

# Modeling language acquisition, change and variation

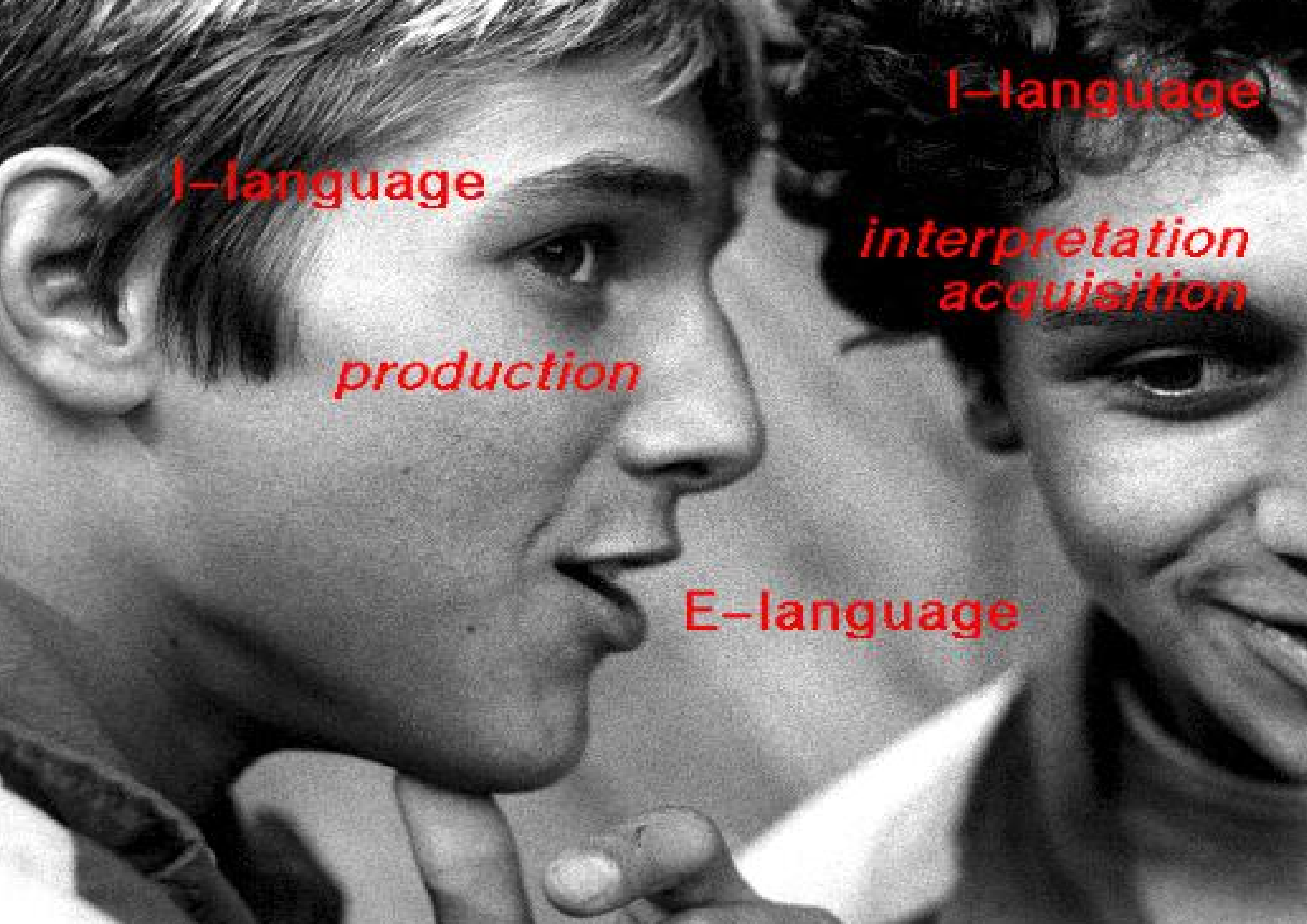
Willem Zuidema

Language Evolution & Computation Research Unit

Institute for Cell, Animal and Population Biology

University of Edinburgh, U.K.





