Efficient Communication and Language Evolution

Paul Kay

Fifth International Conference on Evolutionary Linguistics

17-19 August 2013

The Chinese University of Hong Kong

Collaborative work with Terry Regier and Charles Kemp

The role of communication in the evolution of language is currently a contested issue.

"The use of language for communication might turn out to be a kind of epiphenomenon**…** If you want to make sure that we never misunderstand one another, for that purpose language is not well designed, because you have such properties as ambiguity. If we want to have the property that the things that we usually would like to say come out short and simple, well, it probably doesn't have that property" (**Chomsky** 2002, 107).

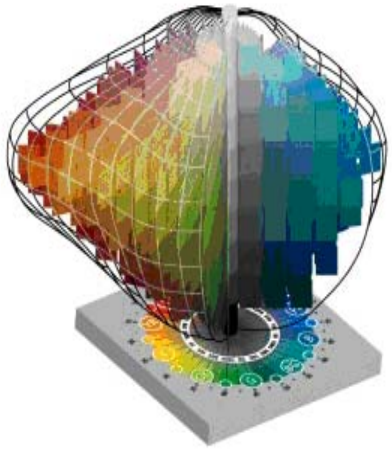
"Here, we argue that this perspective on ambiguity is exactly backwards. We argue, contrary to the Chomskyan view, that ambiguity is in fact a desirable property of communication systems, precisely because it allows for a communication system which is ''short and simple'' **(Piantadosi, Tily, & Gibson** 2011, 281).

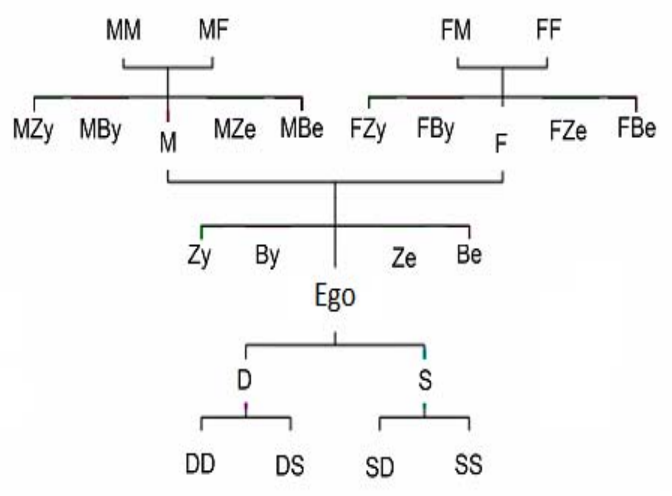
Chomsky, N. (2002) An interview on minimalism. *Noam Chomsky, On Nature and Language*, Cambridge University Press, Cambridge, 92–161.

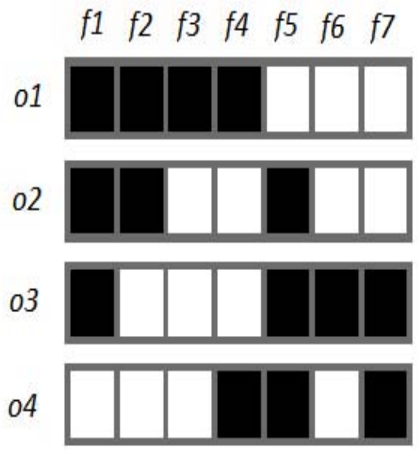
Piantadosi, S. T., Tily, H. & Gibson, E. (2011) The communicative function of ambiguity in language. *Cognition,* 122, 280-291.

Piantadosi et al. present a formal, information-theoretic argument
that a language containing ambiguities that are resolvable in context
supports more ***efficient communication*** than a language without ambiguity.

Plan of this talk*: Efficient communication* in three lexical domains.







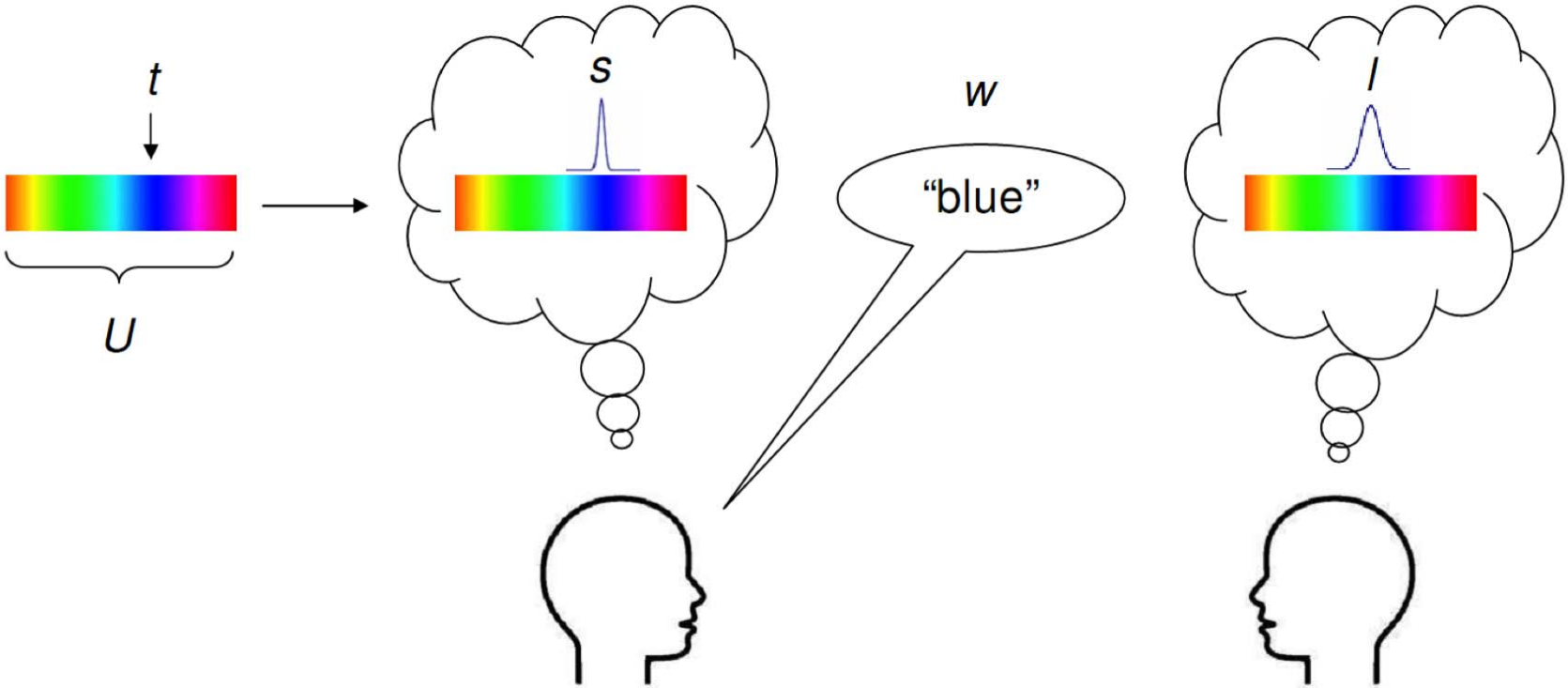
Three domains, with different formal structures: color – continuous space, kinship – discrete relational hierarchy; fruit names – binary feature vectors.

