Simplicity and informativeness in conceptual structure

Jon W. Carr

Centre for Language Evolution
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
Pressures shaping language

Language
Pressures shaping language

Learning

Language

Simplicity
Pressures shaping language

Language

Simplicity
Informativeness

Learning

Interaction
Pressures shaping language

The simplicity–informativeness tradeoff
Kinship terms are simple and informative

Kinship Categories Across Languages Reflect General Communicative Principles

Charles Kemp and Terry Regier

Languages vary in their systems of kinship categories, but the scope of possible variation appears to be constrained. Previous accounts of kin classification have often emphasized constraints that are specific to the domain of kinship and are not derived from general principles. Here, we propose an account that is founded on two domain-general principles: Good systems are simple and they enable informative communication. We show computationally that these two competing principles. We also show that our account explains specific constraints on kin classification proposed previously. Because the principles of simplicity and informativeness are also relevant to other semantic domains, the trade-off between them may provide a domain-general foundation for variation in category systems across languages.

Concepts and categories vary across cultures but may nevertheless be shaped by universal constraints (1–4). Cross-cultural studies have proposed universal constraints that help to explain how colors (5), plants, animals (7, 8), and spatial relations (9, 10) are organized into categories. Kinship has traditionally been a prominent domain for studies of this kind, and researchers have described many constraints that help to predict which of the many logically possible kin classification systems are encountered in practice (11–15). Typically, these constraints are not derived from general principles, although it is often suggested that they are consistent with cognitive and functional considerations (2, 11–13, 15).

Here, we show that major aspects of kin classification follow directly from two general principles: Categories tend to be simple, which minimizes cognitive load, and to be informative, which maximizes communicative efficiency. Principles like these have been discussed in other contexts by previous researchers (16–19). For example, Zipp suggests that word-frequency distributions achieve a trade-off between simplicity and communicative precision (20, 21). Hawkins (22) has suggested that grammars are shaped by a trade-off between simplicity and communicative efficiency, and Rosch has suggested that category systems “provide maximum information with the least cognitive effort” (p. 190 of 23).

Figure 1A shows a simple communication game that helps to illustrate how kin classification systems are shaped by the principles of simplicity and informativeness. The speaker has a specific relative in mind and utters the category label for that relative. Upon hearing this category label, the hearer must guess which relative the speaker had in mind. However, the speaker cannot communicate through gestures or any other nonverbal means in this game, so the hearer must rely solely on the verbal means available to communicate.
Kinship terms are simple and informative
Kinship terms are simple and informative
Kinship terms are simple and informative
Learning and interaction pressures

Informative

Simple
Learning and interaction pressures
Learning and interaction pressures

Informative

Simple

learning
Learning and interaction pressures

Informative

Simple
Learning and interaction pressures

Informative

Simple

learning

interaction
Learning and interaction pressures
Learning and interaction pressures

- Informative
- Simple

- Learning
- Learning & interaction
- Interaction
Learning and interaction pressures

Kirby, Cornish, & Smith (2008)

Kirby, Tamariz, Cornish, & Smith (2015)
The problem of linkage

Communication
(in the broad sense)

Typological distribution of languages
The problem of linkage

Communication
(in the broad sense)

Black box mechanism

Typological distribution of languages
The problem of linkage

Communication
(in the broad sense)

Induction

Interaction

Simplicity

Informativeness

Typological distribution of languages
Induction

as the pressure for simplicity
The Minimum Description Length principle

$$\text{DL}(H|D) = \text{DL}(D|H) + \text{DL}(H)$$
The Minimum Description Length principle

\[
DL(H|D) = DL(D|H) + DL(H)
\]

\[
\text{posterior}(H|D) \propto \text{likelihood}(D|H) \times \text{prior}(H)
\]
The Minimum Description Length principle

\[
DL(H|D) = DL(D|H) + DL(H)
\]

\[
\text{posterior}(H|D) \propto \text{likelihood}(D|H) \times 2^{-DL(H)}
\]
The Minimum Description Length principle

\[ \text{DL}(H|D) = \text{DL}(D|H) + \text{DL}(H) \]

\[ \text{posterior}(H|D) \propto \text{likelihood}(D|H) \times 2^{-\text{DL}(H)} \]