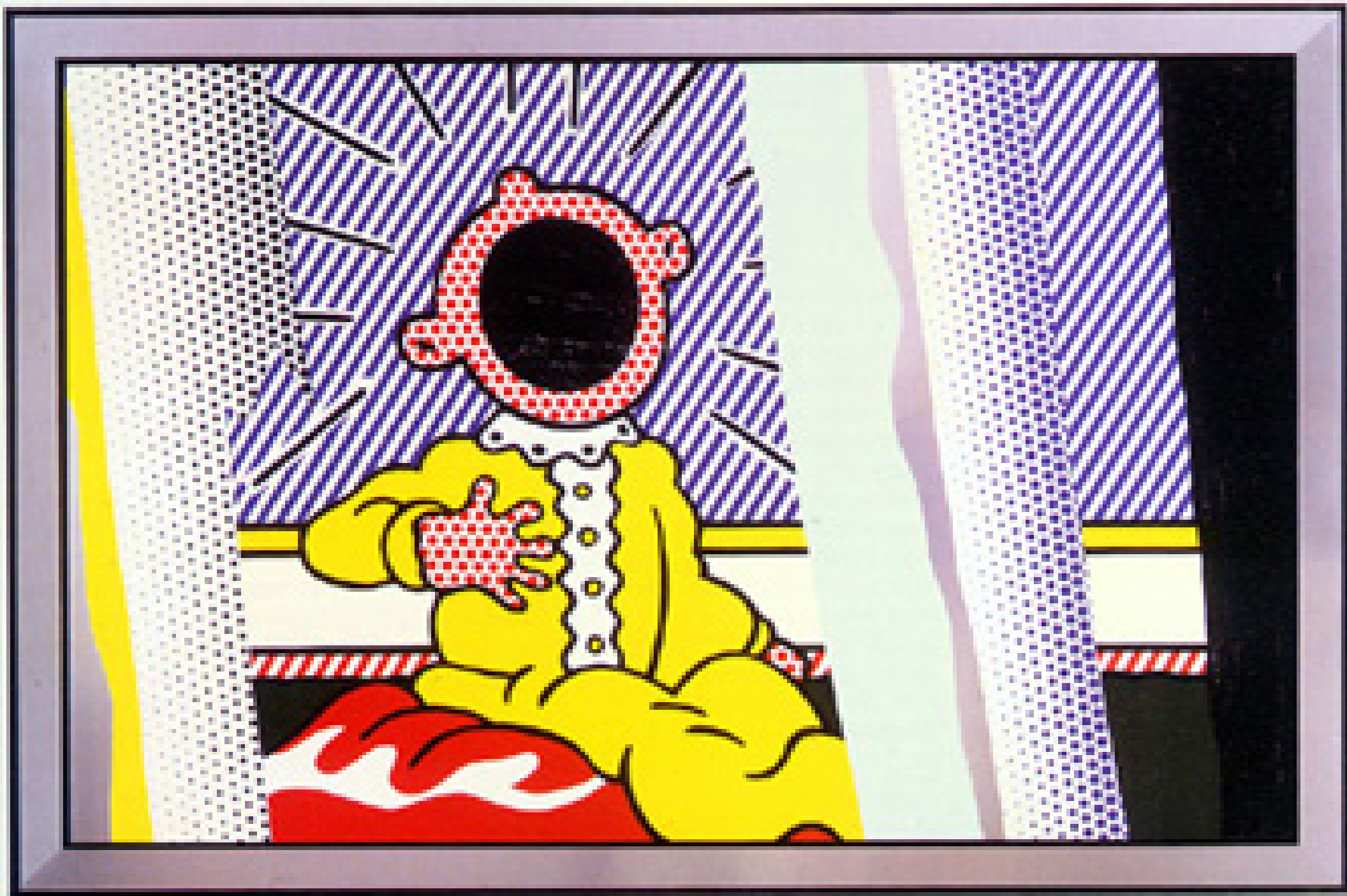
I-language

production

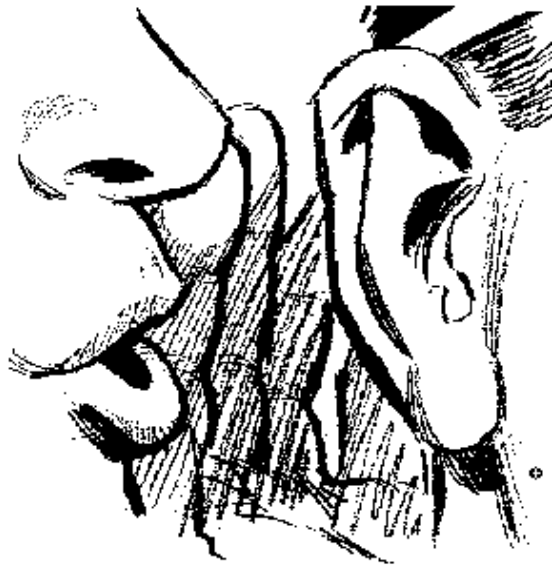
E-language

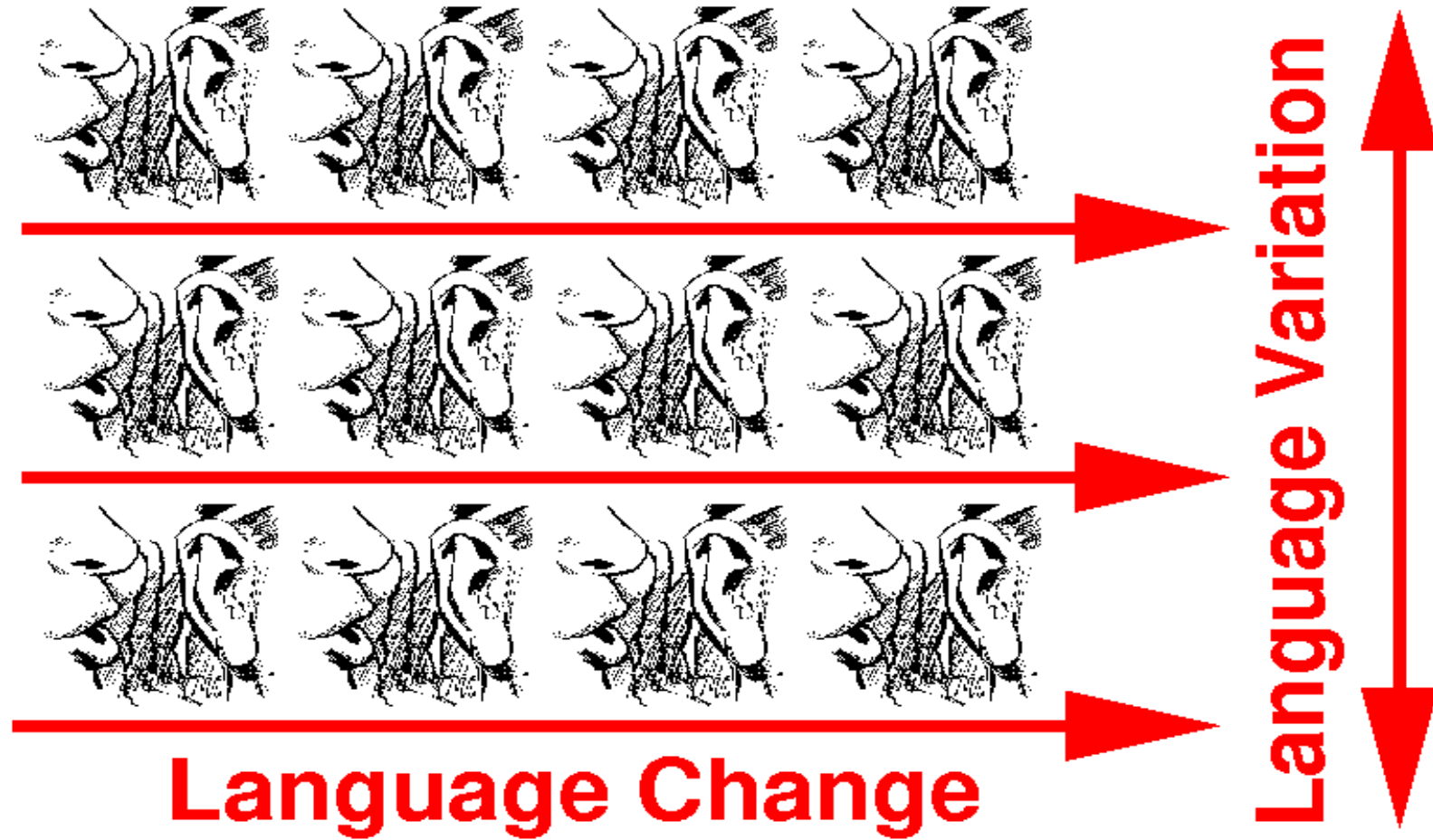
I-language

interpretation  
acquisition



## Cultural Transmission





## Goals of the talk

1. Show that it is possible & *important* to make formal models that relate the production, acquisition, change and variation of language.
2. Emphasize that it is possible & important that evolutionary modeller take psycholinguistic, historical and typological data into account.

## Linguistic Typology

**Isolating languages:** No morphological variation for tense, case or plurality; Each word typically consists of a single morpheme. E.g. Vietnamese:

*Khi tôi đến nhà bạn tôi, chúng tôi bắt đầu làm bài.*  
when I come house friend I PLURAL I begin do lesson  
“When I came to my friend’s house, we began to do lessons.”

**Agglutinating languages:** A word may consist of more than one morpheme; Boundaries between words are always clear-cut; A morpheme has a reasonably invariant shape. E.g. Turkish (*adam*: “man”):

	singular	plural
nominative	<i>adam</i>	<i>adam-lar</i>
accusative	<i>adam-i</i>	<i>adam-lar-i</i>
genitive	<i>adam-in</i>	<i>adam-lar-in</i>
dative	<i>adam-a</i>	<i>adam-lar-a</i>
locative	<i>adam-da</i>	<i>adam-lar-da</i>
ablative	<i>adam-dan</i>	<i>adam-lar-dan</i>

**Fusional languages:** No clear-cut boundary between morphemes; Expression of different categories within the same word is fused together to give a single, unsegmentable morph. E.g. Russian (*stol*: “table”, *lipa*: “lime-tree”):

