you would expect.

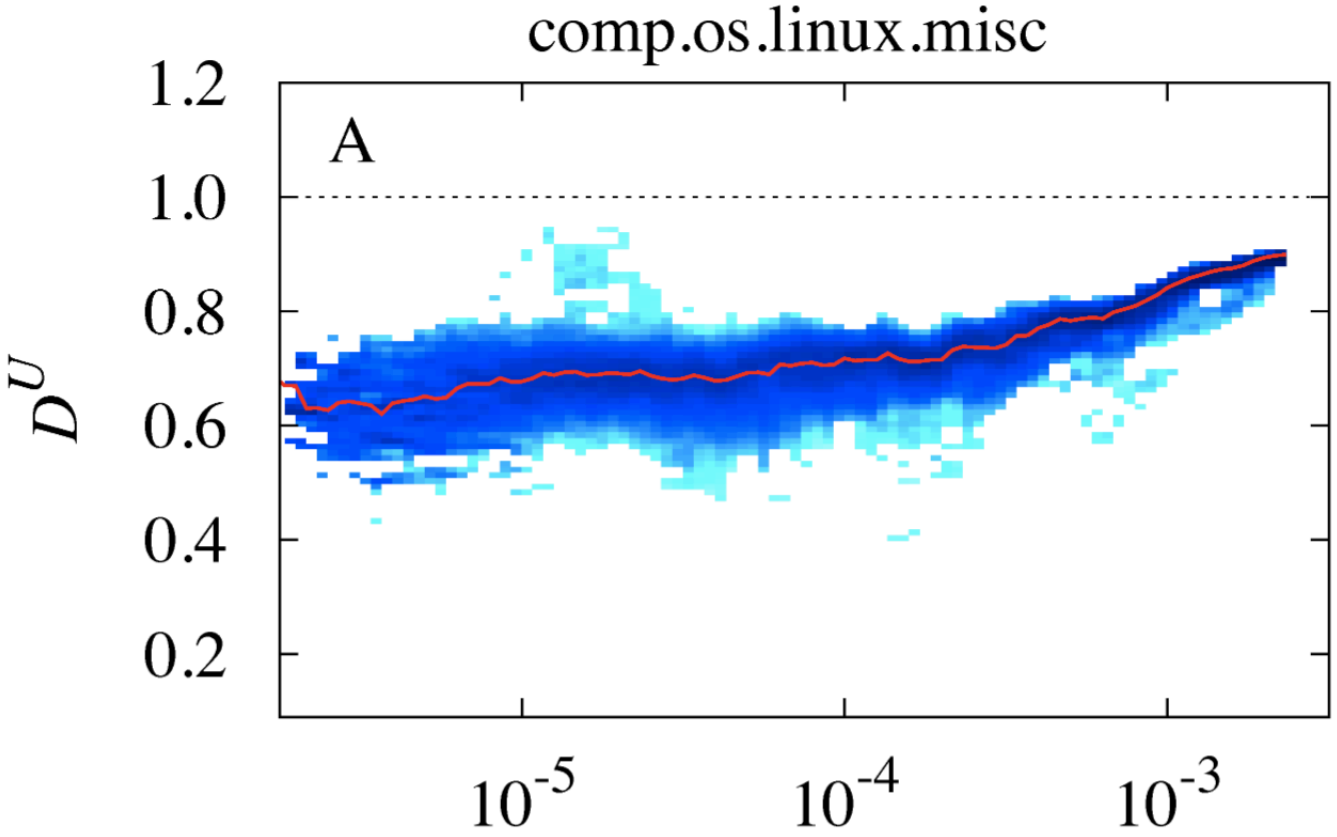
DU << 1: Very few people use the word, compared to what you would expect. (But those people use it very often).

Almost all words are significantly concentrated (by people)



Blue: Density of D^U for words of each given frequency. Red: Median D^U.

... even more in comp.os.linux



More analysis: Different people use different words to talk about the same topic.

Crowdsourced experiments using on-line computer games to explore social associations of words

Roof-jumping game: Guess the right diminutive form for each word.

