Cultural Route to the Emergence of Linguistic Categories

- Motivations and theoretical challenges
- A very short review of the *Naming Game*
- The ***Category Game***
- Discussion and conclusions

A. Baronchelli

Physics Department,
Technical University of Catalonia (UPC),
Barcelona (Spain)

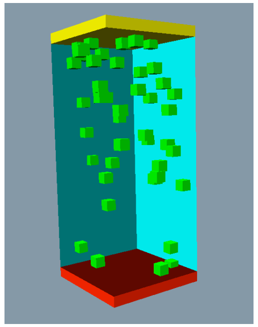
with V. Loreto and A. Puglisi

Physics Department, Università “La Sapienza”, Roma (Italy)

Irvine, March 15 2008

A physicist approach..

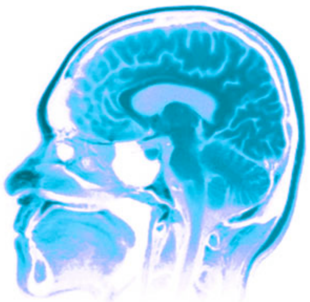
Statistical mechanics and complex system science:

 +  +  + 

emergence
self-organization

+ 

Study of language:



Language as an evolving set of conventions socially (i.e. globally) accepted by a group:
a complex adaptive system [1].

[1] L. Steels,in M. Shoenauer(ed) Proc. of PPNS IV. LNCS; Springer-Verlag (Berlin, 2000).

Semiotic dynamics

View of language as an *evolving* and *self-organizing* system and focus on culture

Motivations

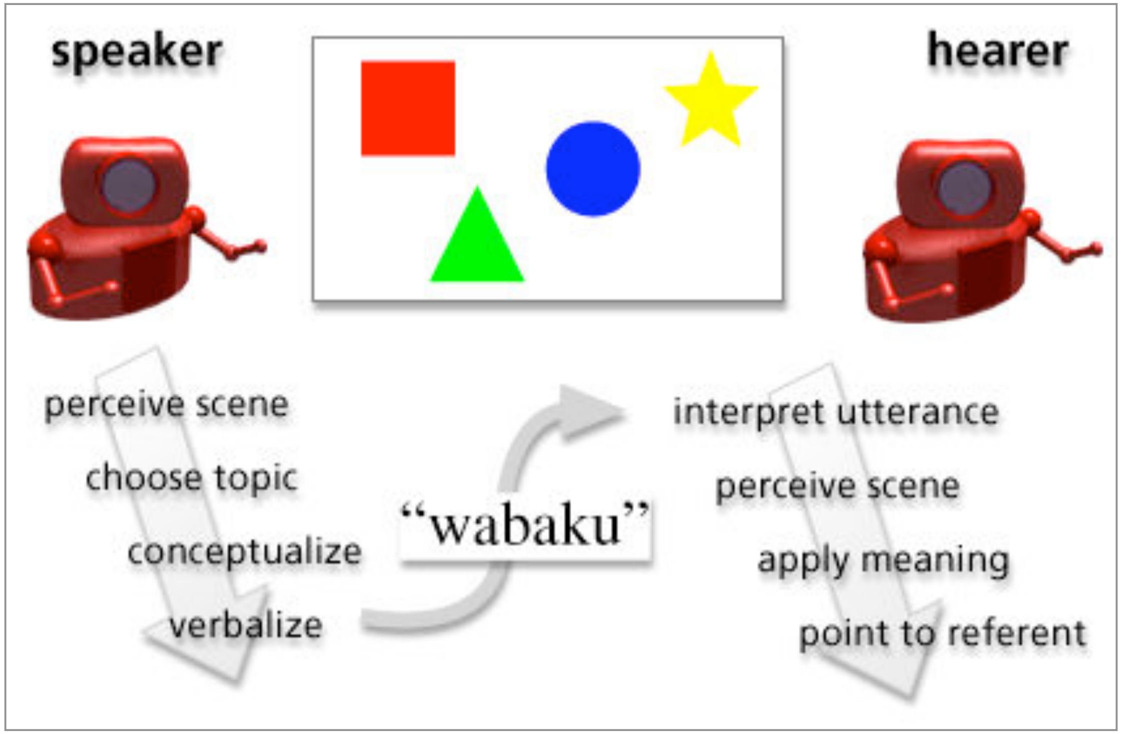
- The Web allows for global monitoring of human communication
- Many social, biological and *technological* systems are made of communicating entities
- Understand how global behaviors emerge out of local interactions

our strategy

✓ Definition of <u>simple</u> models (of increasing complexity)
✓ Quantitative analysis
✓ Analytical approaches (whenever possible)
✓ Connection with real world systems and experiments

The "Talking Heads" experiment

- Artificial robotic agents
- Hidden internal states
- Agents are not given any prior lexicon
- Open ended population and set of meanings



Language Games: Naming, Single property Guessing, Multiple property guessing, etc…

[1] L. Steels, *The Talking Heads Experiment. Vol.1 - Words and Meanings,* Antwerpen, (1999)

The Naming Game

(categories coming soon..)

with A. Barrat, E. Caglioti,
L. Dall'Asta, G. Gosti, V. Loreto,
M. Felici and L. Steels

Theoretical challenges

- What are the minimal requirements for a shared vocabulary to emerge?
- What are the global dynamics that lead to convergence?
- Which features lead to *efficiency*?
- Which is the role of the system size?
- Which is the role of topology?

The Naming Game

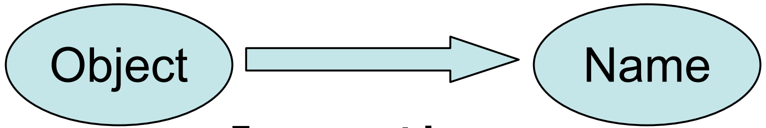
- Population of N agents
- Each agent is characterized by its *inventory* (or lexicon) i.e. a list of name-object associations
- Agents want to build a shared lexicon
- Homonymy is discarded → one single object
- Peer to peer *negotiation*. At each time step two agents (speaker and hearer) are selected

The microscopic interaction rules depend on the particular model and yield to different collective behaviors

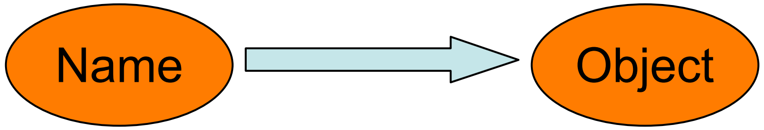
Microscopic Rules

SPEAKER

HEARER



Invention
or
random choice



Interpretation

Failure: the hearer does not know the uttered word; after the interaction he records it

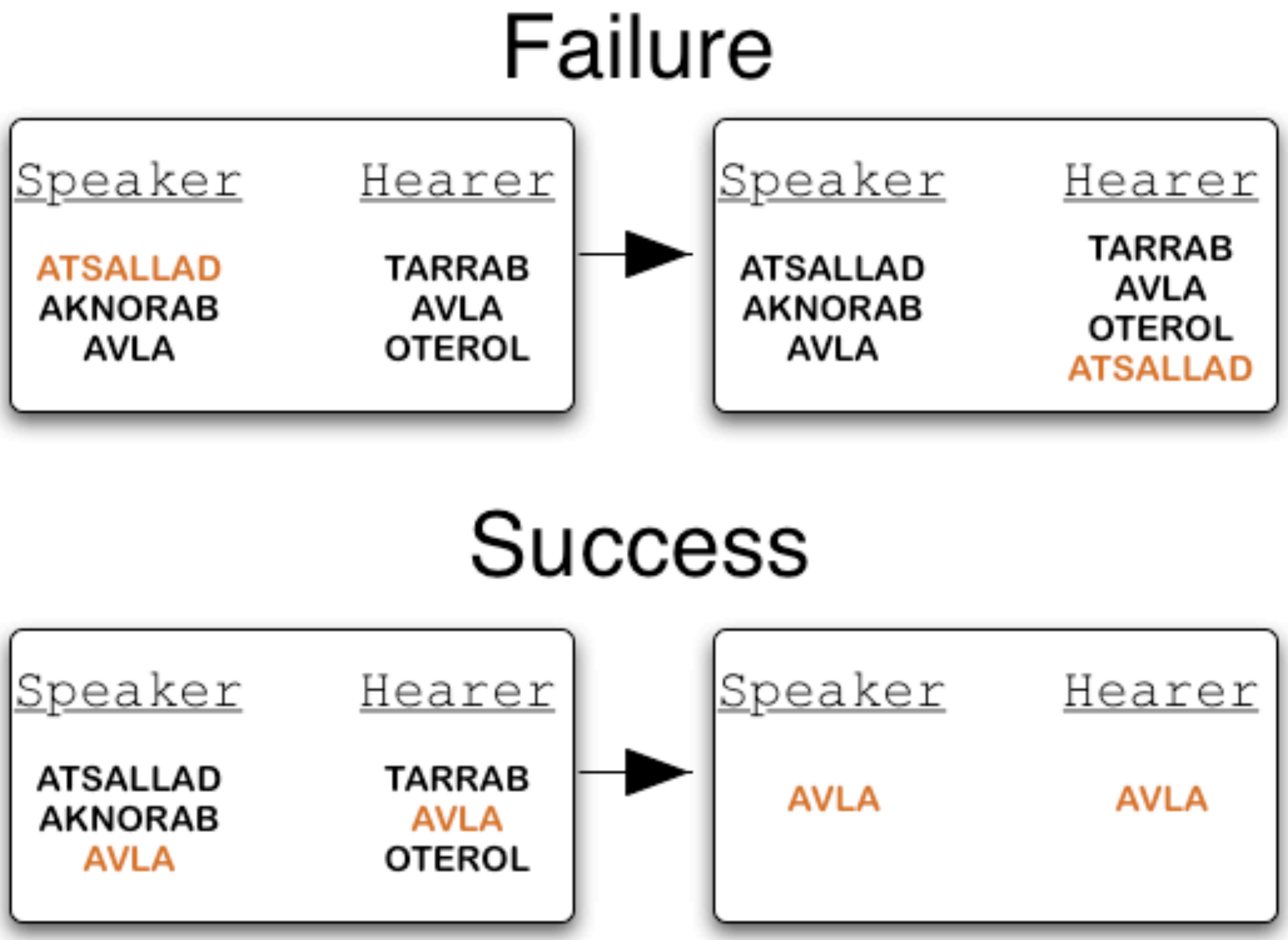
$$(w_1^i, w_2^i, \ldots, w_{n_i}^i) \rightarrow (w_1^i, w_2^i, \ldots, w_{n_i}^i, w_{n_i+1}^i)$$