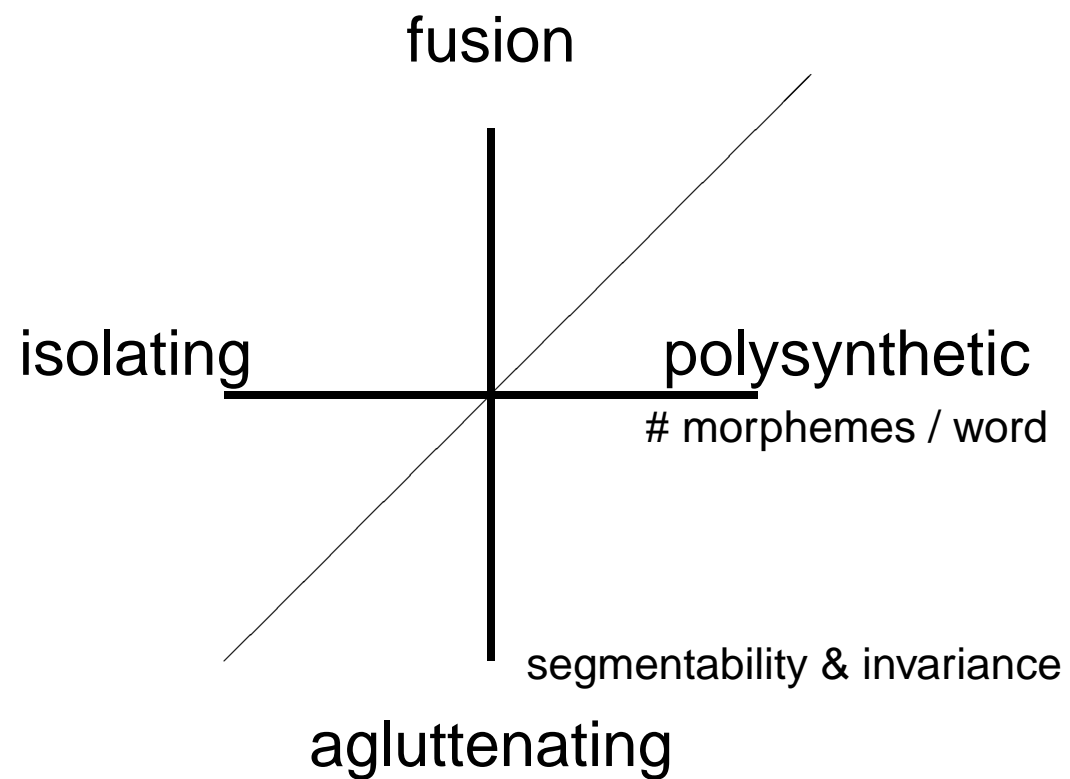
	singular I	plural I	singular II	plural II
nominative	<i>stol</i>	<i>stol-y</i>	<i>lip-a</i>	<i>lip-y</i>
accusative	<i>stol</i>	<i>stol-y</i>	<i>lip-u</i>	<i>lip-y</i>
genitive	<i>stol-a</i>	<i>stol-ov</i>	<i>lip-y</i>	<i>lip</i>
dative	<i>stol-u</i>	<i>stol-am</i>	<i>lipea</i>	<i>lip-am</i>
instrumental	<i>stol-om</i>	<i>stol-ami</i>	<i>lip-oj</i>	<i>lip-ami</i>
prepositional	<i>stol-e</i>	<i>stol-ax</i>	<i>lip-e</i>	<i>lip-ax</i>

**Polysynthetic languages:** Many lexical morphemes combined in a single word. E.g. Chukchi (Siberia):

*tε- meyhε- levte- peγt- erken*  
 great head ache 1stSINGULAR IMPERFECT  
 “I have a fierce head-ache”

(Examples from Comrie, 1981)