EFFICIENT COMMUNICATION: INFORMATIVENESS versus SIMPLICITY

To be efficient a system of categories must be *informative*, but also *simple* – easy to learn and remember. Informativeness and simplicity tend to be negatively related.

Hypothesis: lexical category domains tend toward an optimum compromise between *informativeness* (minimizing information loss) and *simplicity* (minimizing complexity).

Introducing the model



A scenario illustrating informative communication.

The difference between the listener's and the speaker's distribution represents the information lost in communicating using this category system. The Kullback-Leibler (KL) divergence, is a standard measure of the dissimilarity between two probability distributions *s* and *l*:

$$e(t) = \sum_{i \in U} s(i) \log\left(\frac{s(i)}{l(i)}\right) \qquad (1)$$

We are assuming that the speaker has a particular individual domain member *t* in mind. In this case *s*(*i*) = 1 for (*i* = *t*) and 0 for every other domain member.

$$e(t) = \sum_{i \in U} s(i) \log\left(\frac{s(i)}{l(i)}\right) = \log\left(\frac{1}{l(t)}\right) = -\log(l(t)) \qquad (2)$$

Total **COST** of a LEXICAL domain is defined as the sum over the items *t* in the domain of the "need" probability *n*(*t*) that the speaker will wish to communicate about *t* times the information lost when communicating about *t*, *e(t)*.

COST: Sum of product of need probability *n(t)* and the information loss *e*(*t*) = –log(*l*(*t*))

$$E = \sum_{t \in U} n(t)\, e(t) \qquad (3)$$

$$= \sum_{t \in U} n(t)(-\log(l(t))) \qquad (4)$$

COMPLEXITY: Number of terms in domain or a more domain-specific measure of complexity.

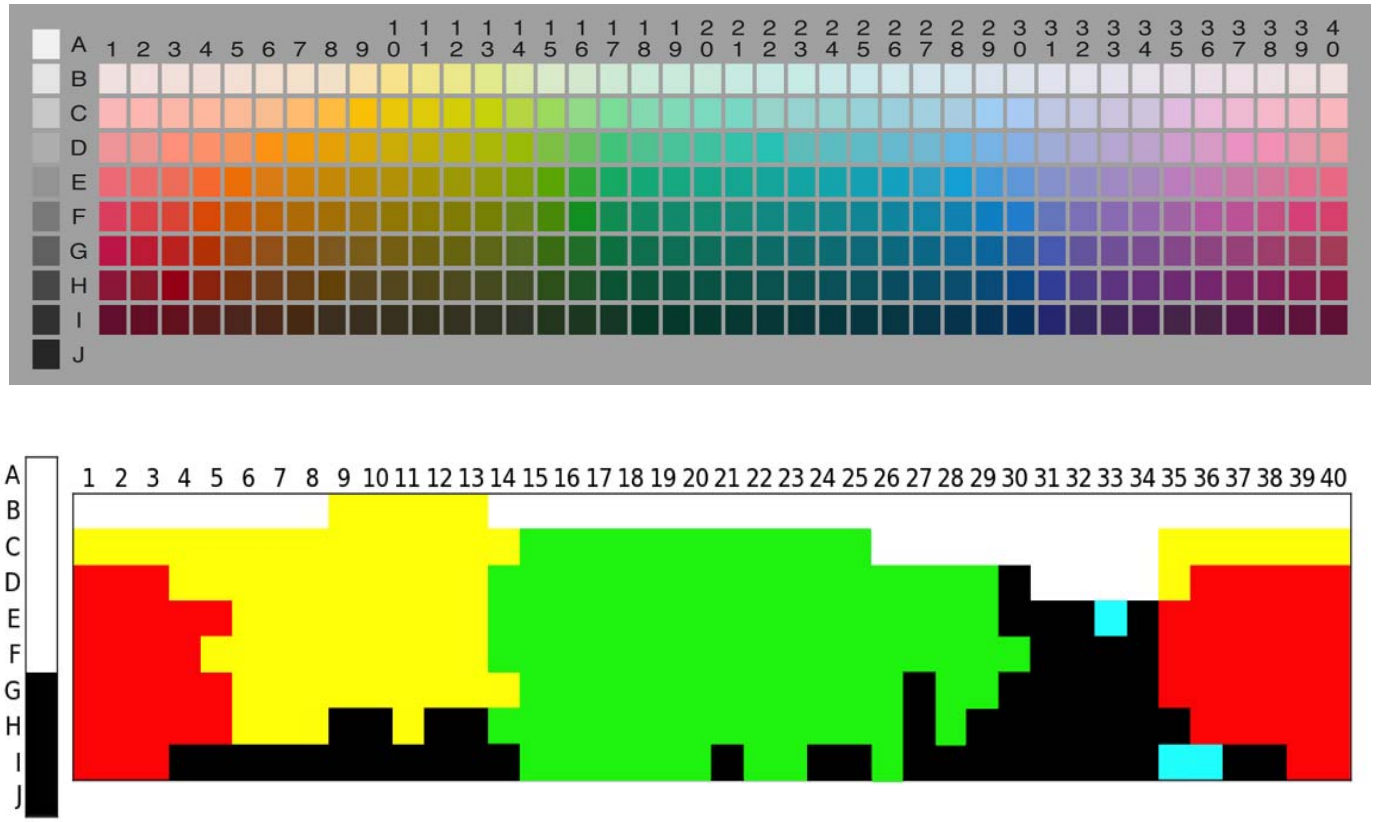
In each of our three case studies, to calculate COST we will have to specify the speaker's and the listener's distributions as well as the need probabilities. We will also have to provide a measure of COMPLEXITY. The general model just developed will apply in each case.

Case study 1: Color, a continuous domain.

World Color Survey*

*Kay, P., Berlin, B. Maffi, L. Merrifield, W.& Cook, R. (2009) *The World Color Survey.* Stanford: CSLI Publications. Data on line at www.icsi.berkeley.edu/wcs/data.html.

Color naming from 110 unwritten languages, average 24 speakers/language.



(Upper panel) Color naming stimulus grid. (Lower panel) *Mode map* for the Iduna language (Austronesian, Papua New Guinea), mapped against the stimulus grid.

Developing the listener's distribution…

$$l(i) \propto \sum_{j \in cat(w)} sim(i, j) \qquad (5)$$

Developing the listener's distribution…

$$l(i) \propto \sum_{j \in cat(w)} sim(i, j) \qquad (5)$$

Assume similarity is a Gaussian function of ΔE distance in CIELAB space.

$$sim(x, y) = \exp(-c \times dist(x, y)^2) \qquad (6)$$

($c = .001$)

Developing the listener's distribution…

$$l(i) \propto \sum_{j \in cat(w)} sim(i, j) \qquad (5)$$

Assume similarity is a Gaussian function of ΔE distance in CIELAB space.

$$l(i) = \exp(-c \times dist(x, y)^2) \qquad (6)$$

($c = .001$)

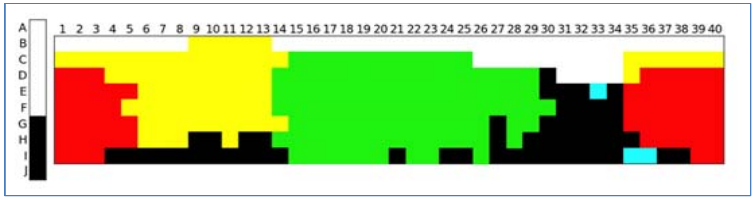
For *i* in category *w*

$$l(i) = \sum_{j \in cat(w)} \exp(-.001 \times dist(i, j)^2) \qquad (7)$$

We now have the information to calculate divergence between the speakers' and listeners' distributions. We assume uniform need probabilities over color space and calculate the COST *E* for each system in the WCS.

