Algorithms for Computational Social Science

Animesh Mukherjee,
Dept. of Computer Sc. and Engg.,
Indian Institute of Technology, Kharagpur

Computational Social Science

- New understanding about the properties that **systems of people and computers together possess**

- **Significance in Engineering:**

  - Develop new technology that leverages on the ever-growing **coupling of the users and the World Wide Web**.

  - **Socially intelligent** computing systems – online games and auctions

  - Employ the **"wisdom of the crowd"** to find precise quantitative answers to fundamental questions in cognitive science

  - Apply the quantitative knowledge back to develop smarter technological systems.

Research Contributions

- **Modeling color categorization**
  - **Origins of the color hierarchy:** PNAS;
    Media Coverage: PNAS Press Highlights, BBC Future, Prospect Magazine, UK, Scientific American, Live Science, MedicalXpress, The Hindu, Business Line, Times of News, Yahoo! News, Deccan Herald, MSN, Wikipedia on "Color Terms"
  - **Metastability and aging in language dynamics:** PLoS One, Journal of Computational Science, Elsevier, Journal of Statistical Mechanics: Theory & Expt.
  - **Empirical analysis of the basic color names:** Advances in Complex Systems
- **Assistive technology development**
  - **Adaptive and predictive virtual peripherals for people with neuro-motor disorders:** Assistive Technology Journal, RESNA, Young Scientist Award from Indian Science Congress Association
- **Delay-tolerant networks**
  - **Message Dissemination:** ACM Mobiopp
- **Self-organization of human speech sound inventories**
  - **Sound systems as bipartite networks:** RRE, EPL, Physica A, Jour. of Quantitative Linguistics, Coling 2008, Coling-ACL 2006
  - **Patterns of co-occurrence of sounds:** Int. Jour. of Mod. Physics C, Advances in Complex Systems, EACL 2009
  - **Fixed redundancy of sound systems:** ACL 2007, EACL 2009, Jour. of Quantitative Linguistics.
- **Natural language processing**
  - **Unsupervised methods for Indian languages:** Coling 2010, ACL 2009, LREC 2008,
  - **Noisy text analytics:** IJDAR Springer
- **Image processing**
  - **Saliency detection:** IET Computer Vision.

Modeling Color Categorization

Language dynamics:

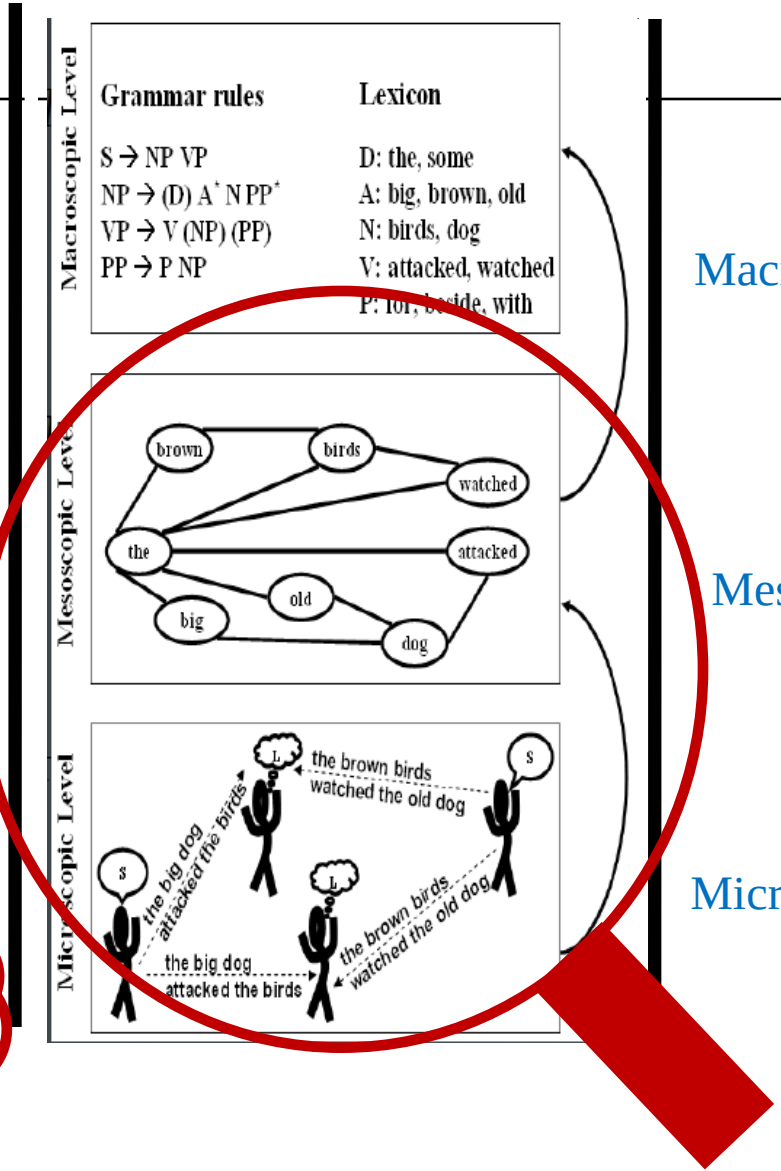
A Physical System Perspective

Language as a whole (grammatical constructs)

Language as a set of interactions among linguistic units

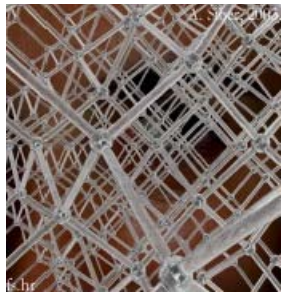
Language as a set of utterances

dynamic

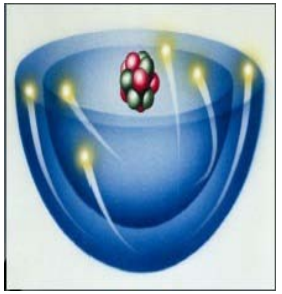


Macroscopic level

Mesoscopic level



Microscopic level



Names for meanings

SPAM !