## Morphological typology

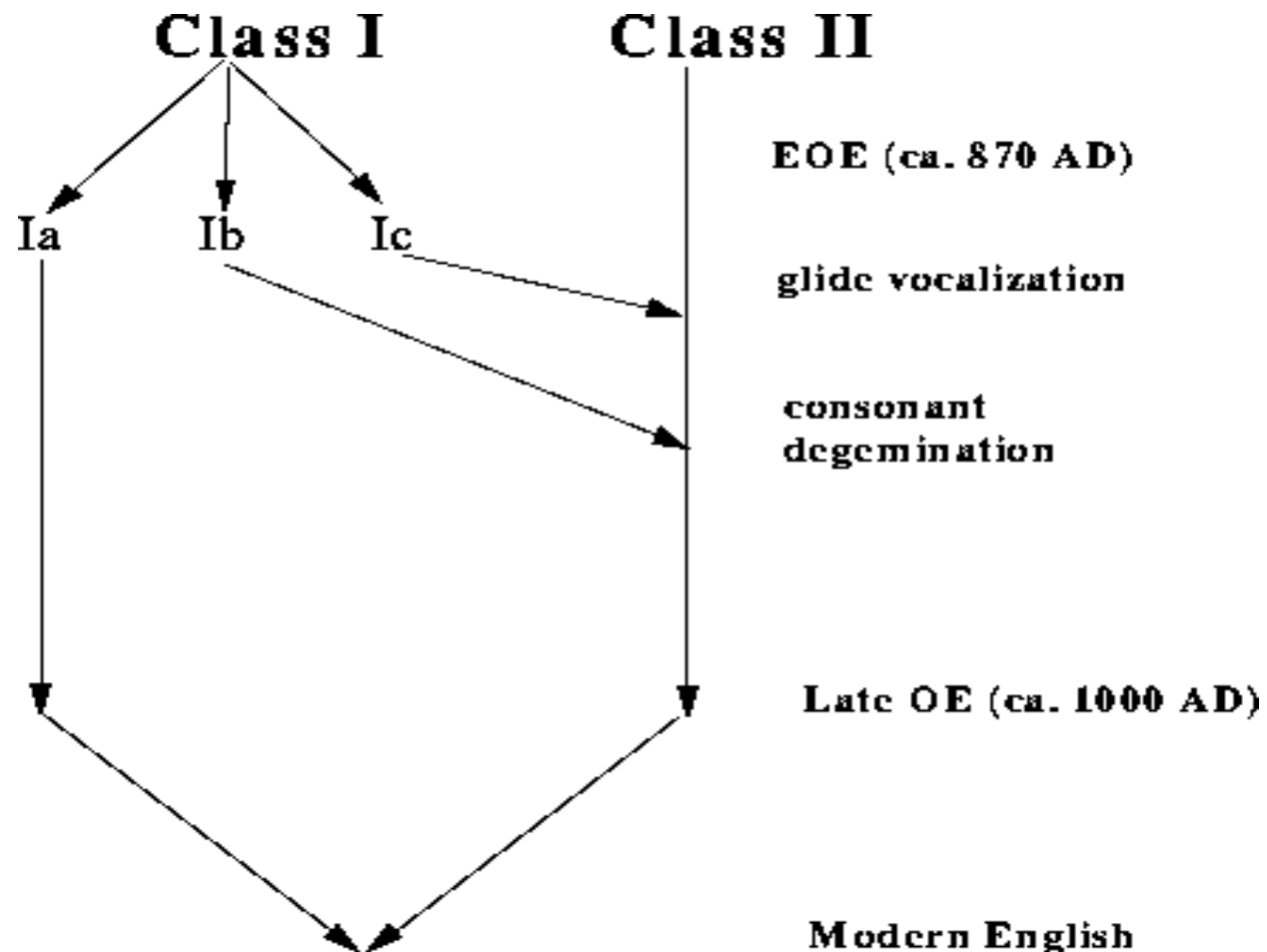


## Linguistic Typology: English past tense

- regular (use-used, look-looked)
- irregular (hit-hit, sing-sang, go-went, deal-dealt)

	regular	irregular
types	86%	14%
tokens	40%	60%

## Language Change: English past tense



## Language Acquisition: English past tense

1. Few past tense forms: said, came, went, took, knew
2. (around 29 months) Overregularization: comed, came, camed
3. Correct use of regular and irregular verbs

## Can we relate acquisition, change and typology?

- Compatibility with historical data is an additional criterion for the validity of a theory of acquisition (Hare & Elman, 1995; Niyogi & Berwick, 1995; Chang, 2000).
- Models of iterated learning might give insights in the origins of (absolute and statistical) language universals (Steels, 1998; Kirby, 1999; De Boer, 2001; Batali, 2002).

## Formalisation

- I-language
- E-language
- Production
- Acquisition
- Population
- Change
- Variation
- Constraints on variation

## I-language

(based on Nowak, Komarova and Niyogi, *Science*, 2001)

a finite space  $\mathcal{G}$  of possible grammars.

$g_1, g_2, g_3, g_4, \dots$

## E-language

languages (string sets) organised such that they fall in a finite number of classes. Assume the similarity between languages is  $a$ .

$$l_1, l_2, l_3, l_4, \dots$$

## Production

a mapping from I- to E-language.

$$S = \left( \begin{array}{c|cccc} & l_1 & l_2 & l_3 & l_4 \\ \hline g_1 & 1 & 0 & 0 & 0 \\ g_2 & 0 & 1 & 0 & 0 \\ g_3 & 0 & 0 & 1 & 0 \\ g_4 & 0 & 0 & 0 & 1 \end{array} \right)$$

## Language Acquisition

a precise algorithm to induce a grammar from text; i.e. a (probabilistic) mapping from E- to I-language.

$$A = \left( \begin{array}{c|cccc} & g_1 & g_2 & g_3 & g_4 \\ \hline l_1 & q & r & r & r \\ l_2 & r & q & r & r \\ l_3 & r & r & q & r \\ l_4 & r & r & r & q \end{array} \right),$$

where  $q$  is the learning accuracy, and  $r = (1 - q)/(N - 1)$ .

E.g. the MEMORY-LESS LEARNER receives examples sentences, and checks if they are consistent with the presently hypothesized grammar. If not, it changes the hypothesis to a random other grammar.

