Computational approaches to the explanation of universal properties of meaning

Fausto Carcassi & Jakub Szymanik

https://thelogicalgrammar.github.io/ESSLLI22_langevo

Tradeoff between pressures

Finding a compromise

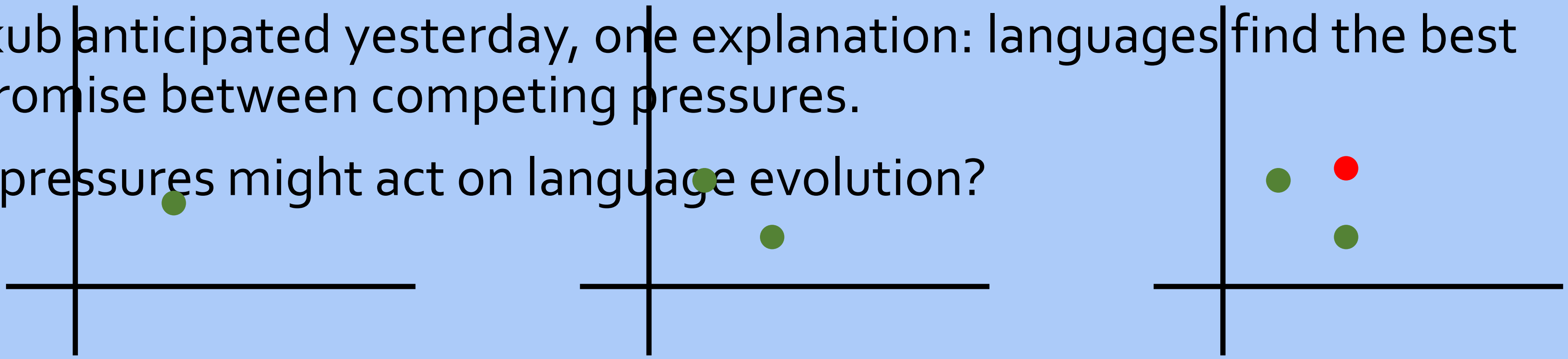
Different pressures act together

- In the previous lectures, we have considered the effect of individual pressures
    - Learnability
    - Complexity
    - Cultural evolution
- However,
    - Learnability and complexity alone predict degenerate languages
    - Cultural evolution alone (w/ assumptions) predicts prior distribution
- More likely: language is the result of a compromise btw competing pressures
    - Some pressures make language simpler / easier to learn
    - Some pressures make language well-adapted for use, e.g., communication
- Two general strategies to study this

Approach I: Finding the Pareto frontier

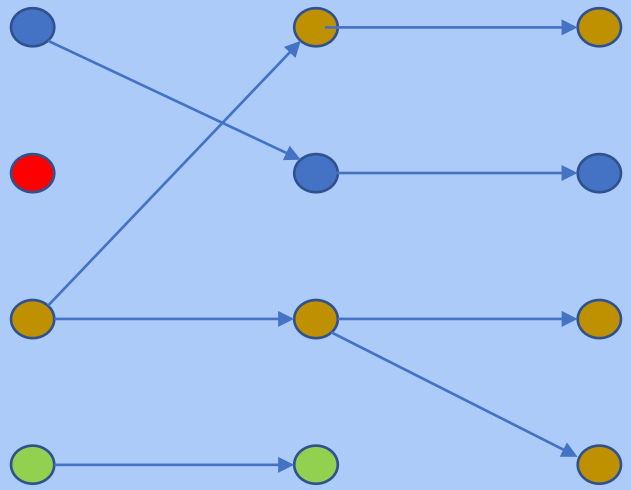
- The *Pareto front* is the set of all *Pareto efficient* solutions to a problem with multiple dimensions to optimize.
- A Pareto efficient solution is one such that there are no other solutions that are better in some dimensions and at least as good in all others.
- In other words, to improve in one dimensions, you have to give up something!

- As Jakub anticipated yesterday, one explanation: languages find the best compromise between competing pressures.
- What pressures might act on language evolution?



Approach II: IL + communication

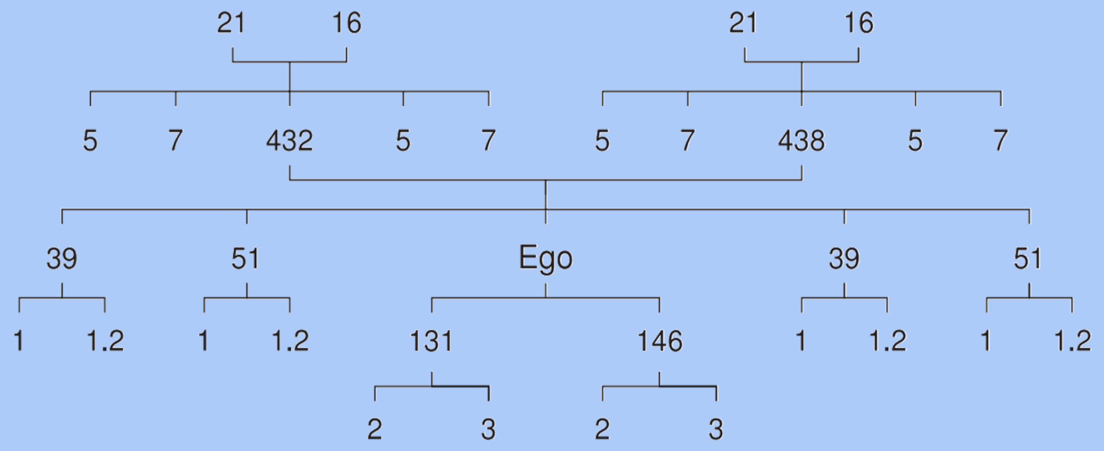
- Second strategy: Use IL in a population of multiple agents, but make it more likely for languages with higher communicative success to become teachers.
- Many modelling details to iron out!