We wish to compare for each level of complexity (=number of color terms) what the cost is compared to what it might have been had the categories been different.

1. “Regularize” all the 110 mode maps by eliminating minor terms, as follows. For all chips labeled by a term that names fewer than 10 chips, reassign that chip to the category of the closest chip of a major term. (Relabeled 1.6% of chips in a mode map on average.) Result: all regularized mode maps have between 3 and 11 terms.

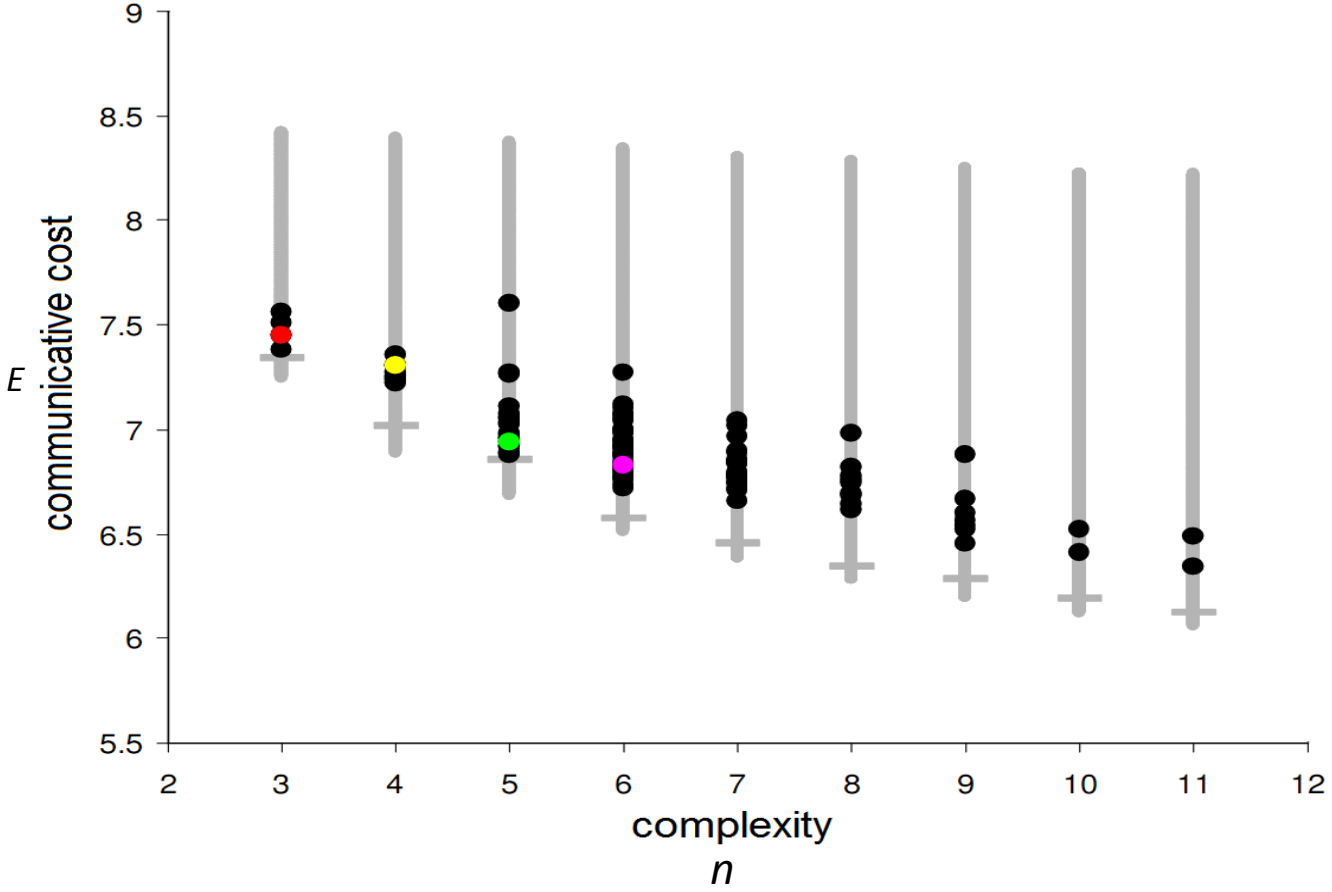


Iduna

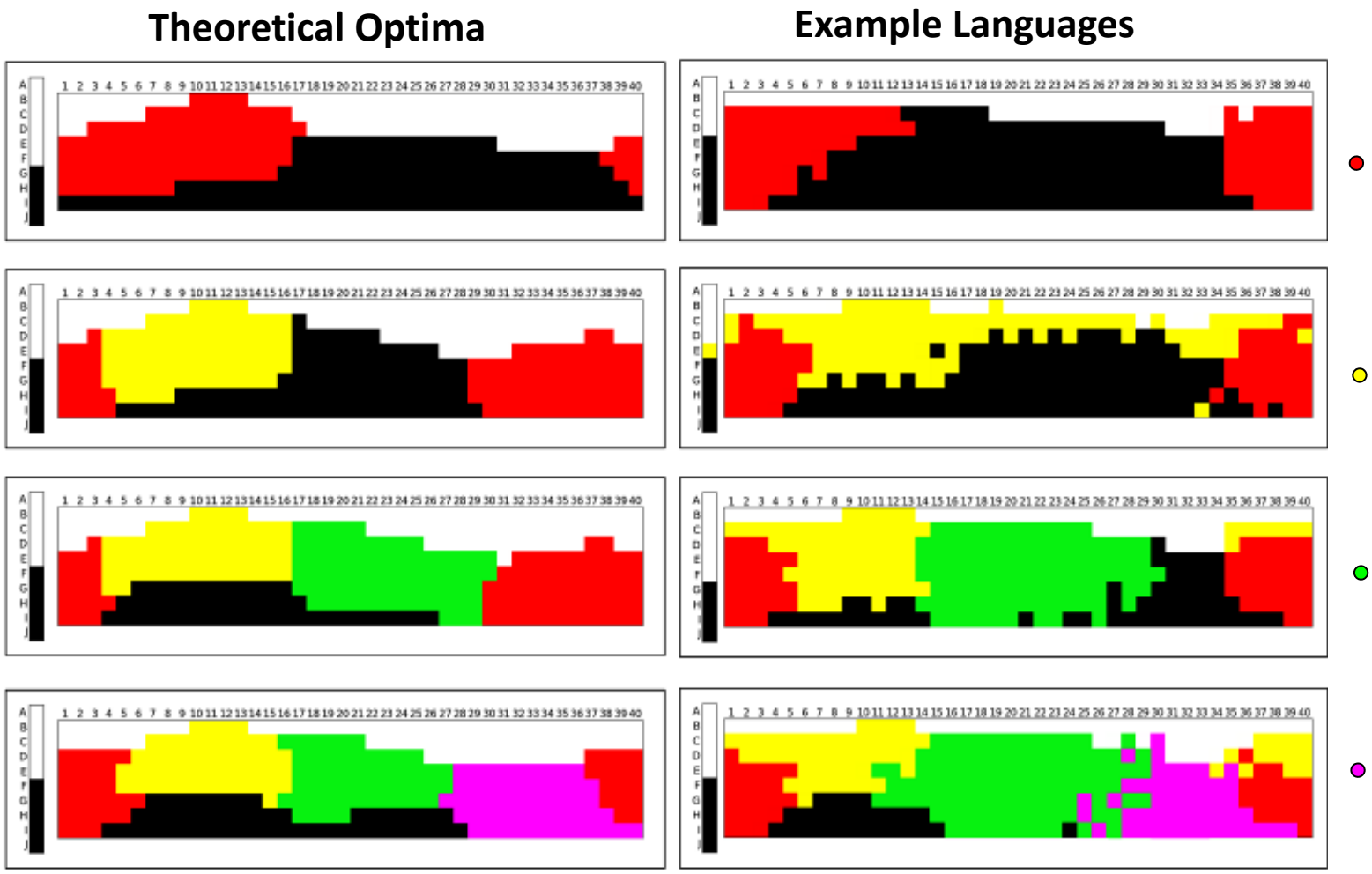

Iduna regularized

2. For each number *n* of categories, perform 20 of the following simulations. Start with a random assignment of chips to the *n* categories Repeatedly reassign category labels to chips to reduce COST *E* until no further reduction in *E* is possible. (“Steepest descent” in *E*)

