

Information processing and cross-linguistic universals

Ted Gibson
Department of Brain and Cognitive Sciences
MIT

<http://tedlab.mit.edu>
Twitter: @LanguageMIT

BAULT symposium, University of Helsinki
December 1, 2016

Ted Gibson: Language processing, language structure & language evolution

Current graduate students:

- Richard Futrell
- Julian Jara-Ettinger
- Alex Paunov

Other recent collaborators:

- Ev Fedorenko
- Steve Piantadosi
- Kyle Mahowald
- Leon Bergen
- Bevil Conway
- Melissa Kline
- Mike Frank
- Roger Levy

Research program

What pressures shape human language?

(1) communication; (2) memory; (3) culture

Evidence: cross-linguistic universals

What is the structure of language? What factors affect the complexity of processing a phrase, sentence or text?

E.g., word frequency; syntactic rules; working memory resources

Methods

- Behavioral experiments (e.g., reading / listening or generation)
 - ▶ Cross-linguistic / cross-cultural experiments
- Corpus analyses
- Computational modeling
- Brain imaging

Information processing and cross-linguistic universals

Words: Language as communication

1. Proposed universal: Contextual predictability predicts word length across languages
2. Information theory applied to the semantic domain of color words: Explaining cross-cultural universals and differences

Syntax: Information processing / memory limitations:

3. Proposed universal: Languages minimize dependency lengths

Dictionaries might be “optimized” for efficient use / communication

Zipf (1949): more frequent words are shorter:

- “Principle of least effort”

High frequency, short words:

act, aid, guy, men, was, war, way, who

Low frequency, long words:

crocheted, phenomenology, stratification, reluctantly, reconfiguration

Piantadosi, Tily & Gibson (2011)



Extension: more *predictable* words should be shorter.

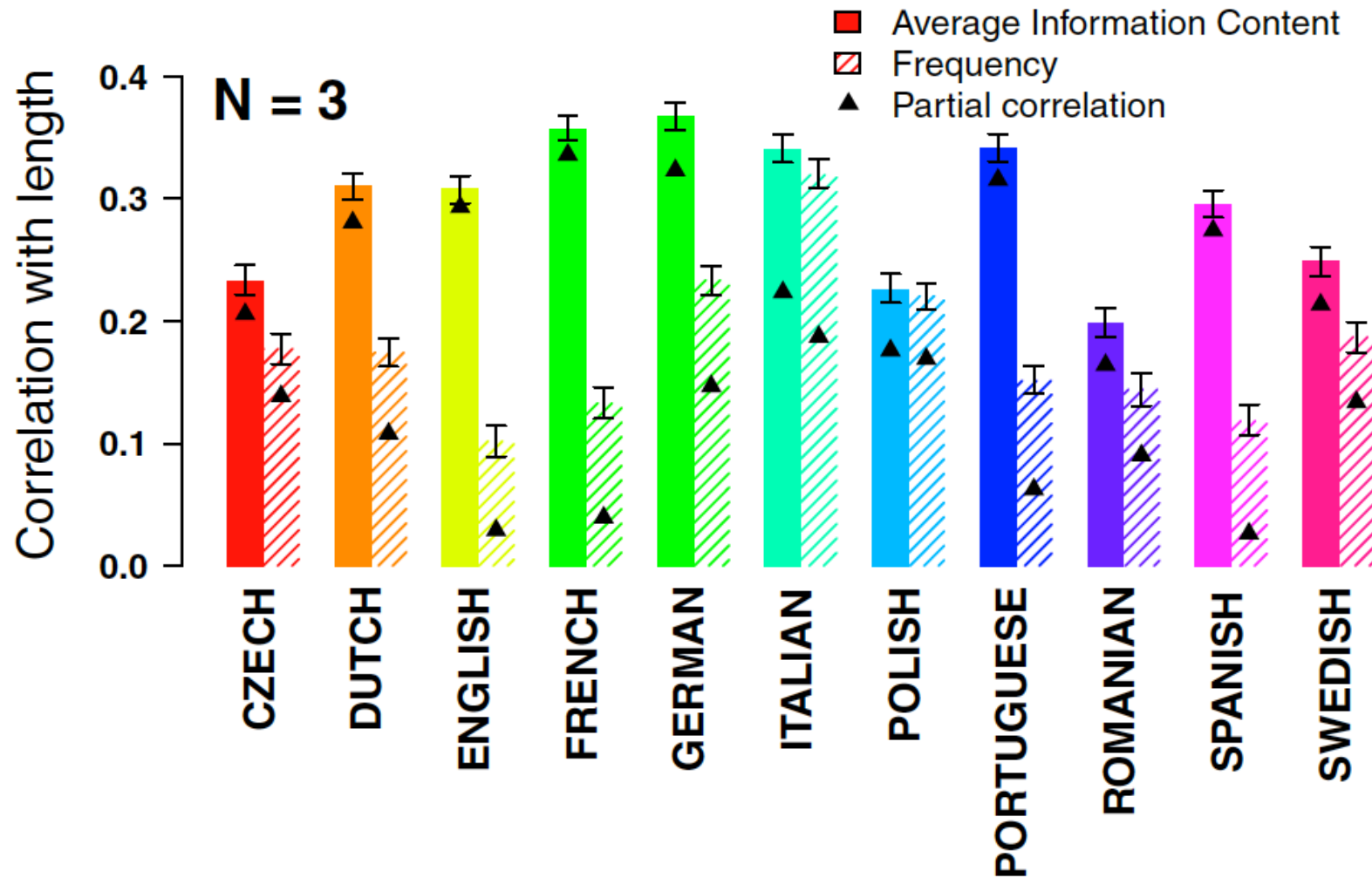
- Estimate of predictability: avg surprisal ($-\log P(w|\text{context})$) from n-grams (3-grams) over large corpora

