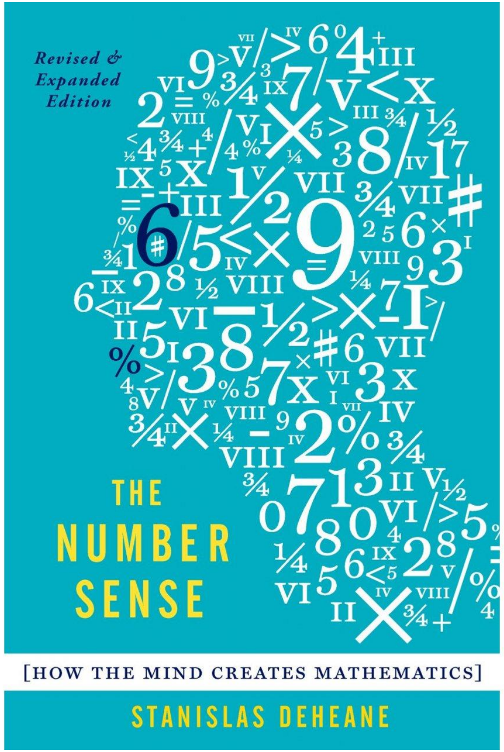
CHALMERS
UNIVERSITY OF TECHNOLOGY

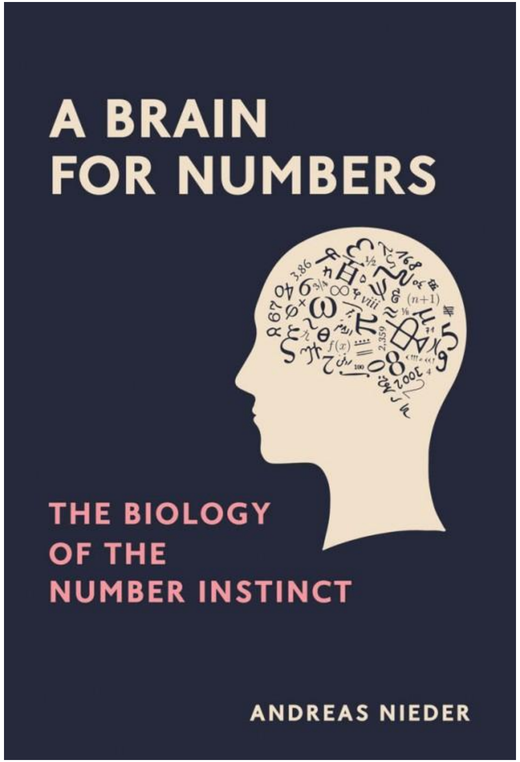
Learning Numeral Systems by Interaction

Devdatt Dubhashi and Emil Carlsson

Division of Data Science and AI

Numbers and the Brain

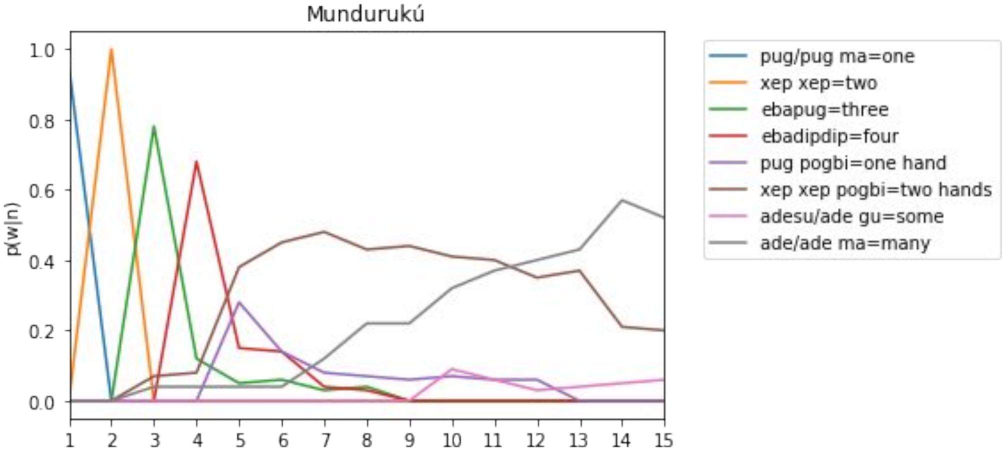


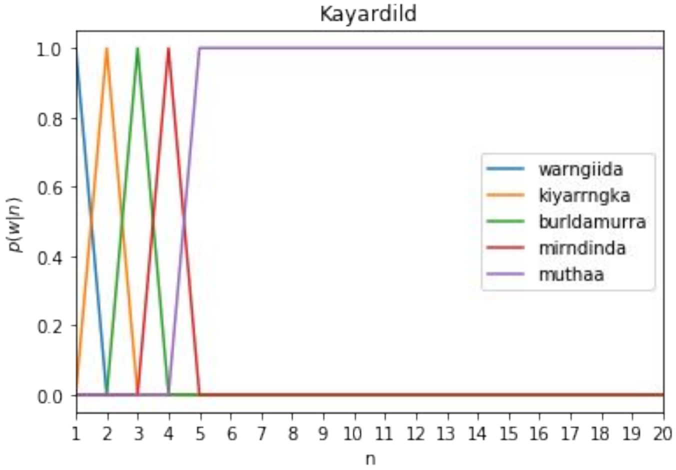


Variation in numeral systems

- Some languages have numeral systems that express only approximate or inexact numerosity
- Other languages have systems that express exact numerosity
- Some only over a restricted range of relatively small numbers
- Other languages have fully recursive counting systems that express exact numerosity over a very large range.

Two Numeral Systems





Munduruku:
Approximate

Kayardild, Exact