Names for meanings

SPAM !

"Spiced Ham"

Monty Python's spam comedy (1970 TV show)

Mr. and Mrs. Bun enter a cheap pub

Mr. Bun: What have you got, then?

Waitress: egg and SPAM; egg, bacon, and SPAM; SPAM, bacon, sausage, and SPAM; SPAM, egg, SPAM, SPAM, bacon, and SPAM; SPAM, egg and SPAM; baked beans, SPAM and SPAM…

Mrs. Bun : Have you got anything without SPAM in it?

Waitress: Well, there's SPAM, egg, sausage, and SPAM. That's not got MUCH SPAM in it.

Mrs. Bun: I don't want any SPAM!

Mr. Bun: Why can't she have egg, bacon,
SPAM, and sausage?

Mrs. Bun: That's got SPAM in it!

Mr. Bun: Not as much as SPAM, egg,
sausage, and SPAM.

Mrs. Bun: Look, could I have egg, bacon,
SPAM, and sausage without the SPAM?

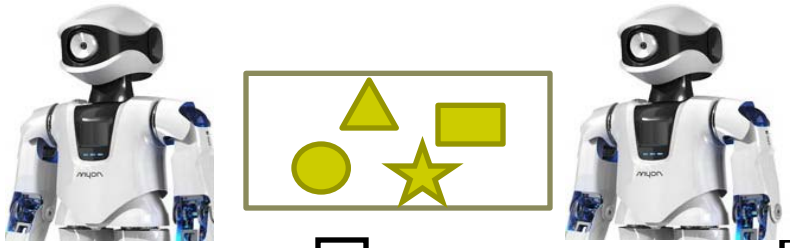
Waitress: Uuuuuuuuuggggggh!

Mrs. Bun: What d'you mean uuugggh!? I don't like SPAM.

Vikings: (singing) SPAM, SPAM, SPAM, SPAM..SPAM, SPAM, SPAM, SPAM... Lovely SPAM,wonderful SPAM....

The Naming Game

Speaker **Hearer**



- Perceive scene
- Choose topic
- Conceptualize
- Verbalize

- Interpret utterance
- Perceive scene
- Apply meaning
- Point to referent

In silico abstraction

- Interaction of **N** artificial agents communicate to agree on the **name** of an object in the environment.
- Each interaction counts as a time step of the dynamics
- Agents can keep in their memory an unlimited number of words
- Games proceed through a series of **success** and **failure** interactions
- One studies the phenomenological properties – total number of words in the system $N_w(t)$, number of unique words $N_d(t)$, maximum memory N_w^{max}, time to reach agreement t_{conv} etc.

Speaker	Hearer
Bottle Apple Tiger Car	Bag Berry Tree

Failure

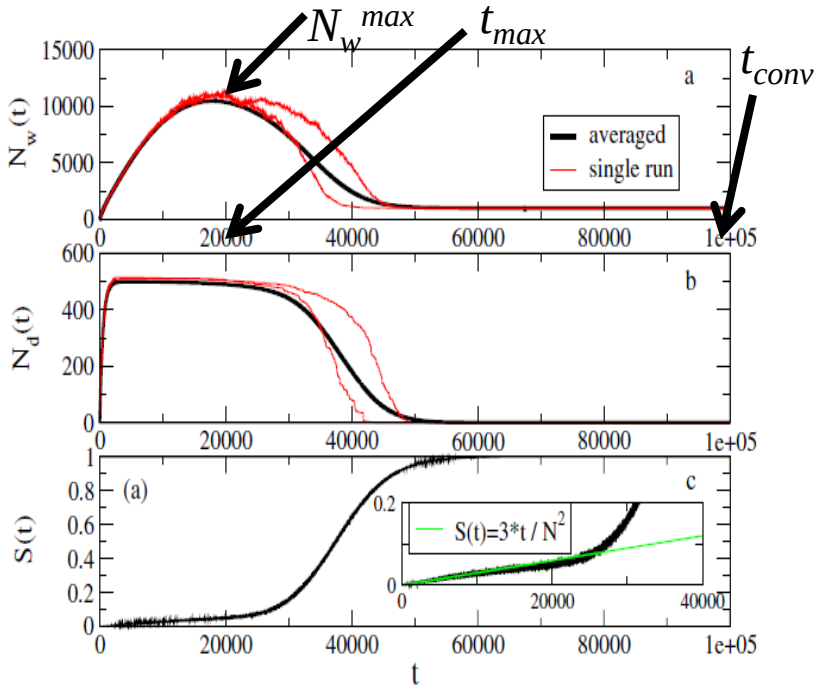
Speaker	Hearer
Bottle Apple Tiger Car	Bag Berry Tree Apple