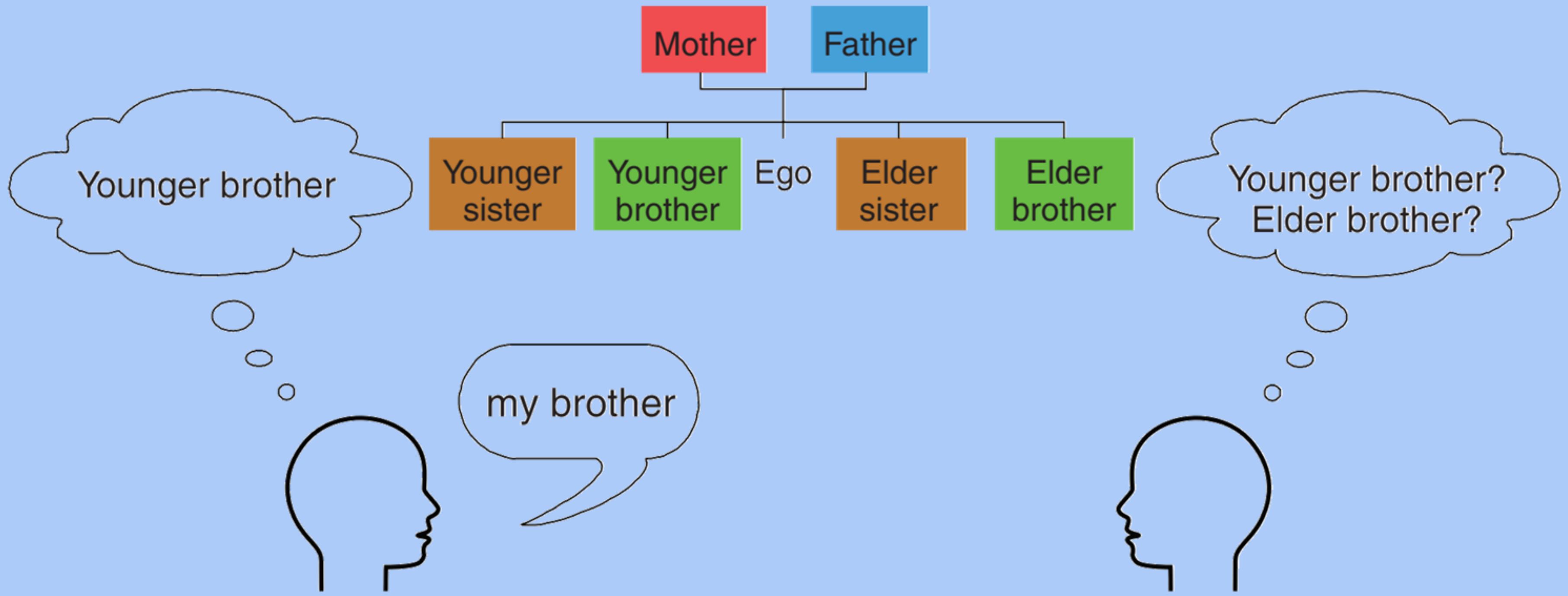
Case Study I: Kinship

General problem

- Kinship systems (Kemp and Regier 2012)
- Murdoch dataset includes kin classification systems for 566 languages.
- Complete systems for 487 languages
- 410 distinct categories (out of 10^55)
- Relative frequencies of kin expressions of the form “my grandmother,” “my mother,” “my daughter,” “my granddaughter,” and the like across corpora for two languages



Modelling communication



- Communicative cost of a system is the expected information-theoretic cost

Modelling simplicity

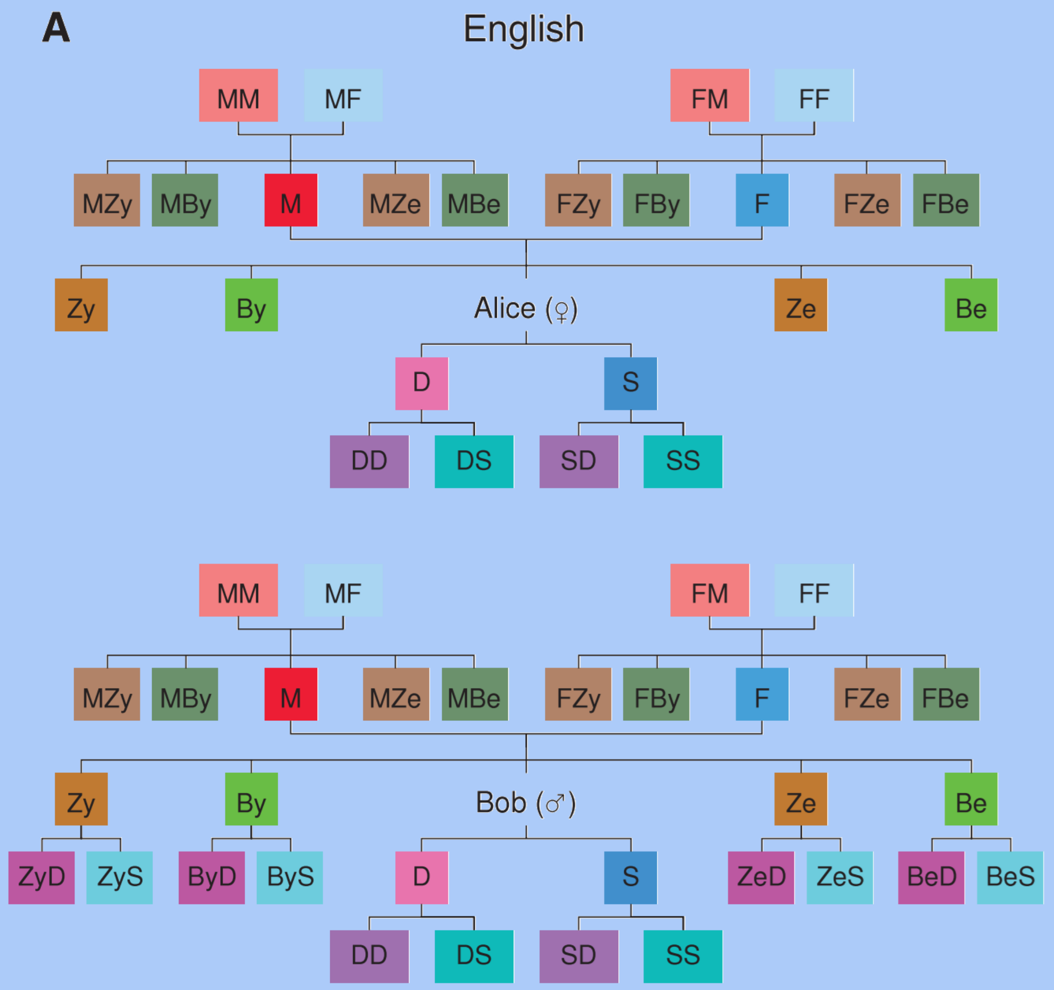
- We need:
  - A semantic space
  - A language to describe it
- Find the MDL in the language of each category
- The complexity of a system is the smallest number of rules needed to define all terms in the system.

$$\begin{aligned}
&\mathtt{FEMALE}(\cdot)\\
&\mathtt{MALE}(\cdot)\\
&\mathtt{PARENT}(\cdot,\cdot)\\
&\mathtt{CHILD}(\cdot,\cdot)\\
&\mathtt{OLDER}(\cdot,\cdot)\\
&\mathtt{YOUNGER}(\cdot,\cdot)\\
&\mathtt{SAMESEX}(\cdot,\cdot)\\
&\mathtt{DIFFSEX}(\cdot,\cdot)
\end{aligned}$$

$$\begin{aligned}
&\mathtt{C}(x,y)\leftrightarrow \mathtt{A}(x,y)\wedge \mathtt{B}(x)\\
&\mathtt{C}(x,y)\leftrightarrow \mathtt{A}(x,y)\wedge \mathtt{B}(y)\\
&\mathtt{C}(x,y)\leftrightarrow \mathtt{A}(x,y)\wedge \mathtt{B}(x,y)\\
&\mathtt{C}(x,y)\leftrightarrow \mathtt{A}(x,y)\vee \mathtt{B}(x)\\
&\mathtt{C}(x,y)\leftrightarrow \mathtt{A}(x,y)\vee \mathtt{B}(y)\\
&\mathtt{C}(x,y)\leftrightarrow \mathtt{A}(x,y)\vee \mathtt{B}(x,y)\\
&\mathtt{C}(x,y)\leftrightarrow \exists z\ \mathtt{A}(x,z)\wedge \mathtt{B}(z,y)\\
&\mathtt{C}(x,y)\leftrightarrow \mathtt{A}(y,x)\\
&\mathtt{C}(x,y)\leftrightarrow \mathtt{A}^{\leftrightarrow}(x,y)\\
&\mathtt{C}(x,y)\leftrightarrow \mathtt{A}^{+}(x,y)
\end{aligned}$$