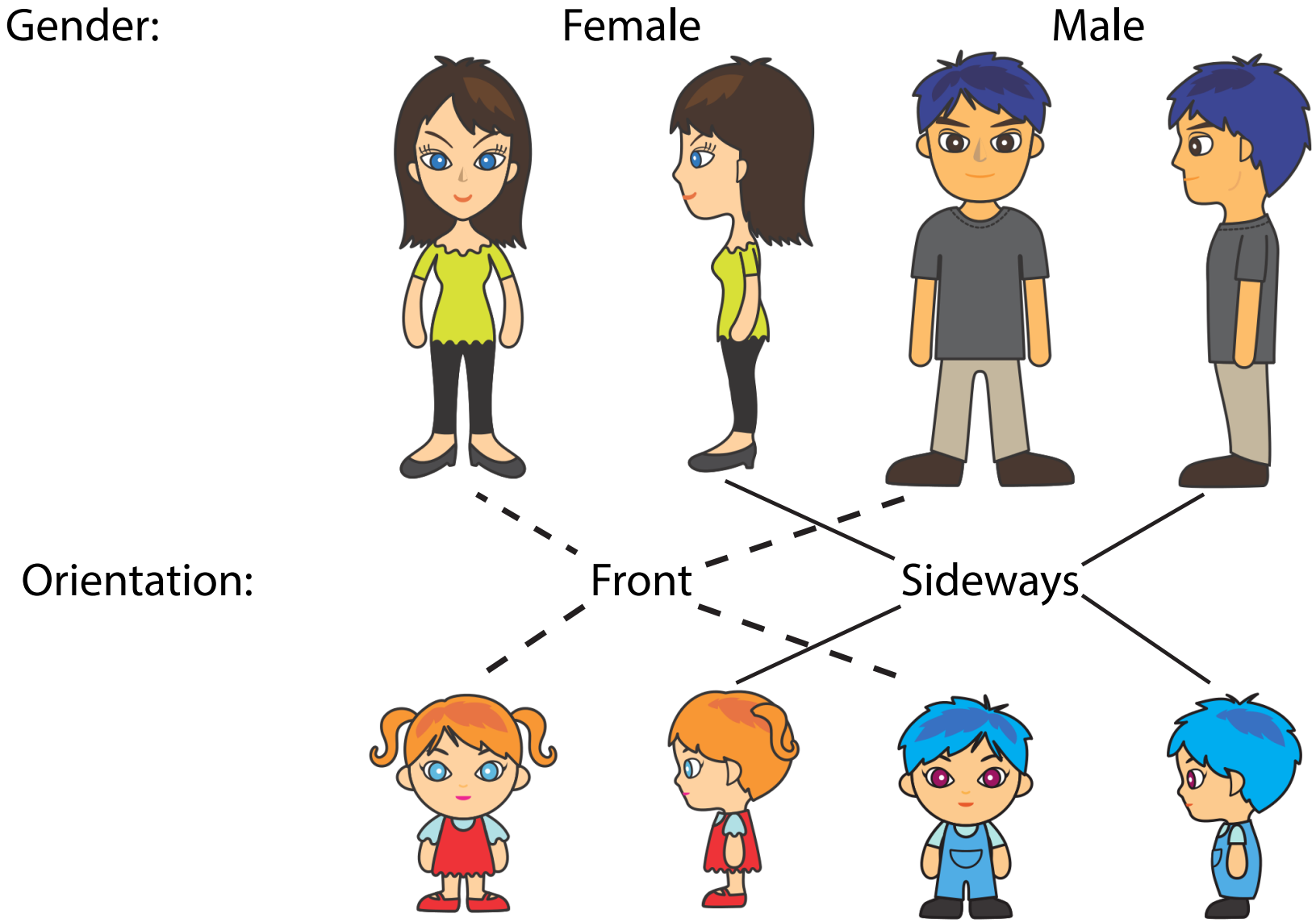
(Racz, Hay and Pierrehumbert, under review).

Game play

- The correct answer depends on some characteristic of the interlocutor.
- E.g, you are supposed to use different words to talk to different people.
- If you guess the right word, you jump forward to the next roof.
- If you guess wrong, the interlocutor pushes you off your roof, you have to flutter up again.
- Test phase includes previously scene and novel items, no feedback.

Test phase is in the dark

Do people generalize to other, similar interlocutors?



Characters for age, ethnicity

Some results

- People learn the different words quite well.
- Nobody generalizes on the basis of the orientation of the speaker.
- Many people are able to generalize by gender, age, or ethnicity.
- People can learn and use different vocabularies for socially-relevant categories of people.

Various experiments in progress

- Social associations for word-formation patterns.
- The player gets a benefit from using the same language as other people.
- The player gets a benefit from using different language from other people (status marking, deception, secrets).
- Influencing player's choices with subtle cues about the social context.

Conclusions

- People have distinctive vocabularies.
- They make and hear new words all the time.
- Recombining meaningful subparts of other words is the most common way to make new ones.
- Learning, remembering and using words involves social factors.
- The Internet helps us explore lexical structure.
  - Very large samples of spontaneous language.
  - Large numbers of people can be recruited to play on-line games.

Thank you.