Low-frequency, short words, that are predictable in context:

aback (taken aback) *yonder* (over yonder)
wasp *lipo* *antler* *bisque*

Language for communication: Words

Piantadosi, Tily & Gibson (2011)



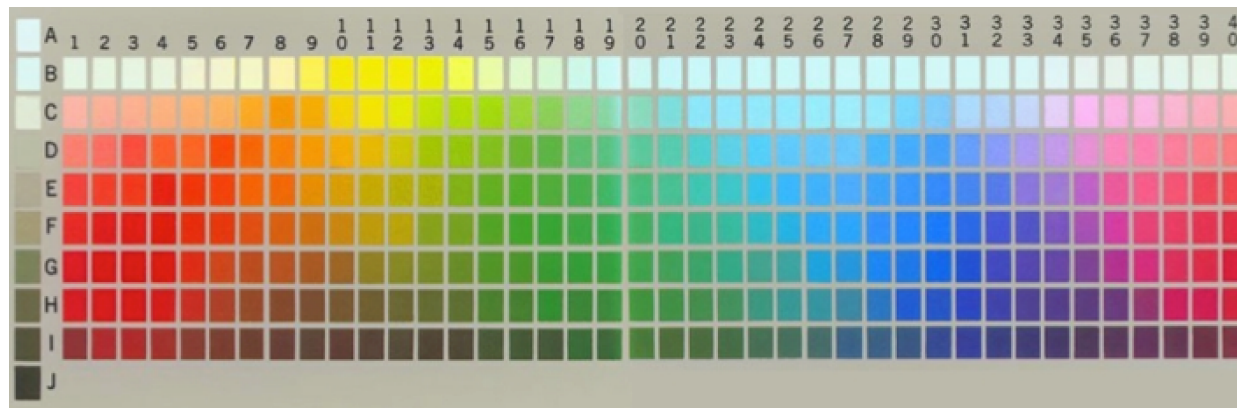
More predictable words are shorter!

Color naming across languages reflects color use

Gibson, Mahowald, Jara-Ettinger, Futrell, Bergen, Piantadosi, Gibson & Conway, 2016



Why do languages have the set of color terms that they do?



English: 11 “basic” color terms: black, white, red, green, yellow, blue, brown, pink, orange, purple, grey

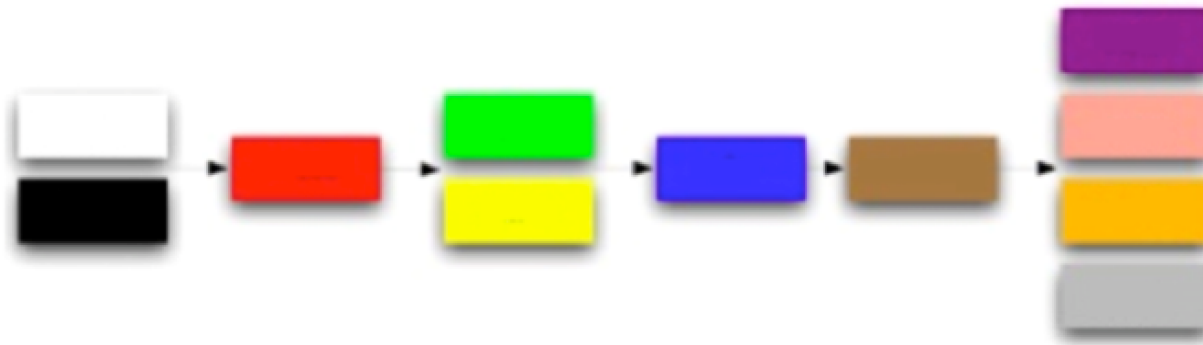
Berino: 5 “basic” color terms (Roberson et al. 2000; Davidoff et al, 1999)

Dani: 2 color terms (Rosch Heider 1972): dark / light or “black” / “white”

Berlin & Kay (1969): The World Color Survey (WCS)

The universalist (perception) hypothesis

330 colors in World Color Survey color grid: Approximately a subset relation among sets of color terms across languages:



Berlin & Kay discuss the distribution of color terms in terms of “basic” color terms: basic color terms are thought to be **visual-perception** based: the most salient colors in the color space (e.g., Kay & Maffi, 1999)

These are the **modal color words** in the WCS

The approximate subset relationship across languages is suggestive evidence for the perceptual hypothesis

Why the wide variability between cultures?

Puzzles for the universalist (perception) hypothesis

Methodological question:

What exactly is a “basic” color term? Is this term well-defined?

The data from the World Color Survey (WCS) was gathered while presupposing “basic” colors: participants were only permitted to provide “basic” terms (Saunders & van Brakel, 1997)

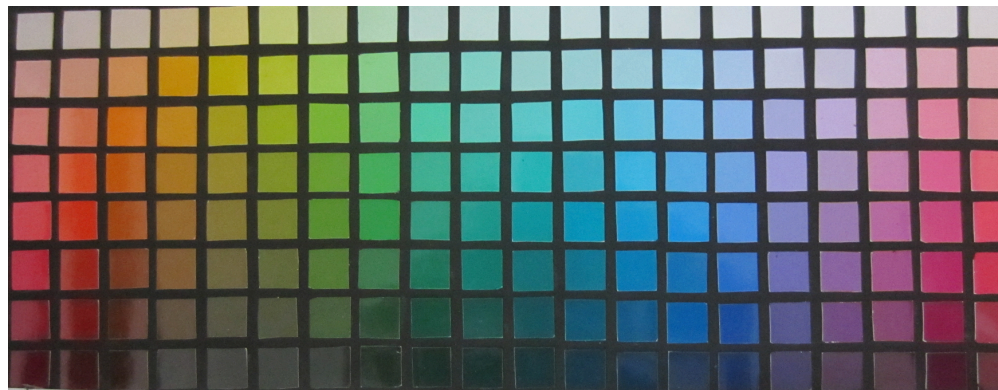
Open research questions:

1. Is there a cross-linguistic universal component to color-naming?
2. If so, what is the cause of a cross-linguistic universal?
3. Why do more industrialized cultures have more color words?