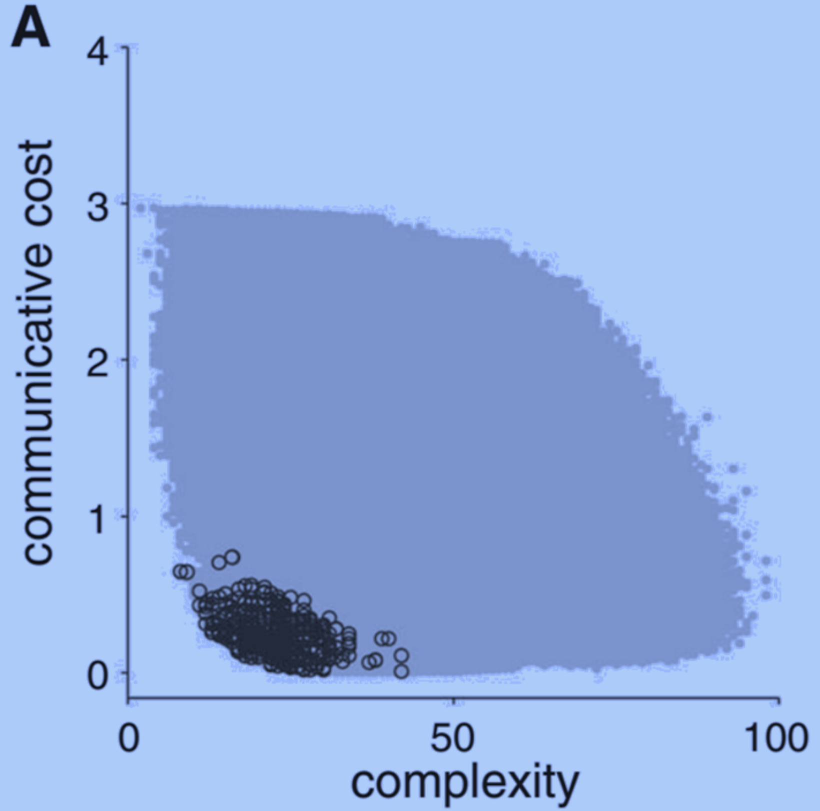
$$\begin{aligned}
\mathtt{mother}(x,y) &\leftrightarrow \mathtt{PARENT}(x,y)\wedge \mathtt{FEMALE}(x)\\
\mathtt{father}(x,y) &\leftrightarrow \mathtt{PARENT}(x,y)\wedge \mathtt{MALE}(x)\\
\mathtt{daughter}(x,y) &\leftrightarrow \mathtt{CHILD}(x,y)\wedge \mathtt{FEMALE}(x)
\end{aligned}$$

An example

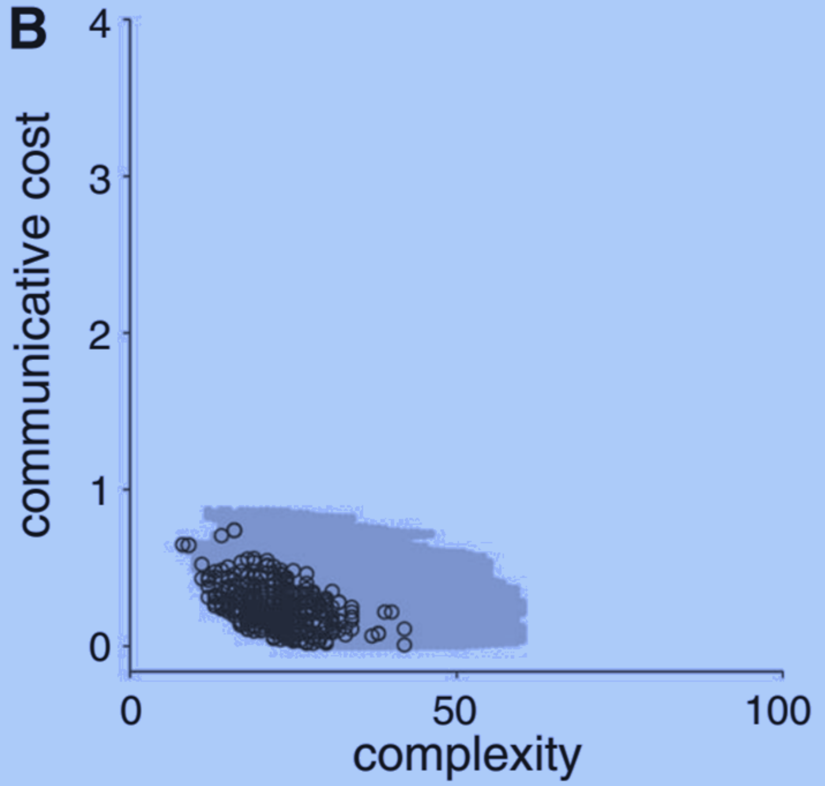


🟥	$\texttt{mother}(x, y)$	$\leftrightarrow \texttt{PARENT}(x, y) \wedge \texttt{FEMALE}(x)$
🟦	$\texttt{father}(x, y)$	$\leftrightarrow \texttt{PARENT}(x, y) \wedge \texttt{MALE}(x)$
🟪	$\texttt{daughter}(x, y)$	$\leftrightarrow \texttt{CHILD}(x, y) \wedge \texttt{FEMALE}(x)$
🟦	$\texttt{son}(x, y)$	$\leftrightarrow \texttt{CHILD}(x, y) \wedge \texttt{MALE}(x)$
🟧	$\texttt{sister}(x, y)$	$\leftrightarrow \exists z\ \texttt{daughter}(x, z) \wedge \texttt{PARENT}(z, y)$
🟩	$\texttt{brother}(x, y)$	$\leftrightarrow \exists z\ \texttt{son}(x, z) \wedge \texttt{PARENT}(z, y)$
	$\texttt{sibling}(x, y)$	$\leftrightarrow \exists z\ \texttt{CHILD}(x, z) \wedge \texttt{PARENT}(z, y)$
🟫	$\texttt{aunt}(x, y)$	$\leftrightarrow \exists z\ \texttt{sister}(x, z) \wedge \texttt{PARENT}(z, y)$
🟩	$\texttt{uncle}(x, y)$	$\leftrightarrow \exists z\ \texttt{brother}(x, z) \wedge \texttt{PARENT}(z, y)$
🟪	$\texttt{niece}(x, y)$	$\leftrightarrow \exists z\ \texttt{daughter}(x, z) \wedge \texttt{sibling}(z, y)$
🟦	$\texttt{nephew}(x, y)$	$\leftrightarrow \exists z\ \texttt{son}(x, z) \wedge \texttt{sibling}(z, y)$
🟥	$\texttt{grandmother}(x, y)$	$\leftrightarrow \exists z\ \texttt{mother}(x, z) \wedge \texttt{PARENT}(z, y)$
🟦	$\texttt{grandfather}(x, y)$	$\leftrightarrow \exists z\ \texttt{father}(x, z) \wedge \texttt{PARENT}(z, y)$
🟪	$\texttt{granddaughter}(x, y)$	$\leftrightarrow \exists z\ \texttt{daughter}(x, z) \wedge \texttt{CHILD}(z, y)$
🟩	$\texttt{grandson}(x, y)$	$\leftrightarrow \exists z\ \texttt{son}(x, z) \wedge \texttt{CHILD}(z, y)$