Numeral Systems: Universal Principles?

- Are there any universal principles common to all numeral systems?

CLASP Guests

Ted Gibson, MIT (2016)

Terry Regier, UC. Berkeley (2019)

Language and Efficient Communication

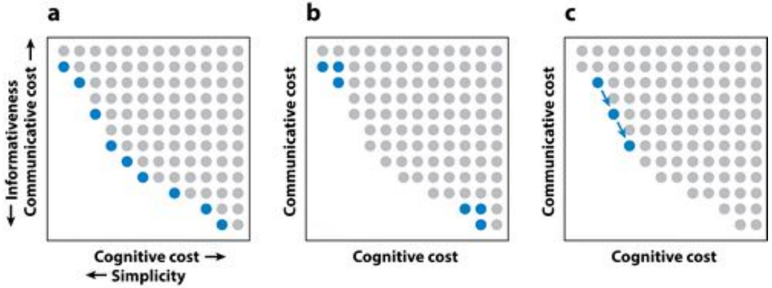
"Languages are under pressure to be simultaneously

- informative (so as to support effective communication) and
- simple (so as to minimize cognitive load)."

Semantic Typology and Efficient Communication

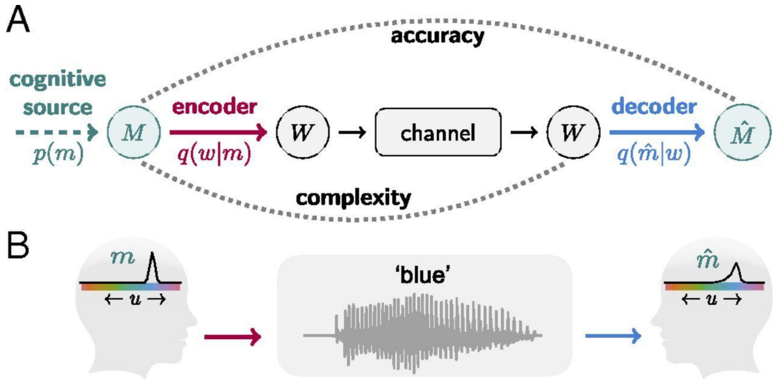
Annual Review of Linguistics
Vol. 4:109-128 (Volume publication date January 2018)
https://doi.org/10.1146/annurev-linguistics-011817-045406

Charles Kemp,[1,*] Yang Xu,[2] and Terry Regier[3,*]



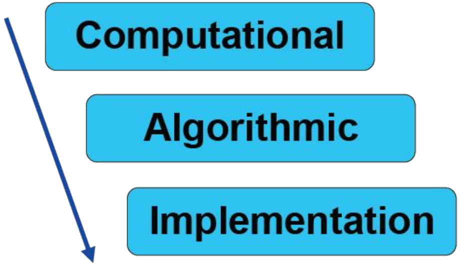
Kemp C, et al. 2018.
Annu. Rev. Linguist. 4:109–28