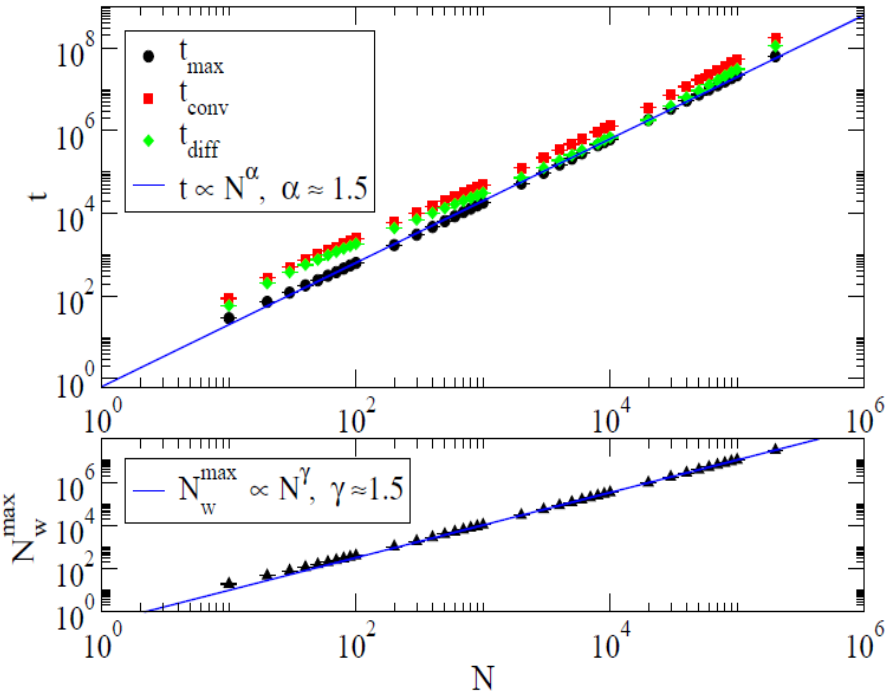
Speaker	Hearer
Bottle Apple Tiger Car	Bag Berry Tree Apple

Success

Speaker	Hearer
Apple	Apple

Phenmenology





$$\frac{dN_w(t)}{dt} \propto \frac{1}{cN^a}\left(1-\frac{2cN^a}{N}\right) - \frac{1}{cN^a}\frac{2cN^a}{N}2cN^a$$

Failure **Success**

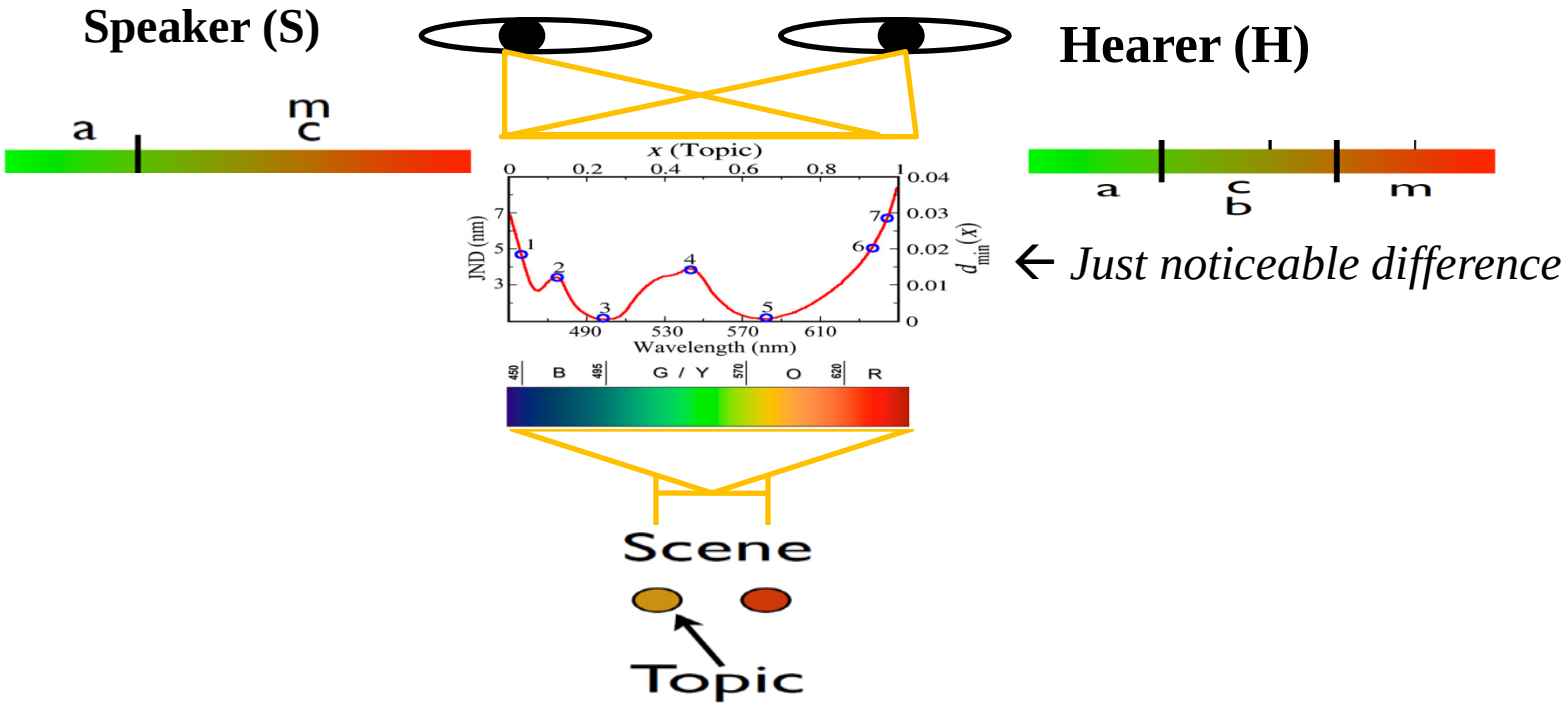
$$\frac{dN_w(t)}{dt} \propto \frac{1}{cN^{1/2}}\left(1-\frac{ct}{N^2}\right) - \frac{1}{cN^{1/2}}\frac{ct}{N^2}2cN^{1/2}$$