Success: the hearer knows the uttered word; after the interaction both agents lexicons contain <u>*only*</u> the winning word

$$(w_1^i, w_2^i, \ldots, w_{n_i}^i) \rightarrow (w_1^{i_{winner}})$$

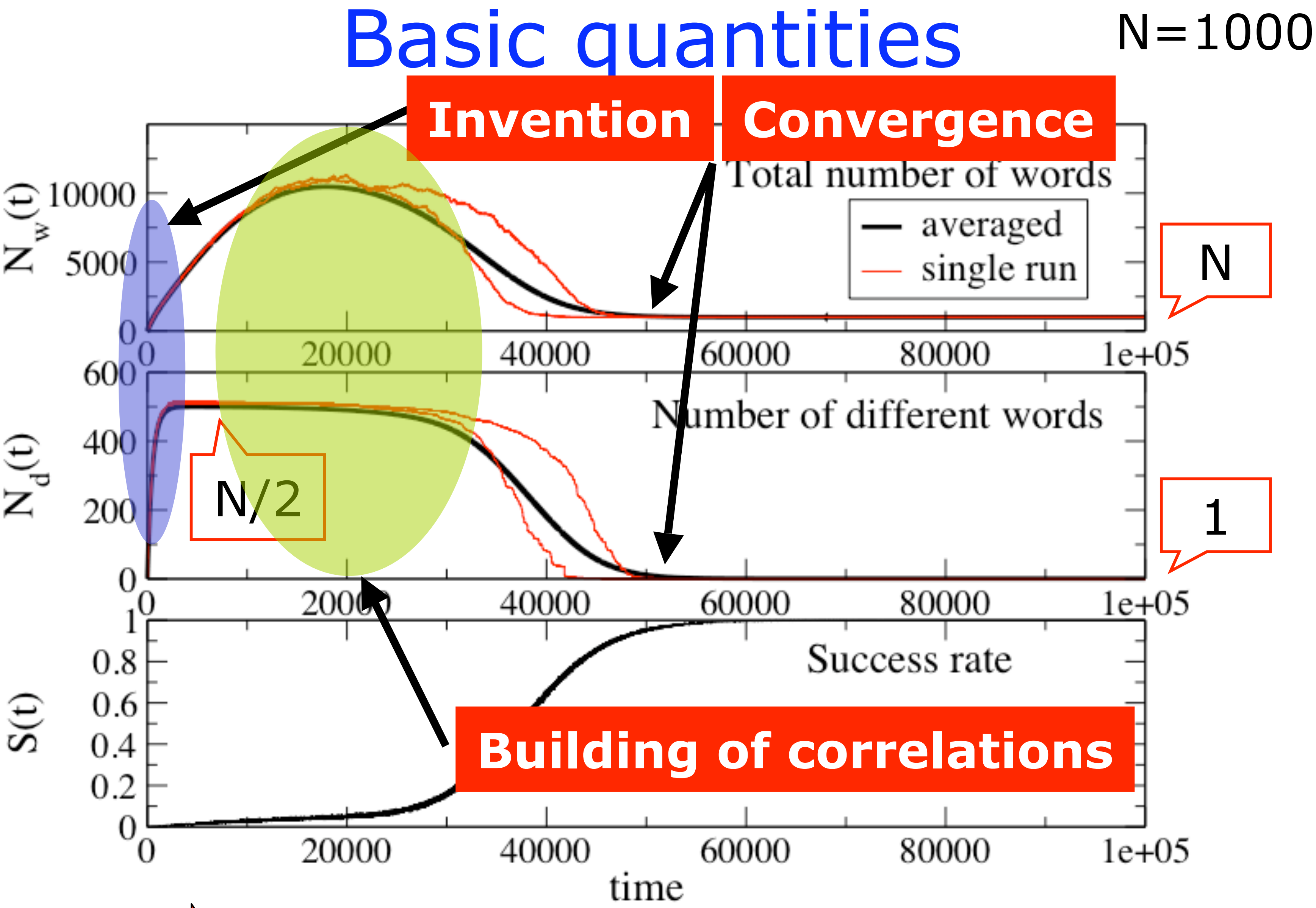
[1] Baronchelli *et al.* J. Stat. Mech. P06014 (2006)

Microscopic Rules



negotiation + memory + dynamic inventories

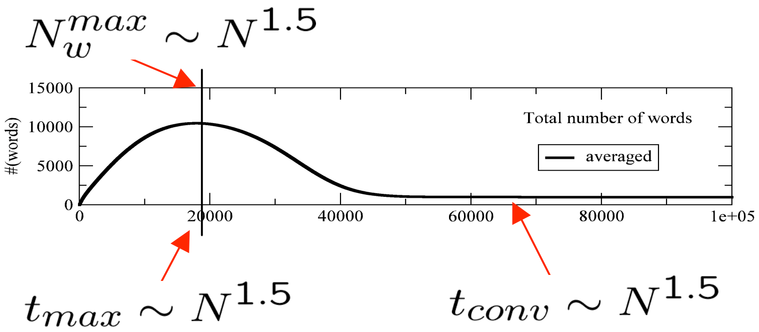
Basic quantities

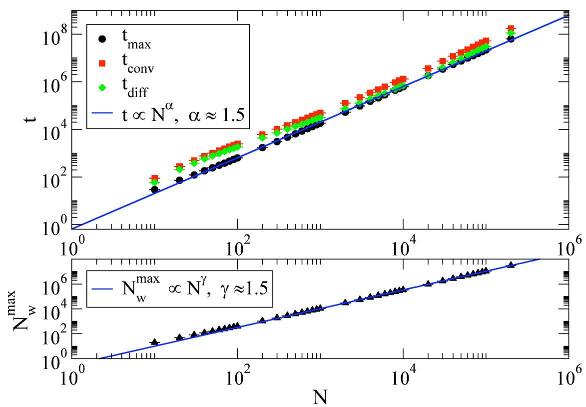


The communication system is efficient

Detailed analysis

- ✓ Convergence dynamics
- ✓ Scaling properties
- ✓ Role of the topology
- ✓ Microscopic activity patterns
- ✓ Generalization of the model (consensus-polarization phase transition)
- ✓ Role of homonymy
- ✓ Etc..





The Category Game

with V. Loreto and A. Puglisi

PNAS (in press)
pre-print: *arXiv:physics/0703164v1*

Theoretical challenges

- How does a population of agents establish and share an effective set of categories? [1]

- Is a macroscopic stationary state always reached starting from <u>very simple</u> microscopic dynamical rules?

- Quantitatively: which is the role of the
    - system size?
    - complexity of the environment?
    - resolution power of the agents (d_{min})?

- The Big Question: categorizing a <u>continuum</u> perceptual space?

[1] L. Steels & T. Belpaeme, *BBS* **28**, 469 (2005)

Linguistic categories

Most natural examples of linguistic categories are "common names",
i.e. words that indicate many different things

Thus, linguistic categories

- allow to quickly point out something without giving too many details (lossy compression)
- are well calibrated to avoid confusion, i.e. to discriminate something among different things
- in brief: must be **not too large nor too small**

The Category Game

- Population of *N* agents
- Individual: set of (perceptual) categories + inventories for them, i.e. a list of name-category associations
- Language-mediated, peer to peer ***negotiation***. At each time step two agents (sp + hea) are selected and presented a scene with different objects (say: colors or real numbers)
- a topic is chosen, the speaker must indicate it through a word
- the hearer must guess which is the topic listening to that word
- discrimination of the topic is implicitly required
- based on **success/failure**: categories, words and their associations are updated

See also: L. Steels & T. Belpaeme, *BBS* **28**, 469 (2005)

A simple scheme

INDIVIDUALS: a simple low-dimensional input channel, such as the complete 3d color channel or just the 1d **hue channel**, temperature sensor, altimeter, etc.

0 Real values on the interval [0, 1] 1

CATEGORIES (perceptual): subsets of the interval
(many ways of defining the subset, not crucial)

SYNONYMY: many words -> one category

HOMONYMY: one word -> many categories

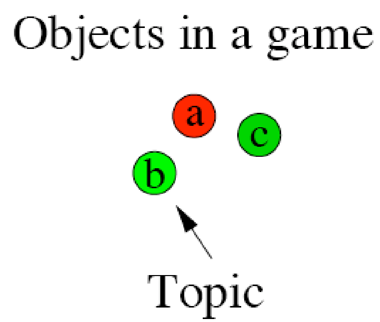
Rules (1/3)

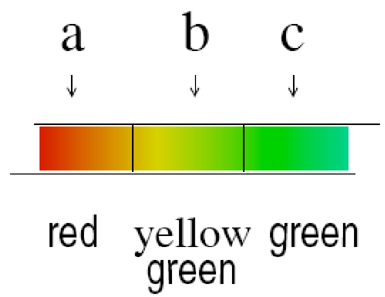
• each agent has a set of non-overlapping categories [subsets of *(0,1)*], defined by boundaries; categories fully cover the interval; at the beginning only the category *(0,1)* exists

• each category comes with an inventory of words; at the beginning a brand new word is associated to each category

• the scene: *M* real numbers in *(0,1)* at a minimal distance d_{min} (resolution power, Just Noticeable Difference or Difference Limen)

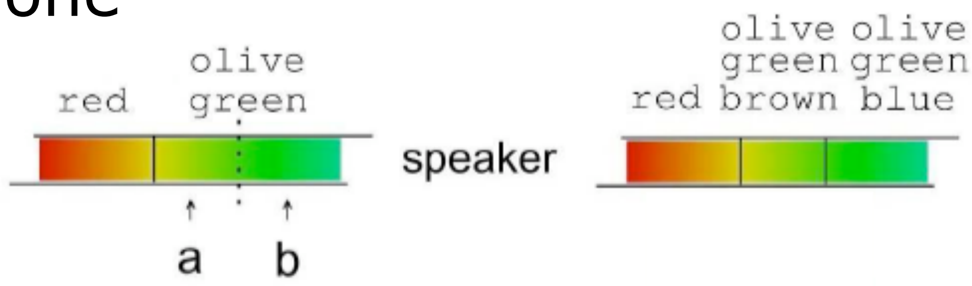
• one of the objects is the topic, known by the speaker only





Rules (2/3)