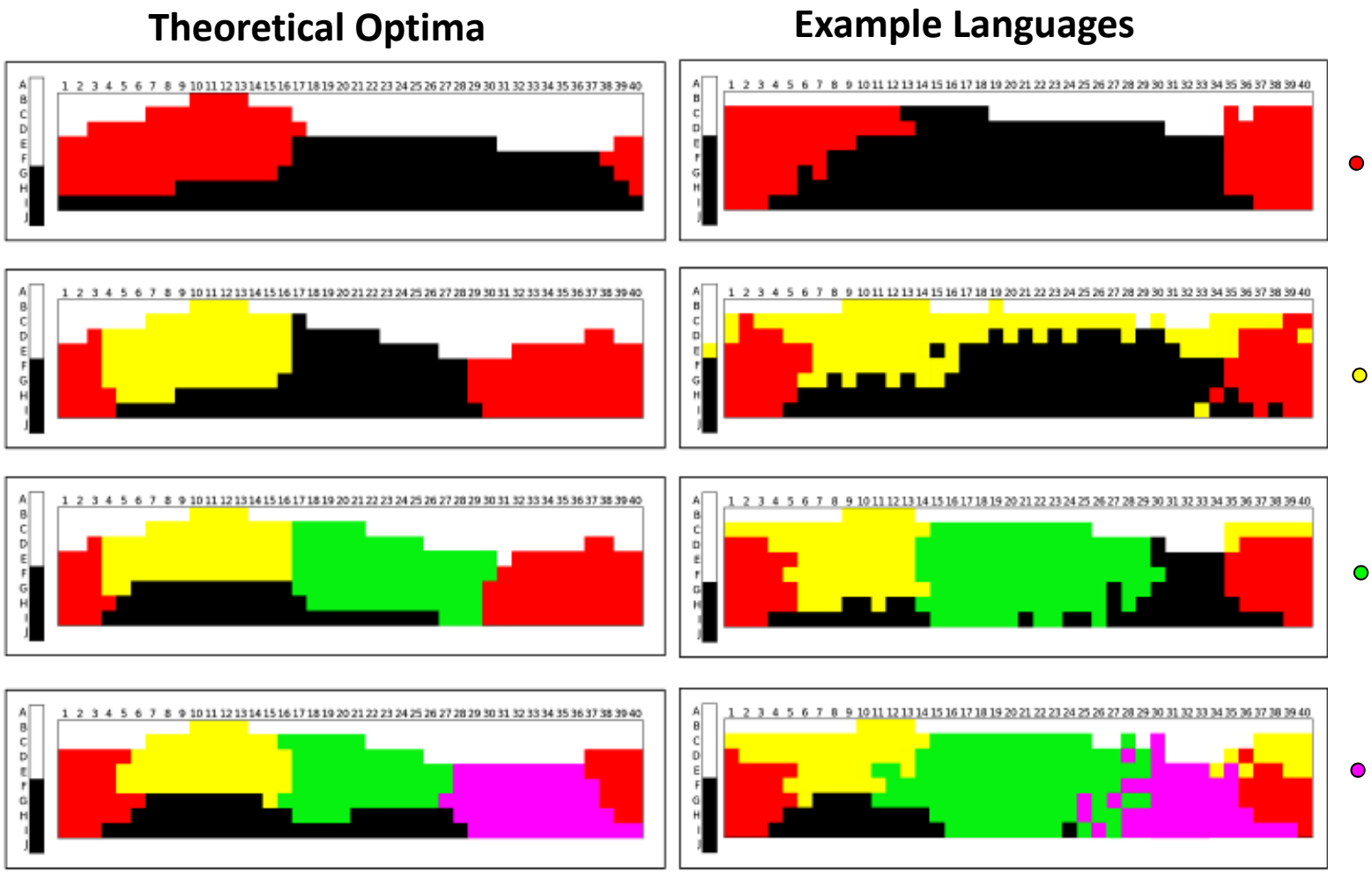
3. Record all values of *E* found in all simulations, also the highest local minimum, i.e., the maximum final state of the simulation.



Communicative cost (expected reconstruction error *E*) vs. complexity (number of color terms *n*) for hypothetical (range shown by gray vertical bars) and all WCS-attested (black and colored dots) color naming systems. Hypothetical systems were those encountered during optimization. Crossbars show the highest local minimum encountered at each level of complexity. Colored dots show specific WCS languages illustrated in detail below.



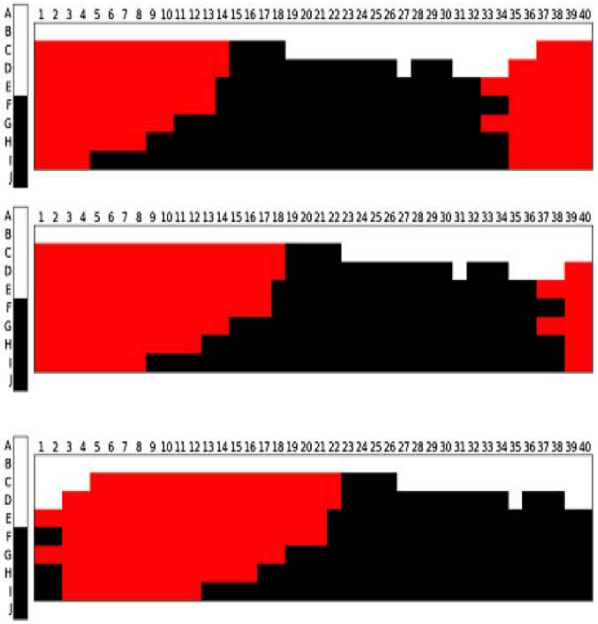
Theoretical optima (left) for n=3,4,5,6 categories, compared with color naming systems (right, top to bottom) of Ejagam (Bantoid, Nigeria/Cameroon), Culina (Arawan, Peru/Brazil), Iduna (Austronesian, Papua New Guinea), and Buglere (Chibchan, Panama)