Color communication and information theory

Task: I name the color of a randomly selected color chip from a set (e.g., 80), assuming a uniform prior..

Let's think of this task in terms of **color communication**. A listener L tries to guess the target. **How many guesses does it take to guess the target?** L can choose any set of chips, and is told “yes” if the target is in the set.

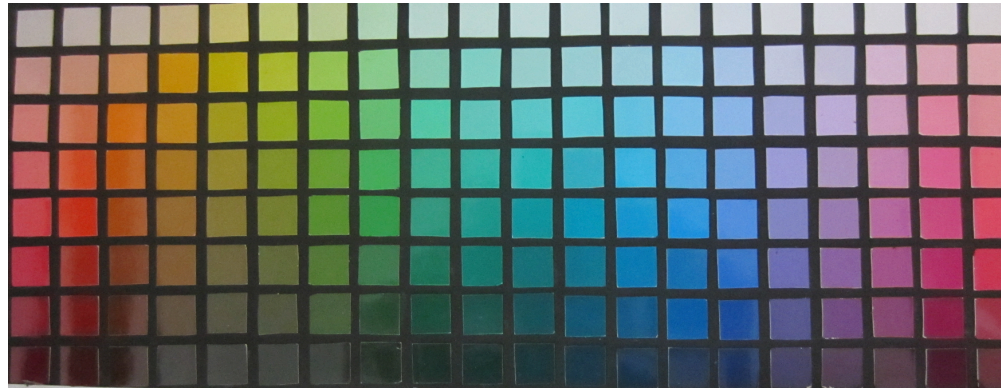


For any **particular color**, the number of guesses you need depends on

- (a) how reliably each color chip is labeled: $P(\text{word} \mid \text{color})$
- (b) how many chips would be labeled with that word: $P(\text{color} \mid \text{word})$

How “easy-to-communicate” is a particular color out of a set of colors?

The Munsell color space is defined by visual perception: *Each chip is equally far from each neighboring chip, in terms of people’s ability to detect differences*



Critical research question: Are all colors equally easy-to-communicate?
Or is the distribution **skewed**, so that some are easier to communicate than others?

Suppose I choose a particular red chip: how much information? How about for a green chip?

How “easy-to-communicate” is a particular color out of a set of colors?

In terms of information (bits), how many guesses on average do you need to correctly guess a color that I am thinking of, out of a fixed set (e.g., 80)?

- The speaker wants to communicate a color chip c .
- The speaker chooses a color word w conditioned on c : $P(w|c)$
 - *Imagine that this is close to 1 for a canonical red, blue, yellow chip etc.*
- The **average surprisal** score in bits for c is $P(w|c) * c$'s surprisal for the listener, given the word:

$$S(c) = \sum_w P(w|c) \log \frac{1}{P(c|w)}$$

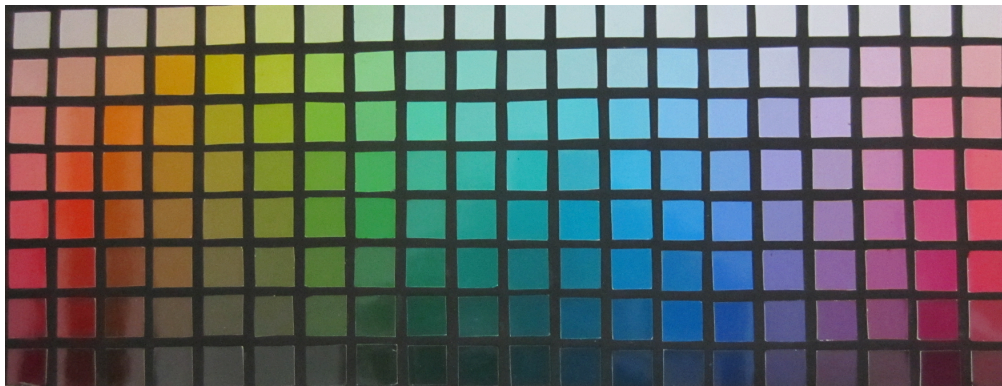
= the optimal number of guesses it takes you to guess which chip I meant

e.g., if there are 8 chips that might be called “red”, then $\log(1/P(c|w)) = 3$ bits

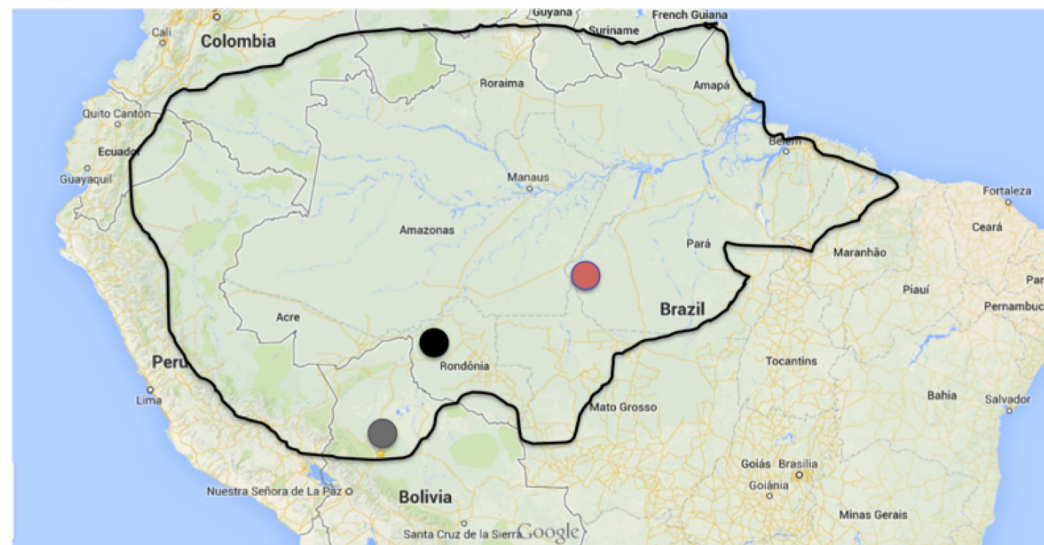
The Tsimane'



- Mundurucú
- Pirahã
- Tsimane'



Task: What color is this chip?