• the speaker discriminates the topic: this may require the creation of new boundaries; each new category inherits the words of the old category, plus a brand new one



• the speaker says the “last-winning” word associated with the discriminating category, or the newly created one if this is the first game played with that category

• the hearer looks at her inventory for that word: a **set of candidate categories** (containing at least one object and that word) is obtained

Rules (3/3)

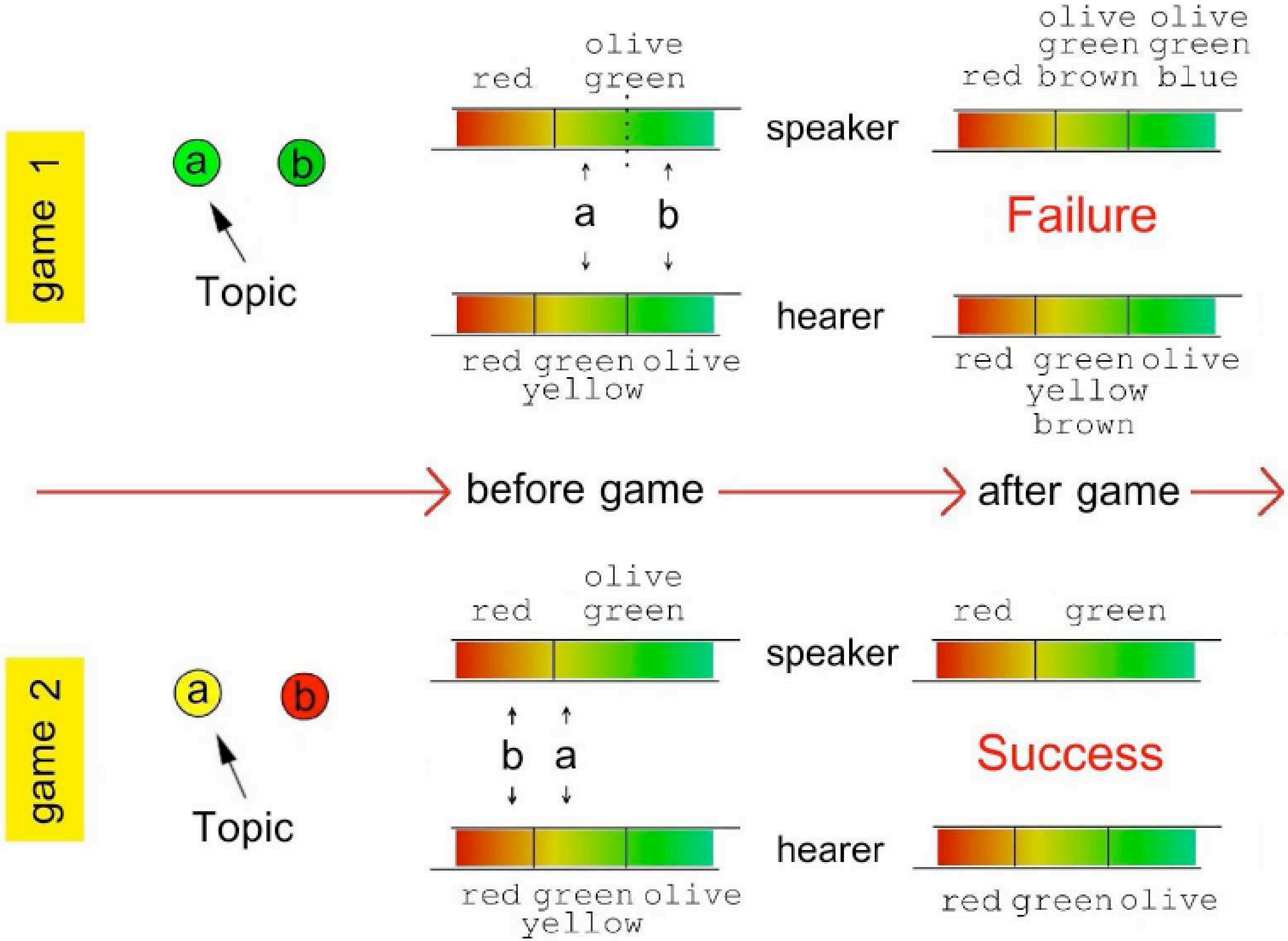
- if the **set is empty**, the game is a failure; the speaker points at the topic and the hearer discriminates it and adds the speaker's word to the correct category

- if the **set is not empty**, the hearer chooses at random the category and the object in it, and finally makes her guess manifest

- if the **guess is correct**, both individuals reduce the inventory associated to the winning category to the winning word only, which is assigned the status of "last-winning" word

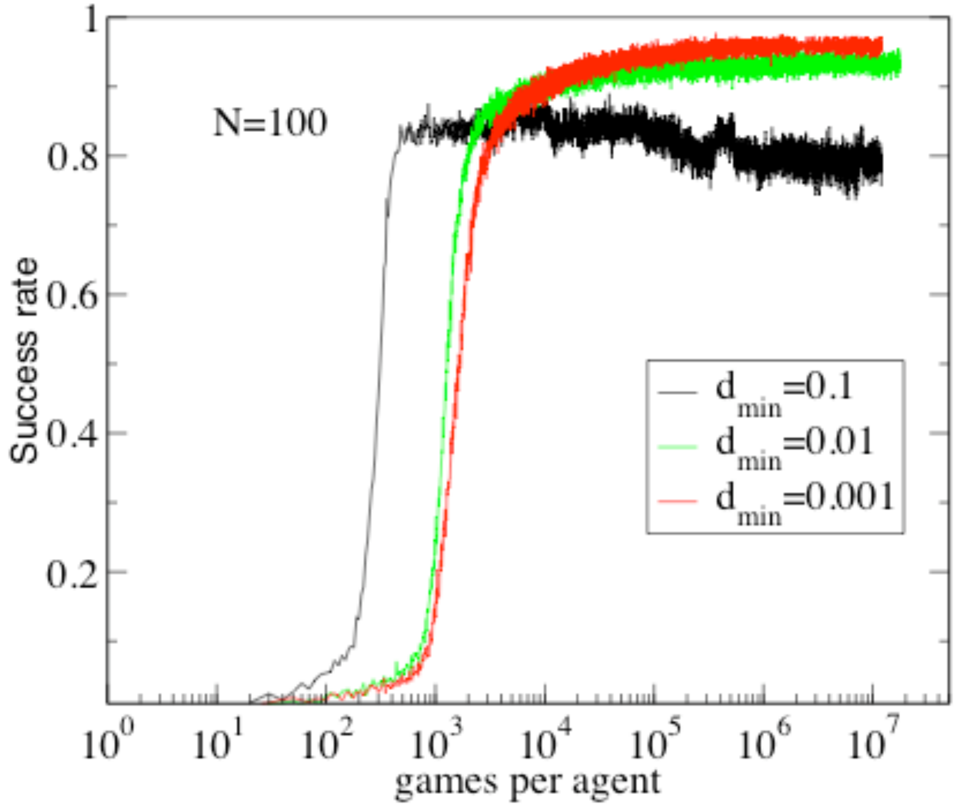
- **otherwise**, the topic is unveiled, the hearer discriminates the scene and (if not present) the uttered word is added to the discriminating hearer's category

Examples


game 1
a
b
Topic
olive
red
green
speaker
a
b
hearer
red green olive
yellow
olive olive
green green
red brown blue
Failure
red green olive
yellow
brown
before game
after game
game 2
a
b
Topic
olive
red
green
speaker
b
a
hearer
red green olive
yellow
red
green
Success
red green olive

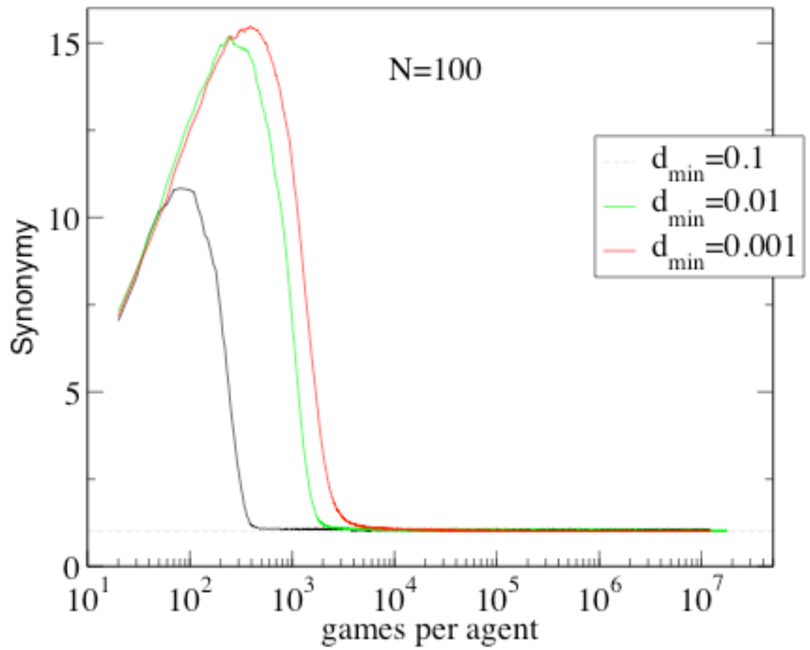
Time evolution

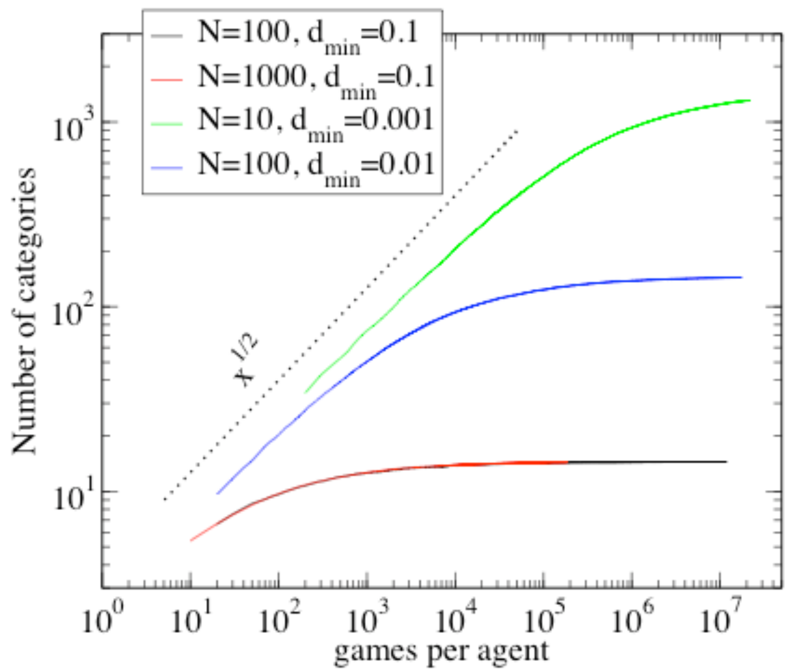
- Initially, most games are unsuccessful
- After *~10^3xN* games a sharp transition: the success rate becomes very high (*~90%* or higher)



