

Information processing and cross-linguistic universals

Ted Gibson
Department of Brain and Cognitive Sciences
MIT

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

CLASP, University of Gothenburg, Sweden
September 12, 2016

Language: structure, acquisition and processing

Ted Gibson

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

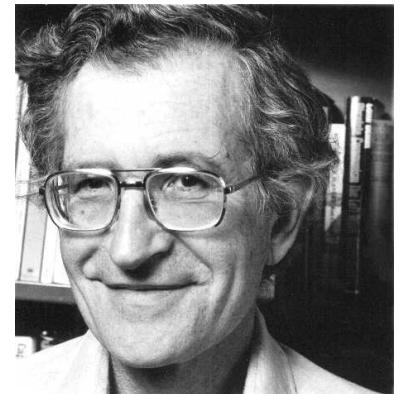
Methods

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

Language for Communication?

- More controversial than some might think...

“The natural approach has always been: Is it well designed for use, understood typically as use for communication? I think that’s the wrong question. 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, p. 107)



Language for Communication?

Contrary to Chomsky, we argue that language approximates an optimal code for human communication (Zipf, 1949).

This can potentially explain:

- the online behavior of language users (Genzel & Charniak, 2002; Aylett & Turk, 2004; Levy, 2005; Jaeger, 2006; Levy & Jaeger, 2007)
- the structure of languages themselves (e.g. Kirby, 1999; Ferrer i Cancho & Sole, 2003; Ferrer i Cancho, 2006; Piantadosi, Tily, & Gibson, 2011; Gibson et al., 2013)

But what about the issue of **ambiguity**?

Ambiguity

Lexicon: run (polysemy); two/to/too (homophony)

Syntax: Frank shot the hunter with the shotgun.

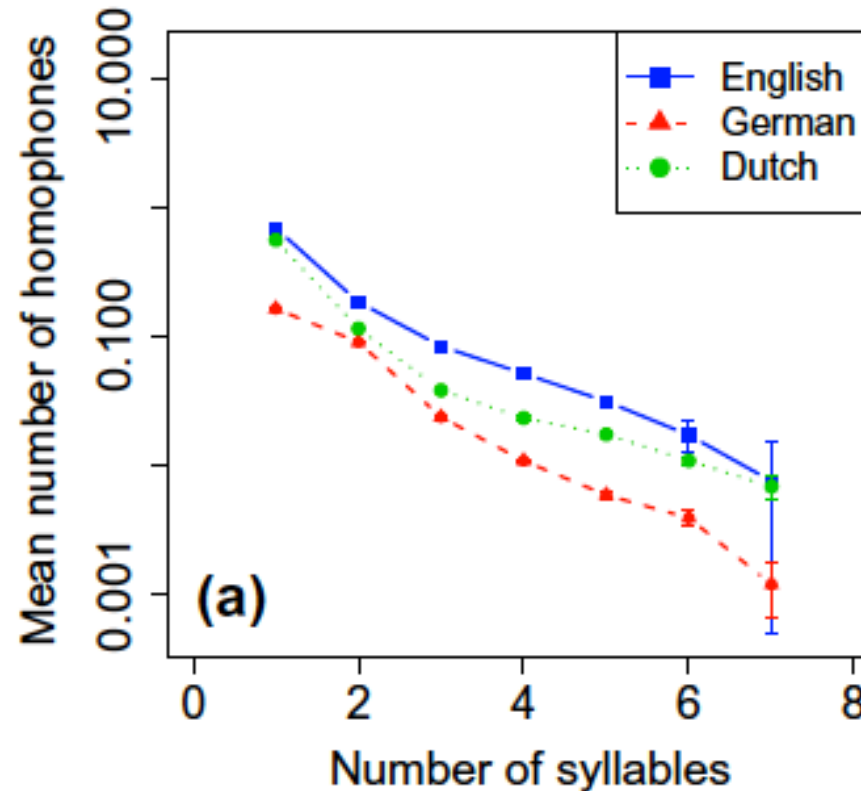
Referential: He said that we should give it to them.