Color naming across languages reflects color use

Gibson et al. (2016)

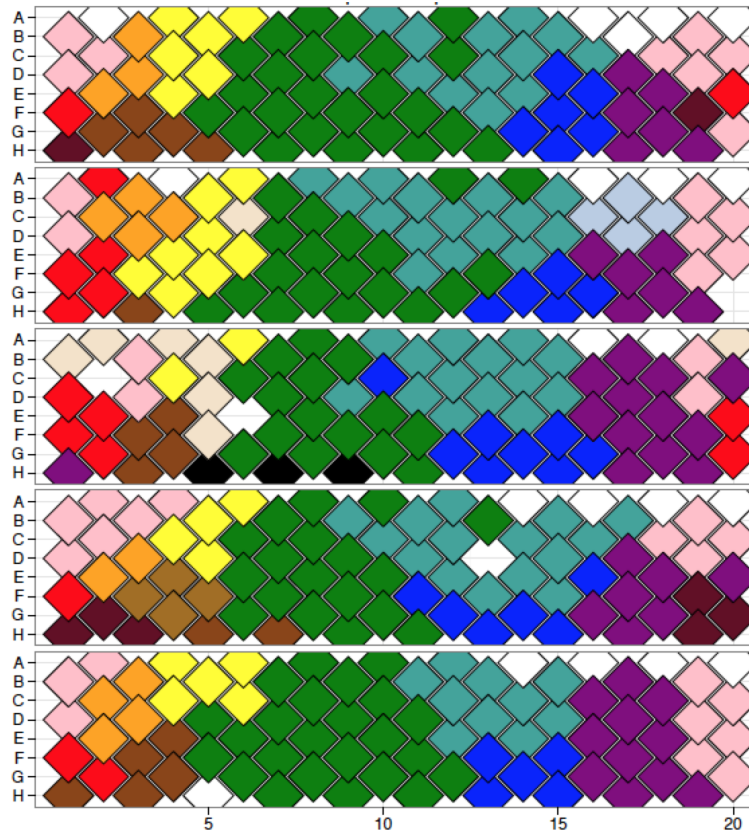
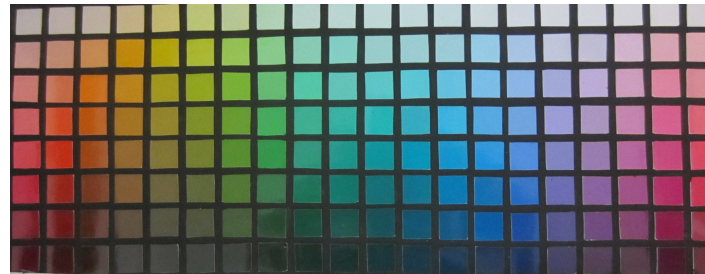
Task: What color is this chip?

This task does not presuppose “basic” color terms.

Munsell chips: the color chips are selected so that they are equally spaced in the perceptual space.

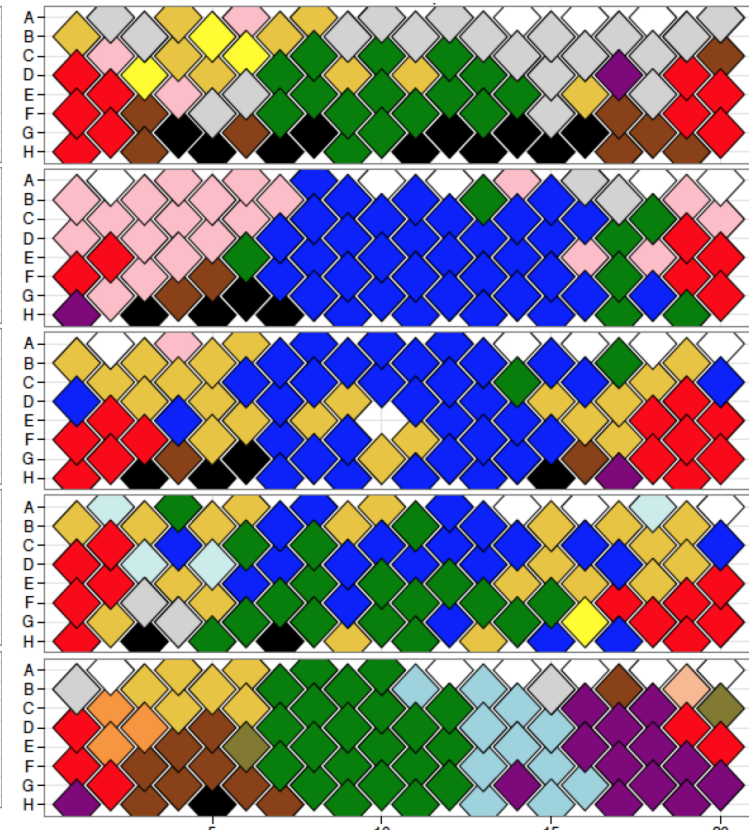
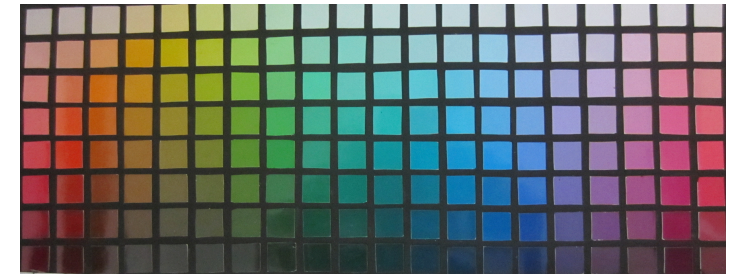
About the same number of terms in each group. A Tsimane’ person is consistent but idiosyncratic in their use of color terms.