1. Naming Game + free categories

- For each category: typical NG synonymy curve
- Number of categories grows as $t^{1/2}$, free discrimination: probability of a new category $1/n_{cat}$
- The growth of category number slows down when $n_{cat} \sim 1/d_{min}$; category number cannot grow above $2/d_{min}$



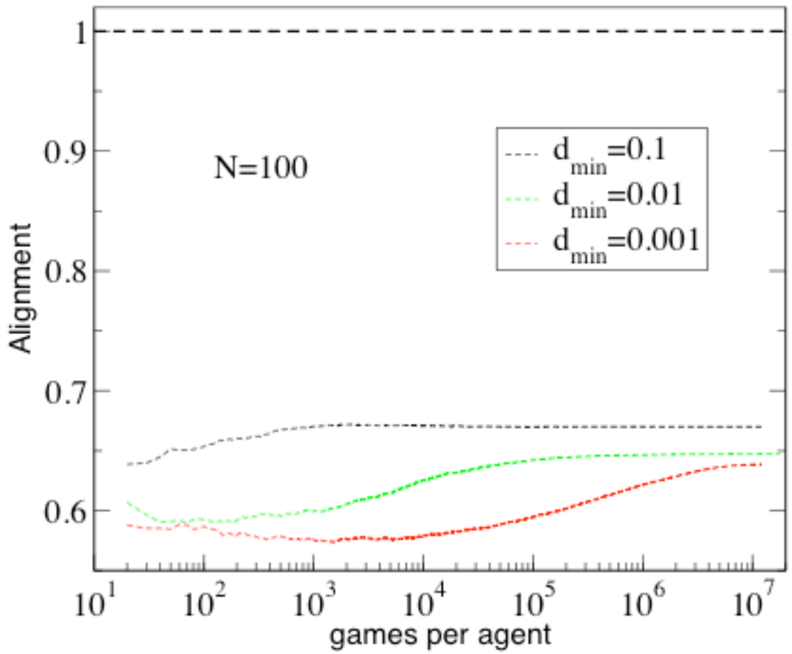


2. Full communicative success

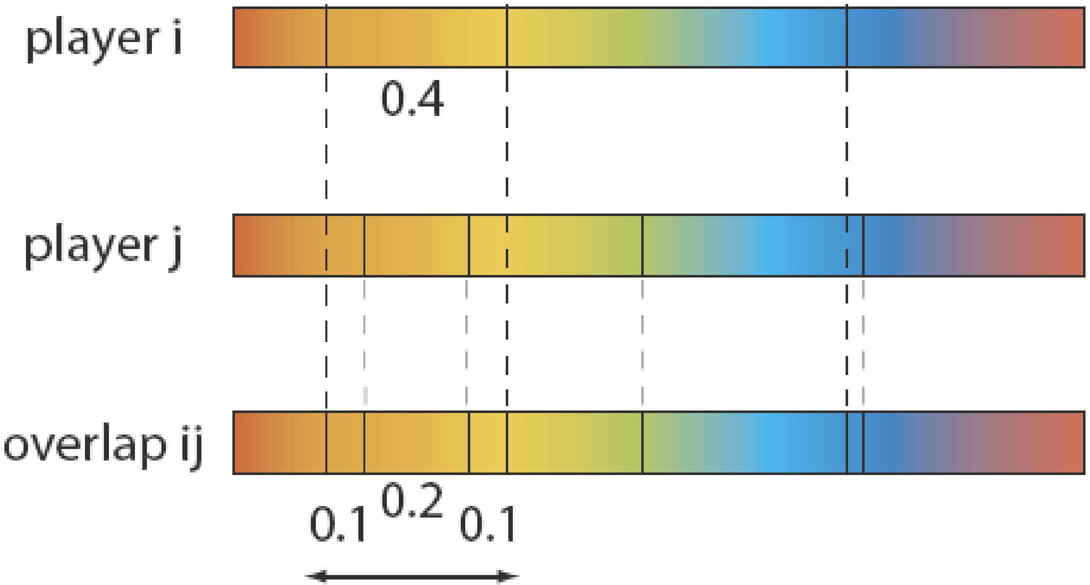
- Success is high (~90%)
- Synonymy is eliminated
- Categories are still evolving (slow refinement)
- Categories are poorly aligned (!)

Pb: How can the success rate be so high?

$$o_{ij} = \frac{2\sum_{c_i^j}(l_{c_i^j})^2}{\sum_{c_i}(l_{c_i})^2 + \sum_{c_j}(l_{c_j})^2}$$



The overlap functional



$$c_i = \cup\, c_i^j$$

$$l_{c_i} = \Sigma\, l_{c_i^j} \qquad l_{c_i}^2 \geqslant \Sigma\, l_{c_i^j}^2$$

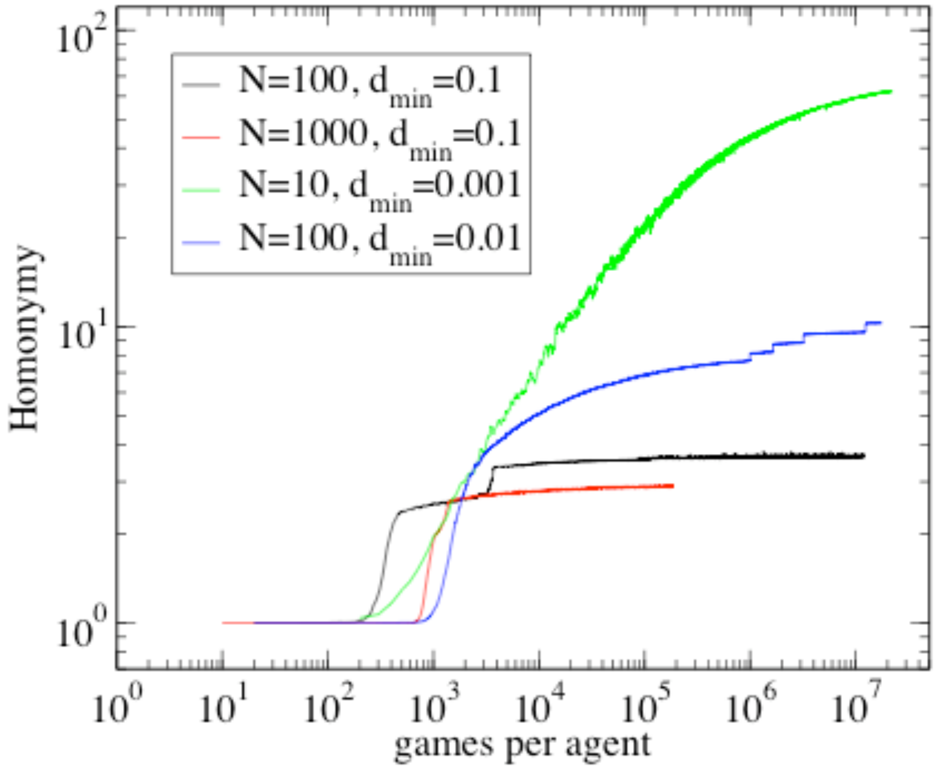
$$O = \frac{2}{N(N-1)} \sum_{i<j} o_{ij}$$

$$o_{ij} = \frac{2\sum_{c_i^j} (l_{c_i^j})^2}{\sum_{c_i} (l_{c_i})^2 + \sum_{c_j} (l_{c_j})^2}$$

The definition of an overlap functional is not trivial since the number of categories is not constant. Luckily, we do not need sophisticated measures.

Emergence of homonymy

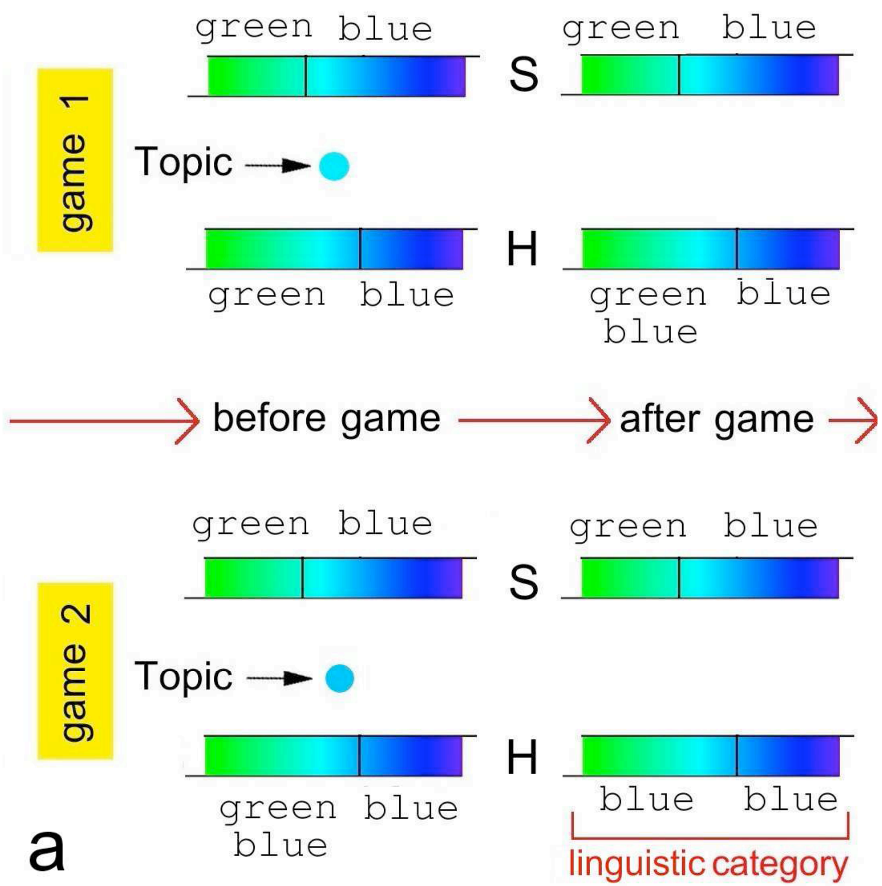
As synonymy disappears
homonymy is growing