Failure **Success**

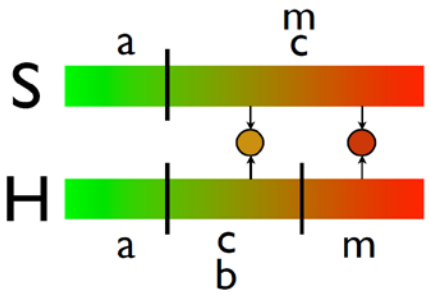
$N_w^{max} \sim N^{3/2}$ $t_{max} \sim N^{3/2}$ $t_{conv} \sim N^{3/2}$

From Naming to Color Naming

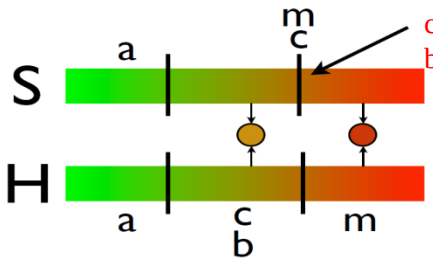
- Color categorization: a central issue both in linguistics and in cognitive science
- Evolution of English color categories [English color terms → gradual semantic shift from largely brightness color concepts (Old English) to almost exclusively hue concepts (Middle English)]



The Category Game

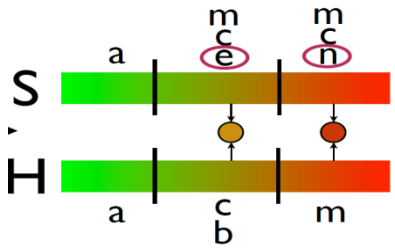


Locate the Stimuli



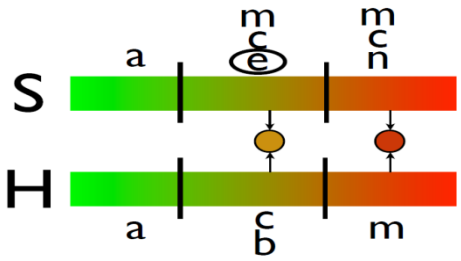
two stimuli colliding on the same perceptual category → a new boundary is created