Any regularities in data can be used to compress that data
The Minimum Description Length principle

\[
DL(H|D) = DL(D|H) + DL(H)
\]

\[
\text{posterior}(H|D) \propto \text{likelihood}(D|H) \times 2^{-DL(H)}
\]

Any regularities in data can be used to compress that data
The more regularities there are, the more the data can be compressed
The Minimum Description Length principle

\[ DL(H|D) = DL(D|H) + DL(H) \]

*For example…*

```
010010111100100011000100010110110001111010001
print('010010111100100011000100010110110001111010001')
```
```
010101010101010101010101010101010101010101010101
print('0101'*12)  or  print('01'*24)
```

Any regularities in data can be used to compress that data.
The more regularities there are, the more the data can be compressed.
The Minimum Description Length principle

\[
\text{DL}(H|D) = \text{DL}(D|H) + \text{DL}(H)
\]

\[
\text{posterior}(H|D) \propto \text{likelihood}(D|H) \times 2^{-\text{DL}(H)}
\]

Any regularities in data can be used to compress that data.

The more regularities there are, the more the data can be compressed.

We equate learning with compression: The more the data can be compressed, the more insight we gain from that data.
The Minimum Description Length principle

\[ DL(H|D) = DL(D|H) + DL(H) \]

\[ \text{posterior}(H|D) \propto \text{likelihood}(D|H) \times 2^{-DL(H)} \]

Any regularities in data can be used to compress that data. The more regularities there are, the more the data can be compressed. We equate learning with compression: The more the data can be compressed, the more insight we gain from that data. In other words, the more regularity we can identify, the more we can predict what the generating process will do next.
The Minimum Description Length principle

\[ \text{DL}(H|D) = \text{DL}(D|H) + \text{DL}(H) \]
The Minimum Description Length principle

\[ DL(H|D) = DL(D|H) + DL(H) \]
The Minimum Description Length principle

$$DL(H|D) = DL(D|H) + DL(H)$$
The Minimum Description Length principle

\[ DL(H|D) = DL(D|H) + DL(H) \]
The Minimum Description Length principle

\[ DL(H|D) = DL(D|H) + DL(H) \]
Bayesian interpretation: MDL is closely related to Bayesian inference

Occam’s razor: MDL trades-off goodness-of-fit with model complexity, embodying Occam’s razor

No overfitting: MDL automatically guards against overfitting noise in data

Predictive performance: Since data compression is formally equivalent to probabilistic prediction, MDL finds models offering good predictive performance on unseen data
Interaction

as the pressure for informativeness
Regier et al.'s informativeness model
Regier et al.’s informativeness model
Regier et al.’s informativeness model
Regier et al.’s informativeness model
Communicative cost

\[ C_j(i) \propto \sum_{c \in C_j} e^{-\gamma d(i,c)^2} \]

\[ K(L) := \sum_{i \in U} P(i) \cdot -\log C(i) \]
Communicative cost

\[ C_j(i) \propto \sum_{c \in C_j} e^{-\gamma d(i,c)^2} \]

\[ K(L) := \sum_{i \in U} P(i) \cdot -\log C(i) \]

**Expressivity** A system of many categories is more informative than a system of few categories
Communicative cost

\[ C_j(i) \propto \sum_{c \in C_j} e^{-\gamma d(i,c)^2} \]

\[ K(L) := \sum_{i \in U} P(i) \cdot -\log C(i) \]

**Expressivity** A system of many categories is more informative than a system of few categories

**Compactness** A system of compact categories is more informative than a system of noncompact categories
Could humans have a learning bias for informativeness?