Lexicalized systems

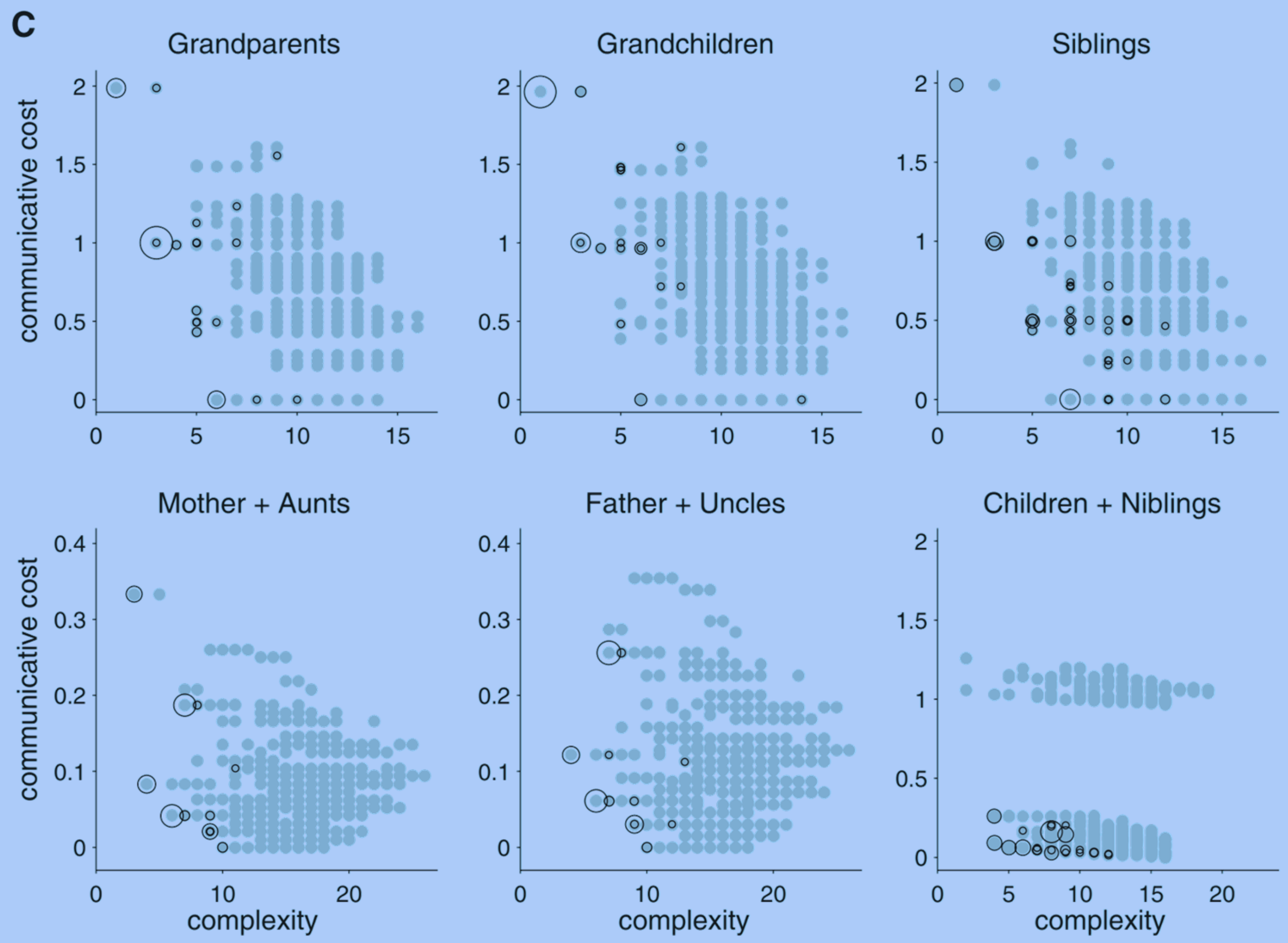


All systems



Systems built from attested categories that appear more than twice in the Murdock data.

Optimality analysis

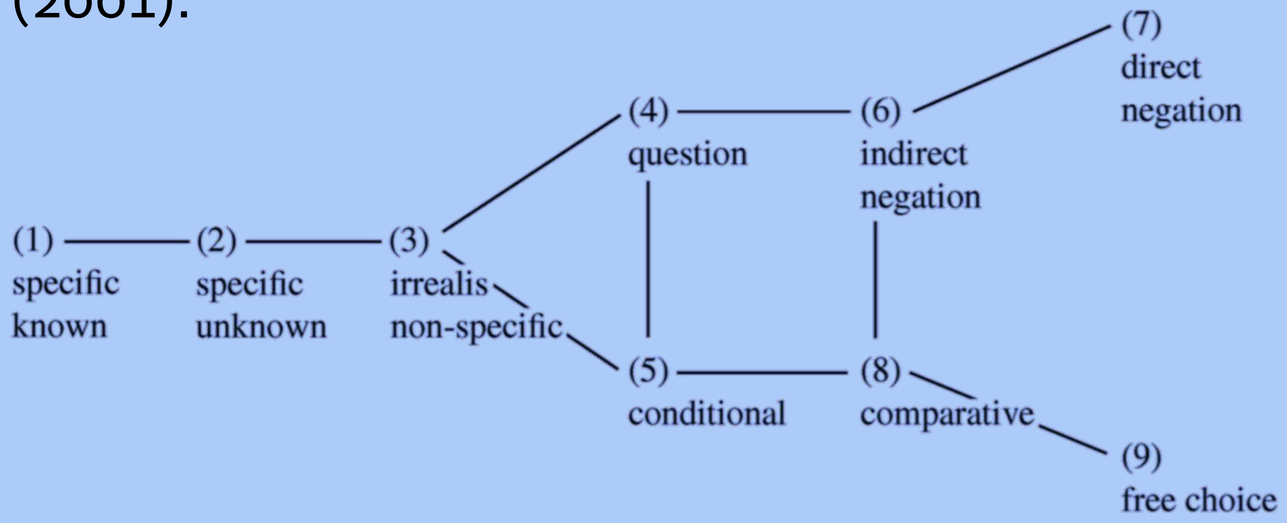


- Black circles are attested
- Size proportional to frequency (Murdoch)

Case Study II:
Indefinite pronouns

Indefinite pronouns

- Denic, Steinert-Threlkeld, & Szymanik (2020)
- Words like *someone*, *something*, *anyone*, *anything*, *no-one*, *nothing*
- Haspelmath (2001):



- Universal: Any indefinite pronoun in any language can only take functions which form a connected area on the map

Six types of indefinite pronouns

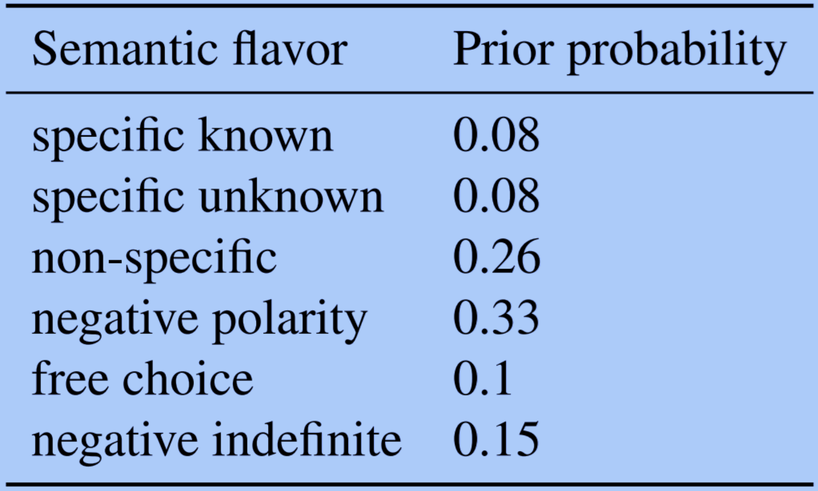
1. Specific known flavor [specific individual that the interlocutors can uniquely identify]:
   - **Someone** managed to mess this up — we all know who!
2. Specific unknown flavor [specific individual that the interlocutors cannot uniquely identify]:
   - I heard that **someone** failed, but I don't know who.
3. Non-specific flavor [existential quantifier over some domain of possible referents]:
   - You should probably talk to **someone** else about this too.
4. Negative polarity flavor [existential quantifier over a widened domain of possible referents]:
   - Less than three companies hired **anyone** this year.
5. Free choice flavor [a wide-scope universal quantifier over some domain of possible referents]:
   - You can hire almost **anyone** here: most of them great.
6. Negative indefinite flavor [a negated existential quantifier over domain of possible referents]:
   - Who went to the party? **No one**.

