Is natural language a good code from an information-theoretic perspective?

Vera Demberg and Matt Crocker

Information Theory Lecture, Universität des Saarlandes

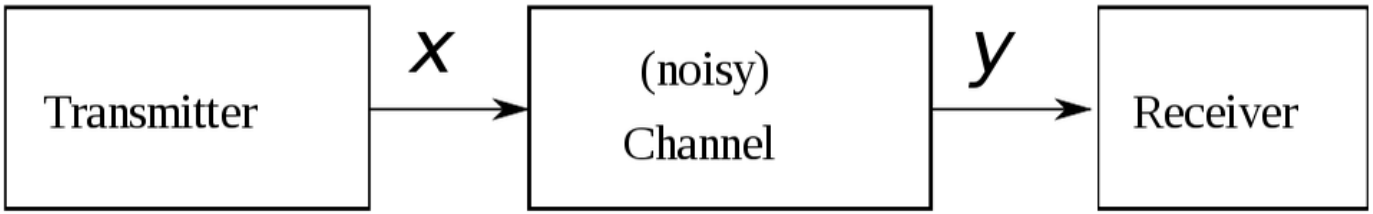
May 8 / 10th, 2017

UNIVERSITÄT
DES
SAARLANDES

Channel Capacity

Channel Capacity is defined as

the **maximal amount of information** that can be conveyed through a channel in a **given amount of time** t, with **arbitrarily small error** (= "without error").



$$\text{Channel Capacity } C = \frac{1}{t} \max_{p_X(x)} I(X;Y)$$

Noisy Channel means that

some of the **input might be corrupted during transmission**, so the output is different from the input. (E.g., for transmitting binary messages, some of the bits may be flipped during transmission.)

Optimal Communication through a Noisy Channel

Channel Capacity is defined as

the **maximal amount of information** that can be conveyed through a channel in a **given amount of time** t, with **arbitrarily small error** (= "without error").

What does this imply for optimal communication?

Optimal Communication through a Noisy Channel

Channel Capacity is defined as

the **maximal amount of information** that can be conveyed through a channel in a **given amount of time** t, with **arbitrarily small error** (= "without error").

What does this imply for optimal communication?

- How can we achieve optimal efficiency? (assuming no error)
- How can one avoid errors in a "noisy channel"?

Example:

we have a butterfly station, and want to transmit a message every day to another butterfly station, saying how far they have developed. Initially we have 5 eggs. States: egg, caterpillar, cocoon, butterfly.

Table of Contents

1. What is a good code?
2. Good Code and Ambiguity
3. Redundancy

Optimal efficiency of channel use in language

How can we address questions about the optimality of language?

Optimal efficiency of channel use in language

How can we address questions about the optimality of language?

1. quantify amount of information conveyed by linguistic events
2. then we can calculate what is the shortest code necessary
3. then we can compare this to how natural languages encode information

Entropy and Surprisal

Step 1: quantifying the amount of information conveyed.

entropy:

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

*(normally, we will talk about conditional probabilities for surprisal, i.e., $p(x|context)$)

Entropy and Surprisal

Step 1: quantifying the amount of information conveyed.

entropy:

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} = -\sum_x p(x) \log_2 p(x)$$

*(normally, we will talk about conditional probabilities for surprisal, i.e., $p(x|context)$)

Entropy and Surprisal

Step 1: quantifying the amount of information conveyed.

entropy:

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} = -\sum_x p(x) \log_2 p(x)$$

surprisal: (entropy = average surprisal for large numbers of events)

$$Surprisal^*(x) = -\log_2 p(x)$$

*(normally, we will talk about conditional probabilities for surprisal, i.e., $p(x|context)$)

Entropy and Surprisal

Step 1: quantifying the amount of information conveyed.

entropy:

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} = -\sum_x p(x) \log_2 p(x)$$

surprisal: (entropy = average surprisal for large numbers of events)

$$Surprisal^*(x) = -\log_2 p(x)$$

- allows us to quantify the information content of linguistic events.
- How many bits are minimally needed to convey a message?
- What's the optimal code for conveying a message?

*(normally, we will talk about conditional probabilities for surprisal, i.e., $p(x|context)$)

Entropy

Example: Simplified Polynesian

p	t	k	a	i	o
1/16	3/8	1/16	1/4	1/8	1/8

Calculating Entropy for Simplified Polynesian

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

$$= 2 * \frac{1}{16} \log_2 16 + 2 * \frac{1}{8} \log_2 8 + \frac{1}{4} \log_2 4 + \frac{3}{8} \log_2 \frac{8}{3}$$

$$= 2.28 \text{ bits per character}$$

Entropy

Example: Simplified Polynesian

p	t	k	a	i	o
1/16	3/8	1/16	1/4	1/8	1/8

optimal code: 2.28 bit/char

Possible Codes

p	t	k	a	i	o
000	001	010	011	100	101

Code for "pato" = 000011001101
(3 bits per character)

Entropy

Example: Simplified Polynesian

p	t	k	a	i	o
1/16	3/8	1/16	1/4	1/8	1/8

optimal code: 2.28 bit/char

Possible Codes

p	t	k	a	i	o
000	001	010	011	100	101

Code for "pato" = 000011001101
(3 bits per character)

p	t	k	a	i	o
100	101	110	111	00	01

Code for "pato" = 10011110101

Entropy

Example: Simplified Polynesian

p	t	k	a	i	o
1/16	3/8	1/16	1/4	1/8	1/8

optimal code: 2.28 bit/char

Possible Codes

p	t	k	a	i	o
000	001	010	011	100	101

Code for "pato" = 000011001101
(3 bits per character)

p	t	k	a	i	o
100	101	110	111	00	01

Code for "pato" = 10011110101
(2.75 bits per character)

average message length for this code:

$$\frac{1}{16} * 3 + \frac{3}{8} * 3 + \frac{1}{16} * 3 + \frac{1}{4} * 3 + \frac{1}{8} * 2 + \frac{1}{8} * 2 = 2.75 \text{ bits}$$

Entropy

Example: Simplified Polynesian

p	t	k	a	i	o
1/16	3/8	1/16	1/4	1/8	1/8

optimal code: 2.28 bit/char

Possible Codes

p	t	k	a	i	o
000	001	010	011	100	101

Code for "pato" = 000011001101
(3 bits per character)

p	t	k	a	i	o
100	101	110	111	00	01

Code for "pato" = 10011110101
(2.75 bits per character)

p	t	k	a	i	o
100	00	101	01	110	111

Code for "pato" = 1000100111
(2.375 bits per character)

(short codes for probable events to reach short average message length)

Entropy

Example: Simplified Polynesian

p	t	k	a	i	o
1/16	3/8	1/16	1/4	1/8	1/8

optimal code: 2.28 bit/char

and we might find an even better code if the letters in simplified Polynesian are not independent of one another, and we use a code that reflects that.

Syllable structure: CV

p	t	k	a	i	o
00	1	01	1	00	01

Code for "pato" = 001101

The CV model is a *better language model* for Polynesian than the one that does not know about syllable structure.

Information theory and natural language

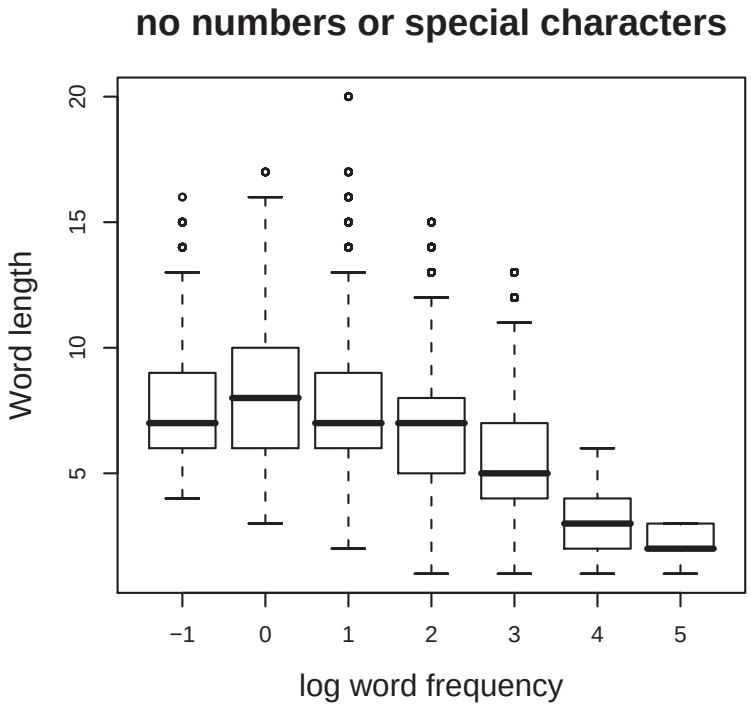
- How good a code is natural language?
- Are humans able to generate encodings of their messages that allow them to communicate optimally*?