Language evolution in the lab tends toward informative communication

Alexandra Carstensen, Jing Xu, Cameron T. Smith, Terry Regier

Department of Psychology, University of California, Berkeley, CA 94720 USA

Department of Neuroscience, Johns Hopkins University, Baltimore, MD 21218 USA

Abstract

Why do language-based human categories tend to be informative? The answer to this question is not yet known. In this study, we tested whether human categories tend to be more informative than random categories. We found that human categories are indeed more informative than random categories. This result suggests that human categories are not random, but rather are shaped by cultural evolution. We also found that human categories tend to be more informative than random categories when they are represented in the brain. This suggests that cultural evolution may have shaped the way that we think about categories. We hypothesize that cultural evolution may have shaped the way that we think about categories, and that this has led to the development of more informative categories. We also found that human categories tend to be more informative than random categories when they are represented in the brain. This suggests that cultural evolution may have shaped the way that we think about categories. We hypothesize that cultural evolution may have shaped the way that we think about categories, and that this has led to the development of more informative categories. We also found that human categories tend to be more informative than random categories when they are represented in the brain. This suggests that cultural evolution may have shaped the way that we think about categories. We hypothesize that cultural evolution may have shaped the way that we think about categories, and that this has led to the development of more informative categories.
Could humans have a learning bias for informativeness?
Could humans have a learning bias for informativeness?

Carstensen, Xu, Smith, Regier (2015)
Could humans have a learning bias for informativeness?

Carstensen, Xu, Smith, Regier (2015)
Bayesian model
Conceptual spaces and convexity
Conceptual spaces and convexity
Conceptual spaces and convexity
Conceptual spaces and convexity
Bayesian inference

\[ \mathcal{L} = \{ \ldots \} \]
Bayesian inference

\[ L = \{ \langle m_1, s_1 \rangle, \langle m_2, s_2 \rangle, \langle m_3, s_3 \rangle, \ldots, \langle m_n, s_n \rangle \} \]

\[ D = [\langle m_1, s_1 \rangle, \langle m_2, s_2 \rangle, \langle m_3, s_3 \rangle, \ldots, \langle m_n, s_n \rangle] \]
Bayesian inference

\[ \mathcal{L} = \{ \cdot \cdot \cdot \} \]

\[ D = [\langle m_1, s_1 \rangle, \langle m_2, s_2 \rangle, \langle m_3, s_3 \rangle, \ldots, \langle m_n, s_n \rangle] \]

\[
\text{likelihood}(D|L) \propto \prod_{\langle m, s \rangle \in D} \frac{1}{|M|} P(s|L, m) = \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]

\[
= \]
Bayesian inference

\[ \mathcal{L} = \{ \ldots \} \]

\[ D = [\langle m_1, s_1 \rangle, \langle m_2, s_2 \rangle, \langle m_3, s_3 \rangle, \ldots, \langle m_n, s_n \rangle] \]

\[
\text{likelihood}(D|L) \propto \prod_{\langle m, s \rangle \in D} \frac{1}{|M|} P(s|L, m)
\]

\[ \text{prior}(L) \propto 2^{-DL(L)} \]
Bayesian inference

\[ \mathcal{L} = \{ \ldots \} \]

\[ D = [\langle m_1, s_1 \rangle, \langle m_2, s_2 \rangle, \langle m_3, s_3 \rangle, \ldots, \langle m_n, s_n \rangle] \]

\[
\text{likelihood}(D|L) \propto \prod_{\langle m, s \rangle \in D} \frac{1}{|M|} P(s|L,m) \\
\text{prior}(L) \propto 2^{-DL(L)} \\
\text{posterior}(L|D) \propto \text{likelihood}(D|L) \times \text{prior}(L)
\]
Computing DL(L): The rectangle code

Categorization Under Complexity: A Unified MDL Account of Human Learning of Regular and Irregular Categories

David Fass
Department of Psychology
Center for Cognitive Science
Rutgers University
Piscataway, NJ 08854
dfass@ruccs.rutgers.edu

Jacob Feldman
Department of Psychology
Center for Cognitive Science
Rutgers University
Piscataway, NJ 08854
jacob@ruccs.rutgers.edu

Abstract

We present an account of human, concept learning—that is, learning of categories from examples—based on the principle of minimum description length (MDL). In support of this theory, we tested a wide range of two-dimensional concept types, including both regular (simple) and highly irregular (complex) matrices, and found the MDL theory to give a good account of subjects’ performance. This suggests that the intrinsic complexity of a concept (that is, its description length) systematically influences its learnability.