Discrimination

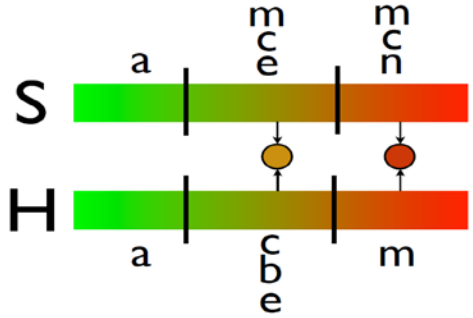


Invention of new words

Failure

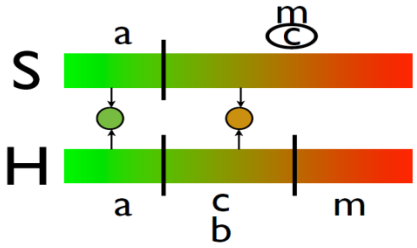


Utterance

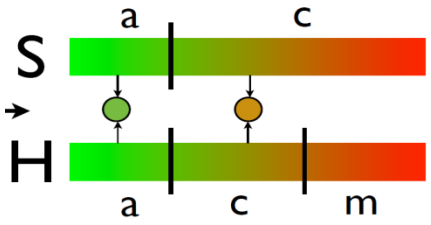


Outcome: Failure

Success

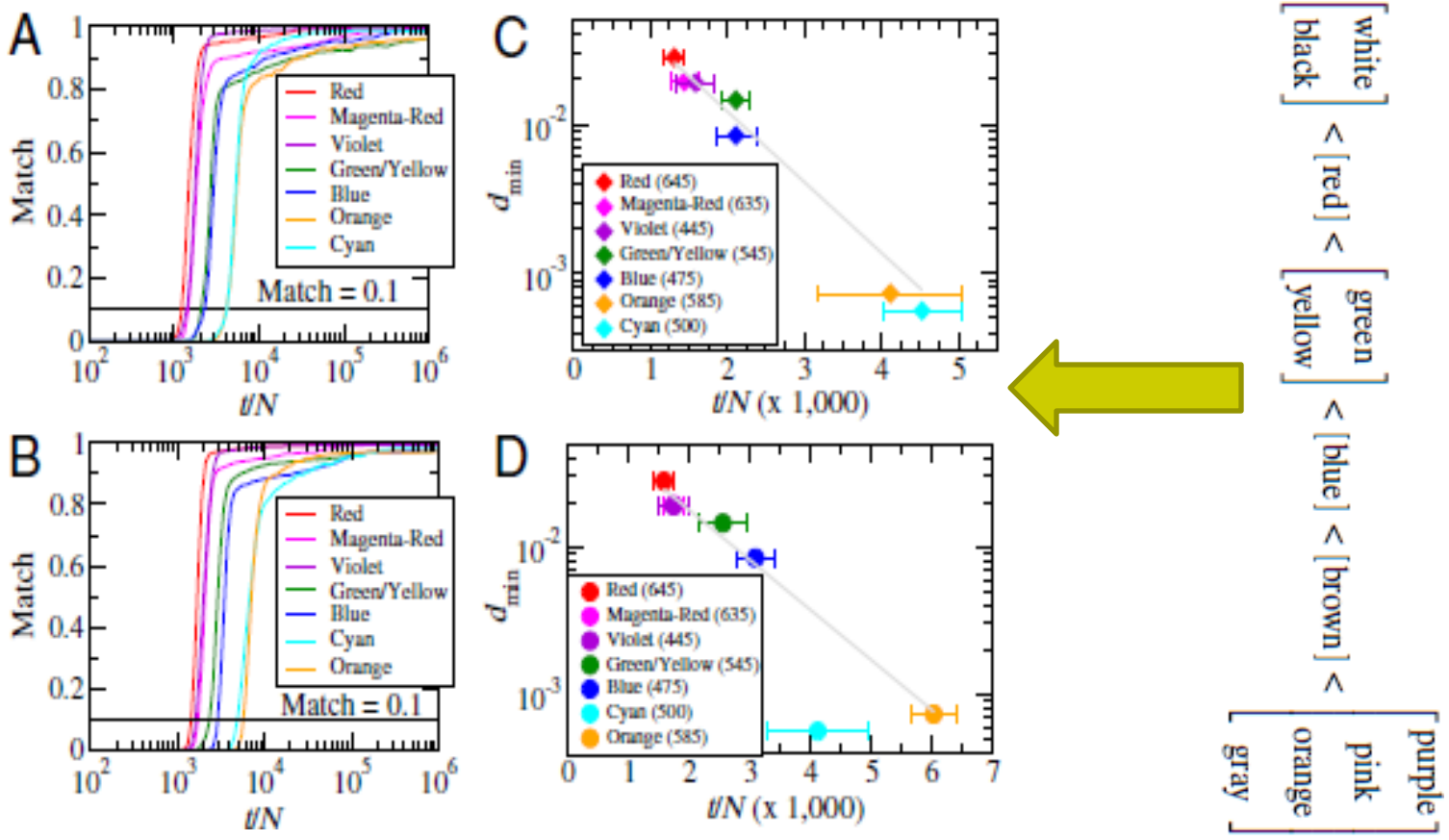


Utterance



Outcome: Success

The Color Hierarchy

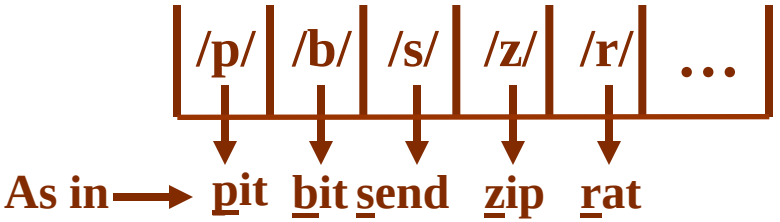


Loreto, Mukherjee and Tria, On the origin of the hierarchy of color names, *PNAS* , May 1, 2012 vol. **109** no. 18 6819-6824

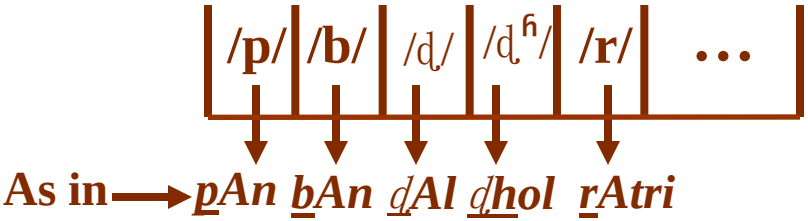
Self-organization of Sound Inventories

Network of Sounds

- A repertoire of unique sounds (aka *phonemes*) that the speakers of a language use for communication