Spanish



5 sample Spanish speakers

Tsimane’



5 sample Tsimane’ speakers

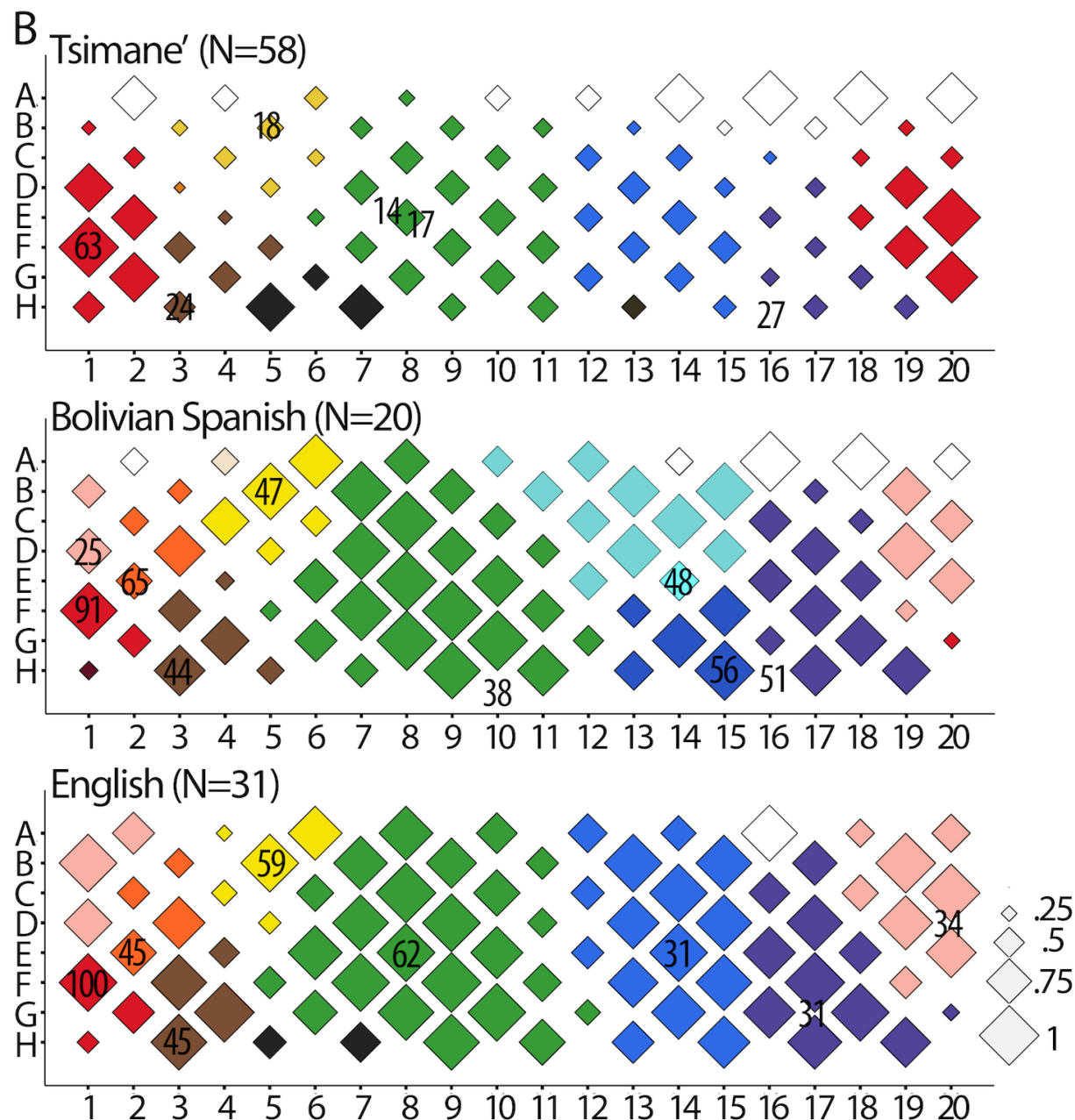
Color naming across languages reflects color use

Gibson et al. (2016)

The Tsimane' are idiosyncratic between individuals (cf. Lindsey et al., 2015).

Color of each diamond corresponds to the color of the chip in the middle of the distribution with that modal name.

Diamond size = % people with modal response.



Evaluating Open-response vs. Fixed-response tasks: About the same amount of information in each

Two versions of task:

(a) open-response; (b) fixed-response, like the World Color Survey (WCS)

High correlation between color-chip average-surprisal across the two tasks:

Tsimane': $\rho=.71$; Bolivian-Spanish: $\rho=.90$; English: $\rho=.92$

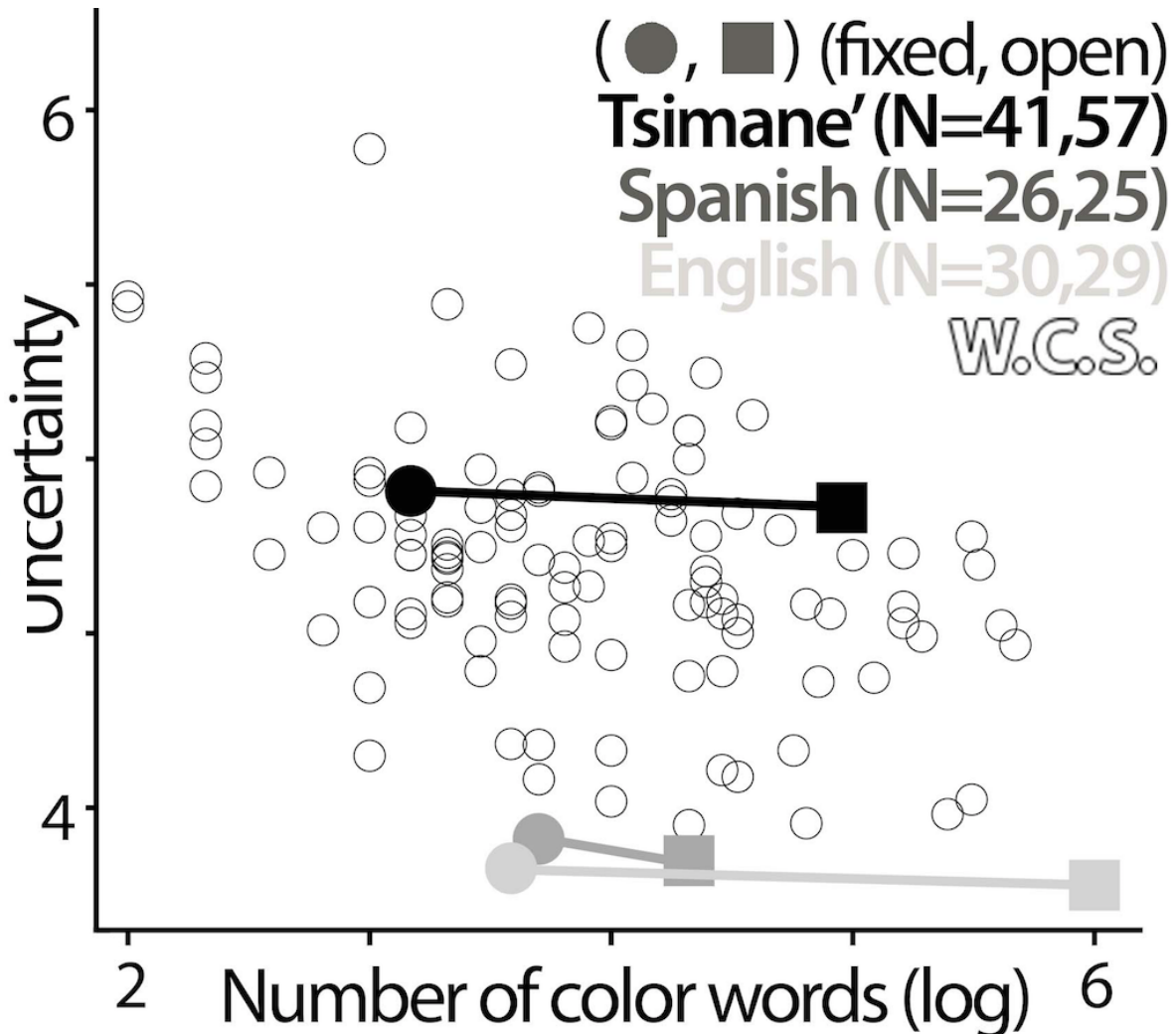
Evaluating Open-response vs. Fixed-response tasks: About the same amount of information in each

Compare uncertainty (average surprisal) between words and color chips vs. total number of color words in a language, for Tsimane', Spanish, English, and World Color Survey.