Learning by Interaction: Information Theoretic Framework

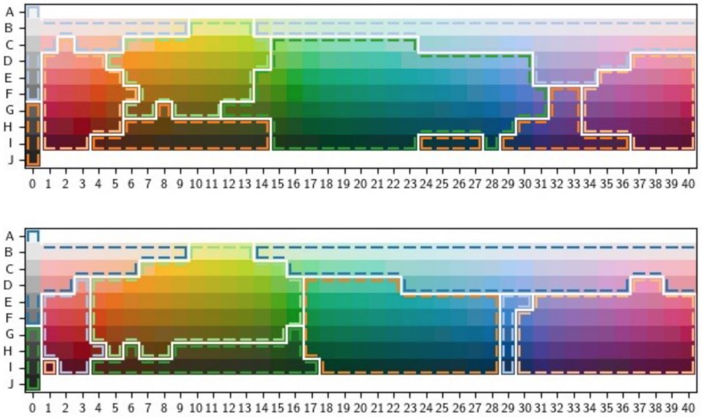


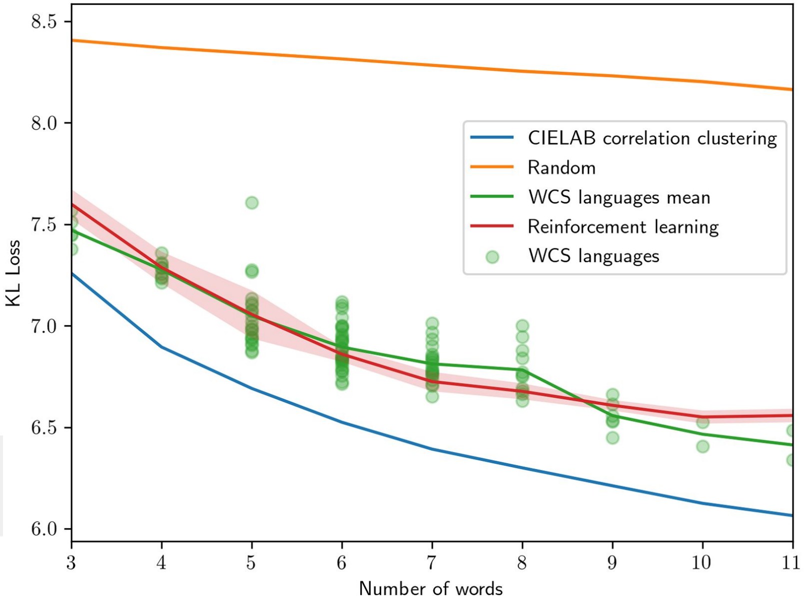
A Learning Perspective

- Human languages are observed to optimize communication efficiency in information-theoretic sense,
- But is there a computational mechanism to explain how?
- Can agents learn an efficient communication scheme from scratch by interacting to solve a shared task?
- Marr's three levels of analysis
- Poggio (afterword to re-release of Marr(1982)):
  "Add learning at the very top level of understanding, above the computational level."



Reinforcement Learning & Efficient Communication





PLOS ONE

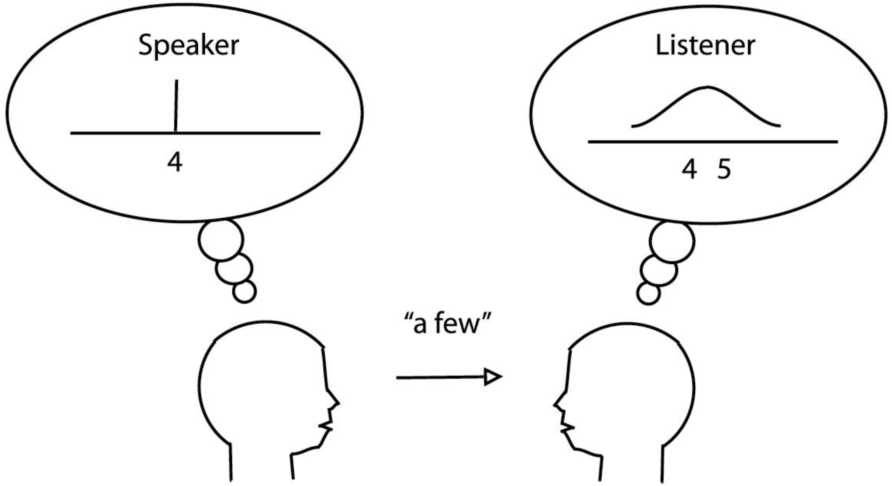
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

A reinforcement-learning approach to efficient communication

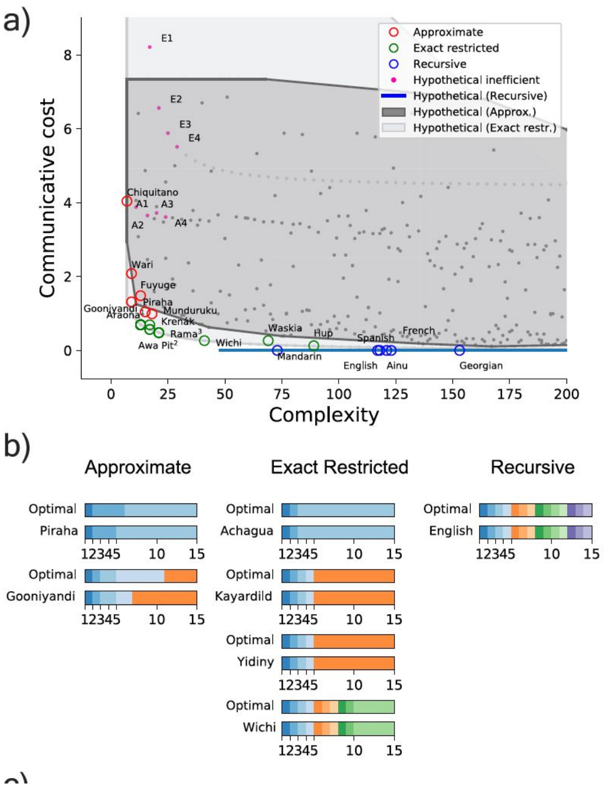
Mikael Kågebäck, Emil Carlsson, Devdatt Dubhashi, Asad Sayeed