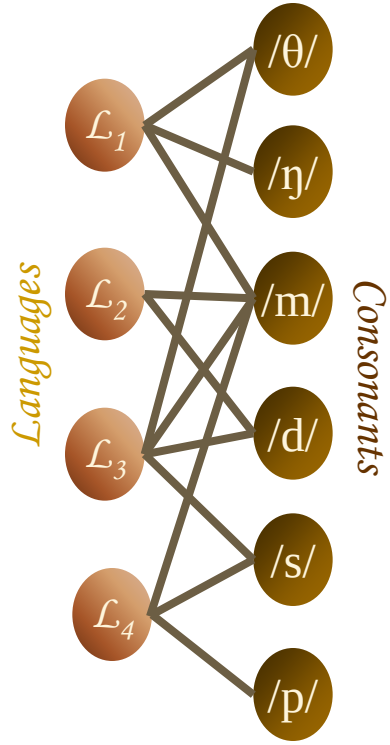


English Consonants



Bangla Consonants

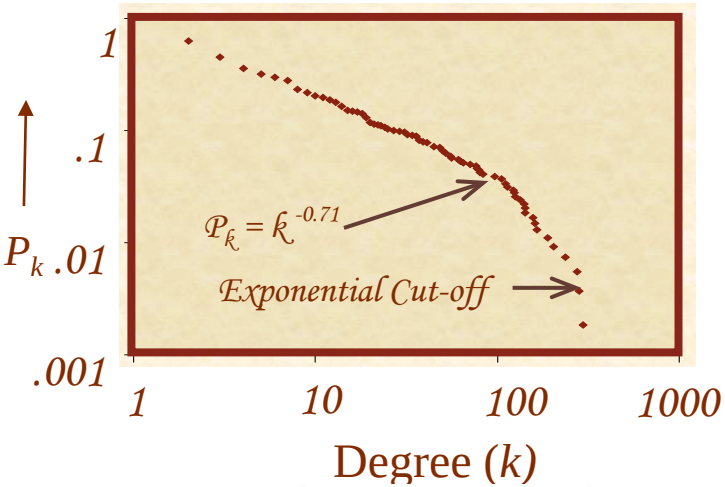
- Bipartite network representation: Phoneme-Language Network (PlaNet)
- V_L (set of nodes in the language partition)
- V_C (set of nodes in the consonant partition)
- Edge $e \in E$ between $v_l \in V_L$ and $v_c \in V_C$ iff the consonant c in present in the language
- Source: UCLA Phonological Segment Inventory Database (UPSID) → 317 languages, 541 unique consonants

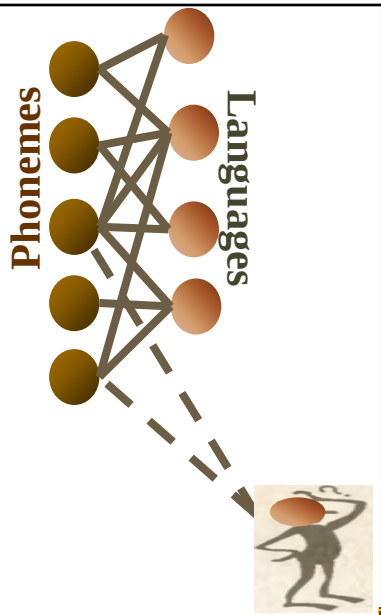


PlaNet

Occurrence Principles of Consonant Sounds

DD of the consonant nodes follows a **power-law with an exponential cut-off**





Rules of the game:

- A new language is born
- Chooses μ distinct phonemes from the set of existing phonemes preferentially based on the degree

$$\frac{\gamma k + 1}{\sum (\gamma k + 1)}$$

all phonemes not already chosen

$$p_{k,t+1} = (1 - \tilde{P}(k,t))p_{k,t} + \tilde{P}(k-1,t)p_{k-1,t}$$

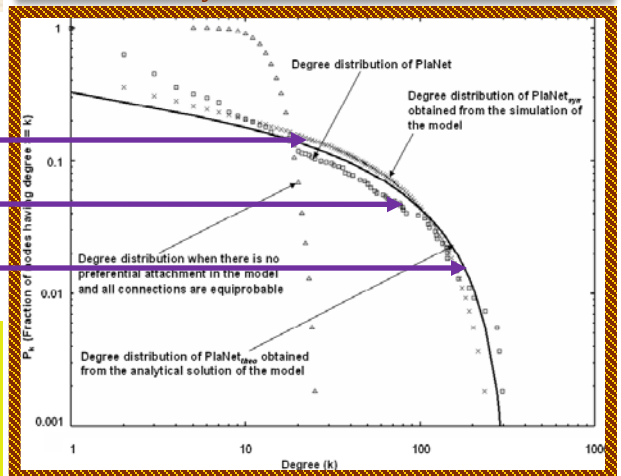
$$\tilde{P}(k,t) = \begin{cases} \frac{\gamma k+1}{\gamma t+N} & \text{for} \quad 0 \le k \le t \\ 0 & \text{otherwise} \end{cases}$$

$$p_{k,t} = \binom{t}{k} \frac{\prod_{i=0}^{k-1}(\gamma i + 1)\prod_{j=0}^{t-1-k}(N - 1 + \gamma j)}{\prod_{m=0}^{t-1}(\gamma m + N)}$$