1 The Structure of Categories

A number of different principles have been advanced to explain the manner in which humans learn to categorize objects. It has been variously suggested that the underlying principles might be the similarity structure of objects [1], the manipulability of decision boundaries [2], and the global structure of experimental conditions, among others [3][4]. While many of these theories are mathematically appealing, it is not clear how or why any of them predicts the comparable success of different categories in the present experiments.

By contrast, the MDL principle is a natural and simple explanation of why certain categories are more learnable than others. This principle states that the information that one obtains from a set of examples can be divided into two parts: the description of the learning system required to generate and test the examples, and the description of the data that the learning system has generated. The MDL principle suggests that the description length of the learning program, not the complexity of the data, is the source of any differences in learnability.
### Computing DL($L$): The rectangle code

<table>
<thead>
<tr>
<th>Class</th>
<th>Positions</th>
<th>Probability</th>
<th>Code(\text{length}) (bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1×1</td>
<td>16</td>
<td>(1/16 \times 1/16)</td>
<td>(-\log 1/256) 8.0</td>
</tr>
<tr>
<td>1×2</td>
<td>12</td>
<td>(1/16 \times 1/12)</td>
<td>(-\log 1/192) 7.58</td>
</tr>
<tr>
<td>1×3</td>
<td>8</td>
<td>(1/16 \times 1/8)</td>
<td>(-\log 1/128) 7.0</td>
</tr>
<tr>
<td>1×4</td>
<td>4</td>
<td>(1/16 \times 1/4)</td>
<td>(-\log 1/64) 6.0</td>
</tr>
<tr>
<td>2×1</td>
<td>12</td>
<td>(1/16 \times 1/12)</td>
<td>(-\log 1/192) 7.58</td>
</tr>
<tr>
<td>2×2</td>
<td>9</td>
<td>(1/16 \times 1/9)</td>
<td>(-\log 1/144) 7.17</td>
</tr>
<tr>
<td>2×3</td>
<td>6</td>
<td>(1/16 \times 1/6)</td>
<td>(-\log 1/96) 6.58</td>
</tr>
<tr>
<td>2×4</td>
<td>3</td>
<td>(1/16 \times 1/3)</td>
<td>(-\log 1/48) 5.58</td>
</tr>
<tr>
<td>3×1</td>
<td>8</td>
<td>(1/16 \times 1/8)</td>
<td>(-\log 1/128) 7.0</td>
</tr>
<tr>
<td>3×2</td>
<td>6</td>
<td>(1/16 \times 1/6)</td>
<td>(-\log 1/96) 6.58</td>
</tr>
<tr>
<td>3×3</td>
<td>4</td>
<td>(1/16 \times 1/4)</td>
<td>(-\log 1/64) 6.0</td>
</tr>
<tr>
<td>3×4</td>
<td>2</td>
<td>(1/16 \times 1/2)</td>
<td>(-\log 1/32) 5.0</td>
</tr>
<tr>
<td>4×1</td>
<td>4</td>
<td>(1/16 \times 1/4)</td>
<td>(-\log 1/64) 6.0</td>
</tr>
<tr>
<td>4×2</td>
<td>3</td>
<td>(1/16 \times 1/3)</td>
<td>(-\log 1/48) 5.58</td>
</tr>
<tr>
<td>4×3</td>
<td>2</td>
<td>(1/16 \times 1/2)</td>
<td>(-\log 1/32) 5.0</td>
</tr>
<tr>
<td>4×4</td>
<td>1</td>
<td>(1/16 \times 1/1)</td>
<td>(-\log 1/16) 4.0</td>
</tr>
</tbody>
</table>