Can we calculate the probability that the algorithm has found the right grammar (out of  $N$  possible ones) after  $b$  example sentences?

$$\begin{aligned} q_{\text{memoryless}} &= 1 - \frac{(N-1)}{N} \left( a + \frac{(N-2)(1-a)}{N-1} \right)^b \\ &= 1 - \frac{(N-1)}{N} \left( 1 - \frac{(1-a)}{N-1} \right)^b \end{aligned}$$

The BATCH LEARNER, in contrast, memorizes all received sentences and finds all grammars from the set of possible ones that are consistent with these sentences.

The probability that the batch learner has found the correct grammar after  $b$  input sentences is found by Nowak et al. (2001) to be

$$q_{batch} = \frac{\left(1 - (1 - a^b)^N\right)}{(Na^b)} \quad (1)$$

## Population

a set of grammars, or a vector of relative frequencies  $\vec{x}$ . E.g.:

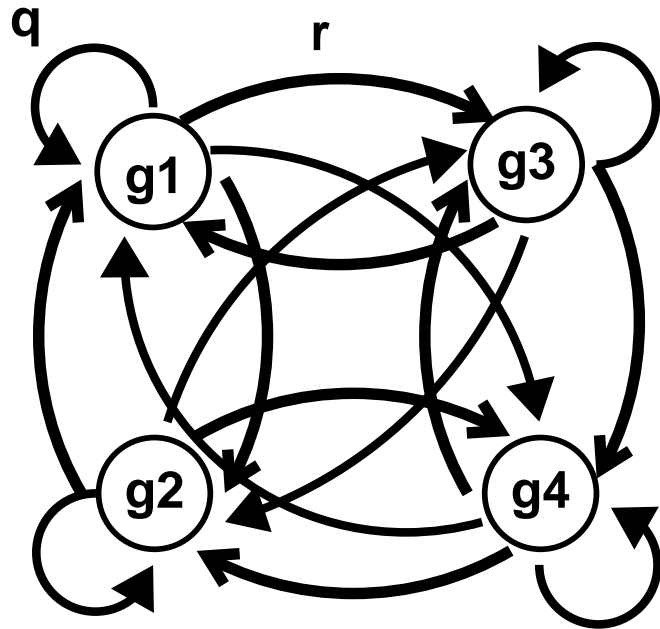
$$\vec{x}(t) = \begin{pmatrix} g_1 & g_2 & g_3 & g_4 \\ .3 & .2 & .4 & .1 \end{pmatrix}$$

## Change

a function of the population's production and acquisition

$$Q = S \cdot A = \left( \begin{array}{c|cccc} & g_1 & g_2 & g_3 & g_4 \\ \hline g_1 & q & r & r & r \\ g_2 & r & q & r & r \\ g_3 & r & r & q & r \\ g_4 & r & r & r & q \end{array} \right)$$
$$\Delta x_i = \sum_{j=0}^N x_j Q_{ji}$$