Number of categories associated to the same (unique) word

The word-contagion effect

Homonymy growth is due to an intra-category word-contagion

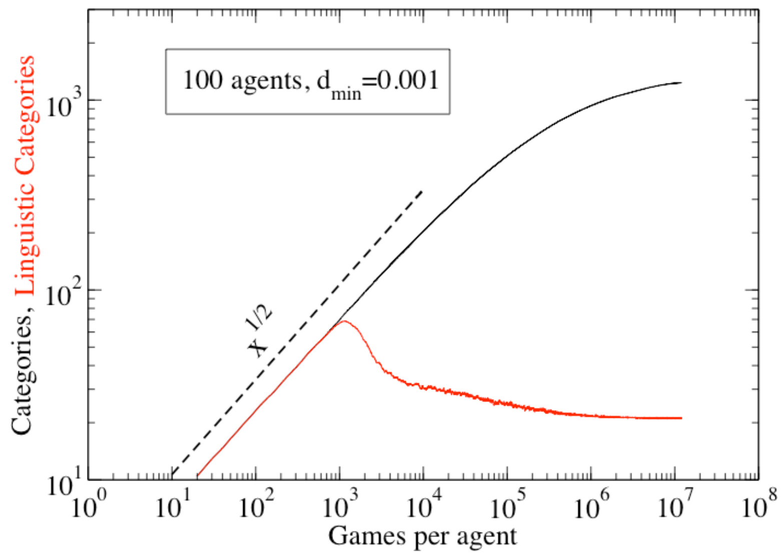


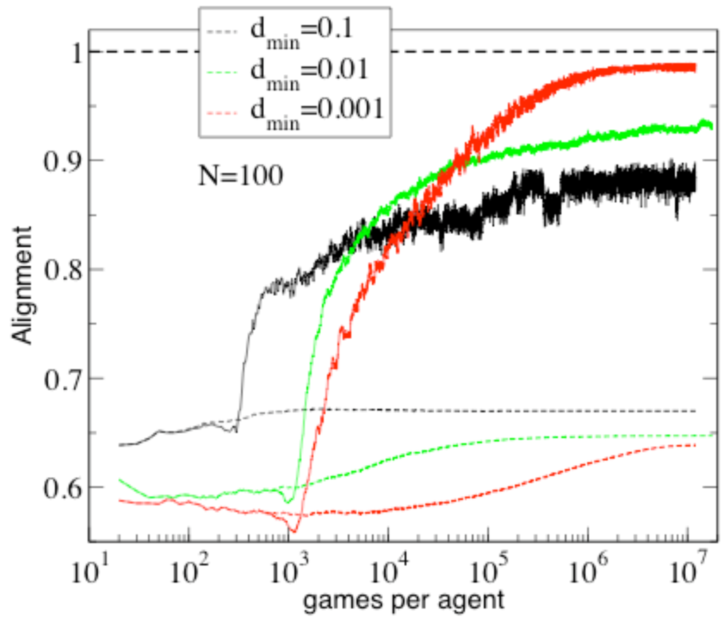
Emergence of Linguistic Categories

adjacent categories identified by the same word can be considered a single linguistic category

Linguistic categories:

- emerge as connected sets
- their number is much lower and becomes stable
- their alignment is much higher (hence the success!)

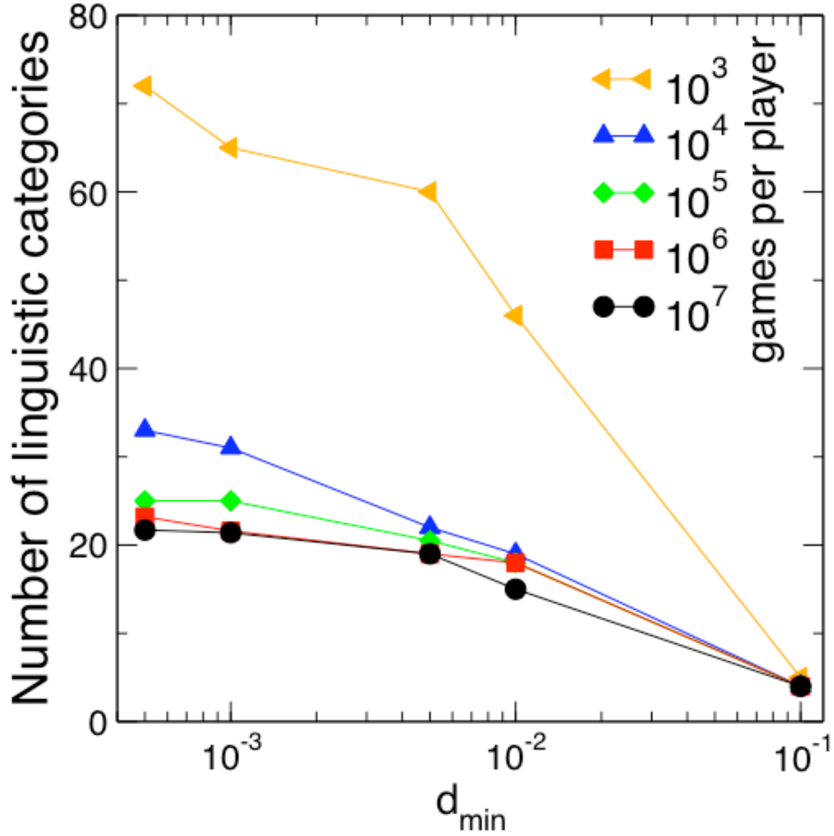


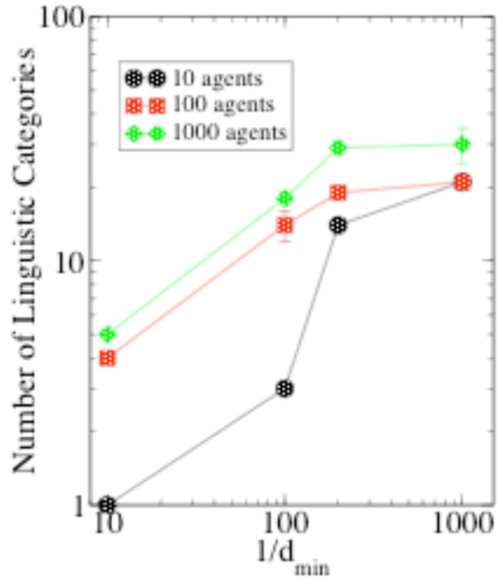


Role of the parameters

The number of linguistic categories:

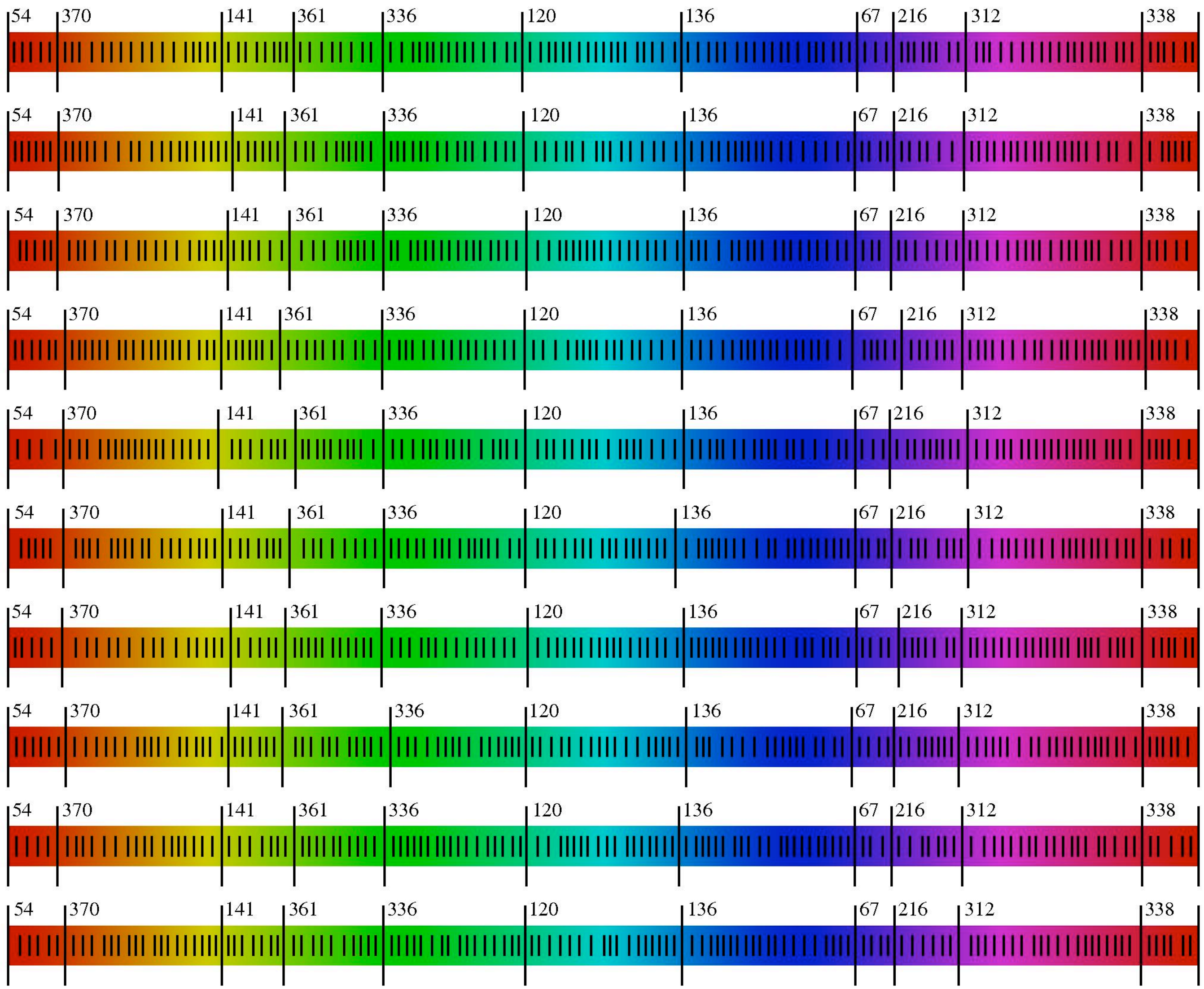
- slightly increases with N and
- most importantly, saturates for small d_{min}