Ambiguity: A communicative *benefit*



- Ambiguity is only a problem *in theory*
 - Ambiguity is not a problem in normal language use, because context disambiguates (Wasow & Arnold, 2003; Wasow et al., 2005; Jaeger, 2006; Roland, Elman, & Ferreira, 2006; Ferreira, 2008; Jaeger, 2010).
- context disambiguates, e.g., word use:
 - *John wanted to run.*
 - *John went to school.*
 - *John wanted two dollars.*
 - *Sam wanted some money too.*
- Piantadosi, Tily & Gibson (2012): An information-theoretic proof that efficient communication systems will necessarily be globally ambiguous when context is informative about meaning (because short / easy items will get re-used in different contexts)

Language as efficient communication: Shorter words are more ambiguous Piantadosi, Tily & Gibson (2012)



- Number of additional meanings each phonological form has, as a function of length.
- Shorter phonological forms have more meanings

Ambiguity out of context: Evidence for information theory

The existence of ambiguity out of context in human language (which is disambiguated by context) is explained by **information theory**.

I.e., why do we re-use words? In part, to keep the code short.

In other approaches, the existence of ambiguity out of context is an *unexplained accident*.

Language for Communication?

As argued in Piantadosi, Tily & Gibson (2012), ambiguity is not a problem for human language codes.

Contrary to Chomsky, we argue that language approximates an optimal code for human communication (Zipf, 1949).

This can potentially explain:

- the online behavior of language users (Genzel & Charniak, 2002; Aylett & Turk, 2004; Levy, 2005; Jaeger, 2006; Levy & Jaeger, 2007)
- the structure of languages themselves (e.g. Kirby, 1999; Ferrer i Cancho & Sole, 2003; Ferrer i Cancho, 2006; Piantadosi, Tily, & Gibson, 2011; Gibson et al., 2013)

Information processing and cross-linguistic universals

Word length and information theory:

- Proposed universal: Shorter words are more ambiguous

- Proposed universal: Contextual predictability predicts word length across languages

- Information theory applied to the semantic domain of color words: Explaining cross-cultural universals and differences

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)

- Cross-linguistic word order universals: SOV and SVO word order

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

Information processing and cross-linguistic universals

Word length and information theory:

- Proposed universal: Shorter words are more ambiguous

- Proposed universal: Contextual predictability predicts word length across languages

Language / Communication: Words

Piantadosi, Tily & Gibson (2011)



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

Language / Communication: Words

Piantadosi, Tily & Gibson (2011)



Extension: more *predictable* words should be shorter.

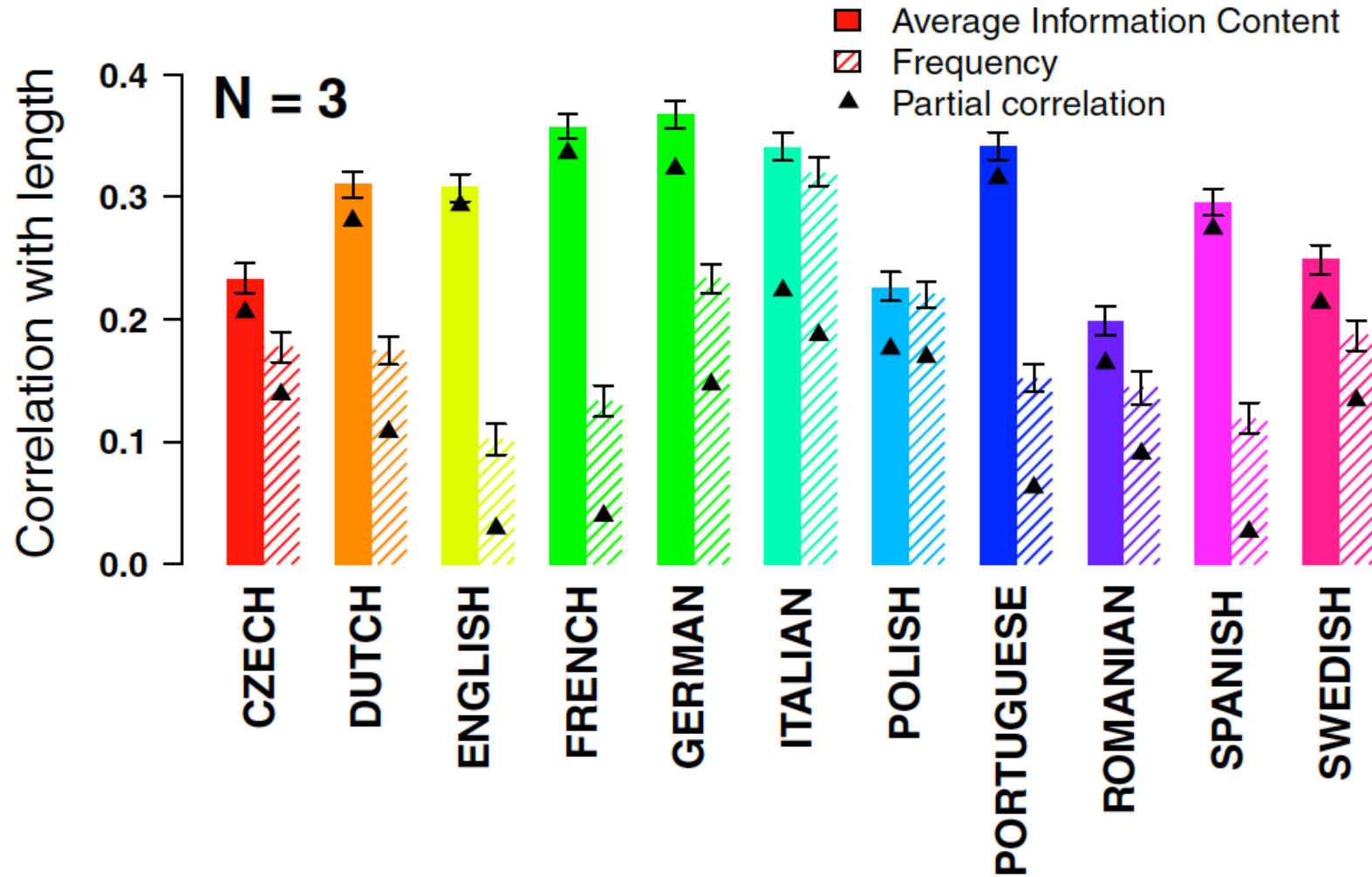
- e.g., to maintain Uniform Information Density (Aylett & Turk, 2004; Jaeger, 2006; Levy & Jaeger, 2007)
- Estimate of predictability: n-grams (3-grams) over large corpora