#### Uniformly sample a class

#### Uniformly sample a position
76.58 bits
Bayesian iterated learning under a simplicity prior
Bayesian iterated learning under a simplicity prior
Bayesian iterated learning under a simplicity prior
Bayesian iterated learning under a simplicity prior
Bayesian iterated learning under a simplicity prior
Bayesian iterated learning under a simplicity prior
Simplicity prior

Informativeness prior
Informativeness prior
Model results

- **Expressivity**
- **Transmission error**
- **Complexity**
- **Communicative cost**

- Simplicity prior
- Informativeness prior
Experiment
Stage 1: Training

15 minutes

You are going to learn a simple language. We will train you on 4 words in the language and we will test how well you are learning the words. Try to learn the language as well as you can and aim to be accurate in your answers. You will receive a 2¢ bonus payment for every correct test answer. If you decide to stop the task, please click the EXIT button so that someone else can take part.

1. Look at the picture
   - This is a triangle

2. Learn the word
   - Look at the picture that you saw before

3. Click on the word to confirm you learned it
   - What is it called?
   - to, tuk, gax, toa

4. Try to recall the correct word
   - What is this called?
   - to, tuk, gax, toa

5. If correct, you get a 2¢ bonus

START
Stage 2: Test

5 minutes

You have now completed the training stage! Next we will test you on the language that you just learned. For each picture, try to click on the correct word. You will get a 2¢ bonus payment for every correct answer. It is therefore possible to earn up to $1.28 in this stage of the task. However, this time we will not tell you if you are correct or incorrect. You will find out at the end how many you got correct.

START
### Experimental stimuli

<table>
<thead>
<tr>
<th>Angle</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>pov</td>
</tr>
<tr>
<td>2</td>
<td>gex</td>
</tr>
<tr>
<td>3</td>
<td>buv</td>
</tr>
<tr>
<td>4</td>
<td>fod</td>
</tr>
</tbody>
</table>
Iterated learning with humans
Iterated learning with humans
Iterated learning with humans
Systems converged on
Experimental results

- Expressivity
- Transmission error
- Complexity
- Communicative cost

Graphs showing trends over generations.

Left: Heatmaps for different categories A to L.

Right: Graphs illustrating expressivity, transmission error, complexity, and communicative cost over generations.
Experimental results

- Expressivity
- Transmission error
- Complexity
- Communicative cost

Graphs showing changes in communicative cost, generation, and complexity.
Model fit
Estimating unknown parameters of the model

Simplicity prior

Informativeness prior
Rerun the model with parameters estimated from the experiment
Two ways of achieving simplicity
Two ways of achieving simplicity

Increase in compactness
Two ways of achieving simplicity

Increase in compactness

Decrease in expressivity
Two ways of achieving simplicity

Decrease in expressivity

Increase in compactness

increases informativeness

Decrease in expressivity
Two ways of achieving simplicity

Increase in compactness

* increases informativeness

Decrease in expressivity

* decreases informativeness
Two ways of achieving simplicity

- Increase in compactness
  - *increases informativeness*

- Decrease in expressivity
  - *decreases informativeness*
Conclusions

Languages are shaped in the simplicity–informativeness tradeoff by pressures from induction and interaction.

For a rational learner, induction contains a simplicity bias to prevent overfitting noise, and to aid reasoning about unseen meanings.

Iterated learning (repeated induction) converges to the prior bias, favouring languages that are as simple as possible:

- **Loss of expressivity**: Loss of words/concepts to aid learning.
- **Compact categories**: Reorganization of the space to aid learning.

In the process, some informativeness may come along for the ride, potentially obscuring the causal mechanism.

Nevertheless, some kind of interactional dynamics (e.g. learning based on communicative success) must restrain languages from total degeneration.
Thanks!