Communicating a number

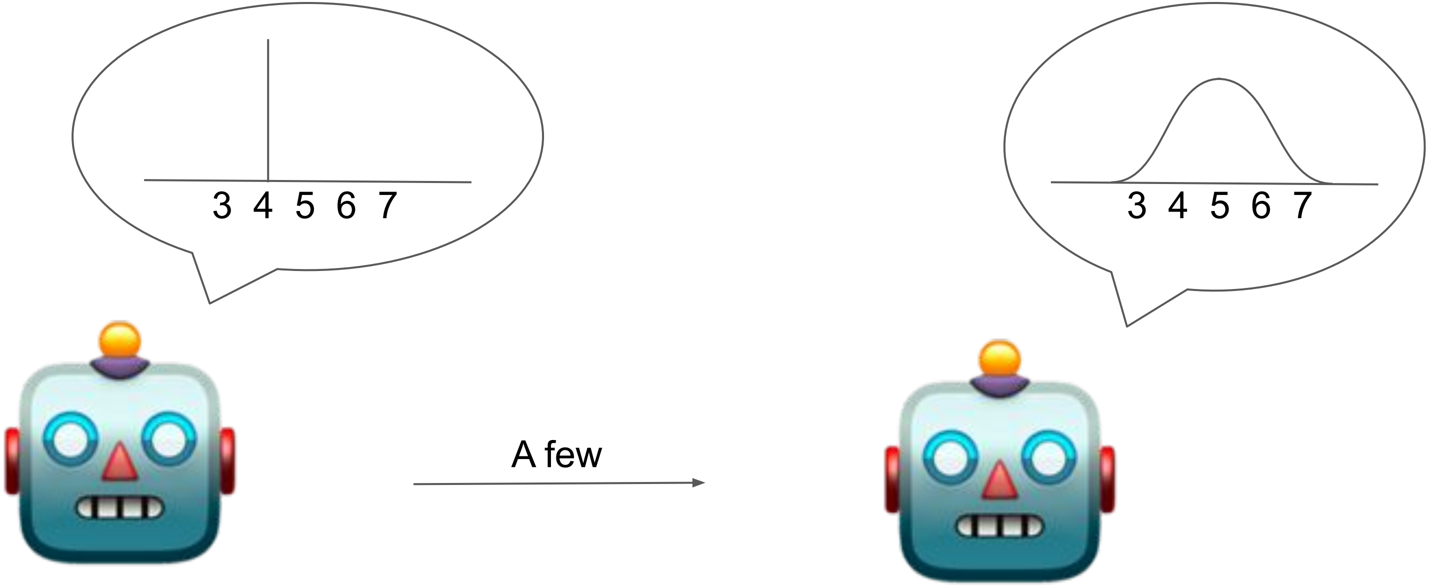


Yang Xu, Emmy Liu, and Terry Regier (2020). Numeral
Open Mind, 4, 57-70.

Efficiency of Numeral Systems



Yang Xu, Emmy Liu, and Terry Regier (2020). Numeral *Open Mind, 4,* 57-70.