Simulation

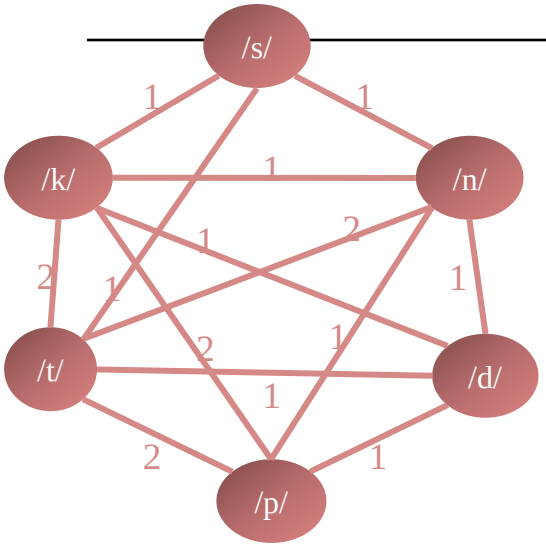
Data

Analytical



Peruani, Choudhury, Mukherjee, Ganguly, Emergence of a non-scaling degree distribution in bipartite networks: a numerical and analytical study, *Euro. Phys. Letters*, **79**(2), 28001

Co-occurrence Principles of Consonant Sounds



Community structure analysis of PhoNet
Calculate strength *S* of each edge

$$S = \frac{w_{uv}}{\sqrt{\sum_{i \in V_c-\{u,v\}}(w_{ui} - w_{vi})^2}} \text{ if } \sqrt{\sum_{i \in V_c-\{u,v\}}(w_{ui} - w_{vi})^2} > 0 \text{ else } S = \infty$$

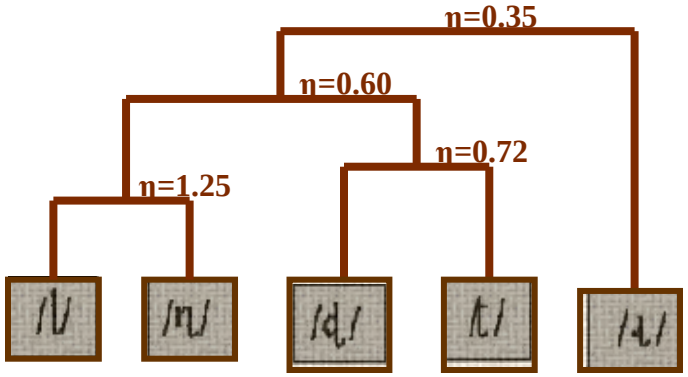
Remove edges with *S* less than a threshold η

PhoNet: Phoneme-Phoneme Network

plosive	*voiced*	*voiceless*
bilabial	/b/	/p/
dental	/d/	/t/

If a language has in inventory

then it will also tend to have



Mukherjee, Choudhury, Basu, Ganguly (2007) Modeling the Co-occurrence Principles of the Consonant Inventories: A Complex Network Approach, *Int. Jour. of Modern Phy. C*, **18**(2), 281—295

Mukherjee, Choudhury, Basu, Ganguly (2008) Rediscovering the Co-occurrence Principles of the Vowel Inventories: A Complex Network Approach, *Advances in Complex Systems*, **11**(3), 371--392

Feature Entropy

p_f – number of consonants in a community (C) in which feature f is present

q_f – number of consonants in C in which feature f is absent

The probability that a consonant chosen at random form C has f is $\frac{p_f}{N}$ and that is does not have f is $\frac{q_f}{N}\left(1-\frac{p_f}{N}\right)$

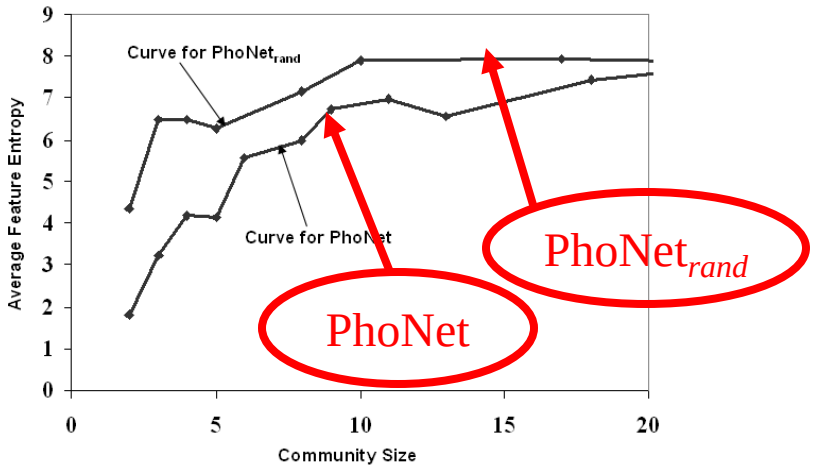
If F denote the set of all features,

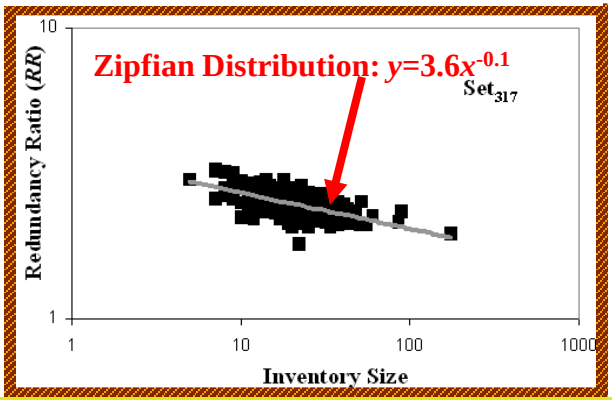
$$F_E = -\sum_{f \epsilon F} \frac{p_f}{N} log_2 \frac{p_f}{N} + \frac{q_f}{N} log_2 \frac{q_f}{N}$$