Role of the environment 1

Inputs uniformly distributed in (0,1)

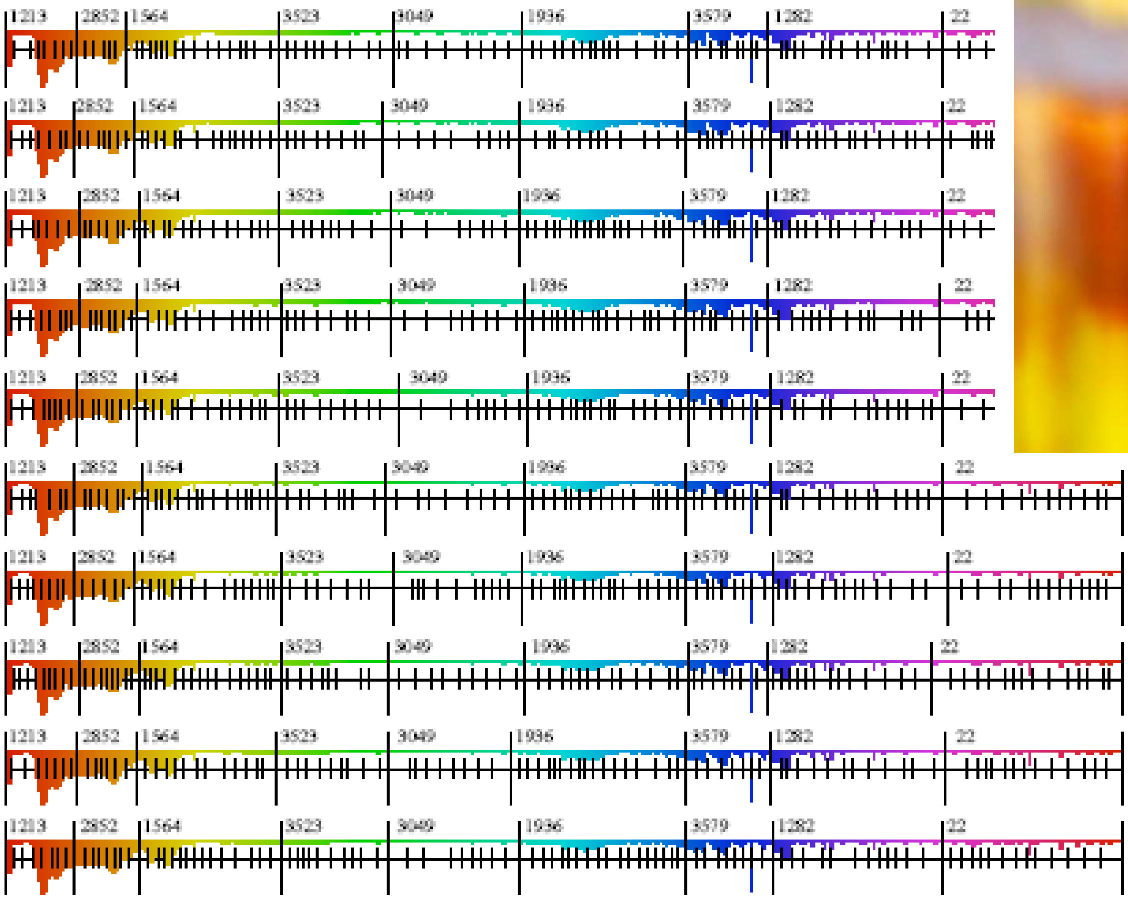


N=100

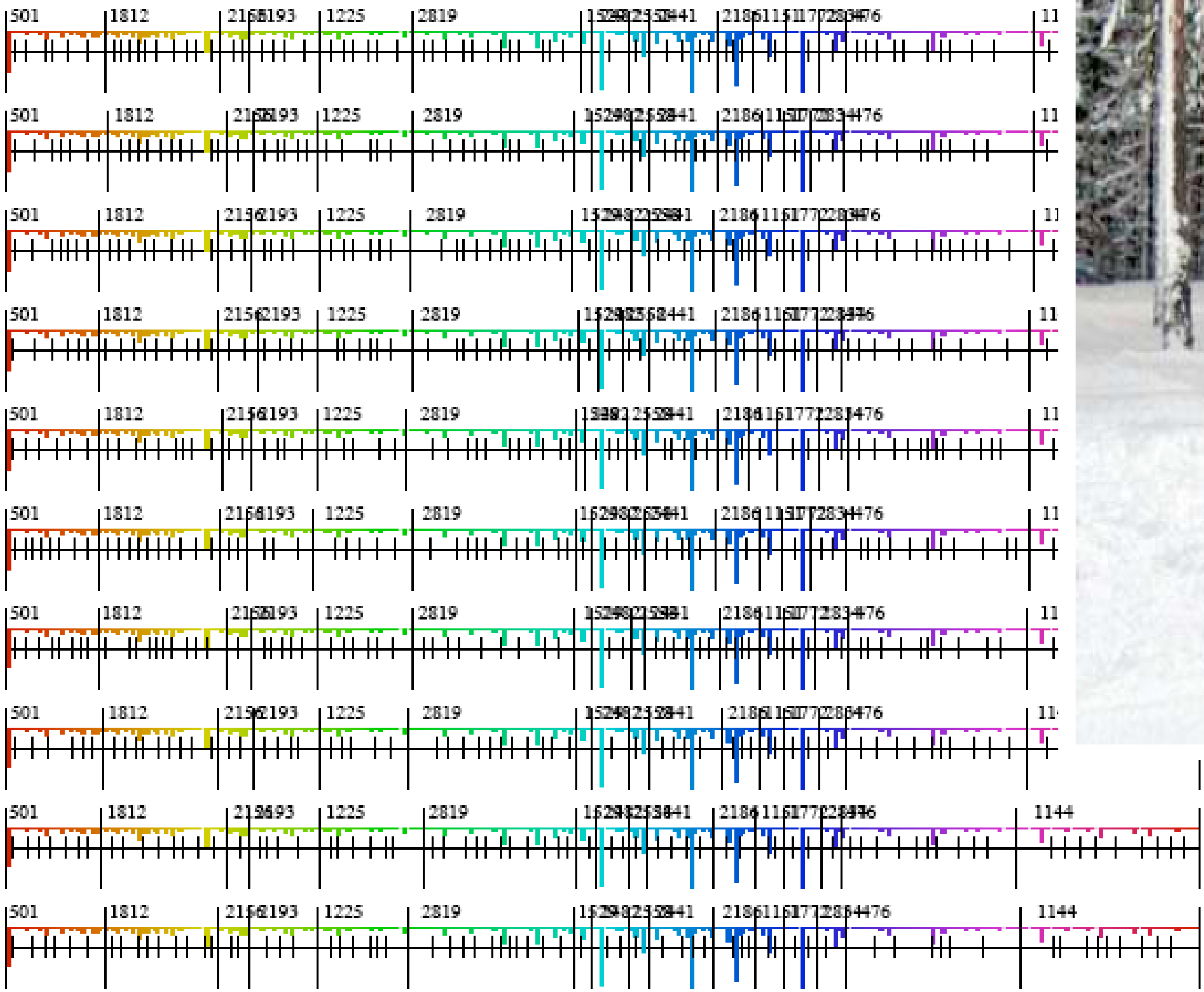
d_{min}=0.01

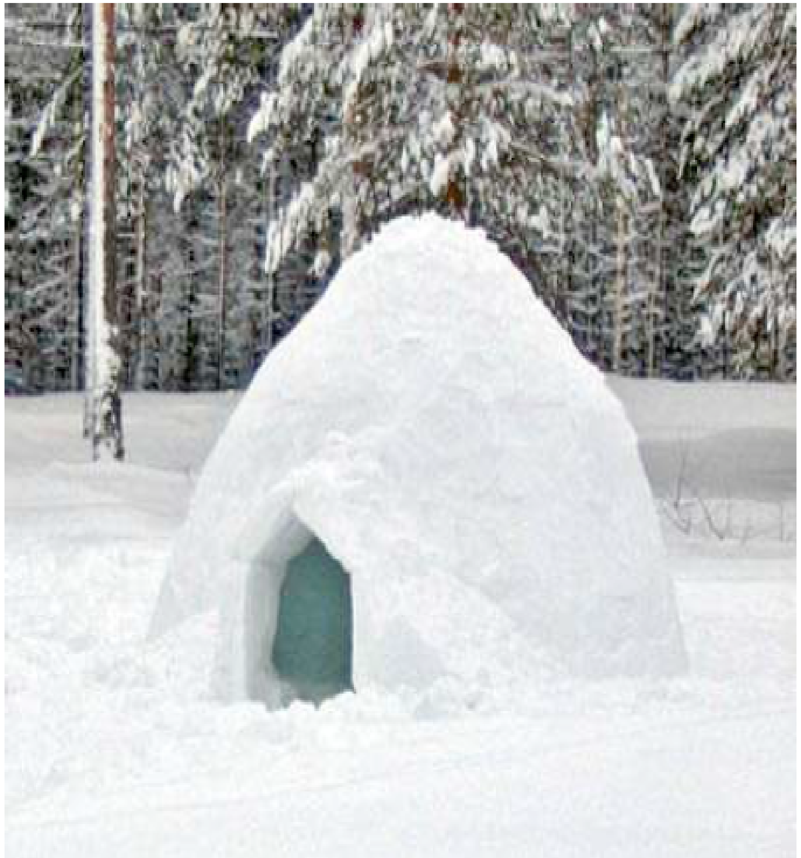
Role of the environment 2

N=100, d_{min}=0.01



Role of the environment 3





N=100

d_{min}=0.01

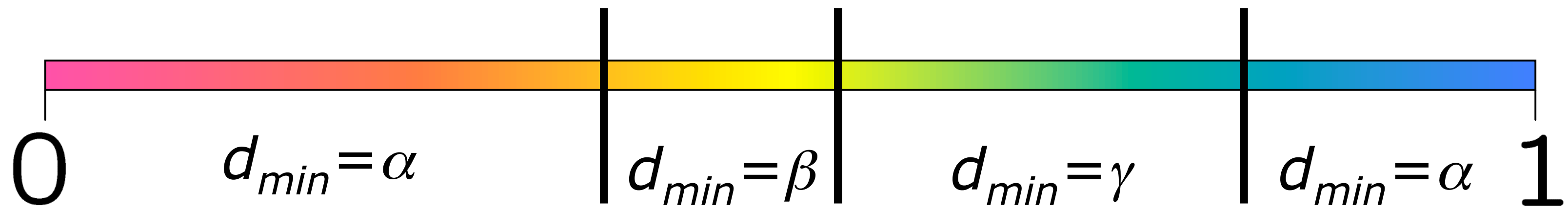
“Genetic” biases

d_{min} is a “genetic” trait of the individuals

In principle it can vary:

1. On the [0,1] axis (different resolution power for different stimuli)
2. From individual to individual (population heterogeneity) [1]

Here we focus on 1.