Complexity

- Semantic features (binary):
  - Known to the speaker (K)
  - Specific (S)
  - Scalar endpoint (SE)
    - Negative polarity and free choice indefinites evoke a pragmatic scale of alternatives ordered by likelihood, and they associate with its lowest endpoint
    - SE+: negative polarity, free choice, negative indefinite
    - SE–: specific known, specific unknown, non-specific
  - Scale reversal (R)
    - Reverse the order of alternatives on the pragmatic scale
  - In the scope of negation (N)
- Complexity is the number of features an indefinite pronoun has
- E.g., specific unknown flavor is 'S+\ K–'
  - Complexity of 2

Informativeness

- Probability that the communication will be successful given
  - The prior over flavors from the set of flavors
  - The conditional probability which reflects the probability that the speaker uses the indefinite pronoun i to communicate f
  - The probability that the listener correctly guesses f upon hearing i
- Corpus estimate for flavour prior probability:

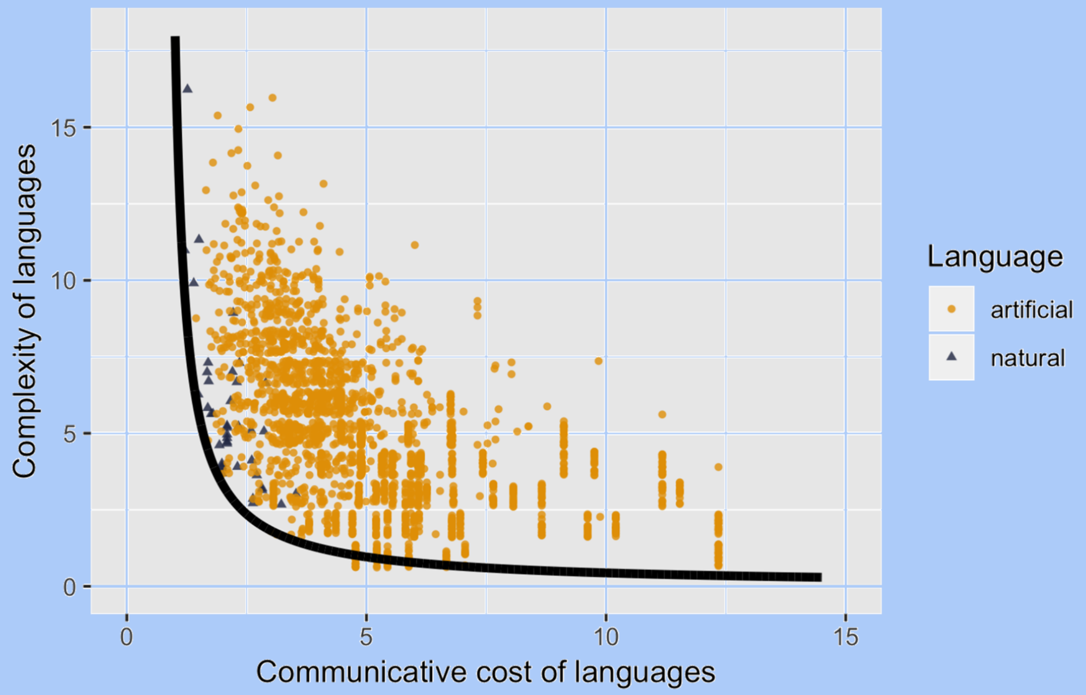


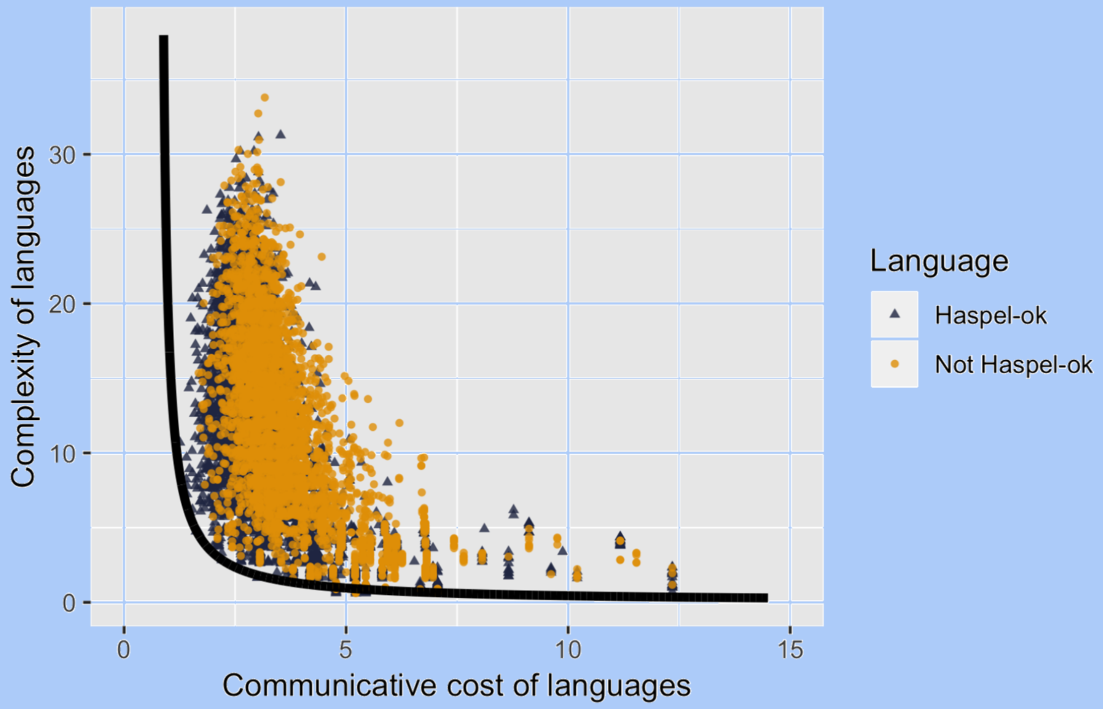
Semantic flavor	Prior probability
specific known	0.08
specific unknown	0.08
non-specific	0.26
negative polarity	0.33
free choice	0.1
negative indefinite	0.15

Some detail (can skip!)

- Complexity and communicative cost measures for each of the 40 languages in Haspelmath’s corpus.
- We artificially generated 10000 languages, which could have between 1 and 7 indefinite pronouns (7 is the maximum number of indefinite pronouns that any natural language has in Haspelmath’s corpus).
- Each indefinite pronoun in each artificial language was randomly assigned one of the 63 logically possible combinations of flavors (26 –1 combination whereby an indefinite pronoun doesn’t convey any of the 6 flavors).
- The artificial languages were then matched to natural languages for the degree of synonymy. The degree of synonymy captures how many different indefinites can be used to express a flavor: if the indefinites in a language have more overlapping meanings, the degree will be higher.
- Matching ensured that for each degree of synonymy d of natural languages, the proportion of artificial languages with d was the same as the proportion of natural languages with d. After matching, 2133 artificial languages remained for comparison to natural languages (mean degree of synonymy in both groups is 0.67)

Results


Complexity of languages
0
5
10
15
Communicative cost of languages
0
5
10
15
Language
● artificial
▲ natural


Complexity of languages
0
10
20
30
Communicative cost of languages
0
5
10
15
Language
▲ Haspel-ok
● Not Haspel-ok

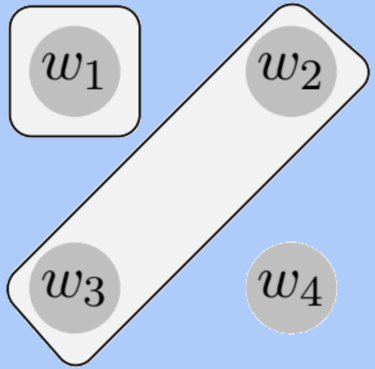
Case Study III: Boolean universals

Boolean universals

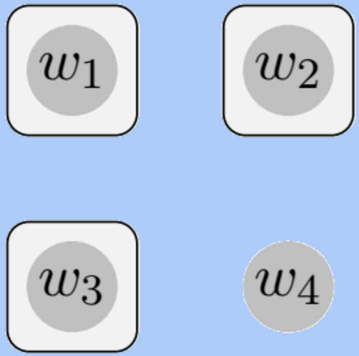
- Model in Uegaki (2021)
- One universal: the absence of a term expressing the negation of the conjunction among attested Boolean connectives.
- Other universal: specific attested inventories of operators.
- Let’s try and see if (all and only) attested inventories lie at the Pareto frontier of simplicity and communicative accuracy.