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

aback (taken aback); *rasa* (*tabula rasa*); *Zappa* (Frank Zappa)
lipo (lipo suction); *bongo* (bongo drums); *chez* (chez moi)

Language for communication: Words

Piantadosi, Tily & Gibson (2011)



More predictable words are shorter!

Language for communication: Words

Piantadosi, Tily & Gibson (2011)

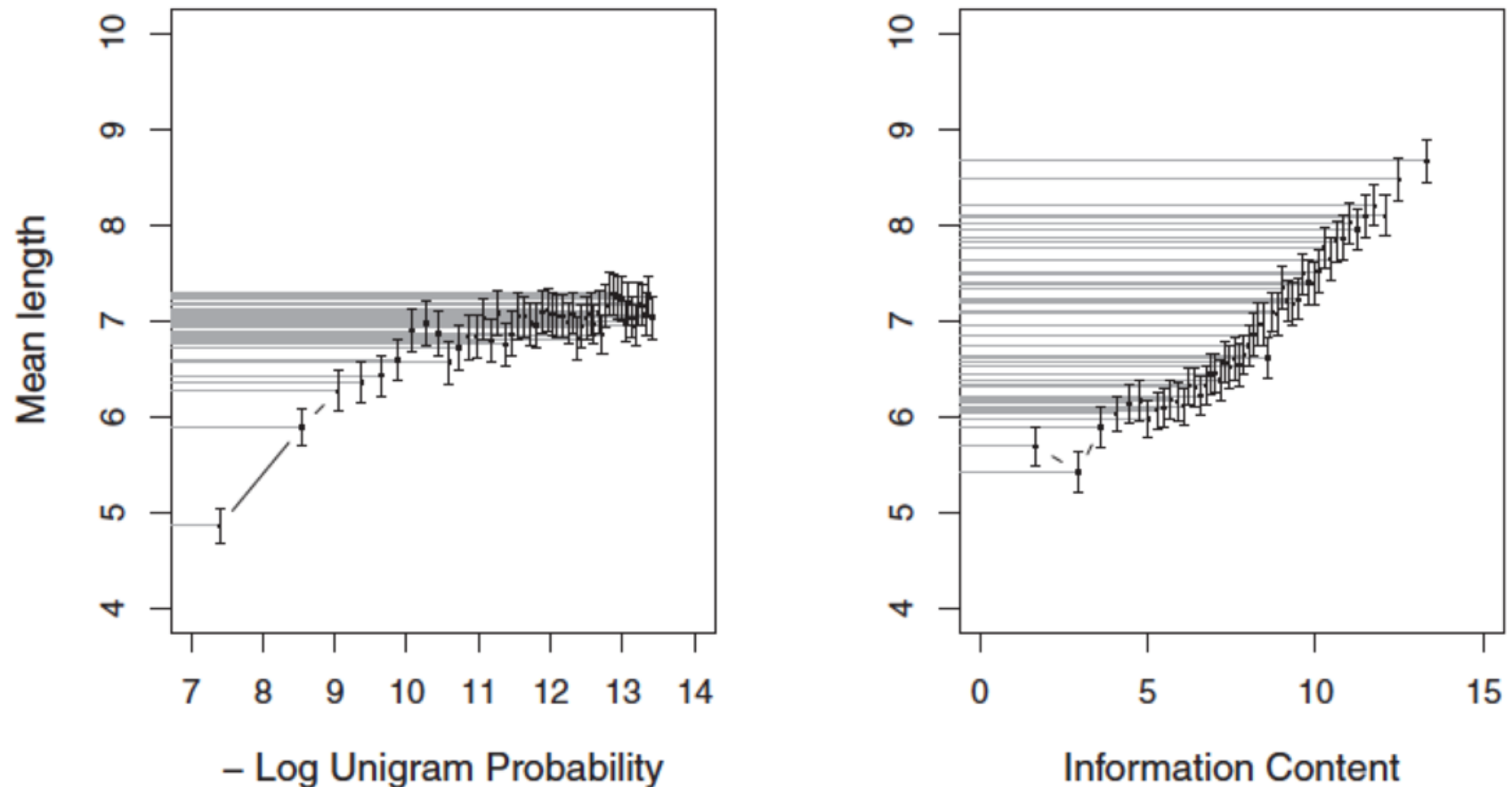