* efficiently and effectively given their environment (= through a noisy channel)

Is natural language an efficient code?

- Efficient code is one that allows us to, on average, encode short messages.
- Observation: frequent events have shorter codes.

(Zipf, 1949) "Principle of least effort"

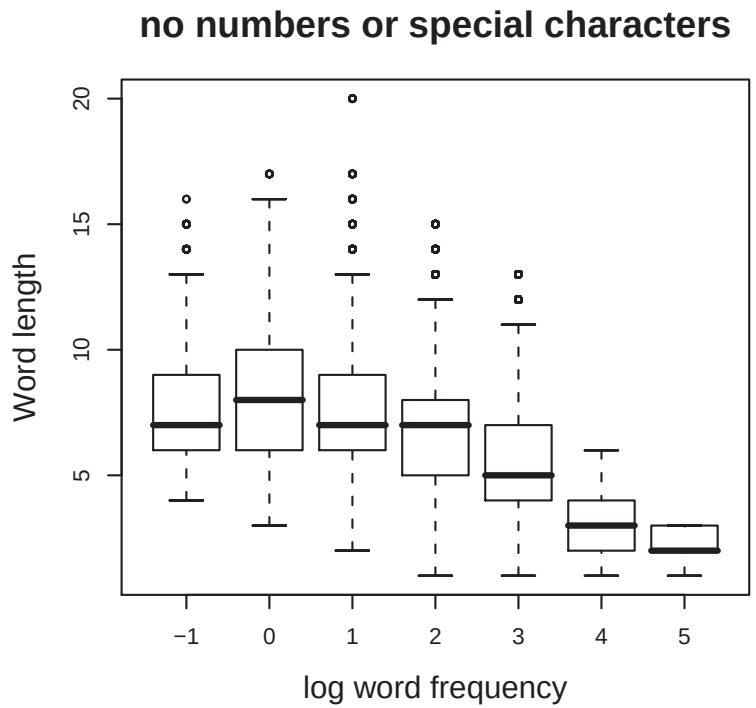


A language where frequent words are short is optimal if words occur independently of one another.

Is natural language an efficient code?

- Efficient code is one that allows us to, on average, encode short messages.
- Observation: frequent events have shorter codes.

(Zipf, 1949) "Principle of least effort"



A language where frequent words are short is optimal if words occur independently of one another.

But: in natural languages, word probabilities change depending on their context.

Entropy and language models

Remember:

- better language model = lower entropy = shorter codes

  Example: simplified Polynesian modelling letter frequencies vs. letter frequencies in syllable contexts.

Entropy and language models

Remember:

- better language model = lower entropy = shorter codes

  Example: simplified Polynesian modelling letter frequencies vs. letter frequencies in syllable contexts.

- Word frequency = very simple (unigram) language model

Entropy and language models

Remember:

- better language model = lower entropy = shorter codes

  Example: simplified Polynesian modelling letter frequencies vs. letter frequencies in syllable contexts.

- Word frequency = very simple (unigram) language model

- If the human language system has good code design, we'd expect a strong (positive) correlation between word length and information content (=surprisal) of a word given its context.

Piantadosi, Tily and Gibson 2011

Test whether surprisal of a word is correlated with word length.

How to test this (given a very large corpus)

- calculate the average surprisal of a word w given its context c:

$$\sum_c P(C = c|W = w) * -\log P(W = w|C = c)$$

Piantadosi, Tily and Gibson 2011

Test whether surprisal of a word is correlated with word length.

How to test this (given a very large corpus)

- calculate the average surprisal of a word w given its context c:

$$\sum_c P(C = c|W = w) * -\log P(W = w|C = c)$$

$$= -\frac{1}{N} \sum_{i=1}^{N} \log P(W = w|C = c_i)$$

(note that Zipf at the time did not have the means to test this.)

Piantadosi, Tily and Gibson 2011

Test whether surprisal of a word is correlated with word length.

How to test this (given a very large corpus)

- calculate the average surprisal of a word w given its context c:

$$\underbrace{\sum_c P(C = c|W = w)}_{\textit{weight}} * \underbrace{-\log P(W = w|C = c)}_{\textit{information content}}$$

(note that Zipf at the time did not have the means to test this.)

Piantadosi, Tily and Gibson 2011

Test whether surprisal of a word is correlated with word length.

How to test this (given a very large corpus)

- calculate the average surprisal of a word w given its context c:

$$\underbrace{\sum_c P(C = c|W = w)}_{\textit{weight}} * \underbrace{-\log P(W = w|C = c)}_{\textit{information content}}$$

- test whether word length negatively correlated with average word predictability
  (= word length positively correlated with average surprisal).

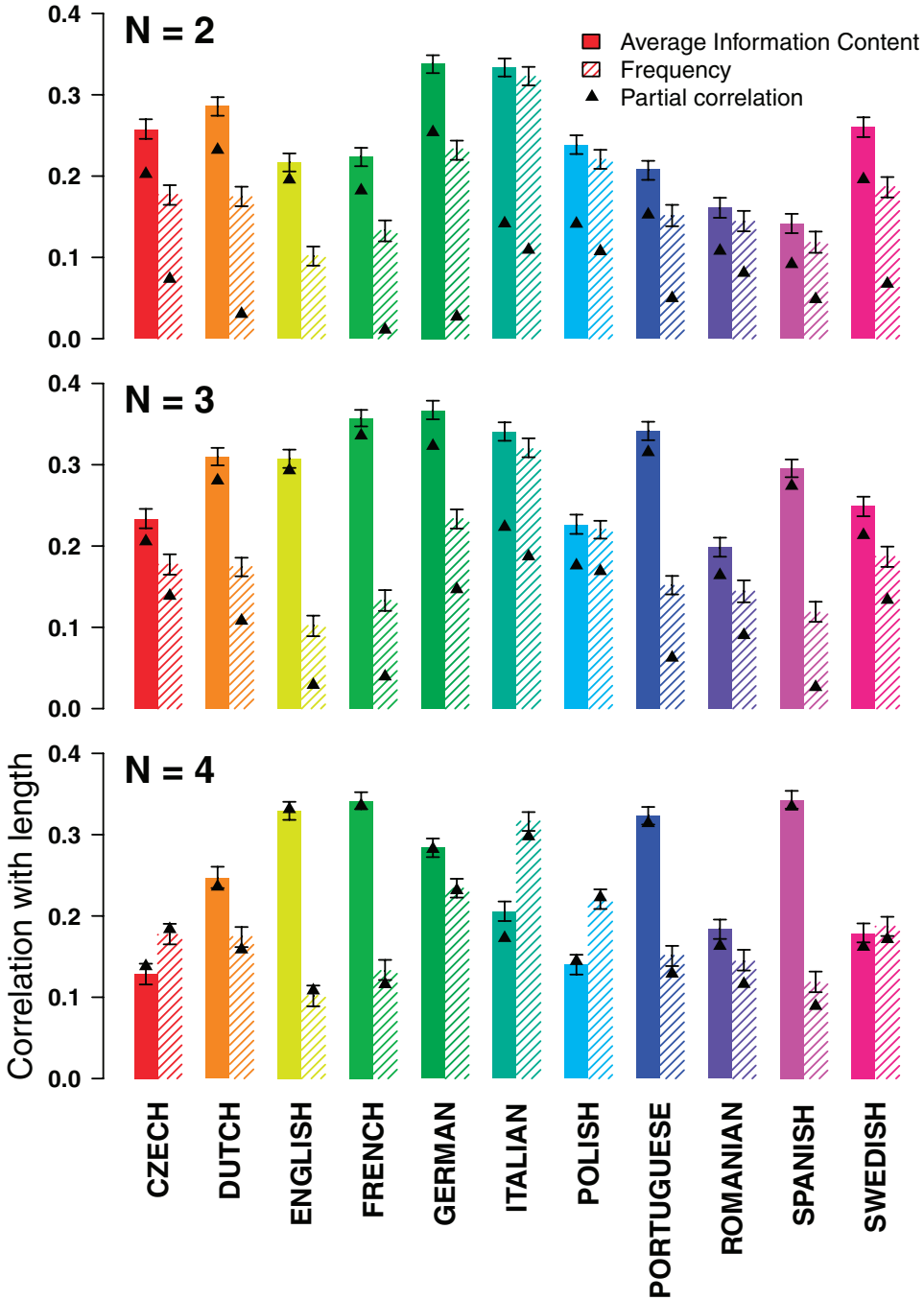
(note that Zipf at the time did not have the means to test this.)