3 4 5 6 7
A few
3 4 5 6 7

$$n \sim p(n)$$

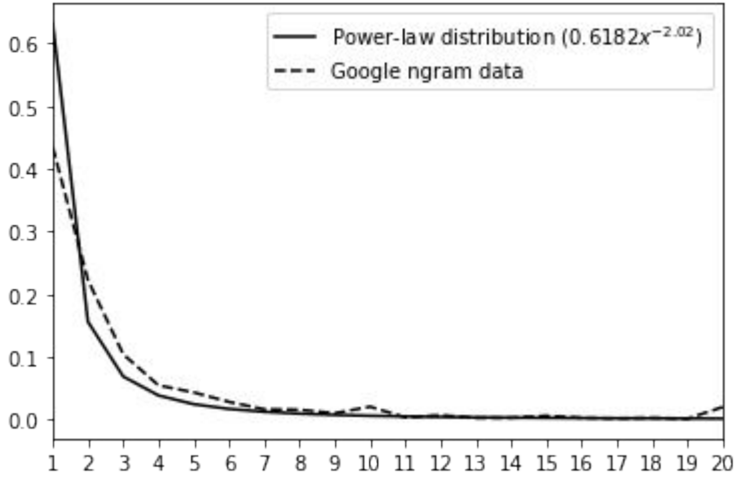
$$w \sim p(w|n)$$

$$w$$

$$\hat{n} \sim p(\hat{n}|w)$$

$$r = 1 - \frac{|n - \hat{n}|}{N}$$

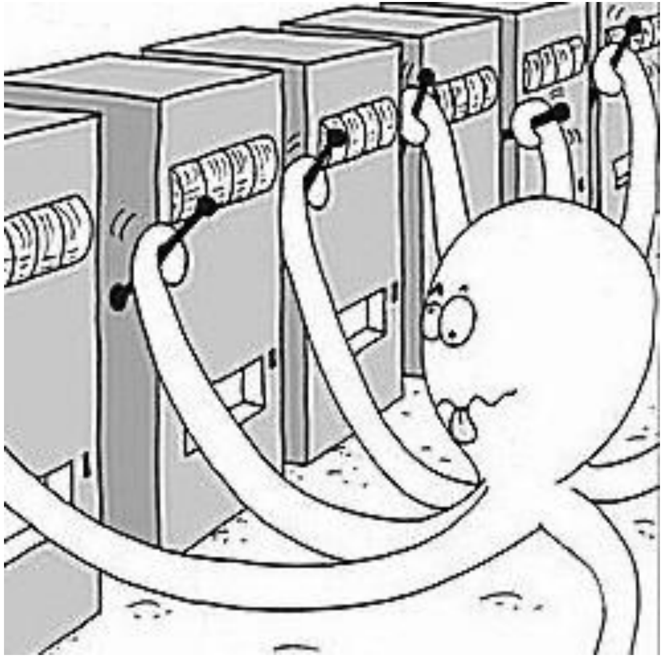
- We restrict our game to the numerals 1 to 20, i.e N=20.
- Agents have access to a small set of tokens.
- *Tabula rasa* agents.
- The meaning of the tokens are created by the agents while playing the game.



Need probability estimated from frequencies of English numerals in Google ngram Corpus (Michel et al. 2011)

Contextual Bandit

- Each agent can be modelled as a Contextual Bandit.
- The agent sees a context and has to pick an action from a set of actions.
- In our case, the context is a numeral and the actions are different tokens/words.



Thompson Sampling

- A common approach to bandit problems.
- The learner has a prior belief over a set of possible environments.

$$p(f)$$

- In each round the learner samples a possible environment from the posterior and acts greedy according to it.

$$f \sim p(f|H)$$

- Given an observation (action and reward) we update the posterior distribution.

Contextual Bandits and Thompson Sampling

- Each agent has a neural network

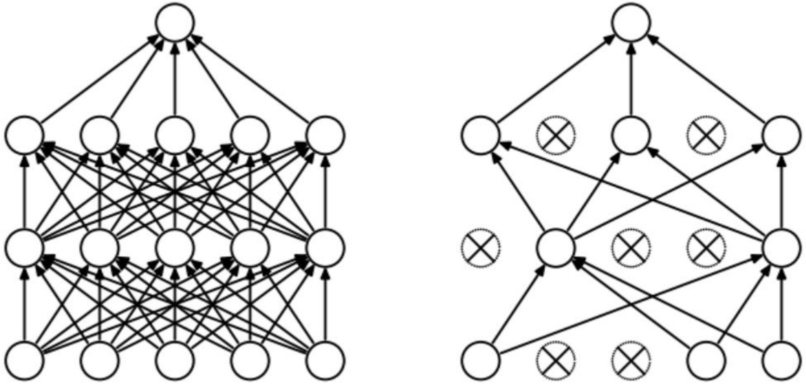
$$f_S(n, w)$$
$$f_L(w, \hat{n})$$

- At each round a smaller network is sampled usi
dropout (Nitish Srivastava et al., 2014).

$$\hat{f}_S(n, w) \sim f_S(n, w)$$
$$\hat{f}_L(w, \hat{n}) \sim f_L(w, \hat{n})$$

- Each agent acts greedy w.r.t smaller network

$$w = \text{argmax}_w \hat{f}_S(n, w)$$
$$\hat{n} = \text{argmax}_{\hat{n}} \hat{f}_L(w, \hat{n})$$



(a) Standard Neural Net (b) After applying dropout.

Contextual Bandits and Thompson Sampling

- The neural networks maps context and action to expected reward.
- Update them using the mean-squared error (MSE):

$$\mathrm{MSE}_S = \frac{1}{M}\sum_i^M (\hat{f}_S(n_i, w_i) - r_i)^2,$$

$$\mathrm{MSE}_L = \frac{1}{M}\sum_i^M (\hat{f}_L(w_i, \hat{n}) - r_i)^2$$

How do we define communication cost and complexity of a numeral system?

Communication cost

- Communication cost or expected surprisal

$$-\sum_n \sum_w p(n)p(w|n) \log L_w(n)$$

- How surprised the listener is by the fact that the sender used word w for numeral n on average.

- Listener computed using Bayes formula

$$L_w(n) \propto p(w|n)p(n)$$

Complexity

- Naive approach: Number of words
- There are relationships between numeral words. We could measure complexity as the number of rules needed to define the numeral system.

Table 2. Grammatical components for representing numeral systems.