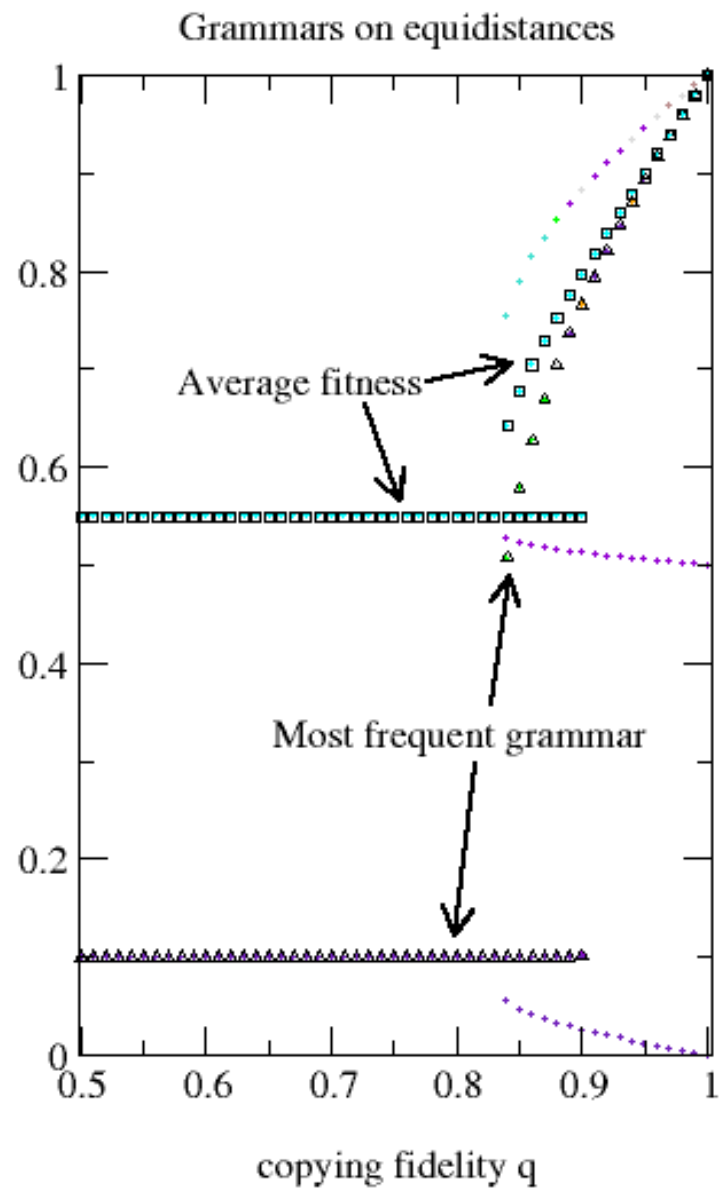
## Variation



## Natural Selection

$$\Delta x_i = \sum_{j=0}^N x_j f_j Q_{ji} - \phi x_i$$
$$f_i = \sum_j x_j F_{ij}$$

# Coherence Threshold



## Constraints on variation

1. There is a minimum learning accuracy  $q_1$  required for coherence in the population;
2.  $q$  is a function of  $N$  and  $b$ ;
3. An upper bound estimate of  $q_0$  implies a minimum  $b$  (poverty of the stimulus) and a maximum  $N$  (Universal Grammar).

## Problems: How to incorporate empirical observations?

**Production:** semantically/pragmatically salient words and interactions with language change;

**Acquisition:** gradual built-up of an infant's linguistic knowledge

**Change:** phonological erosion & interaction with syntactic change; frequency-effects; grammaticalization;

**Variation:** universal tendencies (statistical universals).

## 2nd Formalisation: Niyogi & Berwick, 1995

**I-language:** parametrized grammar, i.e. a set of parameters

**E-language:** set of triggers for each parameter

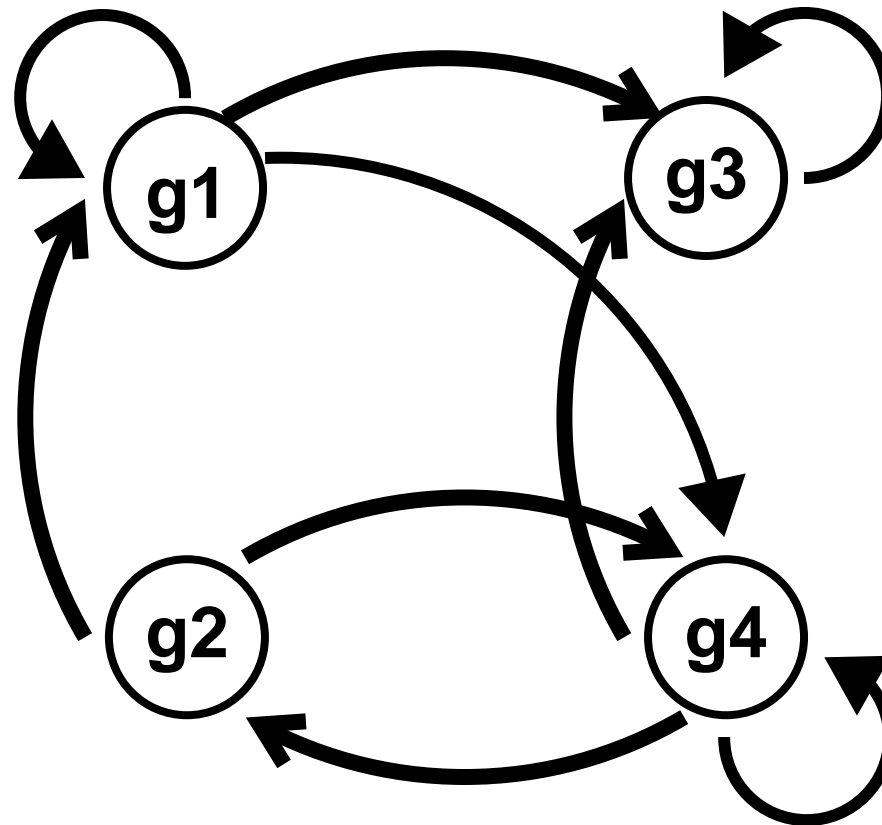
**Production:** specified frequencies of each trigger

**Acquisition:** Trigger Learning Algorithm (Wexler & Culicover, 1981)

## Language change in a 3-parameter model

$$Q = \begin{pmatrix} & \begin{array}{c|cccccccc} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ \hline 000 & 0 & 0.85 & 0 & 0 & 0 & 0.1 & 0 & 0 \\ 001 & 0 & 0.98^* & 0 & 0 & 0 & 0 & 0 & 0 \\ 010 & 0 & 0 & 0 & 0 & 0 & 0.48 & 0 & 0.38 \\ 011 & 0 & 0 & 0 & 0.86^* & 0 & 0 & 0 & 0 \\ 100 & 0 & 0.97 & 0 & 0 & 0 & 0 & 0 & 0 \\ 101 & 0 & 0 & 0 & 0 & 0 & 0.92^* & 0 & 0 \\ 110 & 0 & 0.54 & 0 & 0.35 & 0 & 0 & 0 & 0 \\ 111 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.97^* \end{array} \\ \end{pmatrix}$$