Fig. 2. Relationship between frequency (negative log unigram probability) and length, and information content and length. Error bars represent SEs and each bin represents 2% of the lexicon.

How does the effect arise?

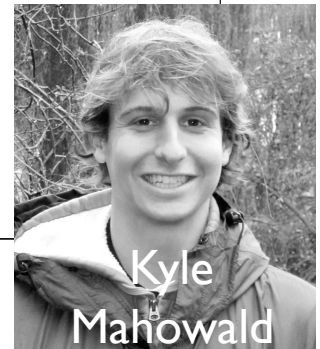
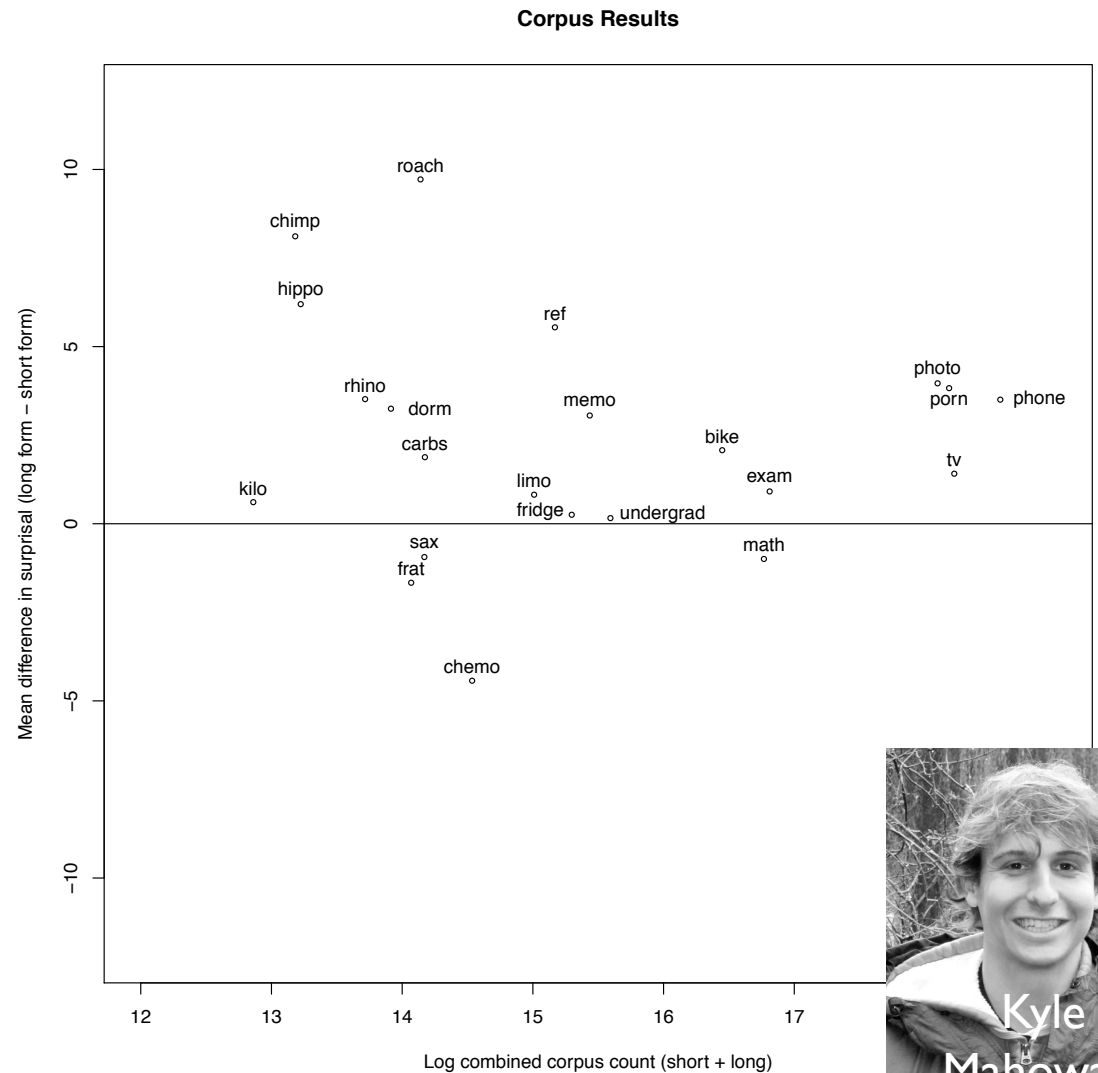
- Is it just differences among broad classes of words like content vs. function words? Or within class too?
- look at long/short pairs (*mathematics* → *math*; *pornography* → *porn*), which differ in length but are controlled for meaning