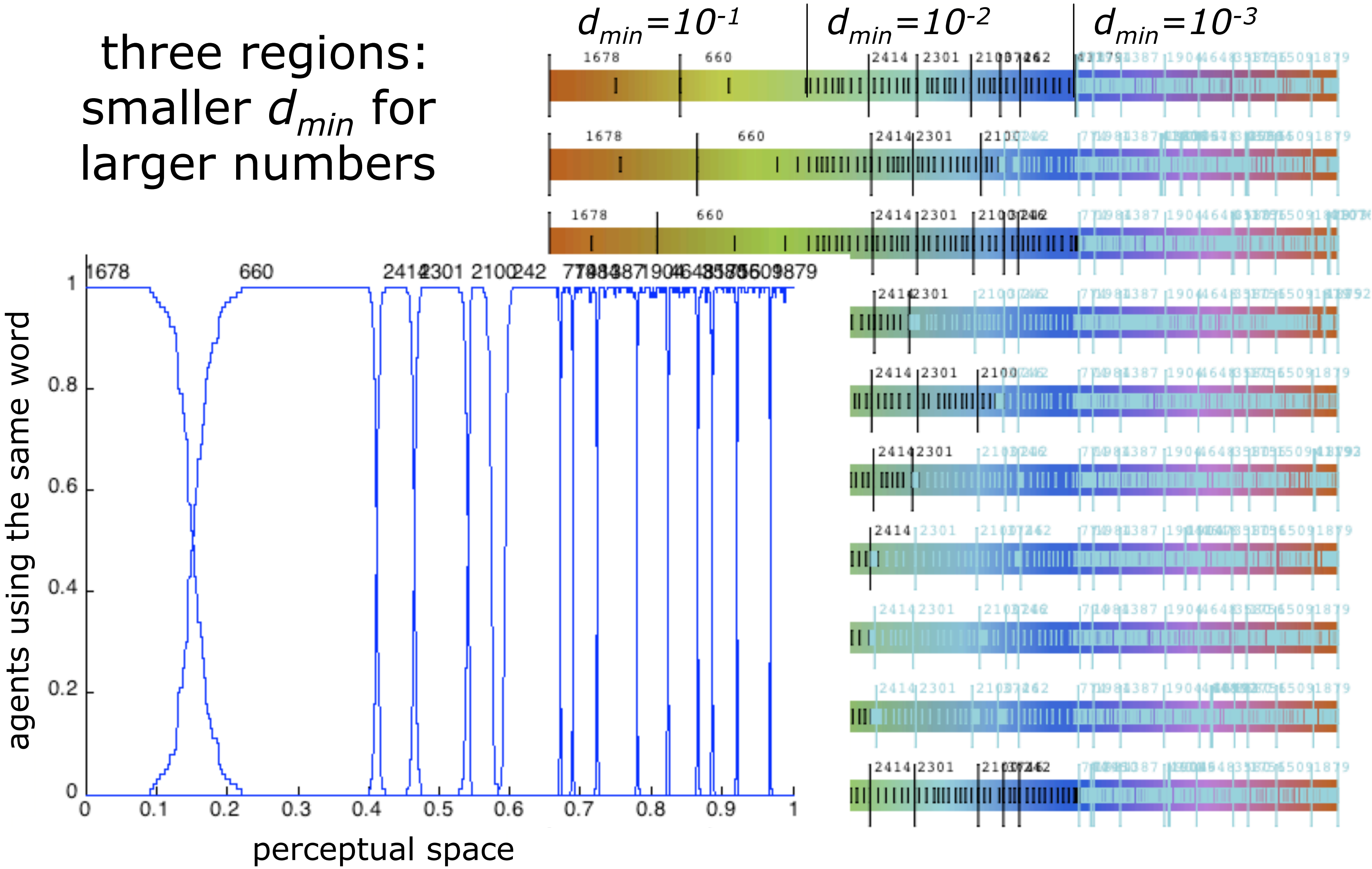


[1] See K. Jameson and N. Komarova (2008): *Agent-based categorization: the role of population and color-stimulus heterogeneity* (for K_{sim} parameter).

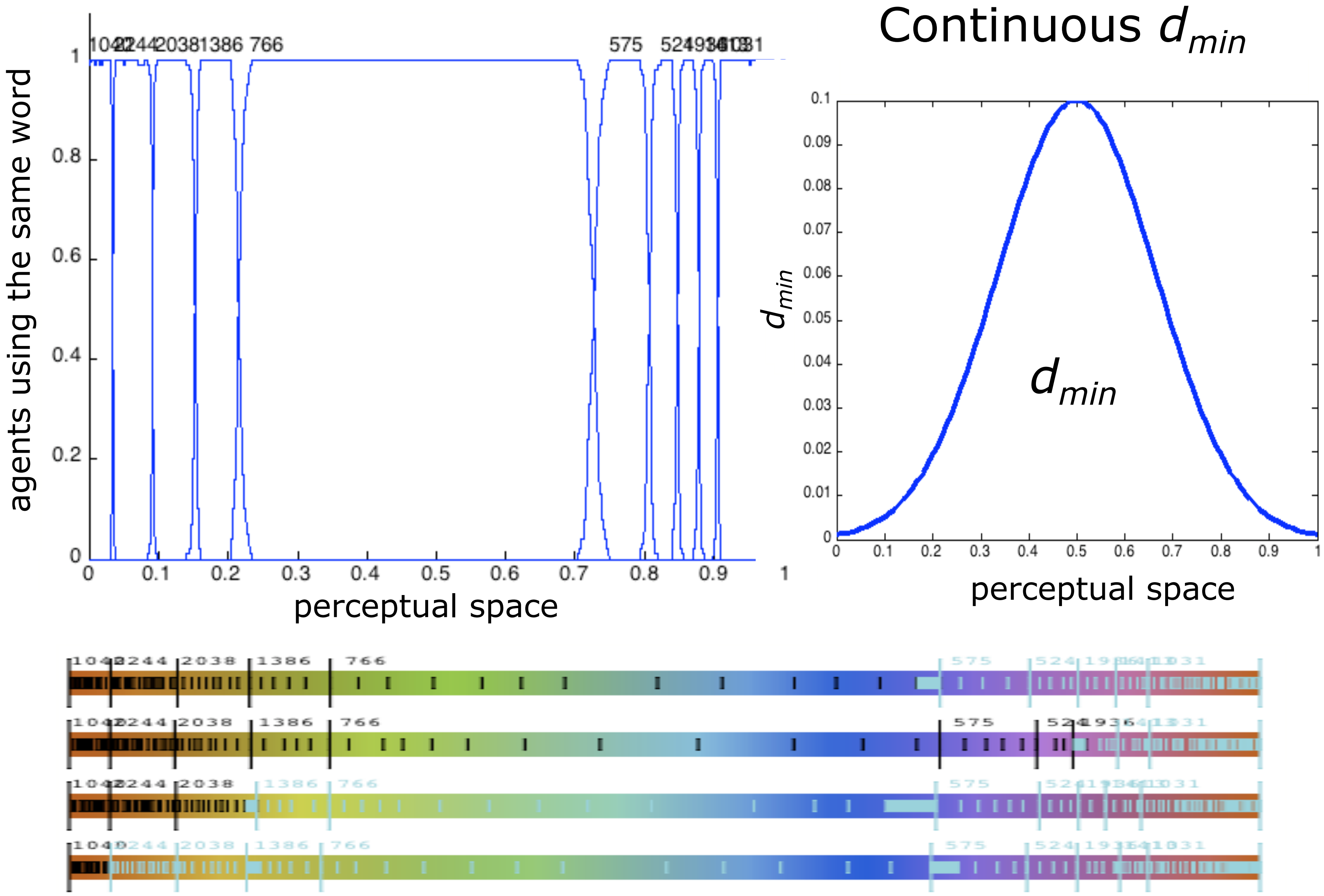
Also with T. Gong, presently at “La Sapienza” University.