Experiments: Piantadosi, Tily, Gibson (2011)

1. Google n-grams
   - 500 billion words for English
   - 11 languages tested (ca 130 billion words for other languages)
   - n-gram model, no smoothing
   - word length in characters
   - on 25k most common words; each data set has ca. 10m different words
2. British National Corpus
   - 100 million words
   - English only
   - n-gram model, with smoothing
   - word length in characters
3. Word length estimate in terms of phones / syllables
   - German, Dutch, English

Experiments: Piantadosi, Tily, Gibson (2011)

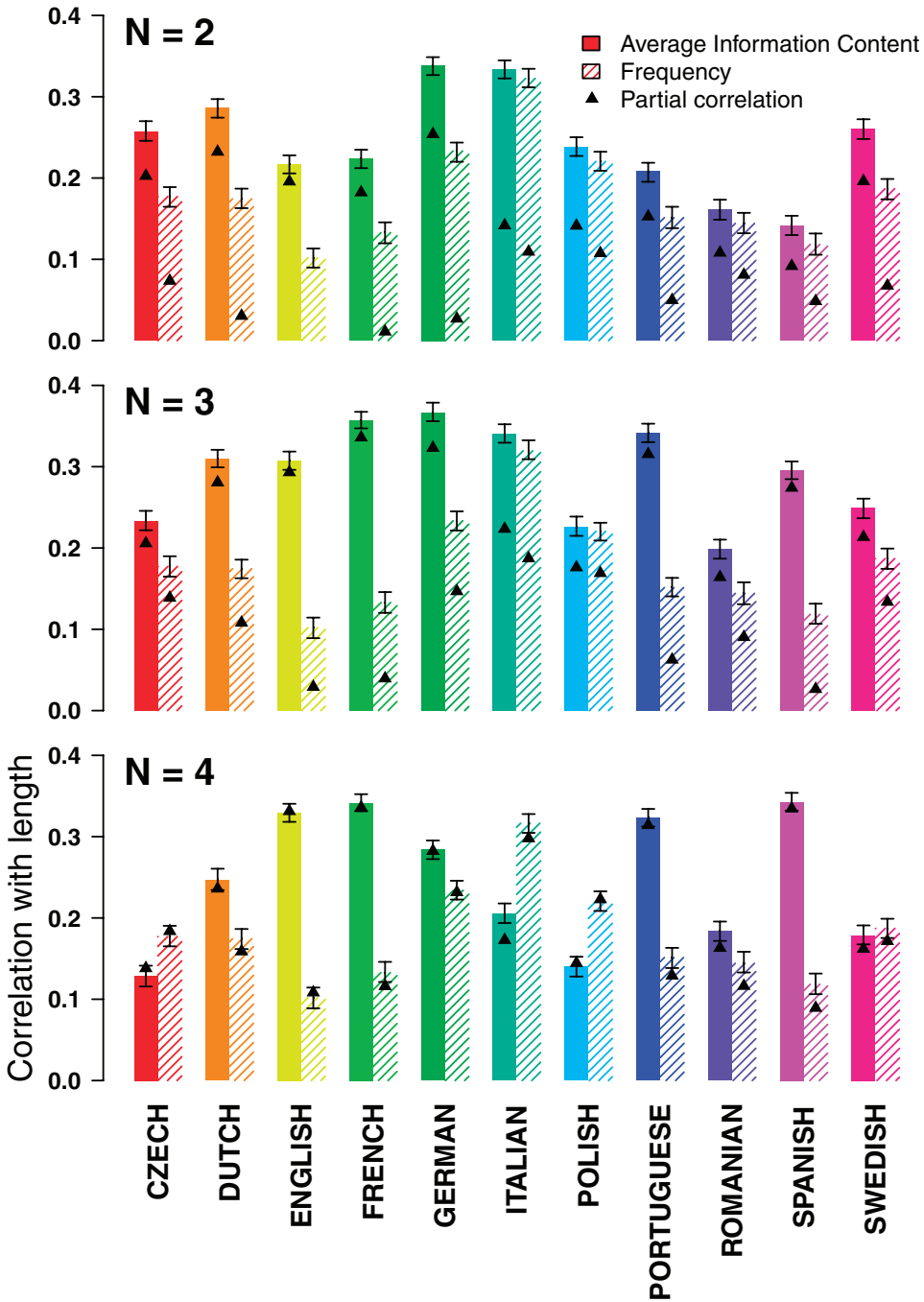
1. **Google n-grams**
   - 500 billion words for English
   - 11 languages tested (ca 130 billion words for other languages)
   - n-gram model, no smoothing
   - word length in characters
   - on 25k most common words; each data set has ca. 10m different words
2. British National Corpus
   - 100 million words
   - English only
   - n-gram model, with smoothing
   - word length in characters
3. Word length estimate in terms of phones / syllables
   - German, Dutch, English