Info/Information theory

Using Google trigrams, we looked at average surprisal for long forms vs. short forms.

Mean surprisal for long forms (9.21) is significantly higher than mean surprisal for short forms (6.90) ($P = .004$ by Wilcoxon signed rank test)

Linear regression shows significant effect of log frequency on surprisal ($t = 2.76$, $P = .01$) even when controlling for frequency.



Info/Information theory

Forced-choice sentence completion
in supportive and neutral contexts:

supportive-context: Bob was very
bad at algebra, so he hated...

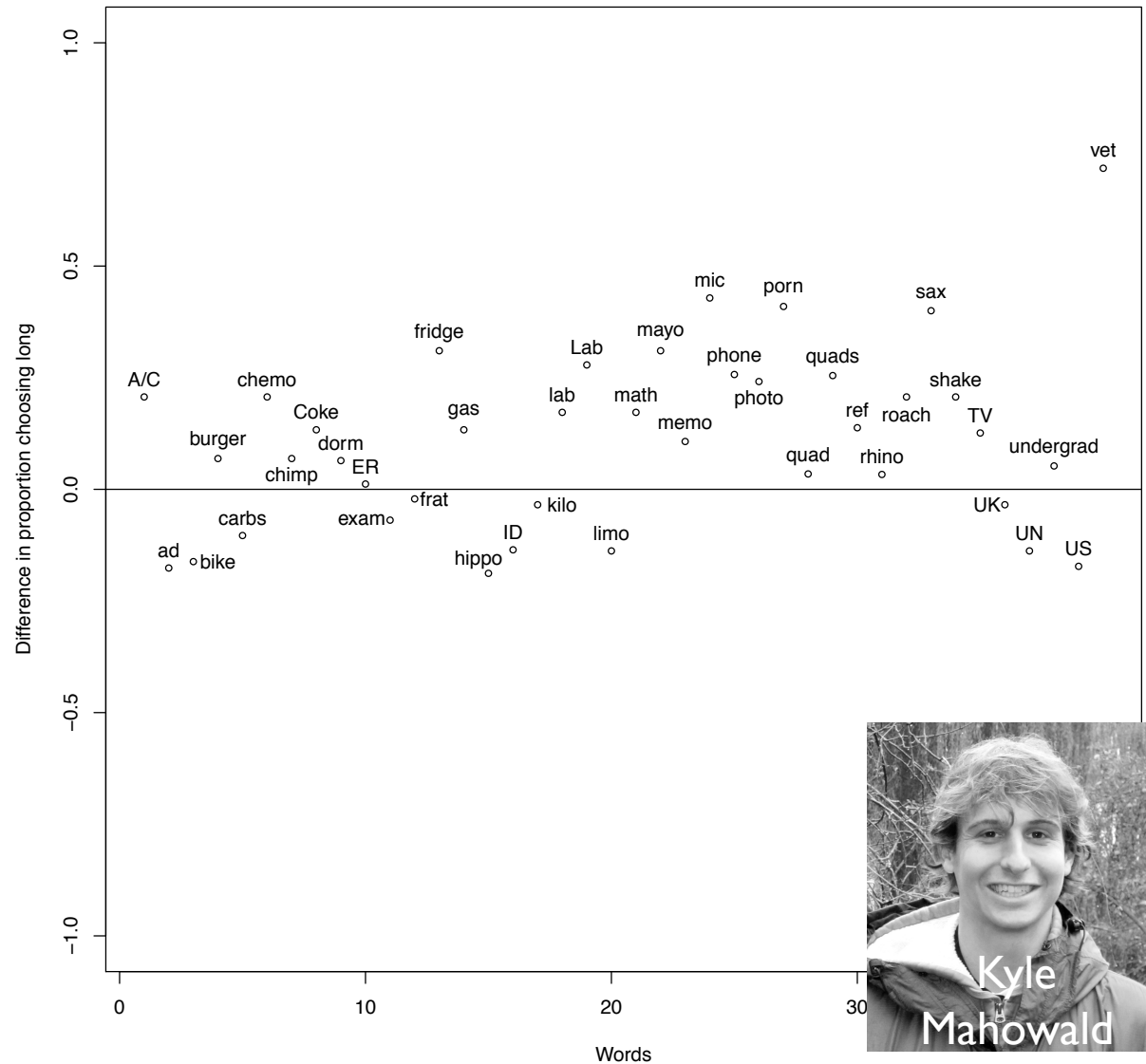
1. math 2. mathematics

neutral-context: Bob introduced
himself to me as someone who
loved...

1. math 2. mathematics

Short form is chosen 67% of the time
in supportive-context sentences vs.
just 56% of the time in neutral-
context sentences.

Significant by maximal mixed effect
logistic regression with both item and
participant slopes and intercepts



Information processing and cross-linguistic universals

Word length and information theory:

- Proposed universal: Shorter words are more ambiguous

- Proposed universal: Contextual predictability predicts word length across languages

- Information theory applied to the semantic domain of color words: Explaining cross-cultural universals and differences

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)

- Cross-linguistic word order universals: SOV and SVO word order

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