Non-uniform d_{min}

three regions:
smaller d_{min} for
larger numbers

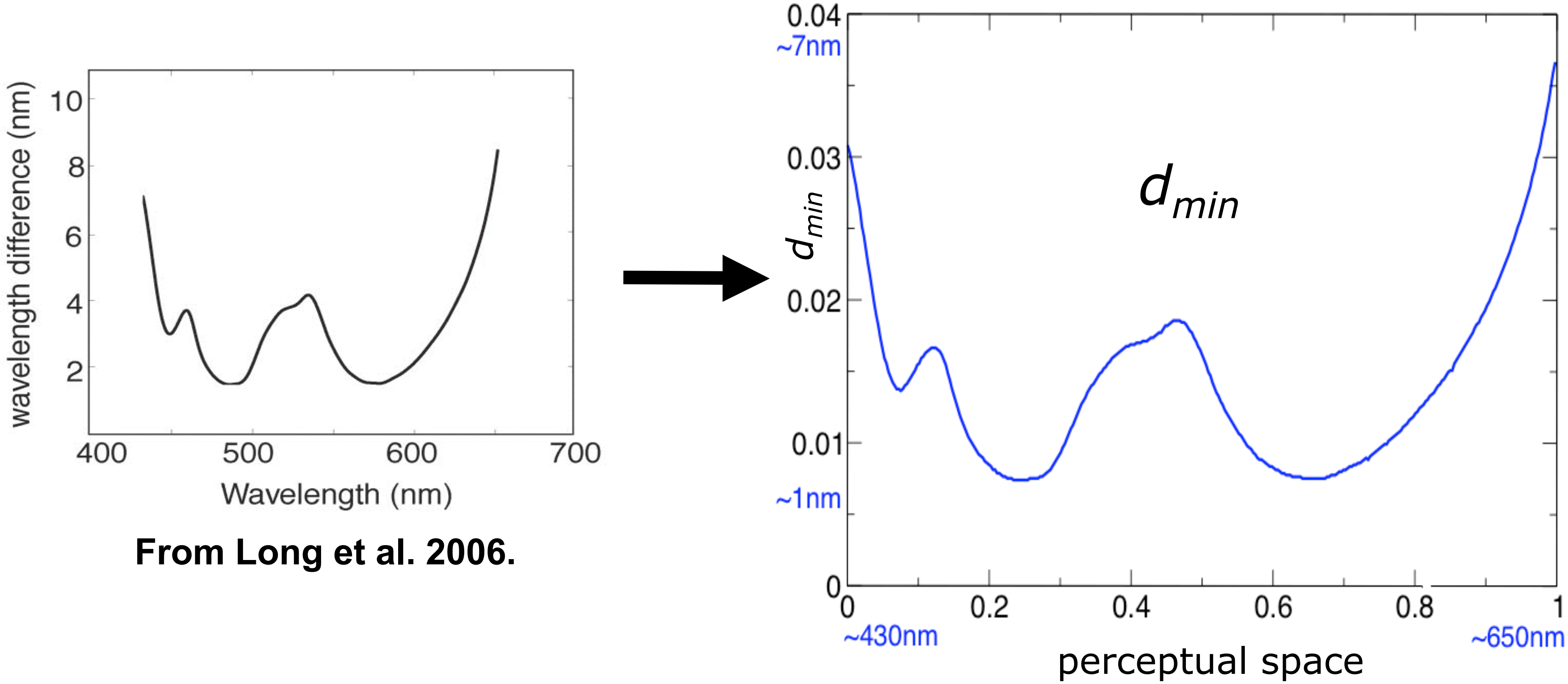


Non-uniform d_{min}



Non-uniform d_{min}

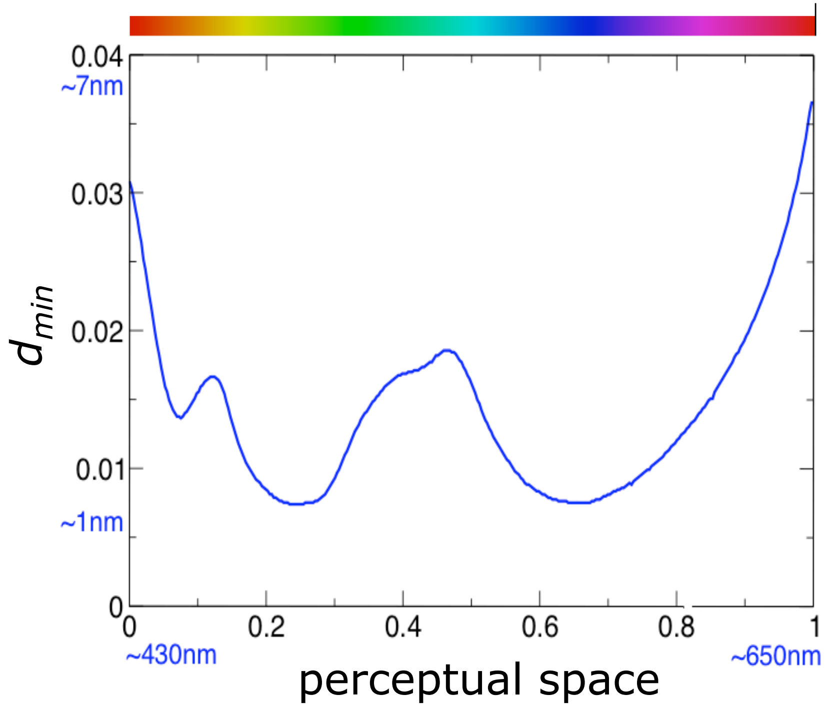
"Human" d_{min}

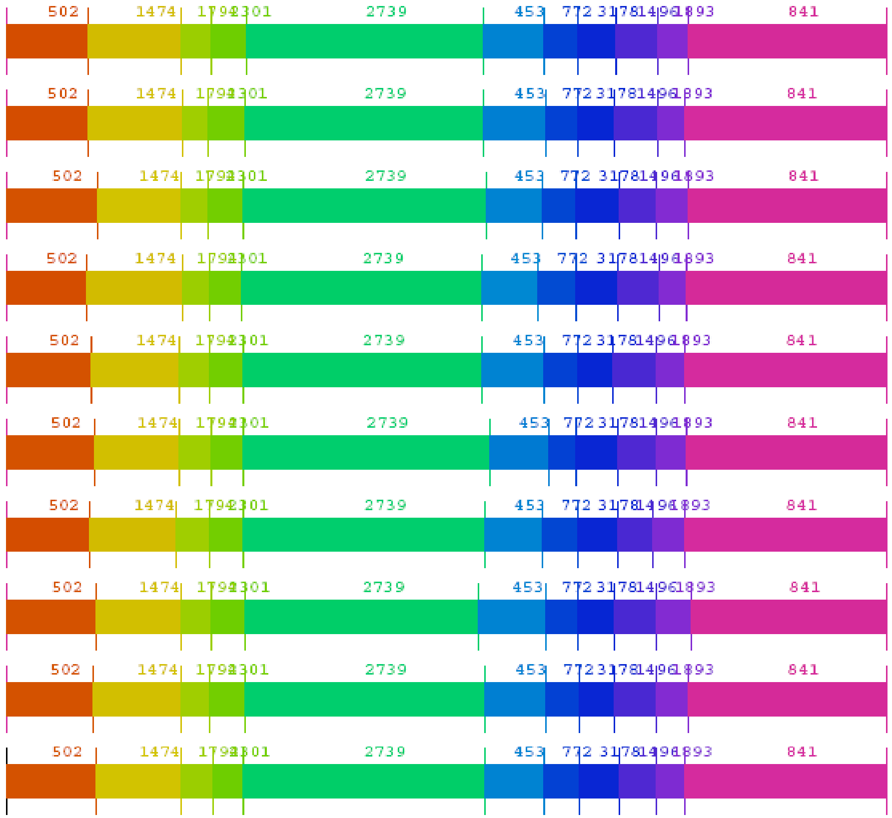


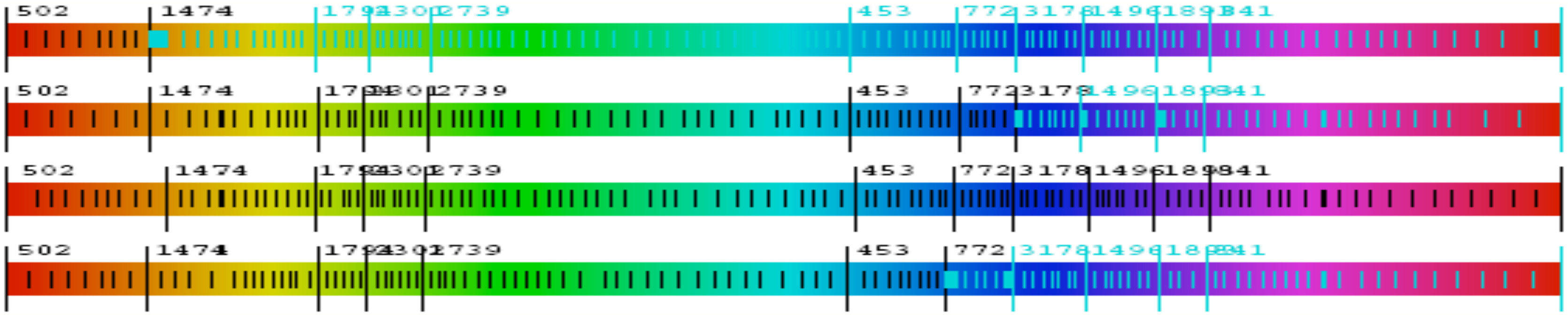
From Long et al. 2006.

Long PH, Yang ZY, Purves D. 2006. Special statistics in natural scenes predict hue, saturation, and brightness. PNAS, 103(15): 6013-6018.

Non-uniform d_{min}





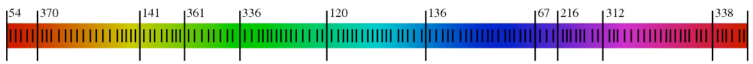


The next step (a roadmap..)

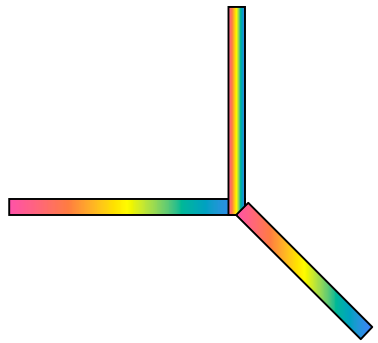
- Linguistic categories work better thanks to compositionality
- If a category is not sufficient to discriminate our topic, we can add further specifications
- We **ignored** this possibility, but this more advanced issue should be taken into account in the next step

<u>A possible path could be continuing in the same spirit</u>

CG agent: made of NG agents (perc. categories). Decides which NG agent must play and in case can create new NG agents.



NEXT agent: made of CG agents (channels), Decides which CG agent must play (or which combination) and in case can create new CG agents.



Conclusions

- The Category Game is simple, can incorporate empirical results, could produce checkable predictions (in progress..)

- Quantitative approach and new discoveries (continuum, role of d_{min}, N, environment, perceptual biases, etc..)

- The systematic appearance of homonymy defines linguistic categories as an emerging layer on top of perceptual categories

- Linguistic categories are much more aligned in the population

Conclusions

- The success rate obviously decreases (slowly) with the number of objects in a scene. In human language we have compositionality

- The number of linguistic categories is kept low by the necessity of alignment, i.e. of comprehension among individuals

- Just to mention:

Some *anomic aphasics*, despite normal color vision, are unable to name color categories (try with few or many)[1], but do not exhibit general categorization problems[2,3]. A possible explanation is that they are unable to produce the necessary[4] non-perceptual discontinuity (e.g. verbal) in the perceptual continuum[4,5].

[1]Goldstein, *Language and language disturbances*, Grune and Stratton (1948). [2]Davidoff & Robertson, *Lang. and Cogn. Processes* **19**, 137 (2004). [3]Luzzatti & Davidoff, *Neuropsych.* **32**, 933 (1994). [4]Dummet, *Synthese* **30**, 301 (1975). [5]Robertson et al., *Cognition* **71**, 1 (1999).

References can be found here

http://andrea.baronchelli.googlepages.com/home

Thank you!