$$(1) \quad S(c) = \sum_w P(w|c) \log \frac{1}{P(c|w)}$$

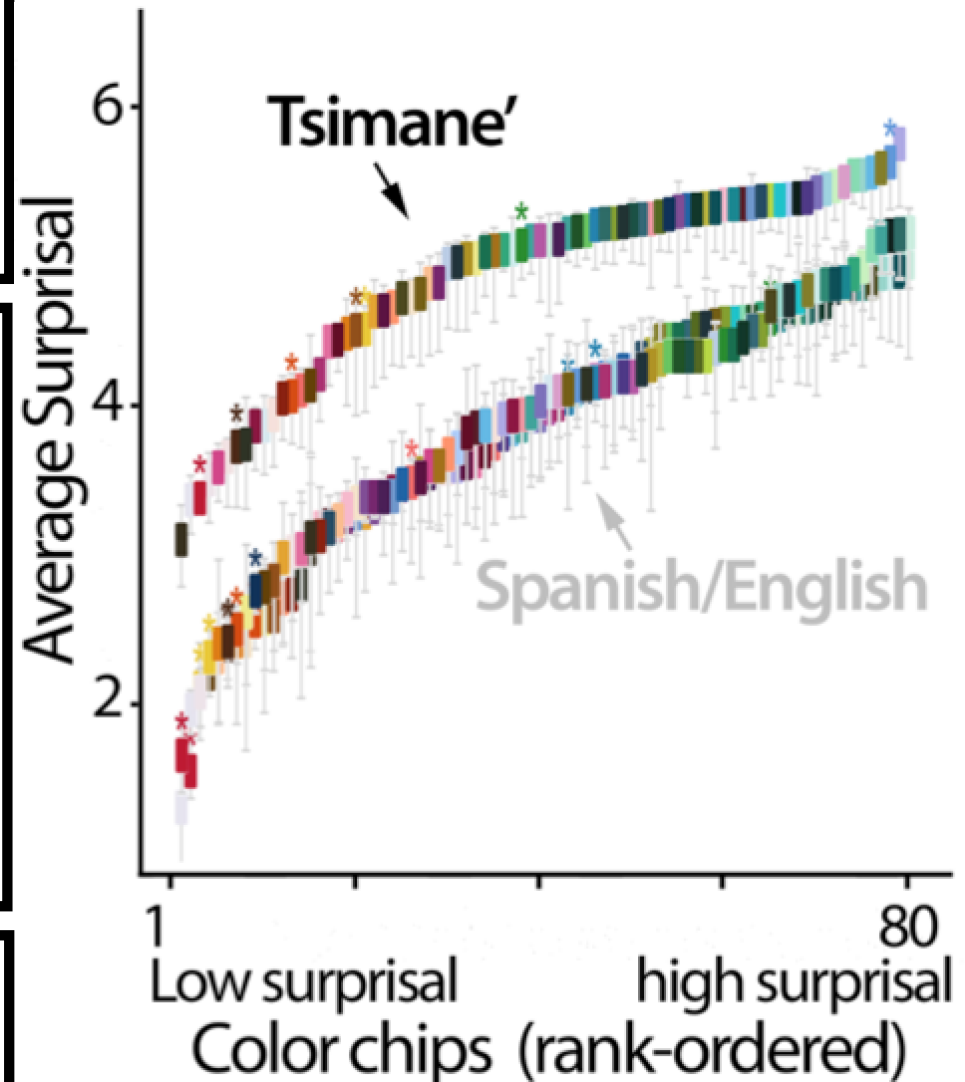
$$(2) \quad \sum_c P(c) S(c)$$

Tsimane' is not an unusual language / culture in the WCS



Cross-linguistic generalization: Correlation among average surprisal scores across languages

- Rank order average surprisal scores for each color chip within a language: highly correlated: Tsimane' vs. Spanish $r = .54$
English vs. Spanish: .87
- Generalization: **Warm colors (red, orange, yellow, pink) are low surprisal; cool colors (blue, green) are high surprisal**
- The average surprisal scores are simply shifted up one bit from English / Spanish to Tsimane!
- The rank ordering is not explained by unique hues or Berlin-Kay ordering: blue / green before pink, orange, brown



Color naming across languages reflects color use

Gibson, Mahowald, Jara-Ettinger, Futrell, Bergen, Piantadosi, Gibson & Conway, 2016

The efficient-communication hypothesis: People vary in their color language depending on the usefulness of color to behavior

1. *The cross-linguistic universal ordering: warm vs. cool colors*
2. *Image statistics, explain this contrast: foreground objects vs. background (not visual perception)*
3. Cross-cultural differences: *Color is less useful for non-industrialized cultures like the Tsimane'.*
 - The Tsimane' are idiosyncratic in their use of color terms
 - The Tsimane' don't use color terms spontaneously
4. Other differences between cultures are likely due to local cultural differences in what objects are relevant (e.g., Berinmo)

Information processing and cross-linguistic universals

The performance-grammar correspondence hypothesis (Hawkins, 2004):
Grammars have conventionalized syntactic structures in proportion to their degree of preference in performance (Haspelmath, 1999; Bybee & Hopper, 2001; Kirby, 1999; Kirby, Cornish & Smith, 2008; Culbertson, Smolensky & Legendre, 2012)

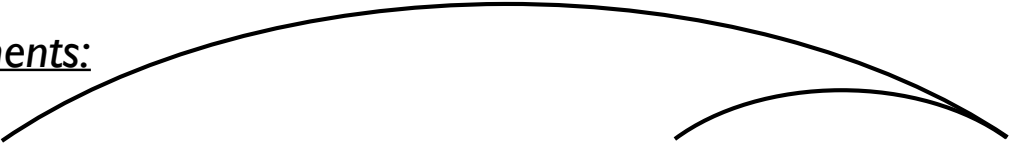
- Information processing / memory limitations: Proposed universal: Languages minimize dependency lengths

Information processing: Working memory

Working memory: Local connections are easier to make than long-distance ones (Gibson, 1998, 2000; Grodner & Gibson, 2005; Warren & Gibson, 2002; Lewis & Vashishth, 2005; Hawkins, 1994)

Ambiguous attachments:

The bartender **told** the detective that the suspect **left** the country **yesterday**.



yesterday is preferred as modifying **left** rather than **told**

(Frazier & Rayner, 1982; Gibson et al., 1996; Altmann et al., 1998; Pearlmutter & Gibson, 2001)

Unambiguous connections:

The **reporter** **wrote** an article.



The **reporter** from the newspaper **wrote** an article.



The **reporter** who was from the newspaper **wrote** an article.



Retrieval / Integration-based theories

Integration: connecting the current word into the structure built thus far: Local integrations are easier than longer-distance integrations


- The Dependency Locality Theory (DLT) (Gibson, 1998; 2000): intervening **discourse referents** cause retrieval difficulty (also in production)
- Activation-based memory theory: similarity-based interference (Lewis & Vasishth, 2005; Vasishth & Lewis, 2006; Lewis, Vasishth & Van Dyke, 2006): **intervening similar elements** cause retrieval difficulty
- Production: Hawkins (1994; 2004): **word**-based distance metric.

Dependency Length Minimization

Futrell, Mahowald & Gibson, 2015, PNAS



- Corpora from 37 languages parsed into dependencies, from NLP sources: the HamleDT and UDT; cf. WALS (Dryer 2013)
- Family / Region
Indo-European (IE)/West-Germanic; IE/North-Germanic; IE/Romance; IE/Greek; IE/West Slavic; IE/South Slavic; IE/East Slavic; IE/Iranian; IE/Indic; Finno-Ugric/Finnic; Finno-Ugric/Ugric; Turkic; West Semitic; Dravidian; Austronesian; East Asian Isolate (2); Other Isolate (1)
- **Result:** All languages minimize dependency distances (c.f. Hawkins, 1994; Gibson, 1998)