The set of possibilities

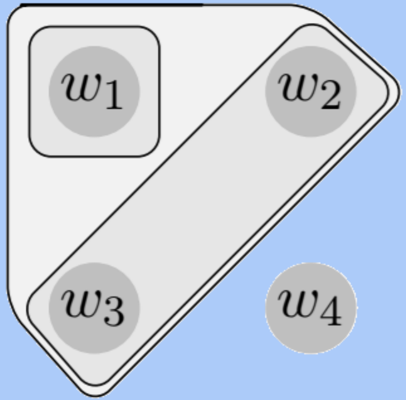
- Set of inventories is the powerset (minus the empty set) of the set of 16 binary Boolean operators (65535)
- We consider strengthened meanings (with EXH)



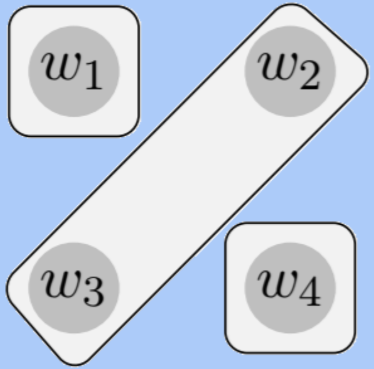
{and,or}



{and,P,Q}



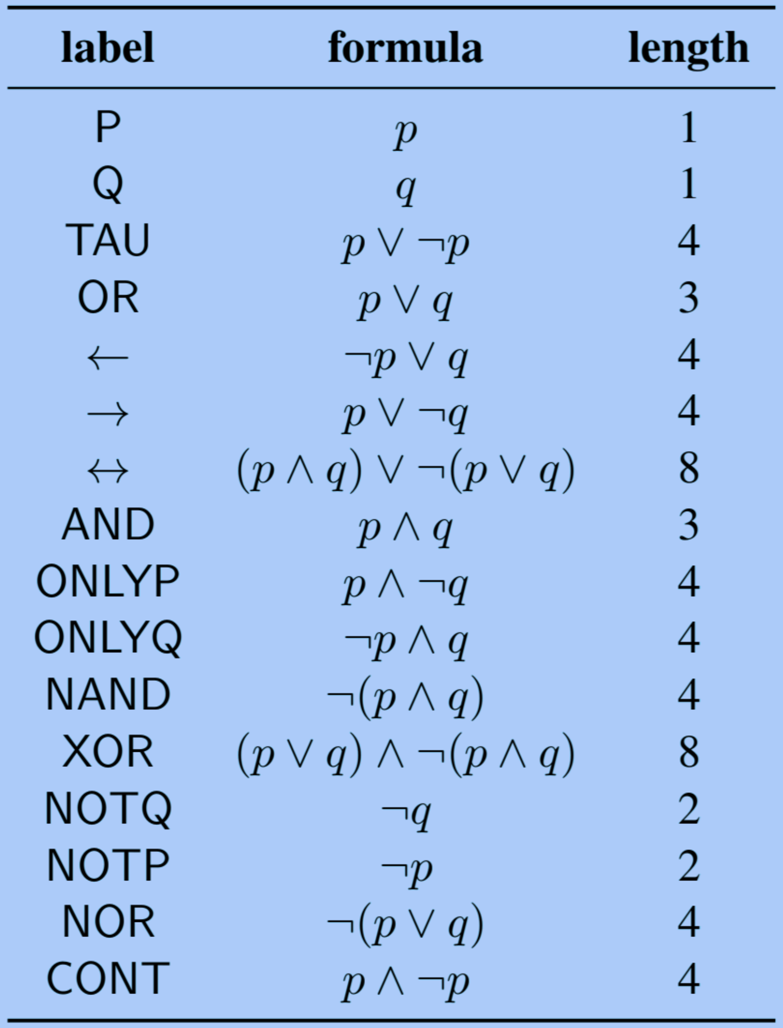
{and,or,xor}



{and,or,nor}
{and,nand, nor}

Complexity

- The meaning of each operator can be encoded in propositional logic with negation, conjunction, and disjunction.
- **Complexity**: the sum of the number of symbols in Propositional Logic containing ¬, ∧, and ∨ necessary to represent all connectives in the inventory.
- Why this choice of logic
  - Circular?



label	formula	length
P	p	1
Q	q	1
TAU	$p \vee \neg p$	4
OR	$p \vee q$	3
$\leftarrow$	$\neg p \vee q$	4
$\rightarrow$	$p \vee \neg q$	4
$\leftrightarrow$	$(p \wedge q) \vee \neg(p \vee q)$	8
AND	$p \wedge q$	3
ONLYP	$p \wedge \neg q$	4
ONLYQ	$\neg p \wedge q$	4
NAND	$\neg(p \wedge q)$	4
XOR	$(p \vee q) \wedge \neg(p \wedge q)$	8
NOTQ	$\neg q$	2
NOTP	$\neg p$	2
NOR	$\neg(p \vee q)$	4
CONT	$p \wedge \neg p$	4

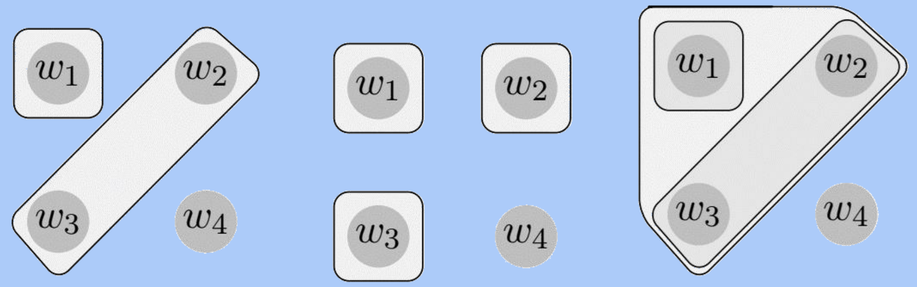
Informativeness

- Informativeness: the likelihood that the meaning intended by the sender is accurately recovered by the addressee, given scalar implicature

$$\sum_{w\in W} \overbrace{P(w)}^{\text{prior prob. of } w} \sum_{c\in L} \overbrace{P(c|w)}^{\text{prob. of uttering } c \text{ given } w} \sum_{w'\in W} \overbrace{P(w'|c)}^{\text{prob. of interpreting } c \text{ as } w'} \cdot \overbrace{u(w,w')}^{\text{utility of } w' \text{ given } w}$$