## Cultural Evolution



### 3d Formalisation: Batali, 2002

**I-language:** explicit exemplar-based grammar formalism, not unlike probabilistic tree grammars with predicate logic;

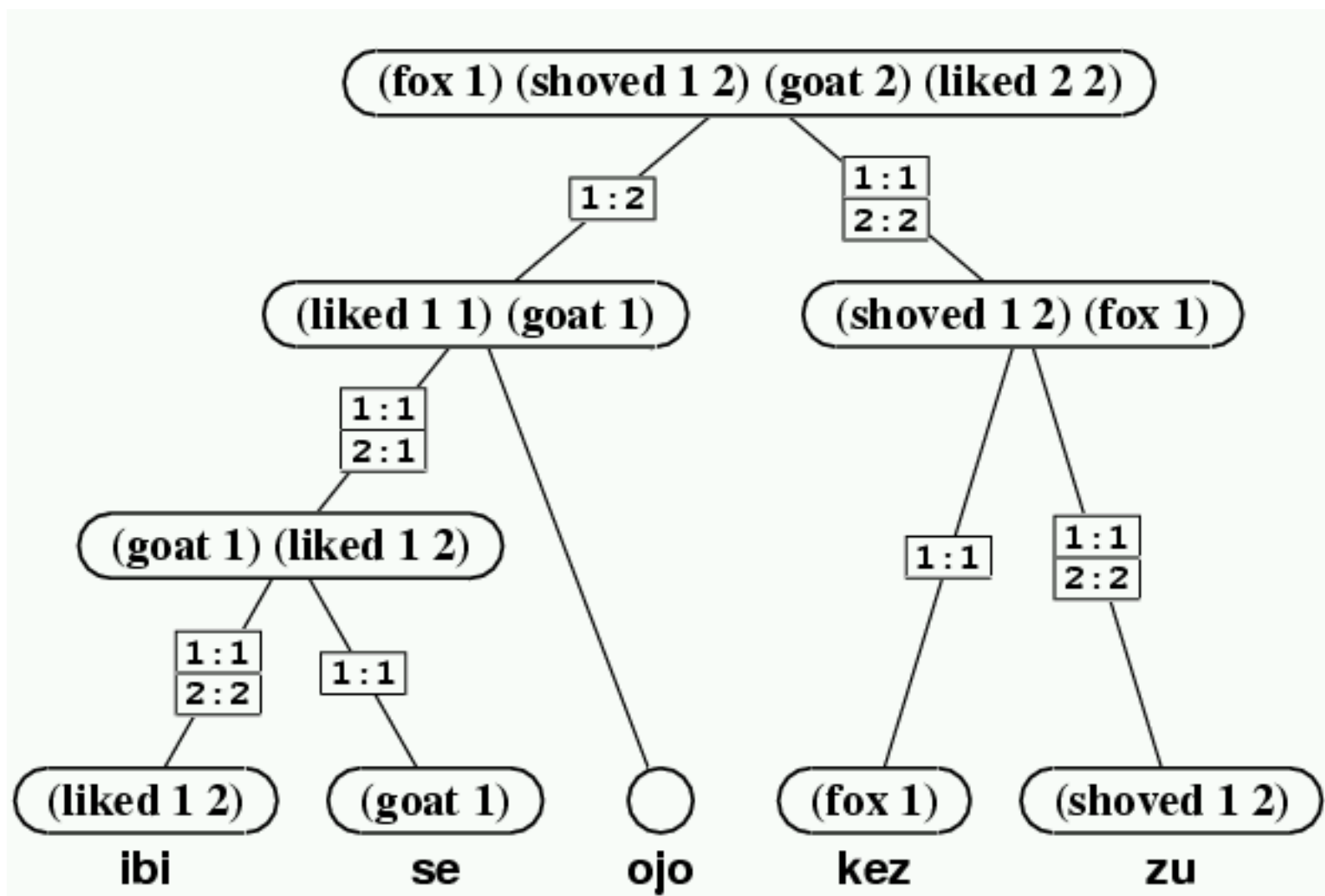
**E-language:** sentence – meaning pairs;

**Production:** most probable derivation for randomly picked meaning.

**Acquisition:** frequency counts; split & merge driven by failures in parsing.

## Results

- The emergence of a complex language, with properties similar to case marking and subordinate clause marking in natural languages.
- The emergent languages are essentially infinite but nevertheless learnable (from meaning–form pairs).



## Conclusions

- Formal models of the production and acquisition of language can be extended to model language change;
- What is Universal Grammar? In asymmetric models:

$$\textit{stable} \subset \textit{learnable} \subset \textit{representable}$$

- Empirical facts from historical linguistics and typology, force us to modify models of language change, and indirectly even models of acquisition and production.