Information processing and cross-linguistic universals

Word length and information theory:

- Proposed universal: Shorter words are more ambiguous

- Proposed universal: Contextual predictability predicts word length across languages

- Information theory applied to the semantic domain of color words: Explaining cross-cultural universals and differences

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)

- Cross-linguistic word order universals: SOV and SVO word order

- 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)



Toronto law to protect squirrels hit by mayor



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)

Toronto law to protect squirrels hit by mayor

You can visit the cemetery where famous Russian composers are buried daily except Thursday

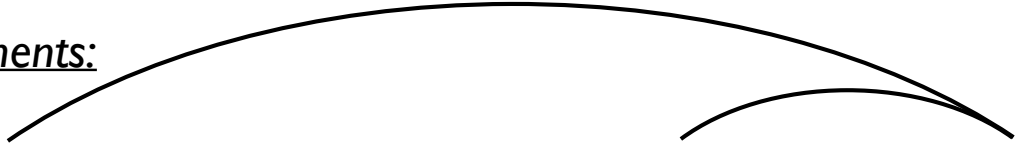
Patient reports pain starting from his penis which goes down to his knee

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.

Consequence:

Nested structures are difficult crosslinguistically

English:

The reporter [who the senator attacked] admitted the error.

The reporter [who the senator [who I met] attacked] admitted the error.

I met the senator who attacked the reporter who admitted the error.

Japanese:

Obasan-wa [bebiisitaa-ga [ani-ga imooto-o ijimeta] to itta] to omotteiru
aunt-top babysitter-nom older-brother-nom younger-sister-acc bullied that said that
thinks

“My aunt thinks that the babysitter said that my older brother bullied my younger sister”