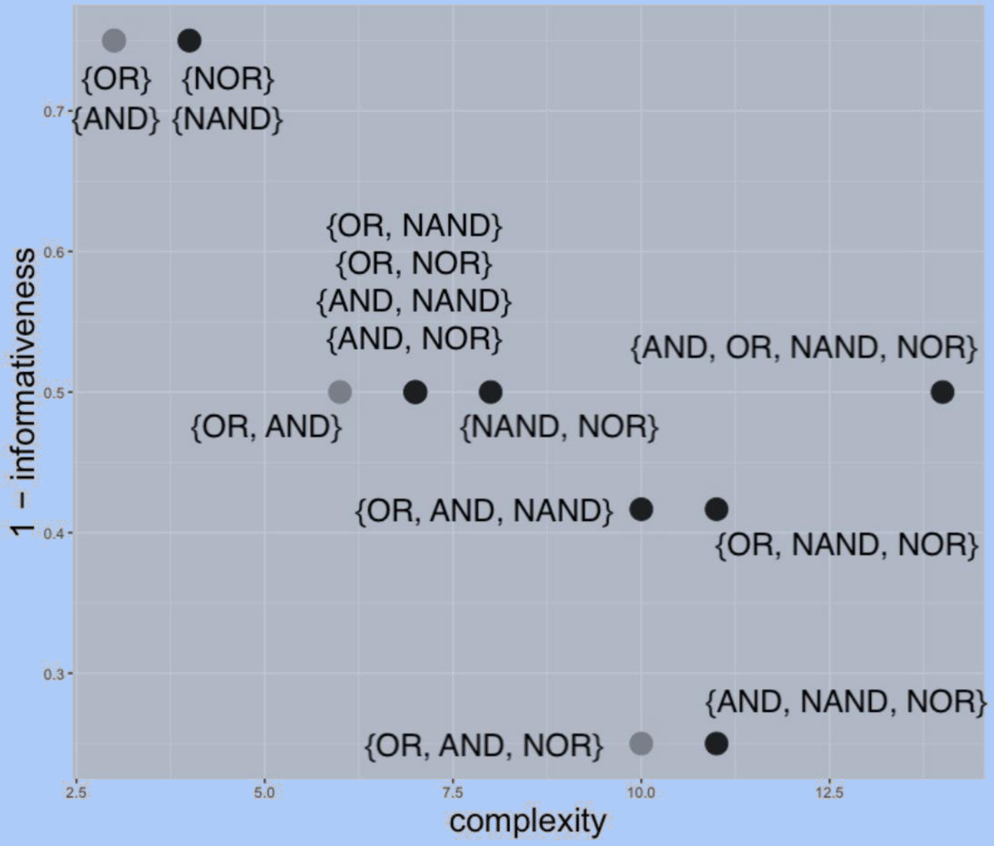
- Utility is binary: 1 if intended message is recovered and 0 otherwise

	W1 (p q)	**W2 (p not-q)**	**W3 (not-q p)**	**W4 (not-p not-q)**	**Total**
AND, OR	1 (AND)	½ (OR)	½ (OR)	0 (None)	0.5
AND, P, Q	1 (AND)	1 (P)	1 (Q)	0 (None)	0.75
AND, OR, XOR	½ 1 (AND) + ½ 1/3 (OR) = 2/3	1/3 (OR)	1/3 (OR)	0 (None)	0.375

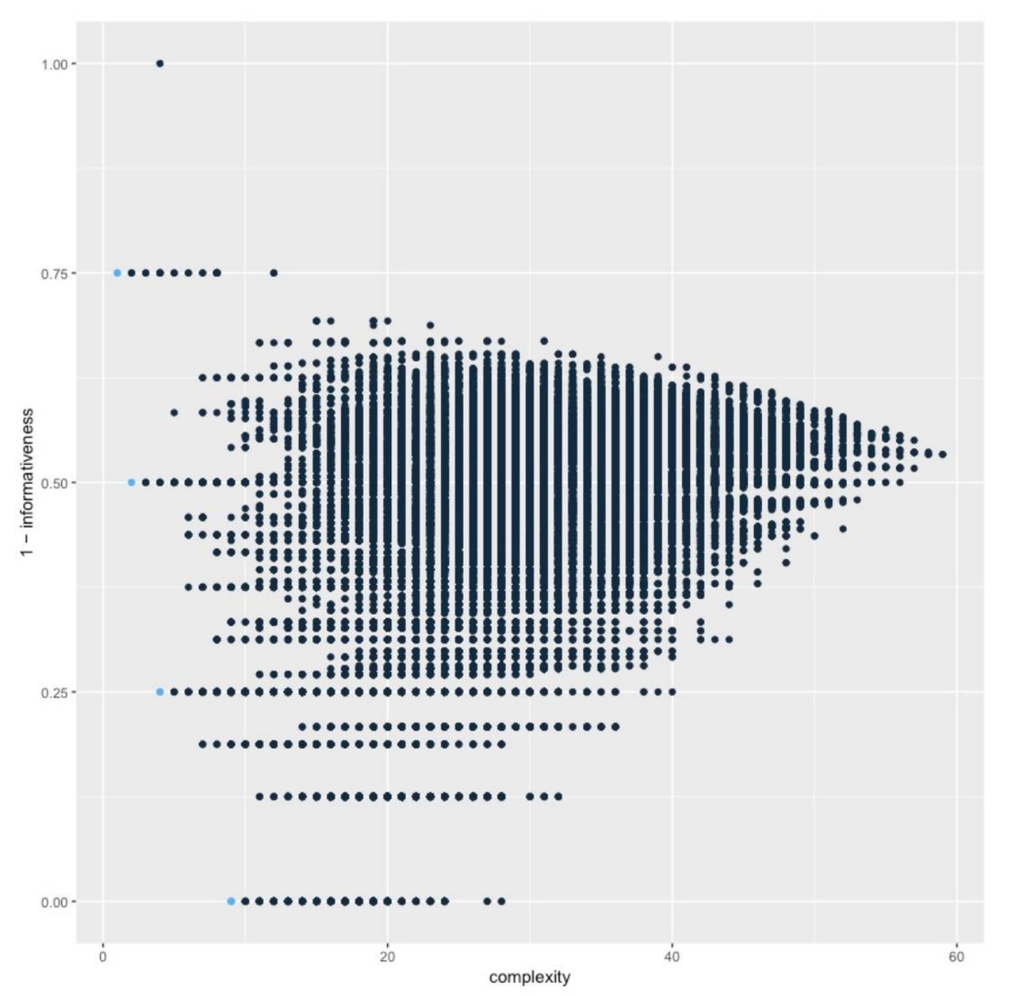


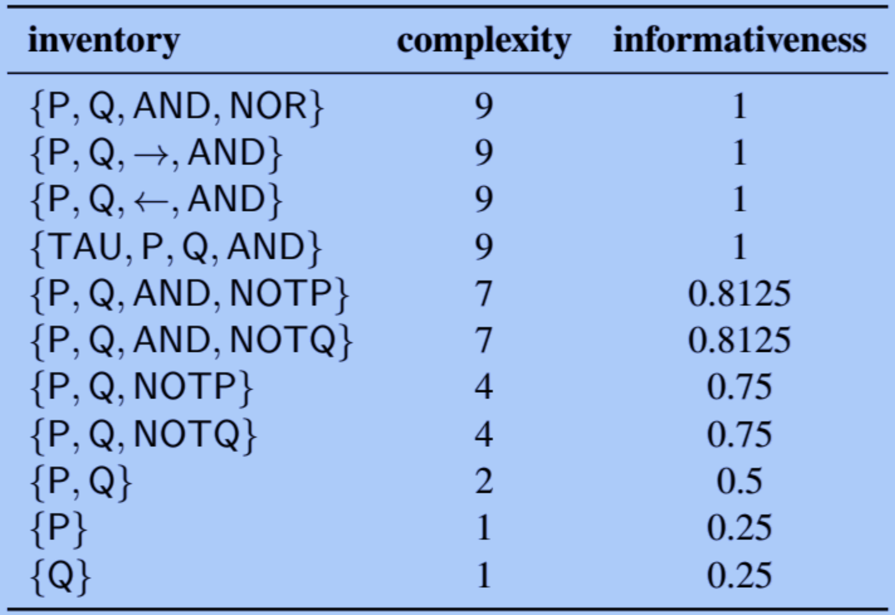
Four-corner inventories

- Let’s start with just the four Aristotelian operators OR, AND, NOR, NAND.
- All *and only* the typologically attested inventories are Pareto-optimal!
- Can we include *all* operators?