F_E → Total discriminative capacity of the features in an inventory

Redundancy Ratio: $RR = \frac{F_E}{log_2 N}$

RR → ratio of excess bits required to represent an inventory



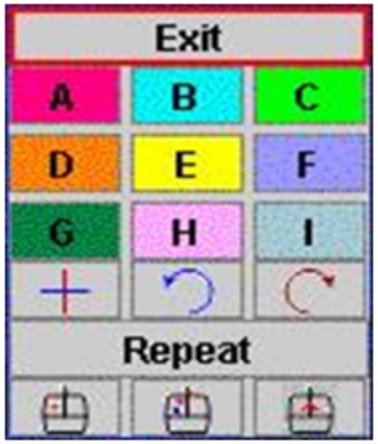


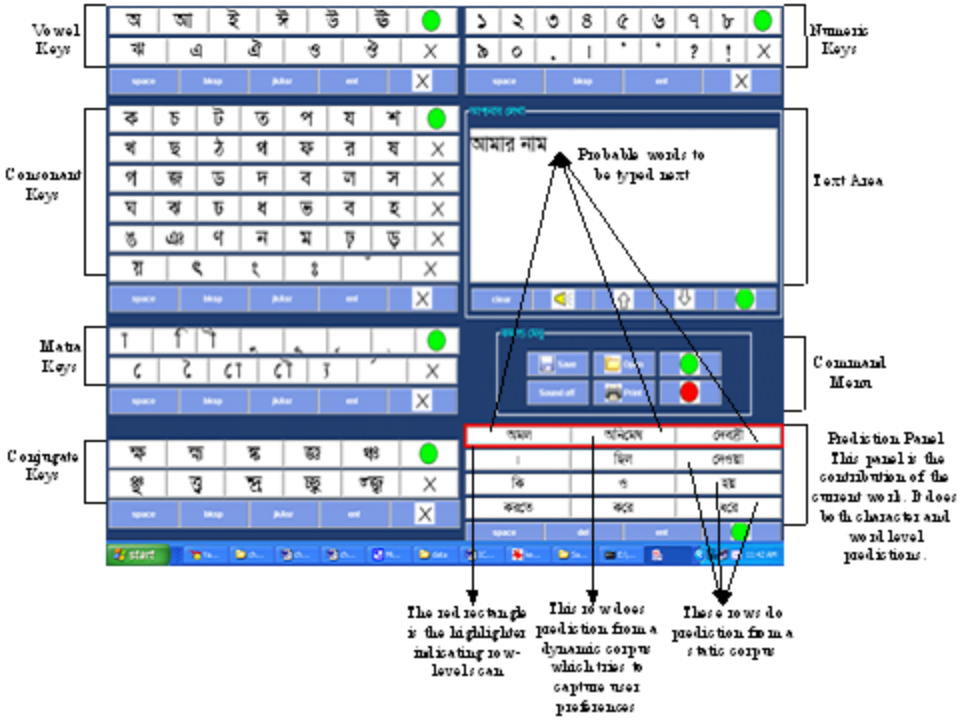
Mukherjee, Choudhury, Basu, Ganguly (2007) Modeling the Co-occurrence Principles of the Consonant Inventories: A Complex Network Approach, *Int. Jour. of Modern Phy. C*, **18**(2), 281—295

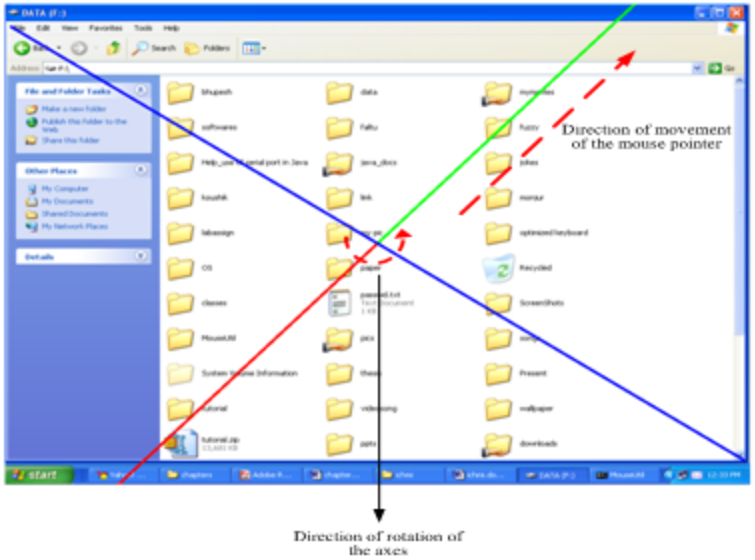
Mukherjee, Choudhury, Basu, Ganguly (2007) Redundancy Ratio: An Invariant Property of the Consonant Inventories of the World's Languages, *ACL, 104*—111, Prague, Czech Republic

Assistive Technology Development

Adaptive Virtual Peripherals







Courtesy IICP, Kolkata

Mukherjee, Chakraborty and Basu (2008), Sweep Sticks - An Adaptive Virtual Mouse for People with Neuro-Motor Disorders, *Assistive Technology Journal of the Rehabilitation Engineering Society of North America*, **20**(2), 111--124

Thank You