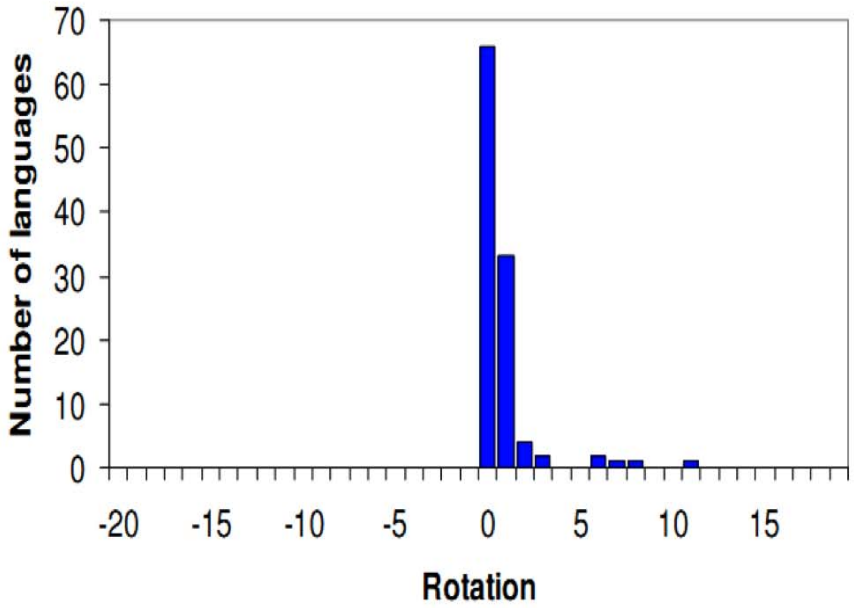


Note: the theoretical optima for 3, 4, 5, and 6 term systems seem to track rather closely the Berlin and Kay evolutionary model, suggesting a motivation for that developmental path.

Many more languages closely approximate the theoretical optima.
But a substantial number of others don't.

This suggests a language-by-language test of how individual languages fare in cost compared to reasonable hypothetical alternatives, i.e. to languages with the same shape map but which are located differently in color space.





Left: The color naming system of Wobé (Niger-Congo, Côte d'Ivoire), shown unrotated (top) and rotated 4 (middle) and 8 (bottom) hue columns. Right: For each amount of rotation, the number of WCS languages exhibiting maximum informativeness (minimum cost) at that rotation. For most languages, the unrotated variant (0 columns rotation) is most informative.

Summary of case study 1.

1. Languages with a given number of color terms tend strongly toward the lower end of the scale of possible costs for that number of terms.

2. Languages with more terms, higher complexity, can and do achieve lower costs, reinforcing the tradeoff hypothesis.

3. WCS languages tend strongly to have lower communicative cost than hypothetical languages with mode maps of the same shape but different locations in color space.

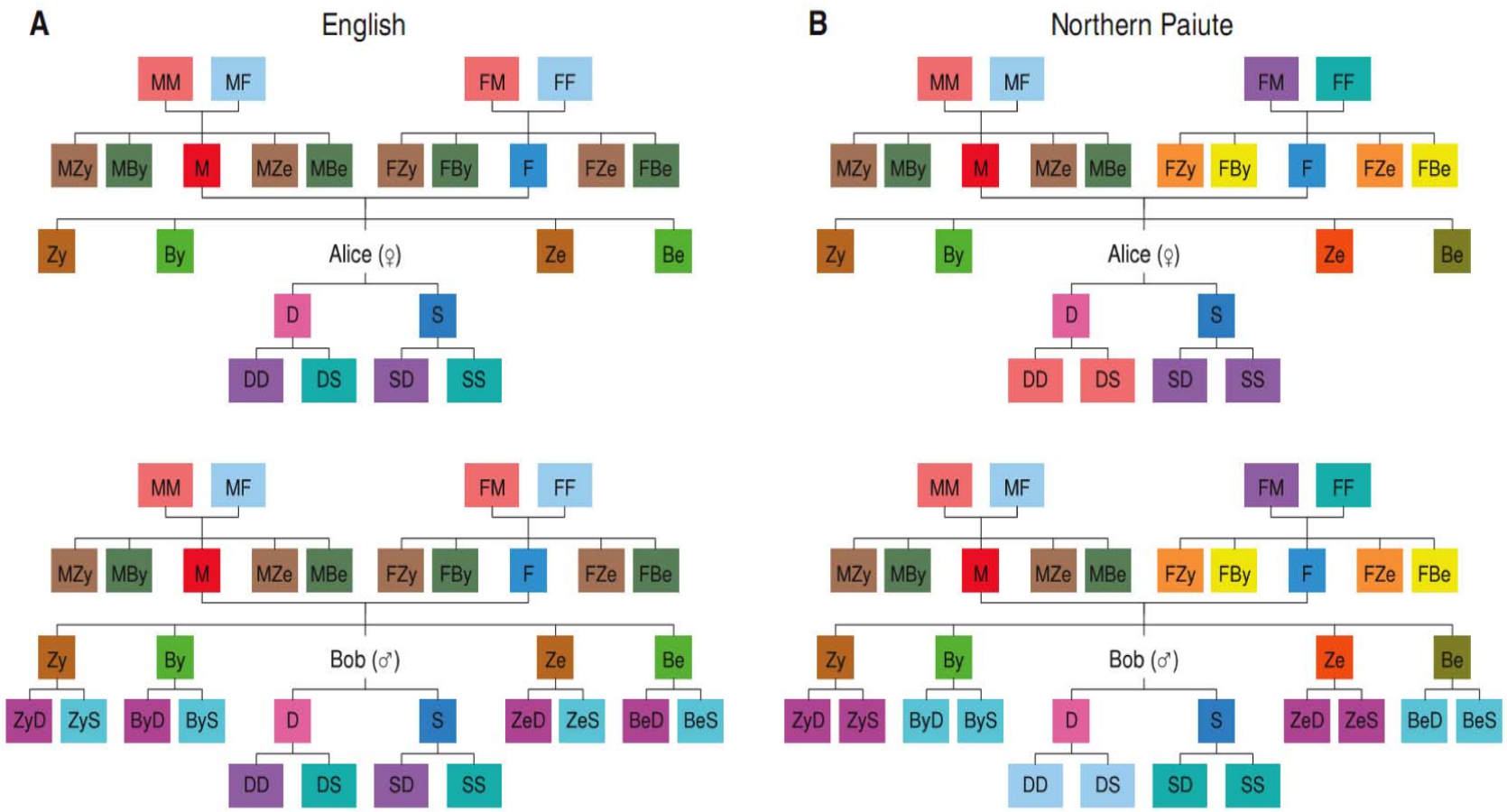
4. Hypothetically cost-optimal languages with 3-6 major terms approximate the ideal mode maps of the (updated) Berlin-Kay evolutionary sequence.