Easier: Bebiisitaa-ga [ani-ga imooto-o ijimeta] to itta] obasan-ga to omotteiru

Dependency Length Minimization

Futrell, Mahowald & Gibson, 2015, PNAS



Richard Futrell

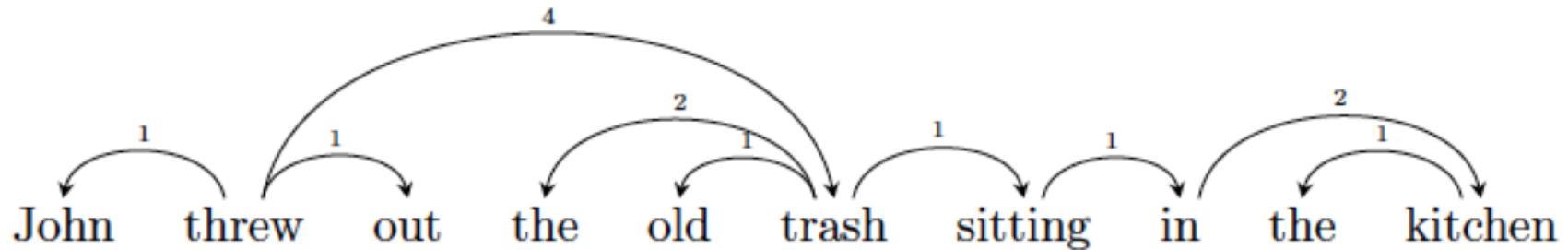


Kyle
Mahowald

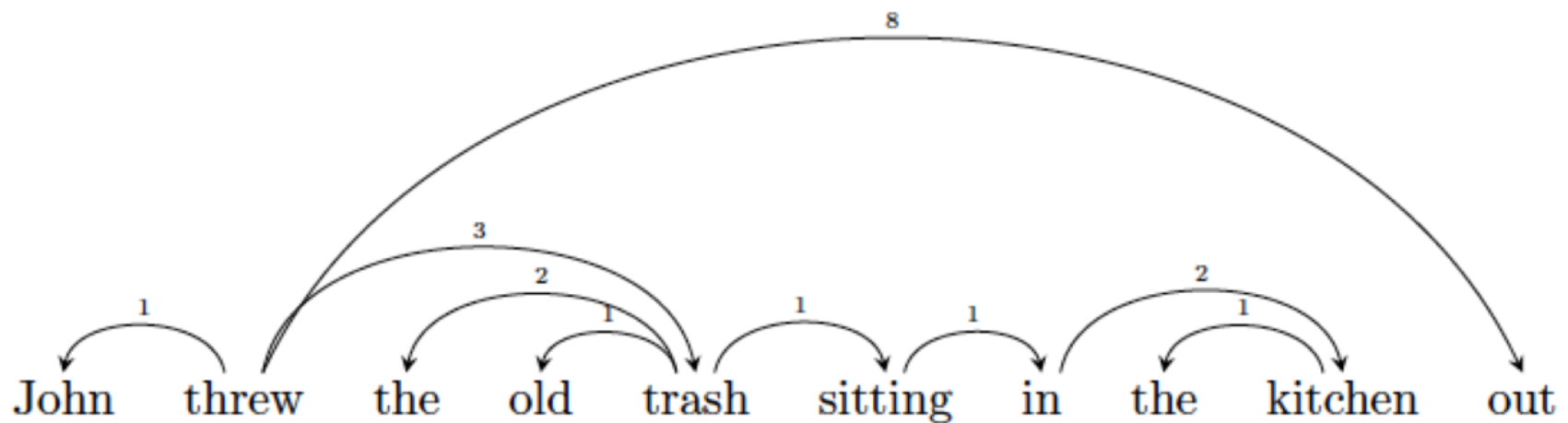
- Corpora from 37 languages parsed into dependencies, from NLP sources: the HamleDT and UDT; cf. WALS (Dryer 2013)
- Family / Region
Indo-European (IE)/West-Germanic; IE/North-Germanic; IE/Romance; IE/Greek; IE/West Slavic; IE/South Slavic; IE/East Slavic; IE/Iranian; IE/Indic; Finno-Ugric/Finnic; Finno-Ugric/Ugric;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