the girl kicks the ball



the girl the ball kicks



the ball the girl kicks



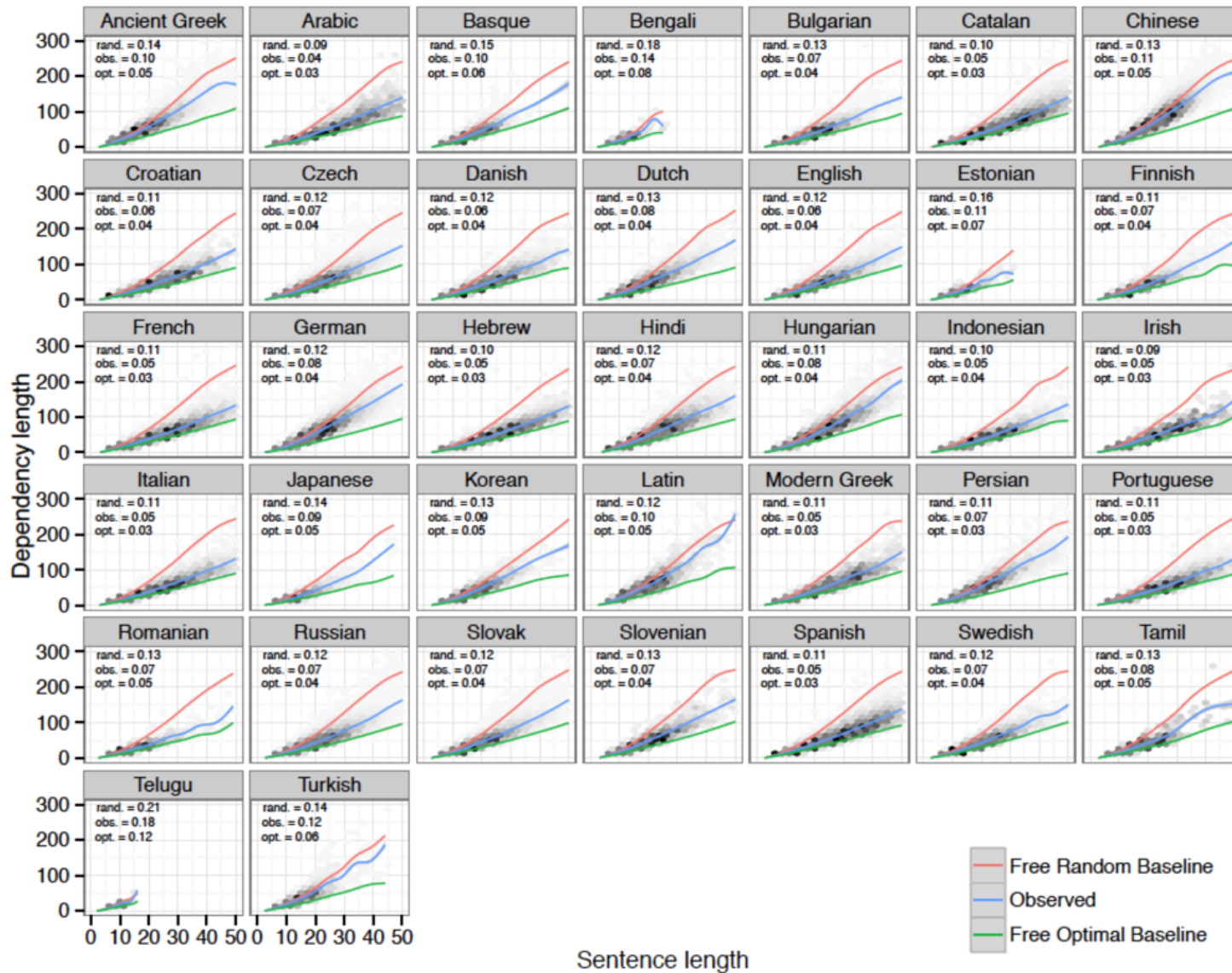
girl the kicks the ball



ball the girl the kicks

Dependency Length Minimization

Futrell, Mahowald & Gibson, 2015, PNAS



Related universals?

Head-direction / Branching direction

Parsed corpora will eventually provide answers to other quantitative questions about word order

- E.g., *language use* vs. *grammars* that minimize dependency length
- **Matching head direction?** Having head-final for some categories and head-initial for others leads to structures with longer-distance dependencies (Gibson, 1998, 2000; Hawkins, 1994; 2004; cf. Greenberg, 1963; Dryer, 1992)

Matching word orders: Head-first V-CP + head-first C-VP: (or both head-final): short



I thought that you would take out the garbage.

The diagram shows two curved arrows above the sentence. The first arrow starts under the word 'thought' and points to the word 'that'. The second arrow starts under the word 'would' and points to the word 'take'.

Distance: V “thought” and C “that” is 1 word; C “that” and Infl “would” is 2 words;

Mismatch word orders: Head-first V-CP + head-final C-VP: long dependencies



I thought you would take out the garbage that.

The diagram shows two long, curved arrows above the sentence. The first arrow starts under the word 'thought' and points to the word 'that' at the end of the sentence. The second arrow starts under the word 'would' and points to the word 'take'.

Distance: V “thought” and C “that” is 7 words; C “that” and Infl “would” is 5 words

Conclusion: Information processing and cross-linguistic universals

Suppose that language approximates an optimal code for information processing. This can potentially explain:

- The evolution of language:
 - Words (Piantadosi, Tily, & Gibson, 2011, 2012; Gibson, et al. 2016)
 - Syntax (Gibson, Piantadosi, Brink, Lim, Bergen & Saxe, 2013; Futrell, Hickey, Lee, Lim, Luchkina & Gibson., 2014; Futrell, Mahowald & Gibson, 2015a, 2015b)
- Language use
 - Sentence interpretation (Gibson, Bergen & Piantadosi, 2013; Bergen & Gibson, 2013; Fedorenko, Stearns, Bergen, Eddy & Gibson, submitted; Gibson, Sandberg, Fedorenko, Bergen & Kiran, 2015)

Acknowledgments

- *National Science Foundation Grants from the linguistics program 0844472 (until 2013); 1534318 (2015-2018)*
- *Collaborators:*
 - ▶ Words: **Steve Piantadosi**, Hal Tily, Kyle Mahowald
 - ▶ Color words: **Bevil Conway**, Kyle Mahowald; Julian Jara-Ettinger, Richard Futrell; Leon Bergen, Steve Piantadosi, Mitchell Gibson
 - ▶ Origin of word order: Richard Futrell, Kyle Mahowald