Google n-grams

Methods

- Spearman's rank correlation



Google n-grams

Methods

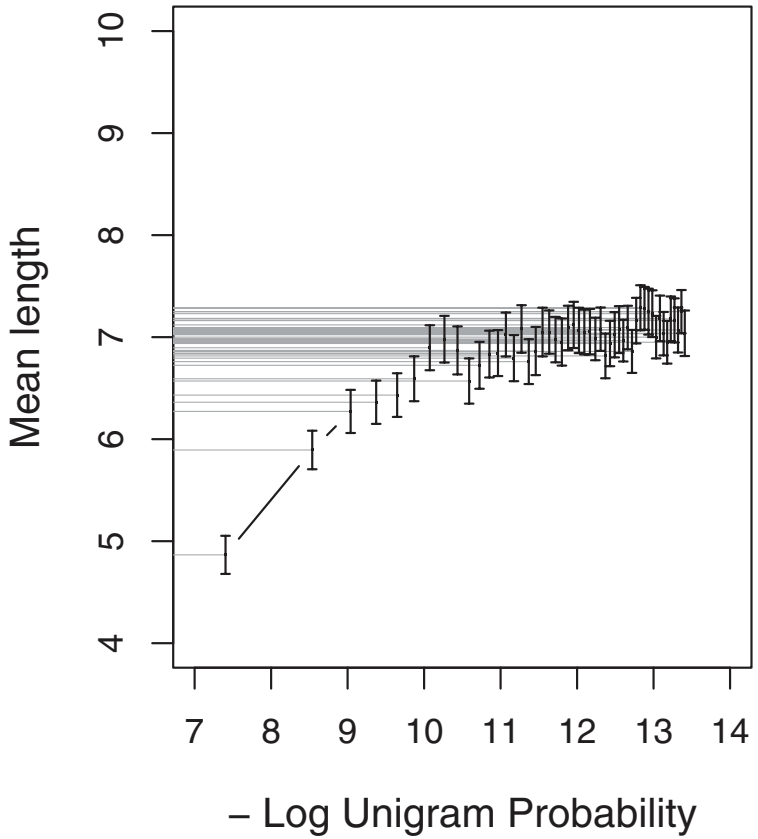
- Spearman's rank correlation

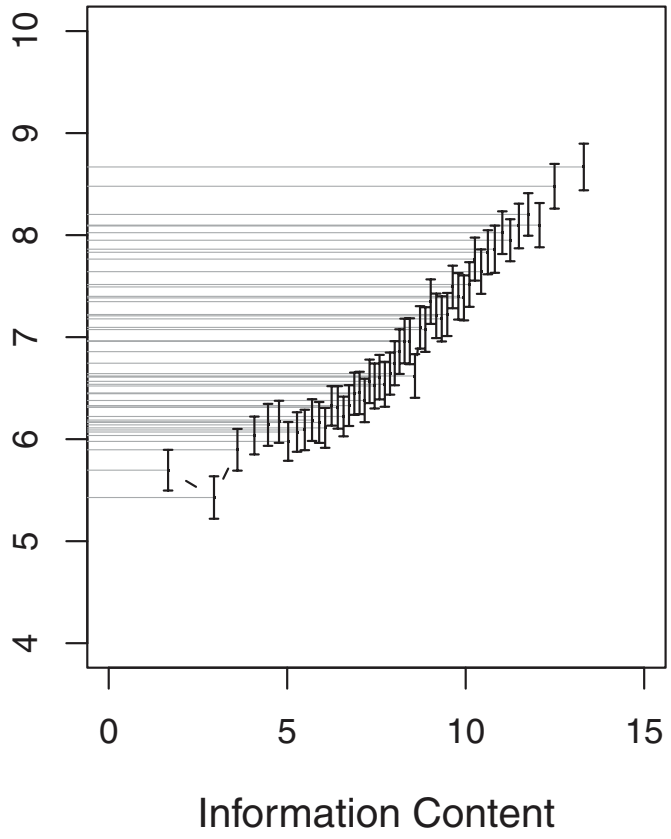
Results

- information content better predictor than frequency
- 2-grams: correct for all languages
- 3-grams: significant for all except Polish
- 4-grams: except Czech, Italian, Polish, Swedish
- Partial correlations provide a similar picture

Predictability vs. Frequency

Correlation of frequency vs. length and predictability vs. length (English)

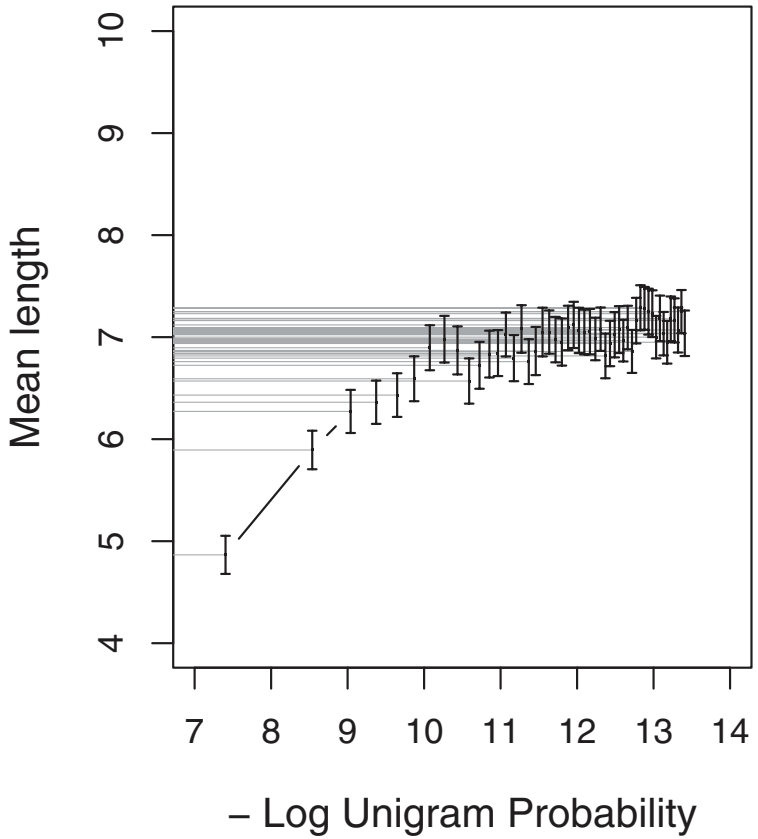


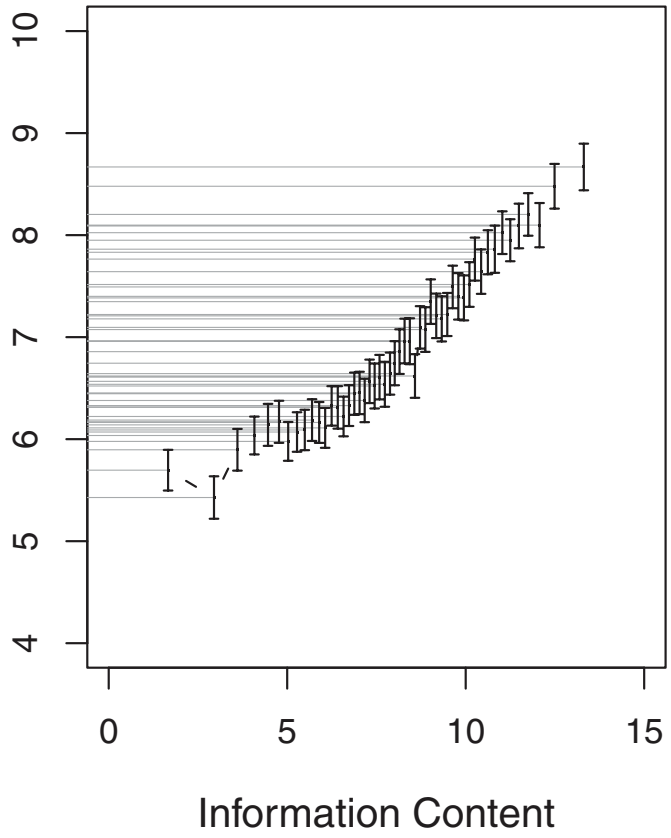


(high frequency = small –Log Unigram Probability)

Predictability vs. Frequency

Correlation of frequency vs. length and predictability vs. length (English)





Frequency doesn't predict word length well for low-frequency words, but average information content does.

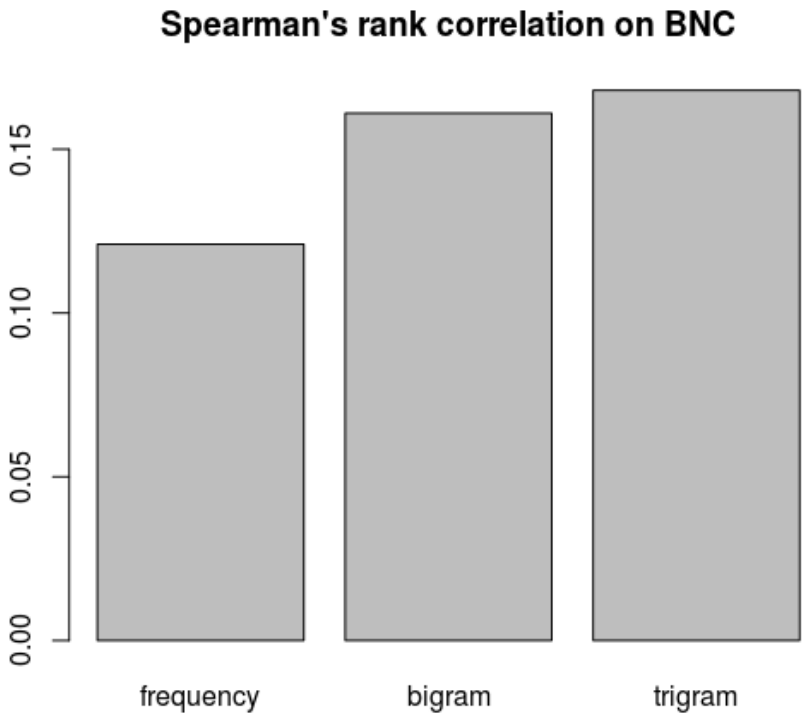
Experiments

1. Google n-grams
   - 500 billion words for English
   - 11 languages tested (ca 130 billion words for other languages)
   - n-gram model, no smoothing
   - word length in characters
   - on 25k most common words; each data set has ca. 10m different words
2. **British National Corpus**
   - 100 million words
   - English only
   - n-gram model, with smoothing
   - word length in characters
3. Word length estimate in terms of phones / syllables
   - German, Dutch, English