Dependency Length Minimization

Futrell, Mahowald & Gibson, 2015, PNAS



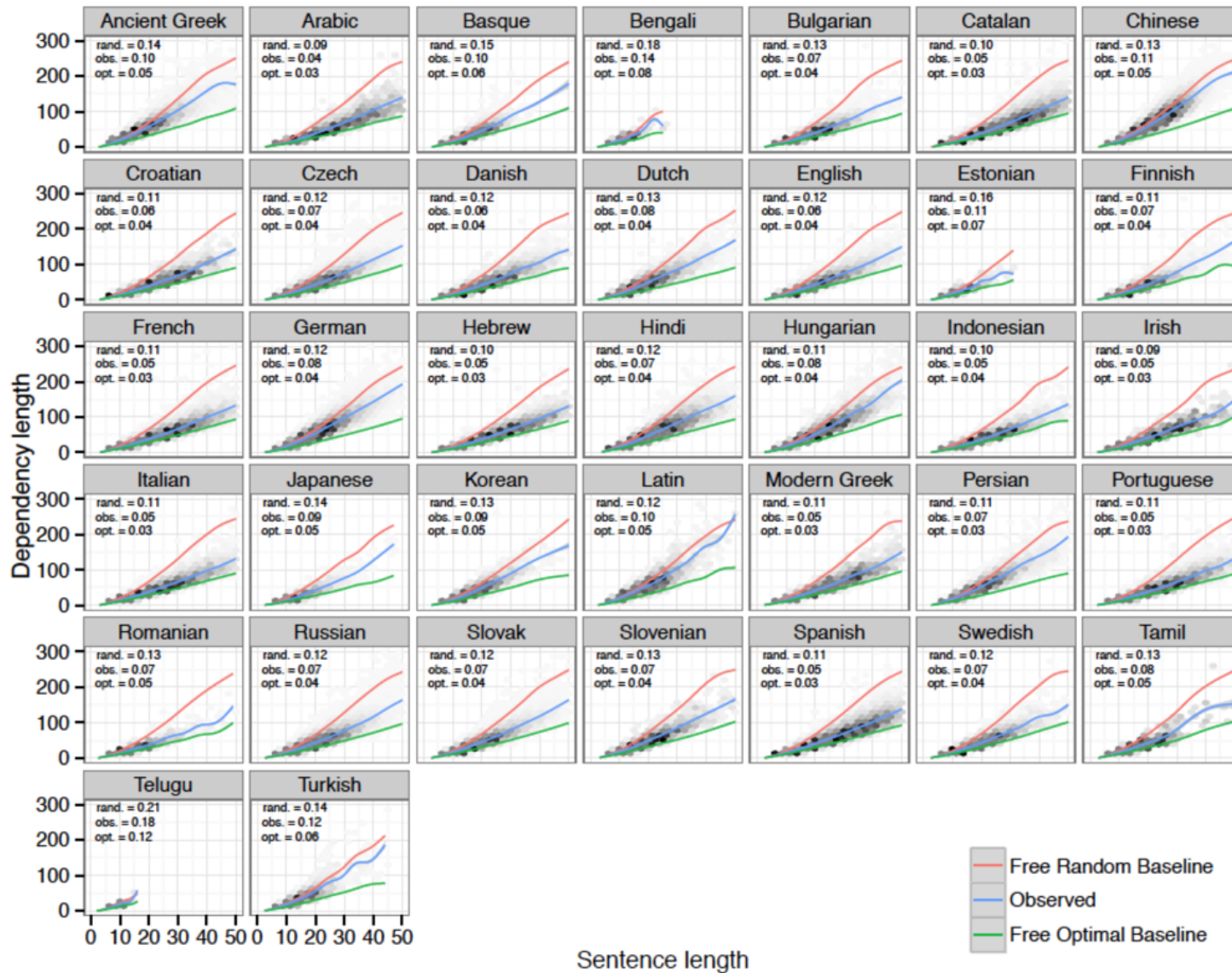
Sentence C: Total dependency length = 14



Sentence D: Total dependency length = 20

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

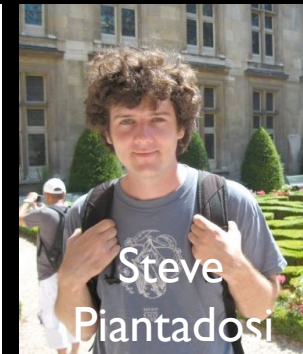
Conclusion: Information processing and cross-linguistic universals

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

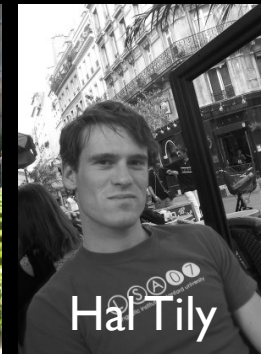
- The evolution of language:
 - Words (Piantadosi, Tily, & Gibson, 2011, 2012; Gibson, et al. 2016)
 - Syntax (Gibson, Piantadosi, Brink, Lim, Bergen & Saxe, 2013; Futrell, Hickey, Lee, Lim, Luchkina & Gibson., 2014; Futrell, Mahowald & Gibson, 2015a, 2015b)
- Language use
 - Sentence interpretation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



Steve
Piantadosi



Hal Tily



Kyle
Mahowald



Leon Bergen



Ev Fedorenko



Roger Levy



Bevil
Conway



Mitchell
Gibson



Julian
Jara-Ettinger



Richard Futrell