Component	Description
$\mathtt{c}$	Primitive concept $\mathtt{c} = 1, 2$ or 3
$\tilde{x}$	Gaussian with approximate mean $\tilde{x}$
$\mathtt{m}(w)$	Meaning of form w
$\mathtt{s}(w,v)$	Successor of w with interval v; $\mathtt{s}(w)=\mathtt{s}(w,1)$
$\mathtt{h}(w)$	Higher than w
$+$	Addition
$-$	Subtraction
$\times$	Multiplication
$\div$	Division
$\mathtt{p}(x,n)$	x to the nth power
$\stackrel{d}{=}$	Form definition
$\in$	Set definition
$\equiv$	Equivalence

Table 4. Grammar for Kayardild (exact restricted) numeral system for the range 1–100.

Number	Rule	Complexity
1	‘warngiida’ $\stackrel{d}{=}$ $\mathtt{1}$	3
2	‘kiyarrngka’ $\stackrel{d}{=}$ $\mathtt{2}$	3
3	‘burldamurra’ $\stackrel{d}{=}$ $\mathtt{3}$	3
4	‘mirndinda’ $\stackrel{d}{=}$ $\mathtt{s}$(‘burldamurra’)	4
5–100	‘muthaa’ $\stackrel{d}{=}$ $\mathtt{h}$(‘mirndinda’)	4
		$\Sigma = 17$

Note. Each rule is composed of symbols, and each symbol adds a unit complexity of 1.

Yang Xu, Emmy Liu, and Terry Regier (2020). Numeral
Open Mind, 4, 57-70.

Table 2. Grammatical components for representing numeral systems.

Component	Description
c	Primitive concept c $= 1, 2$ or 3
$\tilde{x}$	Gaussian with approximate mean $\tilde{x}$
m(w)	Meaning of form w
s(w,v)	Successor of w with interval v; s(w)=s(w,1)
h(w)	Higher than w
$+$	Addition
$-$	Subtraction
$\times$	Multiplication
$\div$	Division
p(x,n)	x to the nth power
$\overset{d}{=}$	Form definition
$\in$	Set definition
$\equiv$	Equivalence

Table 4. Grammar for Kayardild (exact restricted) numeral system for the range 1–100.

Number	Rule	Complexity
1	'warngiida' $\overset{d}{=}$ 1	3
2	'kiyarrngka' $\overset{d}{=}$ 2	3
3	'burldamurra' $\overset{d}{=}$ 3	3
4	'mirndinda' $\overset{d}{=}$ s('burldamurra')	4
5–100	'muthaa' $\overset{d}{=}$ h('mirndinda')	4
		$\Sigma = 17$

Note. Each rule is composed of symbols, and each symbol adds a unit complexity of 1.

Yang Xu, Emmy Liu, and Terry Regier (2020). Numeral
Open Mind, 4, 57-70.

Table 2. Grammatical components for representing numeral systems.

Component	Description
c	Primitive concept c = 1, 2 or 3
$\tilde{x}$	Gaussian with approximate mean $\tilde{x}$
m(w)	Meaning of form w
s(w,v)	Successor of w with interval v; s(w)=s(w,1)
h(w)	Higher than w
$+$	Addition
$-$	Subtraction
$\times$	Multiplication
$\div$	Division
p(x,n)	x to the nth power
$\stackrel{d}{=}$	Form definition
$\in$	Set definition
$\equiv$	Equivalence

Table 4. Grammar for Kayardild (exact restricted) numeral system for the range 1–100.

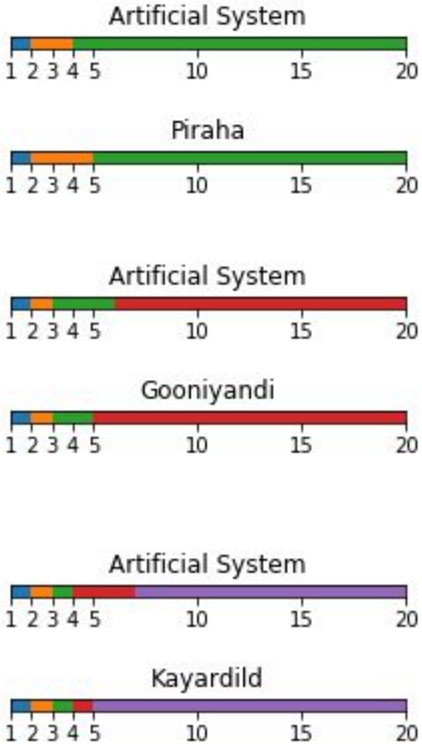
Number	Rule	Complexity
1	'warngiida' $\stackrel{d}{=}$ 1	3
2	'kiyarrngka' $\stackrel{d}{=}$ 2	3
3	'burldamurra' $\stackrel{d}{=}$ 3	3
4	'mirndinda' $\stackrel{d}{=}$ s('burldamurra')	4
5–100	'muthaa' $\stackrel{d}{=}$ h('mirndinda')	4
		$\Sigma = 17$