BNC results

British National Corpus

- predictability of words better than frequency
- after partialing out: no effect of frequency beyond predictability



Experiments

1. Google n-grams
   - 500 billion words for English
   - 11 languages tested (ca 130 billion words for other languages)
   - n-gram model, no smoothing
   - word length in characters
   - on 25k most common words; each data set has ca. 10m different words
2. British National Corpus
   - 100 million words
   - English only
   - n-gram model, with smoothing
   - word length in characters
3. **Word length estimate in terms of phones / syllables**
   - German, Dutch, English

Word length in phones / syllables

Results for word length estimate in terms of phones / syllables

Does **predictability work better than frequency?**

language	2-gram	3-gram	4-gram
English phonemes	yes	yes	yes
English syllables	yes	yes	yes
Dutch phonemes	yes	yes	tendency
Dutch syllables	yes	yes	tendency
German phonemes	tendency	no	no
German syllables	yes	tendency	no

Discussion

1. How does Piantadosi et al's account differ from Zipf's account?
2. How does the lexicon end up having the property of more predictable words being encoded with shorter codes?
3. Ways of measuring information content given the context?

Discussion

1. How does Piantadosi et al's account differ from Zipf's account?
2. How does the lexicon end up having the property of more predictable words being encoded with shorter codes?
3. Ways of measuring information content given the context?

Discuss ideas and speculations on these questions with your neighbors

Discussion

1. **How does Piantadosi et al's account differ from Zipf's account?**
2. How does the lexicon end up having the property of more predictable words being encoded with shorter codes?
3. Ways of measuring information content given the context?

Discuss ideas and speculations on these questions with your neighbors

Differences Zipf's account vs. information content

- Lexicon

Differences Zipf's account vs. information content

- Lexicon
  - Zipf (principle of least effort): best lexicon is the most concise one
  - Information content: highly informative word should not necessarily be shortened even if shorter codes are still available.

Differences Zipf's account vs. information content

- Lexicon
  - Zipf (principle of least effort): best lexicon is the most concise one
  - Information content: highly informative word should not necessarily be shortened even if shorter codes are still available.
- Message

Differences Zipf's account vs. information content

- Lexicon
  - Zipf (principle of least effort): best lexicon is the most concise one
  - Information content: highly informative word should not necessarily be shortened even if shorter codes are still available.
- Message
  - Information content: Easily predictable information has a short code, lots of information has a long code.
  - Using such a code, information is conveyed at an almost constant rate.
  - Good usage of channel capacity.
  - Zipf's account does not smooth out information content over time.

Zipf vs. Piantadosi

Relationship of the two accounts

- predictability account as a **refinement** of Zipf's account.
- assigning word length by information content is
  - least effort if one assumes a superlinear relationship between effort and information content.
  - optimal given a constraint on channel capacity.

Discussion

1. How does the Piantadosi et al's account differ from Zipf's account?
2. **How does the lexicon end up having this property?**
3. Ways of measuring information content given the context?

How come that natural language has developed this property?

Everyday language usage

- information content influences pronunciation (duration, clarity)
- abbreviations are used for long words which are predictable / familiar to the listener

Language evolution

- through lexicalization, predictable words end up being shorter.
- over time, abbreviations can become the new lexical item.

Discussion

1. How does the Piantadosi et al's account differ from Zipf's account?
2. How does the lexicon end up having this property?
3. **Ways of measuring information content given the context?**

Estimating Predictability

The role of context:

- discourse context
- local linguistic context
- syntactic context
- world knowledge

→ n-gram model estimation clearly a rough simplification
(might not work so well for languages with freer word order)

Questions:

- **Could the effect of frequency be explained away completely if we had better models of predictability in context?**

Estimating Predictability

The role of context:

- discourse context
- local linguistic context
- syntactic context
- world knowledge

→ n-gram model estimation clearly a rough simplification
(might not work so well for languages with freer word order)

Questions:

- **Could the effect of frequency be explained away completely if we had better models of predictability in context?**
- What about effects of a word's meaning on length and difficulty beyond its predictability? (denotation, connotation, significance)

Estimating Predictability

The role of context:

- discourse context
- local linguistic context
- syntactic context
- world knowledge

→ n-gram model estimation clearly a rough simplification (might not work so well for languages with freer word order)

Questions:

- **Could the effect of frequency be explained away completely if we had better models of predictability in context?**
- What about effects of a word's meaning on length and difficulty beyond its predictability? (denotation, connotation, significance)
- N-gram estimation impoverished but mathematically precise meaning of quantifying the number of bits needed to encode the message

Reactions to Piantadosi et al., 2011

Response Letter by Reilly and Kean

Jamie Reilly and Jacob Kean wrote a letter which was published in PNAS (2011) in response to Piantadosi et al.'s article.

Response Letter by Reilly and Kean

Jamie Reilly and Jacob Kean wrote a letter which was published in PNAS (2011) in response to Piantadosi et al.'s article.

Main objections:

- information content is just one aspect of a word
- analysis is confounded by syntactic dimensions such as grammatical class (e.g., nouns vs. verbs) and other dimensions of word meaning (e.g., word concreteness):
  - verbs are longer than nouns in many languages
  - abstract words tend to be longer than concrete ones.
    (due to inflection, e.g.: friend vs. friendliness)

Response Letter by Reilly and Kean

Jamie Reilly and Jacob Kean wrote a letter which was published in PNAS (2011) in response to Piantadosi et al.'s article.

Main objections:

- information content is just one aspect of a word
- analysis is confounded by syntactic dimensions such as grammatical class (e.g., nouns vs. verbs) and other dimensions of word meaning (e.g., word concreteness):
  - verbs are longer than nouns in many languages
  - abstract words tend to be longer than concrete ones.
    (due to inflection, e.g.: friend vs. friendliness)
- **Do verbs and abstract nouns convey more information content than concrete nouns?**

Reply by Piantadosi, Tily and Gibson (2011)

Do verbs and abstract nouns convey more information content than concrete nouns?

- Yes.

Corpus study

verbs / abstract nouns are **longer** than concrete nouns
(verbs and abstract nouns: 7.24 chars; concrete nouns: 5.99 chars)

verbs and abstract nouns convey **more information** than concrete nouns:
(verbs and abstract nouns: 8.23 bits; concrete nouns: 7.52 bits)

- concreteness ratings from MRC psycholinguistic database
- selected only unambiguous nouns/verbs

Reply by Piantadosi, Tily and Gibson (2011)

Do verbs and abstract nouns convey more information content than concrete nouns?

- Yes.

Corpus study

verbs / abstract nouns are **longer** than concrete nouns
(verbs and abstract nouns: 7.24 chars; concrete nouns: 5.99 chars)

verbs and abstract nouns convey **more information** than concrete nouns:
(verbs and abstract nouns: 8.23 bits; concrete nouns: 7.52 bits)

- concreteness ratings from MRC psycholinguistic database
- selected only unambiguous nouns/verbs

Information content correlates negatively with concreteness ($R = -0.24$). Why?

Reply by Piantadosi, Tily and Gibson (2011)

Do verbs and abstract nouns convey more information content than concrete nouns?

- Yes.

Corpus study

verbs / abstract nouns are **longer** than concrete nouns
(verbs and abstract nouns: 7.24 chars; concrete nouns: 5.99 chars)

verbs and abstract nouns convey **more information** than concrete nouns:
(verbs and abstract nouns: 8.23 bits; concrete nouns: 7.52 bits)

- concreteness ratings from MRC psycholinguistic database
- selected only unambiguous nouns/verbs

Information content correlates negatively with concreteness ($R = -0.24$). Why?

$\rightarrow$ predictable words, which tend to be concrete, will over time get shortened, while the abstract ones will then be longer in comparison.

Only offline or also online?

The results from Piantadosi et al. indicate that language developed such that information is conveyed at a uniform rate.

Only offline or also online?

The results from Piantadosi et al. indicate that language developed such that information is conveyed at a uniform rate.