Case study 2: Kinship, a discrete and hierarchically structured domain.

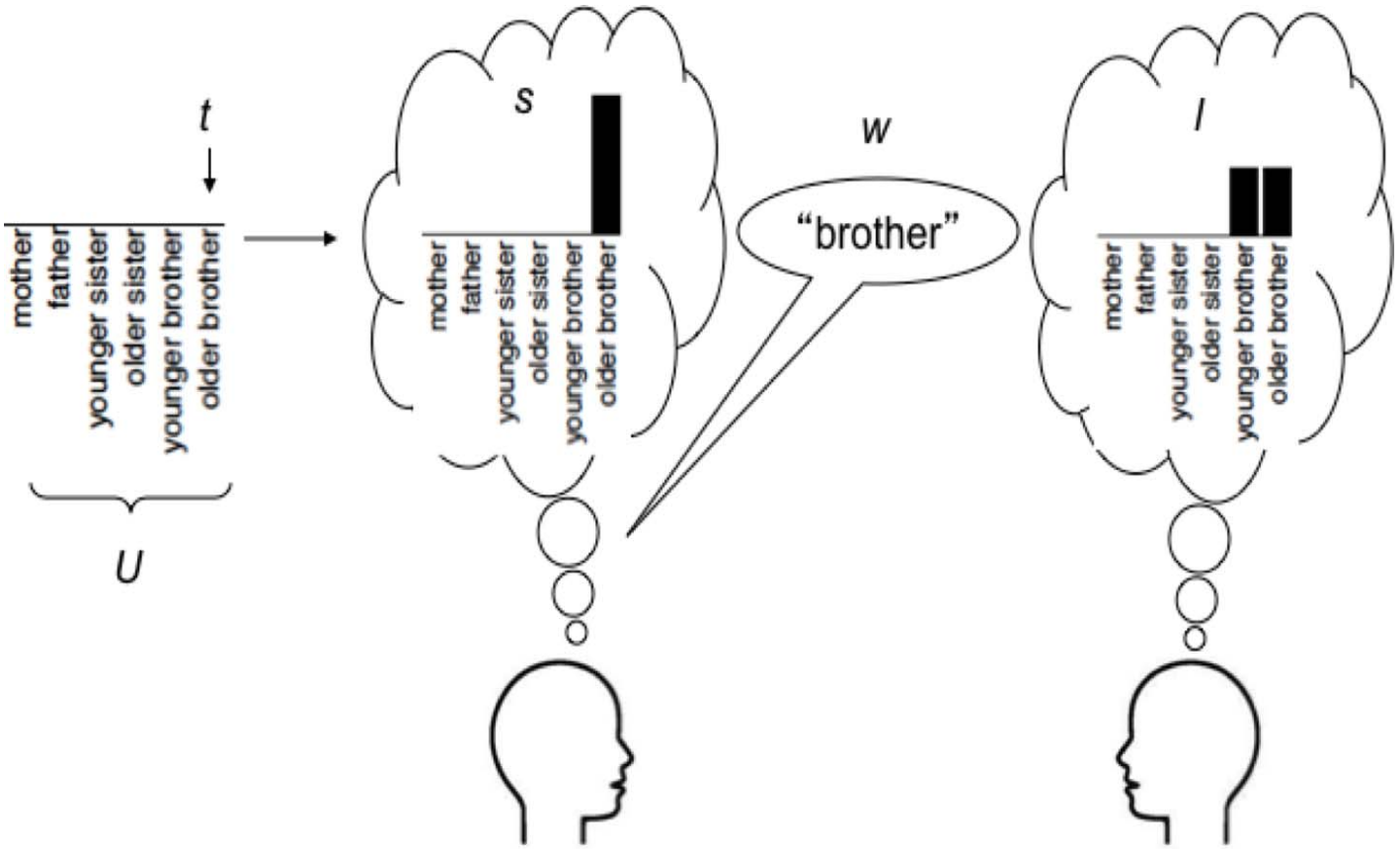
Earlier, related study: Kemp, C. & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science, 336*, 1049-1054.

Finding: Kinship terminology systems tend to maximize similarity of within-category exemplars.

Data: 487 kinship terminology systems described Murdock, G. (1970) Kin term patterns and their distribution, *Ethnology 9*, 165-208.



The kin naming systems of (a) English and (b) Northern Paiute. Color codes denote the extensions of kin terms in these languages. Adapted from Kemp & Regier (2012).



A scenario illustrating communication about a kin type

COST: Sum of product of need probability *n(t)* and the information loss *e*(*t*) = –log(*l*(*t*))

$$E = \sum_{t \in U} n(t)\, e(t) \qquad (3)$$

$$= \sum_{t \in U} n(t)(-\log(l(t))) \qquad (4)$$

The listener knows the category the speaker has in mind so he or she only has interest in the probability weights for the kintypes of that category (here *brother*). Accordingly the listener's probability mass *l*(*i*) for kintype *i* in category *w* is assumed equal to the proportion of the total need probability for *w* that is assigned to *i*.

$$l(i) = \frac{n(i)}{\sum_{j \in cat(w)} n(j)}$$

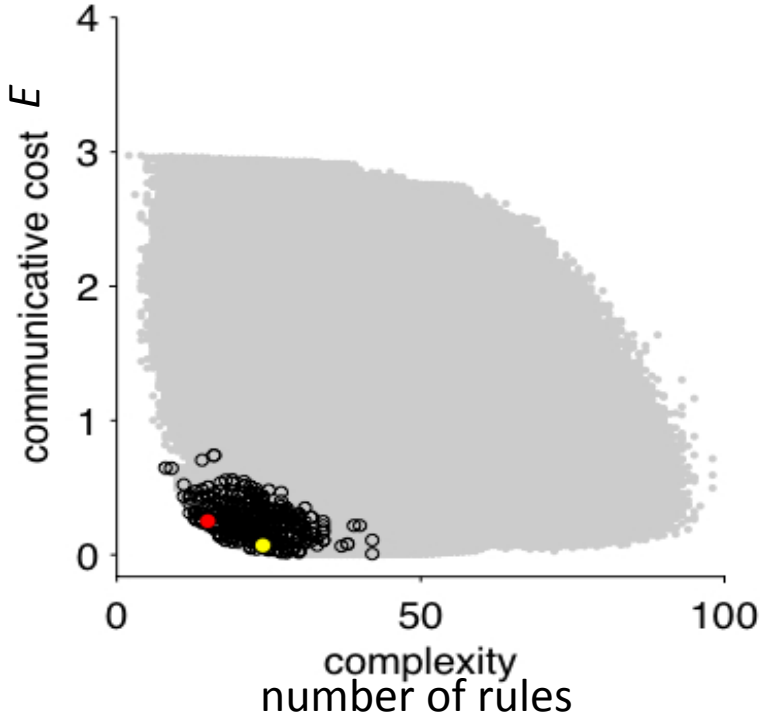
(The need probability *n*(*i*) for referring to a given kin type *i* was estimated through corpus counts for kin terms in English and German.)

COMPLEXITY: Number of terms in domain or a more domain-specific measure of complexity.