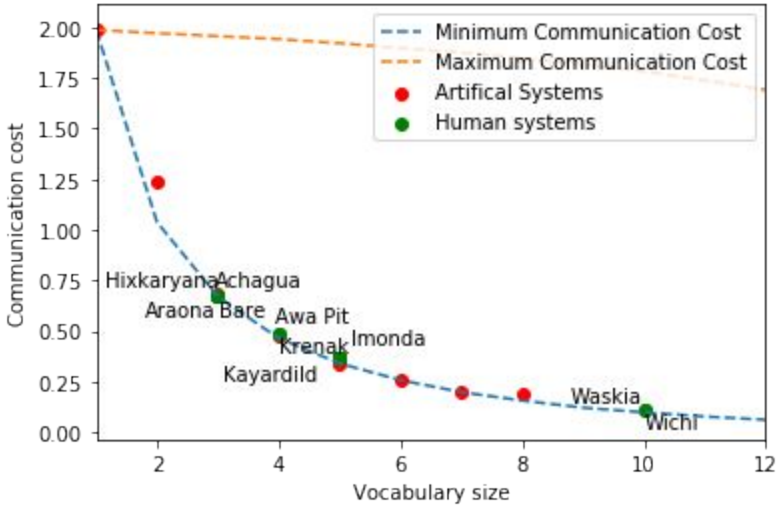
Note. Each rule is composed of symbols, and each symbol adds a unit complexity of 1.

Yang Xu, Emmy Liu, and Terry Regier (2020). Numeral
Open Mind, 4, 57-70.

Results

- We perform 3000 experiments with a vocabulary size of 10.
- It is up to the agents to decide how many of the 10 tokens that will be used.
- Always results in an exact numeral system.
- We grouped the experiments together based complexity and computed the consensus numeral system.
- Compared to 24 languages from non-industrial societies (Xu et al. 2020)

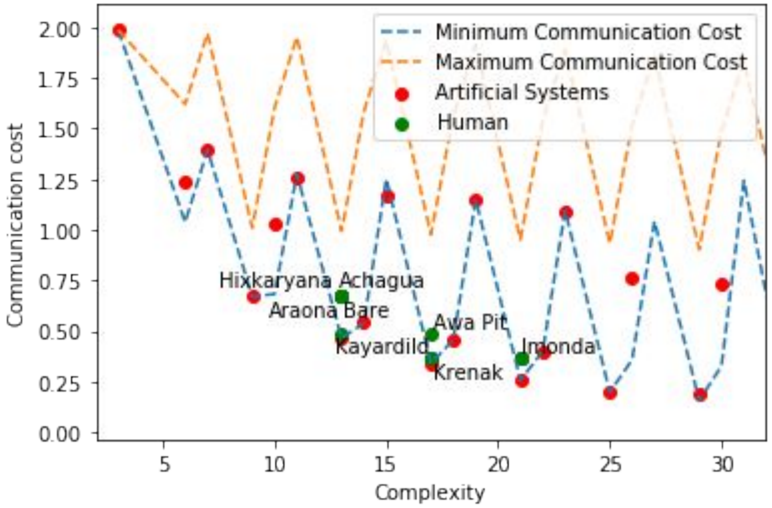

Artificial System
1 2 3 4 5 10 15 20
Piraha
1 2 3 4 5 10 15 20
Artificial System
1 2 3 4 5 10 15 20
Gooniyandi
1 2 3 4 5 10 15 20
Artificial System
1 2 3 4 5 10 15 20
Kayardild
1 2 3 4 5 10 15 20


Minimum Communication Cost
Maximum Communication Cost
Artifical Systems
Human systems
Communication cost
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Hixkaryana
Achagua
Araona
Bare
Awa Pit
Krenak
Imonda
Kayardild
Waskia
Wichi
2
4
6
8
10
12
Vocabulary size

- Both human systems and artificial systems are near-optimal.