But is this just a long-term effect, or do people try to convey information uniformly during online production?

Controlling information rate online

Uniform Information Density hypothesis:
Speakers make use of the variability present in language during online production to achieve a uniform rate of conveying information.

Controlling information rate online

Uniform Information Density hypothesis:
Speakers make use of the variability present in language during online production to achieve a uniform rate of conveying information.

(Next week, we will look at work showing this more generally.)

Controlling information rate online

Uniform Information Density hypothesis:
Speakers make use of the variability present in language during online production to achieve a uniform rate of conveying information.

(Next week, we will look at work showing this more generally.)

But at the level of the lexicon, it seems hard for people to manipulate length.

Controlling information rate online

Uniform Information Density hypothesis:
Speakers make use of the variability present in language during online production to achieve a uniform rate of conveying information.

(Next week, we will look at work showing this more generally.)

But at the level of the lexicon, it seems hard for people to manipulate length.
$\rightarrow$ Is there any evidence that people do this?

Wasn't this already shown in Piantadosi (2011)?

Do people manipulate word length during production to achieve a uniform rate of information transmission?

Wasn't this already shown in Piantadosi (2011)?

Do people manipulate word length during production to achieve a uniform rate of information transmission?

Not really:

- no control for meaning
- relationship could arise from differences between classes of words
  - e.g.: function words shorter and less informative than content words
  - e.g.: difference between nouns and verbs, or other regularities

Wasn't this already shown in Piantadosi (2011)?

Do people manipulate word length during production to achieve a uniform rate of information transmission?

Not really:

- no control for meaning
- relationship could arise from differences between classes of words
  - e.g.: function words shorter and less informative than content words
  - e.g.: difference between nouns and verbs, or other regularities

Mahowald et al. study

Test for correlation between length and information content while keeping the meaning constant.

chimp / *chimpanzee*
math / *mathematics*
chemo / *chemotherapy*

Expected Pattern

Predictions:

predictive context	$\rightarrow$	exam
non-constraining context	$\rightarrow$	examination

If this holds: People actively choose forms during production.

Otherwise: Effect in Piantadosi et al. likely to simply arise from differences in word classes
or long-term pressure for linguistic efficiency,
but not active speaker choice.

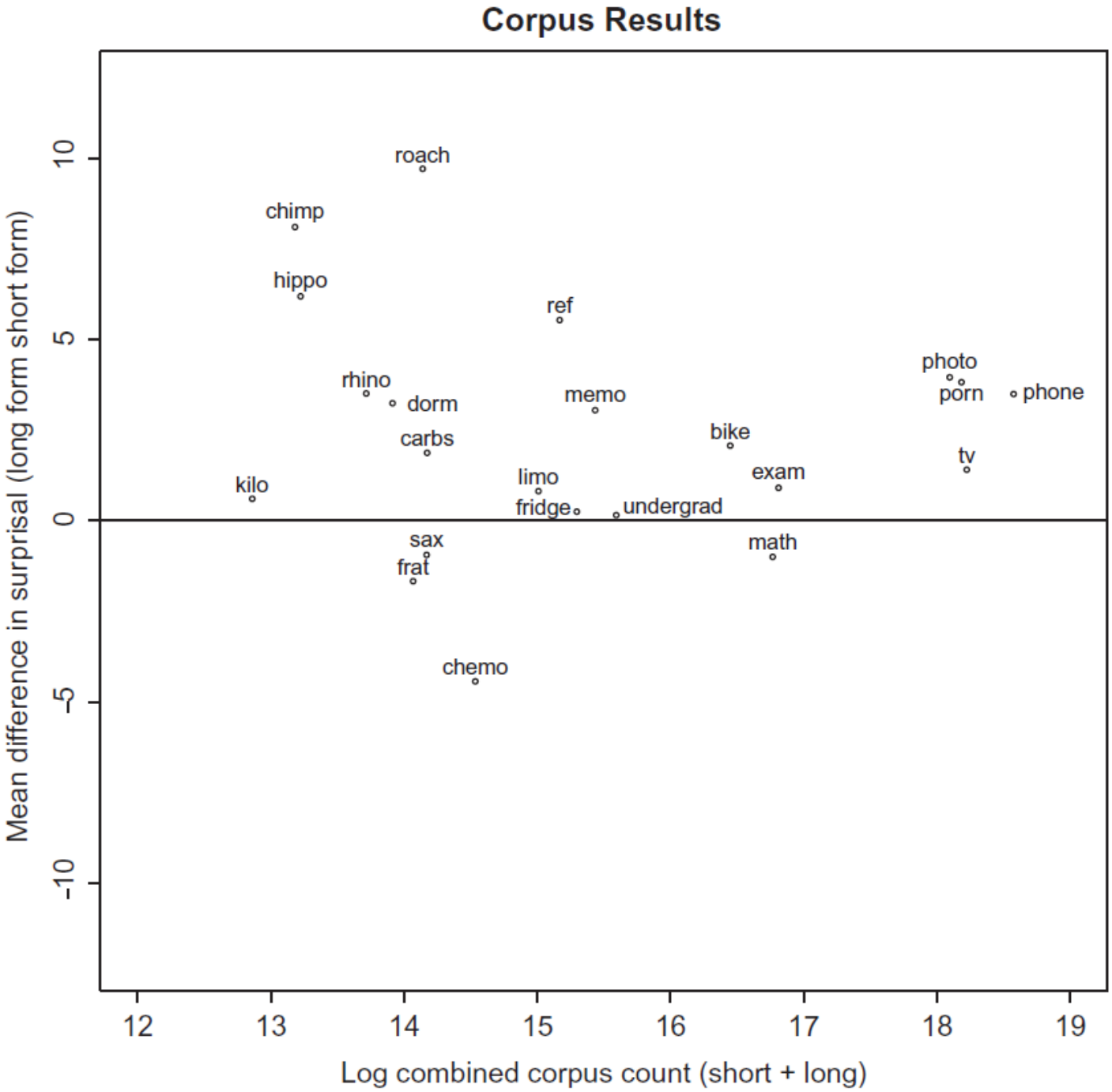
Corpus Study

Compare unambiguous forms that are near-synonyms.

$$-\frac{1}{N}\sum_{i=1}^{N}\log P(W = w|C = c_i)$$

avg. information content of shorter forms (exam): 6.90
avg. information content of longer forms (examination): 9.21

- possible confounding factor: short word forms are also more frequent
- but effect of information content also hold after accounting for frequency effect


Corpus Results
Mean difference in surprisal (long form short form)
Log combined corpus count (short + long)
10
5
0
-5
-10
12
13
14
15
16
17
18
19
roach
chimp
hippo
ref
photo
porn
phone
rhino
dorm
memo
bike
carbs
tv
exam
kilo
limo
fridge
undergrad
sax
math
frat
chemo

Completion Experiment (Mechanical Turk)

Forced choice completion test with neutral vs. predictive context:

Example item: math / mathematics

Supportive: Susan was very bad at algebra, so she hated ...

Neutral: Susan introduced herself to me as someone who loved ...

Experimental control

- no common phrases like "final exam"
- cloze completion task to test supportive / neutral context (supportive: 52.4% cloze prob, 1.6% cloze prob)

Completion Experiment (Mechanical Turk)

Forced choice completion test with neutral vs. predictive context:

Example item: math / mathematics

Supportive: Susan was very bad at algebra, so she hated ...

Neutral: Susan introduced herself to me as someone who loved ...

Results

- short form chosen more often in predictive context (67%) than neutral context (56%).
- short form chosen more often overall.