Complexity in kinship: number of rules

$$
\begin{aligned}
&\texttt{mother}(x,y) && \leftrightarrow \texttt{PARENT}(x,y) \wedge \texttt{FEMALE}(x)\\
&\texttt{father}(x,y) && \leftrightarrow \texttt{PARENT}(x,y) \wedge \texttt{MALE}(x)\\
&\texttt{daughter}(x,y) && \leftrightarrow \texttt{CHILD}(x,y) \wedge \texttt{FEMALE}(x)\\
&\texttt{son}(x,y) && \leftrightarrow \texttt{CHILD}(x,y) \wedge \texttt{MALE}(x)\\
&\texttt{sister}(x,y) && \leftrightarrow \exists z\, \texttt{daughter}(x,z) \wedge \texttt{PARENT}(z,y)\\
&\texttt{brother}(x,y) && \leftrightarrow \exists z\, \texttt{son}(x,z) \wedge \texttt{PARENT}(z,y)\\
&\texttt{sibling}(x,y) && \leftrightarrow \exists z\, \texttt{CHILD}(x,z) \wedge \texttt{PARENT}(z,y)\\
&\texttt{aunt}(x,y) && \leftrightarrow \exists z\, \texttt{sister}(x,z) \wedge \texttt{PARENT}(z,y)\\
&\texttt{uncle}(x,y) && \leftrightarrow \exists z\, \texttt{brother}(x,z) \wedge \texttt{PARENT}(z,y)\\
&\texttt{niece}(x,y) && \leftrightarrow \exists z\, \texttt{daughter}(x,z) \wedge \texttt{sibling}(z,y)\\
&\texttt{nephew}(x,y) && \leftrightarrow \exists z\, \texttt{son}(x,z) \wedge \texttt{sibling}(z,y)\\
&\texttt{grandmother}(x,y) && \leftrightarrow \exists z\, \texttt{mother}(x,z) \wedge \texttt{PARENT}(z,y)\\
&\texttt{grandfather}(x,y) && \leftrightarrow \exists z\, \texttt{father}(x,z) \wedge \texttt{PARENT}(z,y)\\
&\texttt{granddaughter}(x,y) && \leftrightarrow \exists z\, \texttt{daughter}(x,z) \wedge \texttt{CHILD}(z,y)\\
&\texttt{grandson}(x,y) && \leftrightarrow \exists z\, \texttt{son}(x,z) \wedge \texttt{CHILD}(z,y)
\end{aligned}
$$

The shortest description of the English kin naming system in the representation language of Kemp & Regier (2012)



Communicative cost (expected reconstruction error, *E*) vs. complexity (number of rules) for hypothetical (gray mass) and attested (black and colored circles) kin naming systems. Colored circles show the kin naming systems of English (red) and Northern Paiute (yellow).

Summary of Case Study 2:

1. Attested kinship terminology systems tend to be minimal in cost for their level of complexity.

2. In attested systems cost tends to decrease as complexity increases.

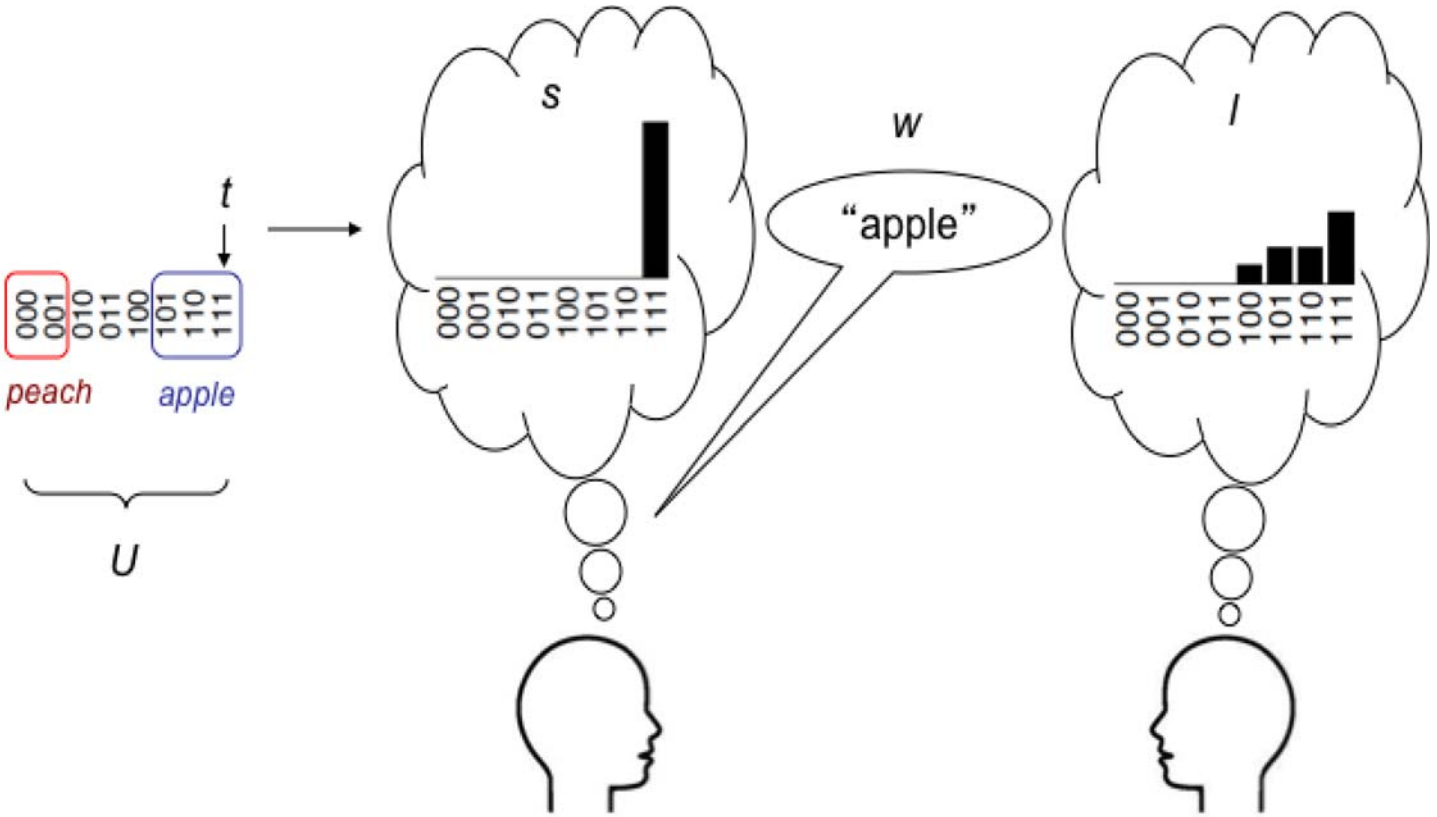
3. Kinship terminology systems demonstrate again the efficient communication optimization of *informativeness* (low information loss) and *simplicity* (low complexity).

Case study 3: Binary feature vectors.

Many accounts of meaning have used feature-based representations to capture knowledge in various semantic domains. But there are no large-scale, cross language studies of this kind to our knowledge.

Will analyze a single-language dataset collected by E. Rosch, et al.* The dataset includes six fruits that are defined in terms of 25 features.

* Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382-439. Also discussed by Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*(2), 169-193 and by Corter, J. & Gluck, M. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin, 111*, 291-303.