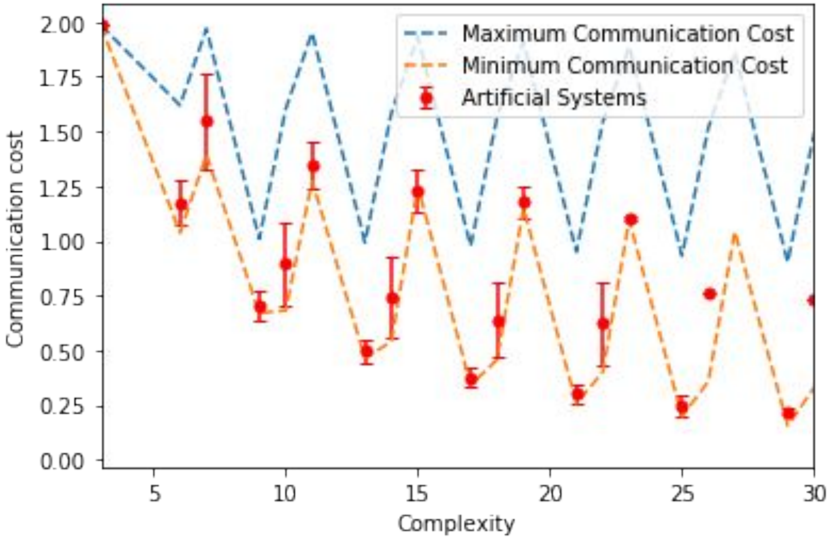
- Jacked shape comes from

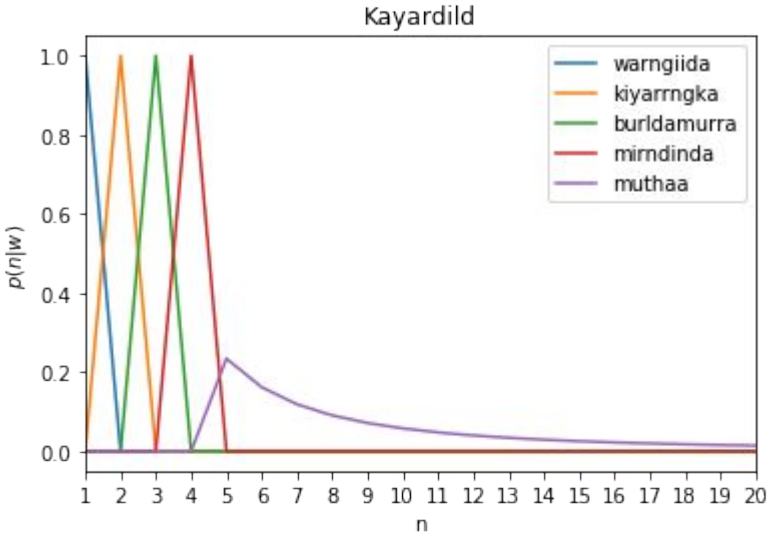
$$3c + 4s$$

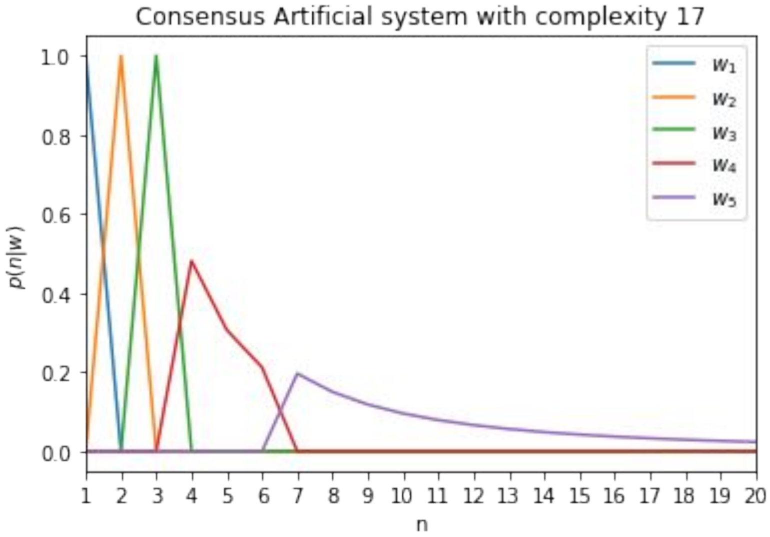


- mean ±1 standard deviation.

- Process is stable. Almost all 3000 experiments gives a numeral system which is near-optimal.




Kayardild
warngiida
kiyarrngka
burldamurra
mirndinda
muthaa
p(n|w)
n


Consensus Artificial system with complexity 17
w_1
w_2
w_3
w_4
w_5
p(n|w)
n

Conclusions

- Our computational mechanism leads to near-optimal exact numeral systems.
- Similar to human systems with same complexity.

Possible extensions

- Many numeral systems are recurrent and we can express any number. Can such systems be learned?

- Exact and approximate arithmetic (Pica et al. 2004).

References

Numeral Systems Across Languages Support Efficient Communication: From Approximate Numerosity to Recursion. Yang Xu, Emmy Liu and Terry Regier.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.

Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov; 15(56):1929−1958, 2014

Supplementary slides

Training details

- 10 000 epochs
- 100 batch size
- Optimizer Adam with learning rate 0.001
- Dropout 0.3
- Hidden neurons 50.

Communication cost

- We define the communication cost as the Kullback-Leibler divergence between the sender (S) and listener distribution (L):

$$C_w(n) = \mathrm{KL}(S||L) = \sum_{i=1}^{20} S(i) \log \frac{1}{L_w(i)}$$

- In the case of speaker certainty this reduces to the surprisal

$$C_w(n) = -\log L_w(n)$$

- Listener computed using Bayes formula

$$L_w(n) \propto p(w|n)p(n)$$