Completion Experiment (Mechanical Turk)

Forced choice completion test with neutral vs. predictive context:

Example item: math / mathematics

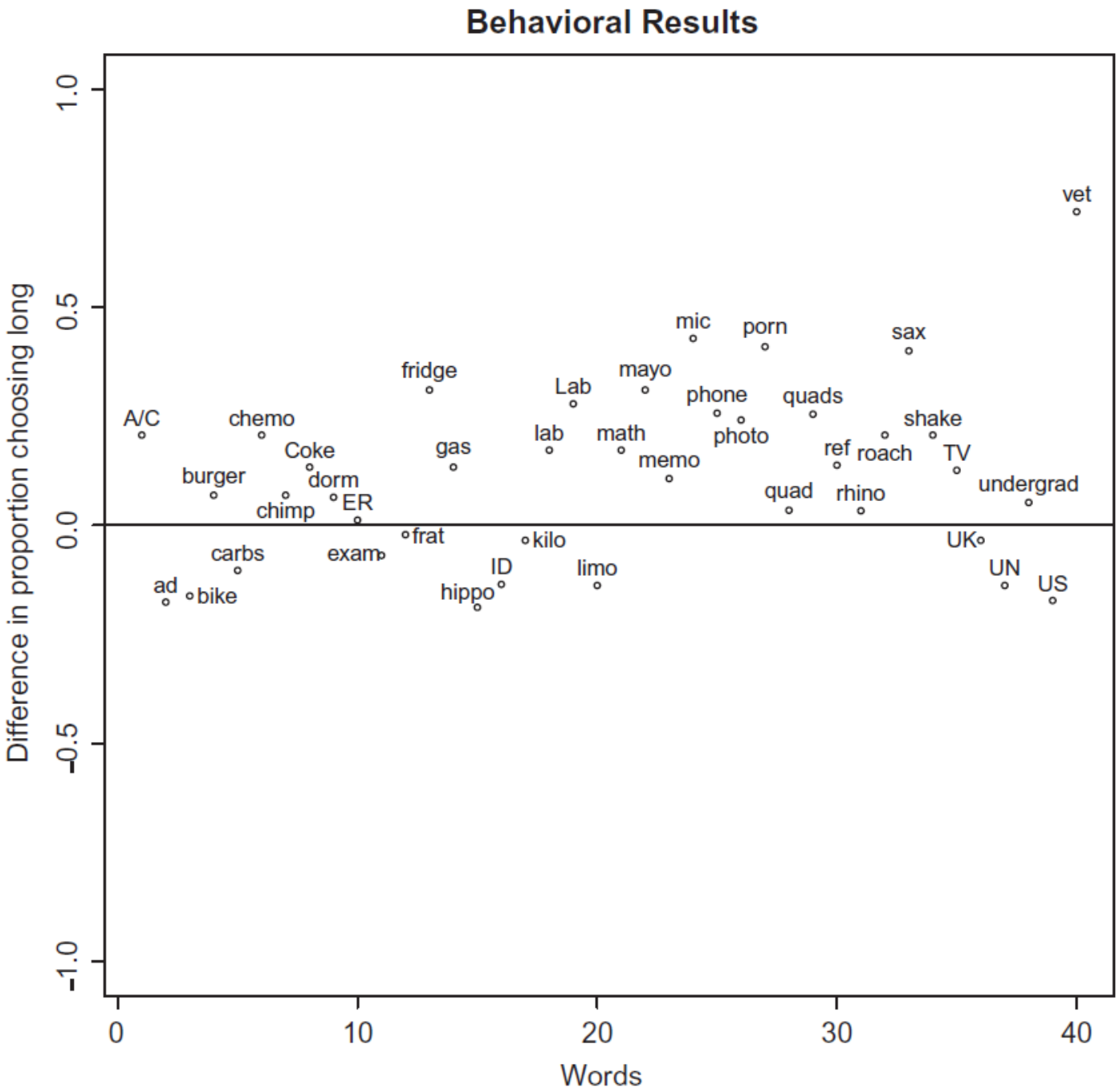
Supportive: Susan was very bad at algebra, so she hated ...

Neutral: Susan introduced herself to me as someone who loved ...

Results

- short form chosen more often in predictive context (67%) than neutral context (56%).
- short form chosen more often overall.

Speakers actively select shorter word forms in more predictive contexts.


Behavioral Results
Difference in proportion choosing long
Words
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
vet
mic
porn
sax
fridge
mayo
Lab
phone
quads
A/C
chemo
lab
math
photo
shake
Coke
gas
memo
ref
roach
TV
burger
dorm
quad
rhino
undergrad
chimp
ER
frat
kilo
UK
carbs
exam
ID
limo
UN
ad
bike
hippo
US

Discussion

Information content given context

- corpus study: predictability given n-gram model
- forced choice completion study: many linguistic factors determine predictability
- $\rightarrow$ results hold across both ways of estimating information content

At what level of abstraction does this knowledge exist?

- association between word length and informativity?
- learned preferences for different types of context?

Relation between online choice and long-term lexicon evolution

- a word's information content decreases $\rightarrow$ abbreviation

Some questions for discussion

- For the corpus study on chimp/chimpanzee, why not calculate predictability of {chimp,chimpanzee}?
- Lexical items can be shortened as they get more predictable on average, can words that get less common also become longer?
- We've been talking about word lengths — does this mean that Finnish/Turkish etc. are badly-designed languages?

Table of Contents

1 What is a good code?

2 Good Code and Ambiguity

3 Redundancy

Ambiguity (Piantadosi et al., 2012)

- Ambiguity exists at all levels of linguistic analysis.
  - words (homonymy, polysemy)
  - morphemes (-s in work*s*, ball*s*)
  - syntactic ambiguity (e.g., attachment)
  - semantic ambiguity (e.g., scopus)
- Wouldn't it be much better for communication, if language was unambiguous???

Ambiguity

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).

Ambiguity

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).

Good code and ambiguity

Remember that we already saw a very good ambiguous code earlier today:

Syllable structure: CV

p	t	k	a	i	o
00	1	01	1	00	01

Code for "pato" = 001101

1 → ambiguous between t and a *when out of context*
00 → ambiguous between p and i *when out of context*
01 → ambiguous between k and o *when out of context*

Why ambiguity?

Hypothesis:
Ambiguity makes it possible to reuse short and simple codes.

- context (situation, world knowledge, linguistic context) very often disambiguates meaning
- an unambiguous signal would then be redundant.
- ambiguity allows for multiple use of speech signals that are easy to
  - produce
  - comprehend
  - remember

The extremes

- no ambiguity: different word for every meaning, marking of every attachment ambiguity etc.
- full ambiguity: just one totally ambiguous word, e.g. "ba".

Is ambiguity optimal?

An optimally efficient communication system will not convey unnecessary (redundant) information.

Therefore, ambiguity is beneficial IF:

- speakers can effectively use the context for disambiguation
- AND ambiguous words usually occur in disambiguating contexts
- AND ambiguous words are the ones that are easy to produce