A simplified scenario illustrating communication about an object represented as a vector of three binary features. Of the 8 possible feature vectors, 2 have been assigned to the category “peach” and 3 have been assigned to the category “apple”. (The Rosch et al. study had 6 categories and 25 binary features.)

COST

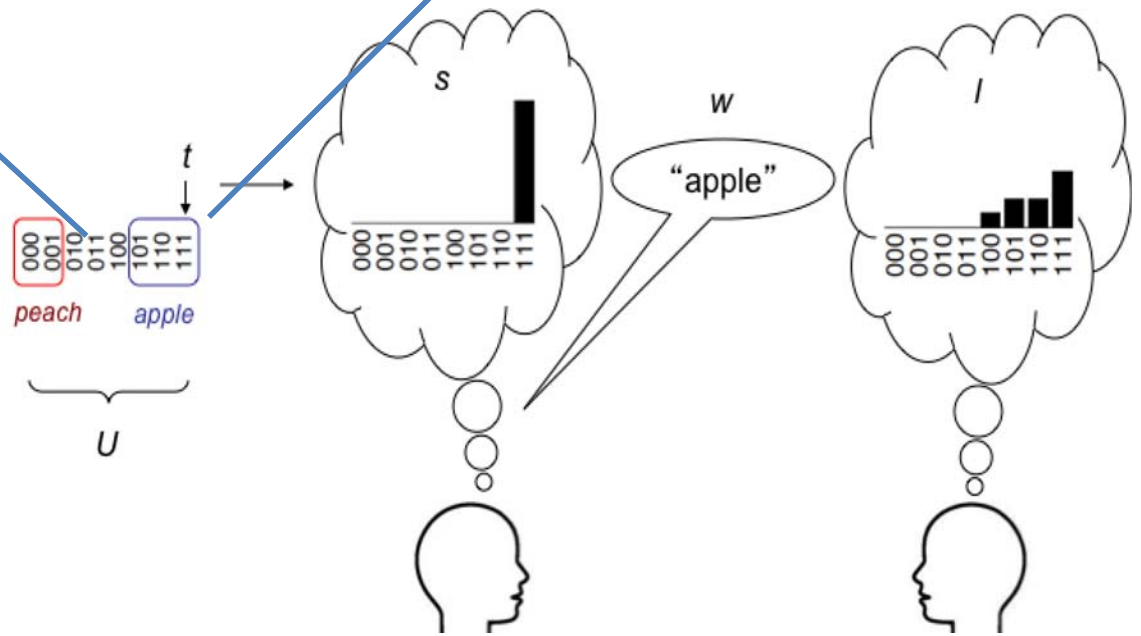
Recall that the information lost *e*(*t*) in communicating about an individual domain item *t* is determined by the listener distribution's value for *t.*

$$e(t) = \sum_{i \in U} s(i) \log\left(\frac{s(i)}{l(i)}\right) = \log\left(\frac{1}{l(t)}\right) = -\log(l(t))$$

We want to define the listener distribution here so as to reflect the degree to which a given item in category *w* is similar to items known to be in *w* . So we define the listener distribution *l*(*i*) for category *w* and domain individual *i* as the product across features of the relative frequencies of the value of *i* among members of *w*.

$$l(i) = p(i \mid w) = p(f_1 \ldots f_n \mid w) = p(f_1 \mid w) \ldots p(f_n \mid w)$$

	item w	item x	item	item y	item z
f1	1	0	1	0	1
f2	1	0	0	1	1
f3	0	1	1	1	1
p(v(f1))|	2/3	1/3	2/3	1/3	2/3
p(v(f2))|	2/3	1/3	1/3	2/3	2/3
p(v(f3))|	0	1	1	1	1
p(item)	0	1/9	2/9	2/9	4/9



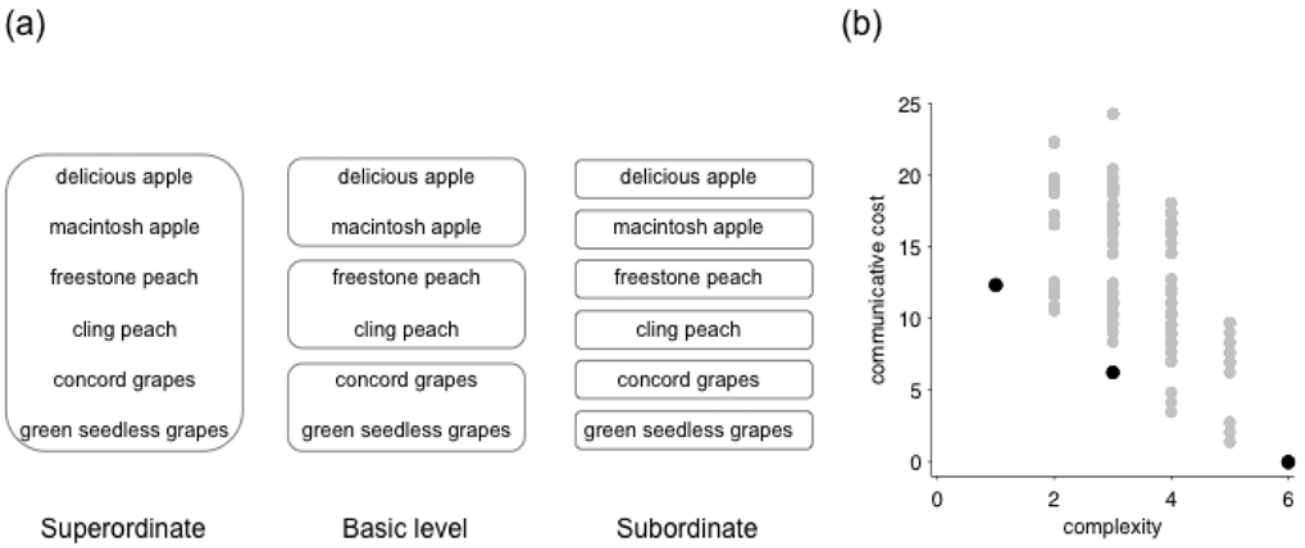
Assume need probability $n(t)$ is the same for every domain element. We can now calculate communicative cost E and compare it to complexity (number of categories).

COST: Sum of product of need probability $n(t)$ and the information loss $e(t) = -\log(l(t))$

$$E = \sum_{t \in U} n(t)\, e(t) \qquad (3)$$

$$= \sum_{t \in U} n(t)(-\log(l(t))) \qquad (4)$$

Comparison of **COST** and **COMPLEXITY**



(a) The 6 fruits in the Rosch et al. data set, here organized into superordinate ("fruit"), basic level ("apple", "peach", "grapes"), and subordinate categories. (b) Reconstruction error *E* versus complexity for all possible systems that organize the 6 fruits into categories. Black dots show the superordinate, basic level, and subordinate categories of (a).

CONCLUSION

- Analyzed three domains of lexical categories with very different structures in terms of *efficient communication*.

- Two of the analyses were based on large cross-language databases.

- All three lexical domains show strong evidence of *efficient communication,* optimizing the tradeoff between INFORMATIVENESS and SIMPLICITY (i.e., optimizing low COST and low COMPLEXITY).

- To the degree that properties of existing languages can inform us about language evolution, these findings provide no support for the Chomskyan position that minimizes the role of communication in the evolution of language.