Experiment Piantadosi et al., 2012

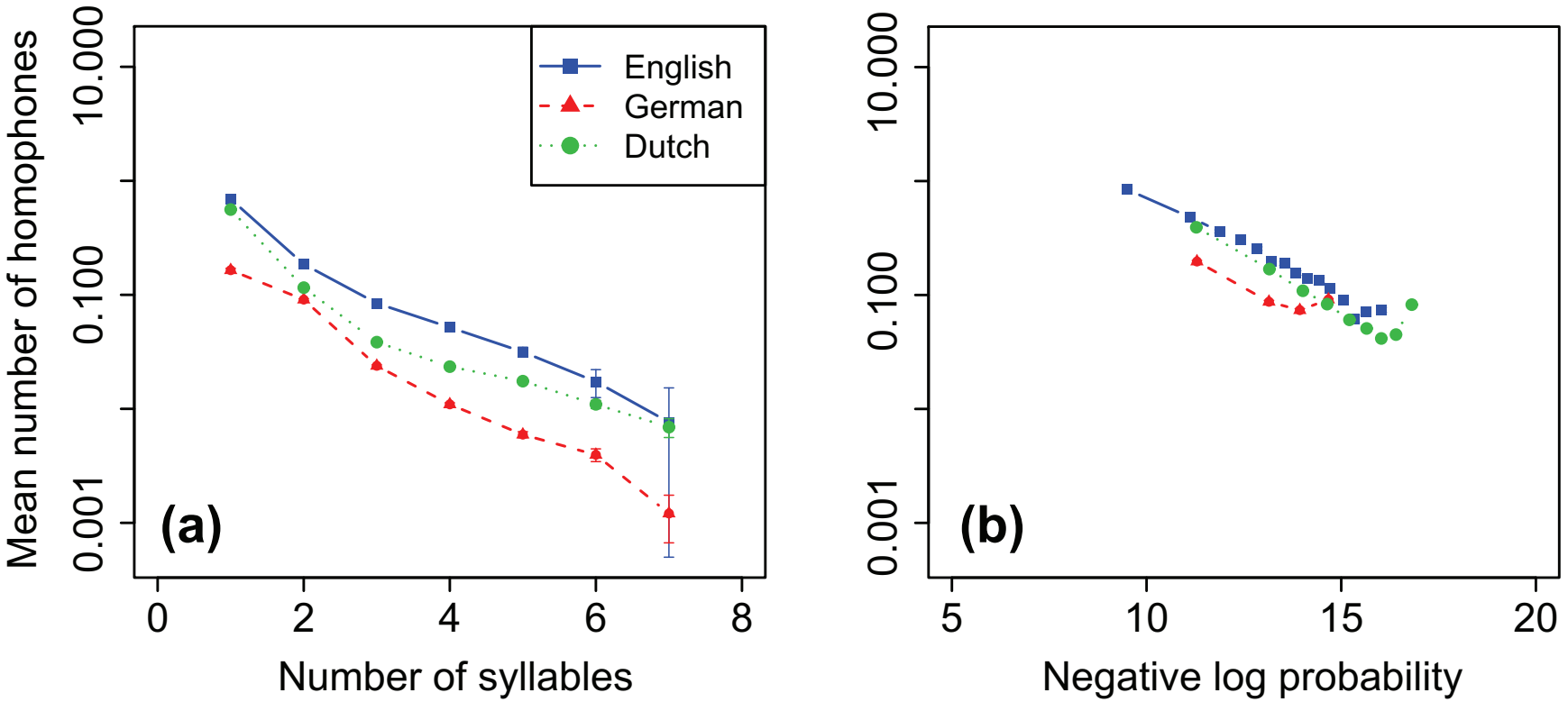
Question: How is ambiguity distributed?

- **If ambiguity is a random bug**
  short and common words, syllables and phonemes should not be more likely to be ambiguous than long and rare ones.

- **If ambiguity is a means for optimal code**
  we should be able to observe that short words / syllables / phonemes are more ambiguous, so these easy and short units can be reused.

Piantadosi et al., 2012

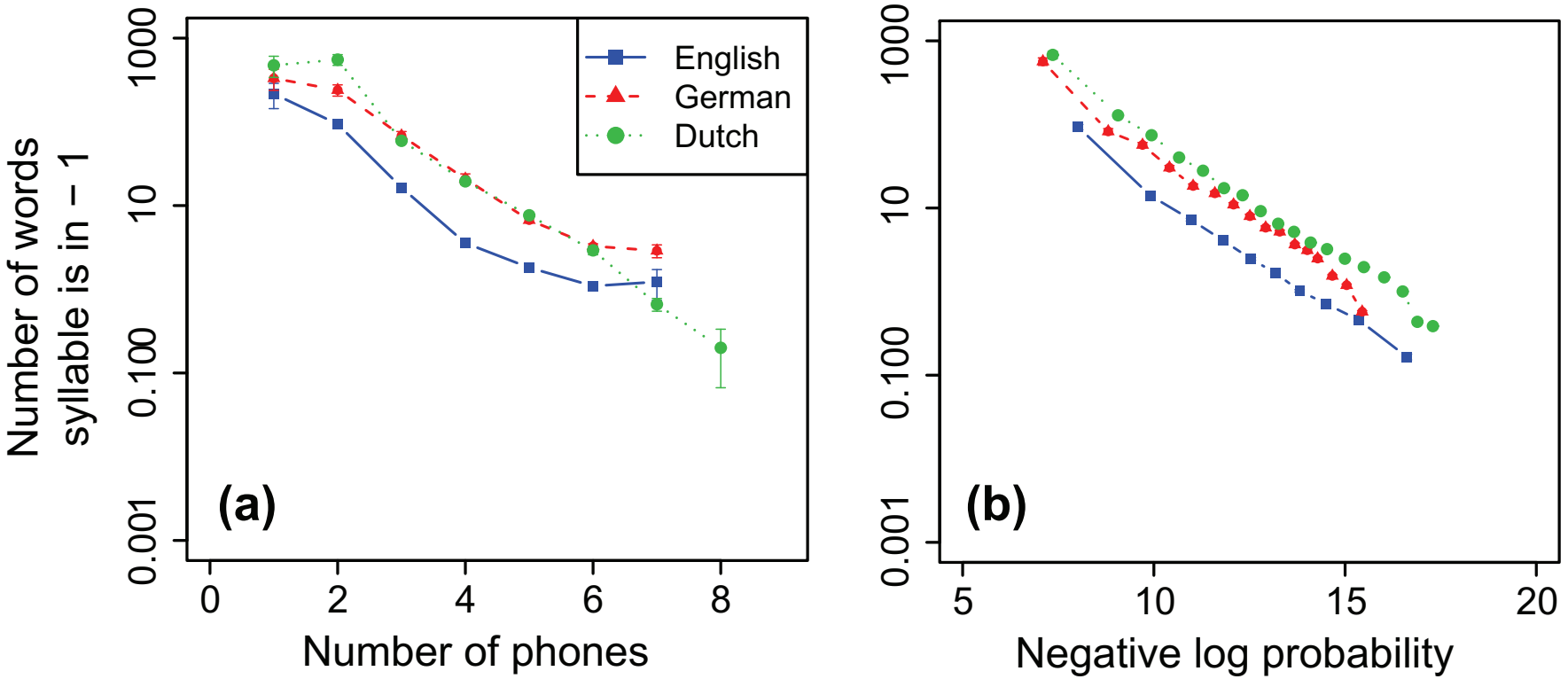
Homophones: (short and frequent words should be more ambiguous)



(low negative log probability = high frequency)

Piantadosi et al., 2012

Syllables (short and frequent syllables should be more ambiguous)



(low negative log probability = high frequency)

Conclusion on Ambiguity

For word homophones and syllables, the predictions for *ambiguity as a way to optimize code* were borne out:
Short and frequent words and syllables were more ambiguous than long / rare ones.

Relation to UID

- Good codes achieve more uniform information density on average
- but there are many short codes available in our languages which are not used
- information density / code length focusses on information per unit of language / unit of time, and does not take into account "production ease".

Table of Contents

1 What is a good code?

2 Good Code and Ambiguity

3 Redundancy

Noisy Channel

Getting back to the **noisy channel**:

- What if transmission isn't error-free?
- If we have a maximally efficient code, how does a noisy channel affect message transfer?
- How can we avoid this?

Noisy Channel

Getting back to the **noisy channel**:

- What if transmission isn't error-free?
- If we have a maximally efficient code, how does a noisy channel affect message transfer?
- How can we avoid this?

simple idea: **Redundancy**

Noisy Channel

Getting back to the **noisy channel**:

- What if transmission isn't error-free?
- If we have a maximally efficient code, how does a noisy channel affect message transfer?
- How can we avoid this?

simple idea: **Redundancy** (e.g., repeat each word three times)

Noisy Channel

Dealing with a noisy channel: the NATO phonetic alphabet.

system of naming letters which is used by the military and pilots:
A is Alpha
B is Bravo
C is Charlie
etc.

Channel Capacity again

We said at the beginning:
Channel capacity is defined as the maximal amount of information that can be conveyed through a channel in a given amount of time, with arbitrarily small error.

Channel capacity depends on speed of transmission and level of noise.
or
Speed of transmission depends on channel capacity and level of noise.

We can use these notions to derive predictions about language processing, in a noisy environment and/or with a smaller channel.

Discussion topics

- What's the channel in language comprehension? Input? Output? What's the information flow?

Discussion topics

- What's the channel in language comprehension? Input? Output? What's the information flow?
- What does the general knowledge about the fact that the channel may be noisy entail for rational communication?

Discussion topics

- What's the channel in language comprehension? Input? Output? What's the information flow?
- What does the general knowledge about the fact that the channel may be noisy entail for rational communication?
- What happens if we try to transmit information at a rate higher than channel capacity?
  Does this imply anything about the uniformity of transmission rate?