All inventories



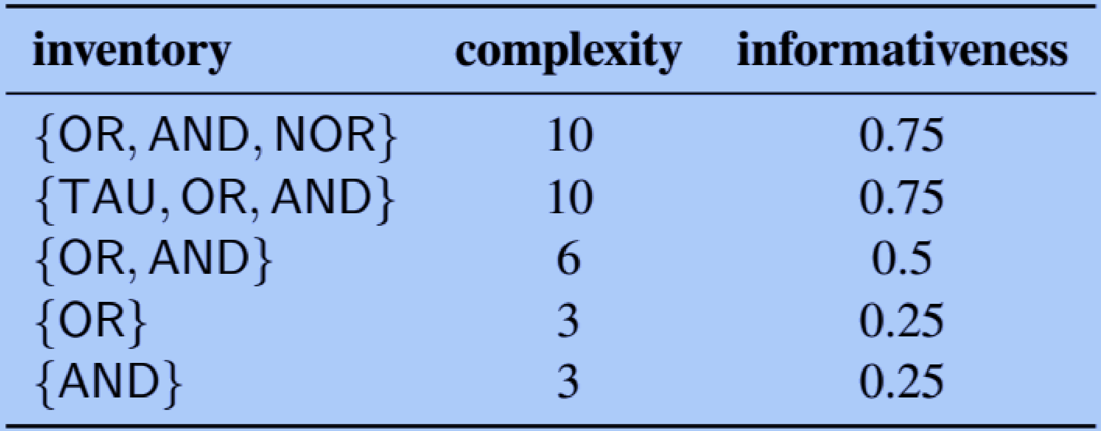


inventory	complexity	informativeness
{P, Q, AND, NOR}	9	1
{P, Q, →, AND}	9	1
{P, Q, ←, AND}	9	1
{TAU, P, Q, AND}	9	1
{P, Q, AND, NOTP}	7	0.8125
{P, Q, AND, NOTQ}	7	0.8125
{P, Q, NOTP}	4	0.75
{P, Q, NOTQ}	4	0.75
{P, Q}	2	0.5
{P}	1	0.25
{Q}	1	0.25

- No ‘or’ in Pareto frontier
- Unattested inventories
- Bad situation!

All inventories

- Natural restriction: Commutative operators!
- Operator O is commutative iff $O(x, y) = O(y, x)$
  - Non commutative: P, Q, NOTP, NOTQ, →, ←, ONLYP, ONLYQ
  - Commutative: TAU, CONT, OR, ↔, AND, NAND, XOR, NOR
- Previous work: only commutative operators are lexicalizable (Gazdar and Pullum 1976; Gazdar, 1979: 74-78)
  - No clear explanation yet!



inventory	**complexity**	**informativeness**
{OR, AND, NOR}	10	0.75
{TAU, OR, AND}	10	0.75
{OR, AND}	6	0.5
{OR}	3	0.25
{AND}	3	0.25

- Only {TAU, OR, AND} is not attested
- Independent work shows that trivial meanings are ungrammatical
- (Barwise and Cooper, 1981; von Fintel, 1993; Gajewski, 2002; Fox and Hackl, 2007; Chierchia, 2013; Del Pinal, 2019)

Case Study IV:
Adjectival monotonicity

Modelling systems of gradable adjectives

$D = \{d_1, d_2, d_3\}$

$S = \{s_1, s_2, s_3\}$

L_1	d_1	d_2	d_3
s_1	1	1	0
s_2	0	1	0
s_3	1	0	0

$\notin \Lambda$

L_2	d_1	d_2	d_3
s_1	1	1	0
s_2	0	1	0
s_3	0	1	1

$\in \Lambda$

$$\Lambda = \{L \subset S \times D \mid \forall d \in D. \exists s \in S. (s, d) \in L\}$$

Call a language “monotonic” iff all its meanings are monotonic

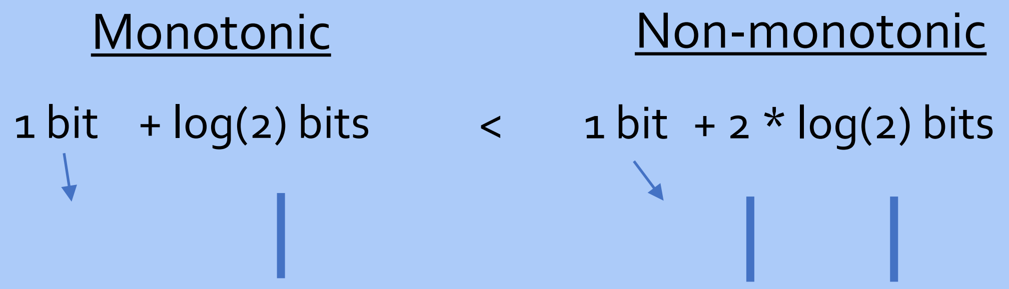
Prior over languages

Prior:

$$p(L) \propto 2^{-\lambda \mathrm{DL}(L)}$$

Step coding:

$$\mathrm{DL}(L) = \sum_{m \in L} 1 + \#\ \text{changes}(m) \log_2(\#\ \text{degrees - } 1)$$



Monotonic

1 bit + log(2) bits

<

Non-monotonic

1 bit + 2 * log(2) bits

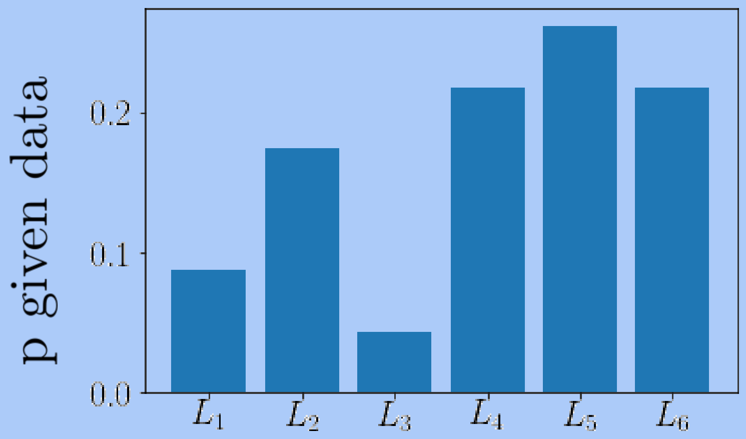
Monotonic meanings get greater prior probability than non-monotonic ones

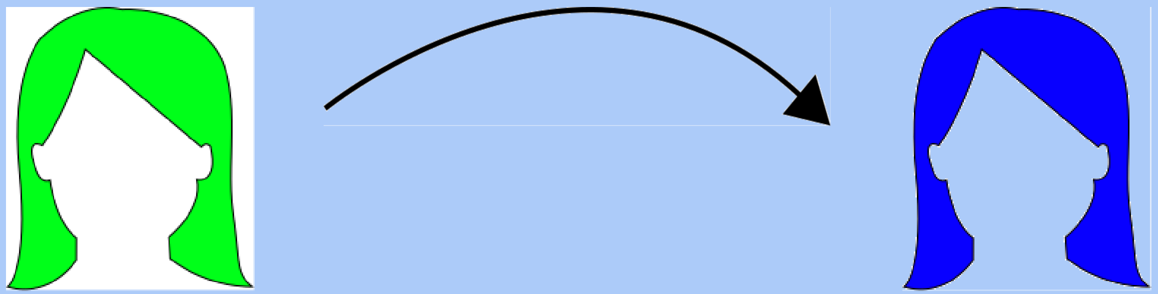
Learning

$$p(L_i \mid \text{data}) \propto p(\text{data} \mid L_i)p(L_i)$$

$$L_n = \begin{array}{c|ccc} & d_1 & d_2 & d_3 \\ \hline s_1 & 1 & 0 & 0 \\ \hline s_2 & 1 & 1 & 0 \\ \hline s_3 & 0 & 1 & 1 \end{array}$$

$$\text{data} =$$



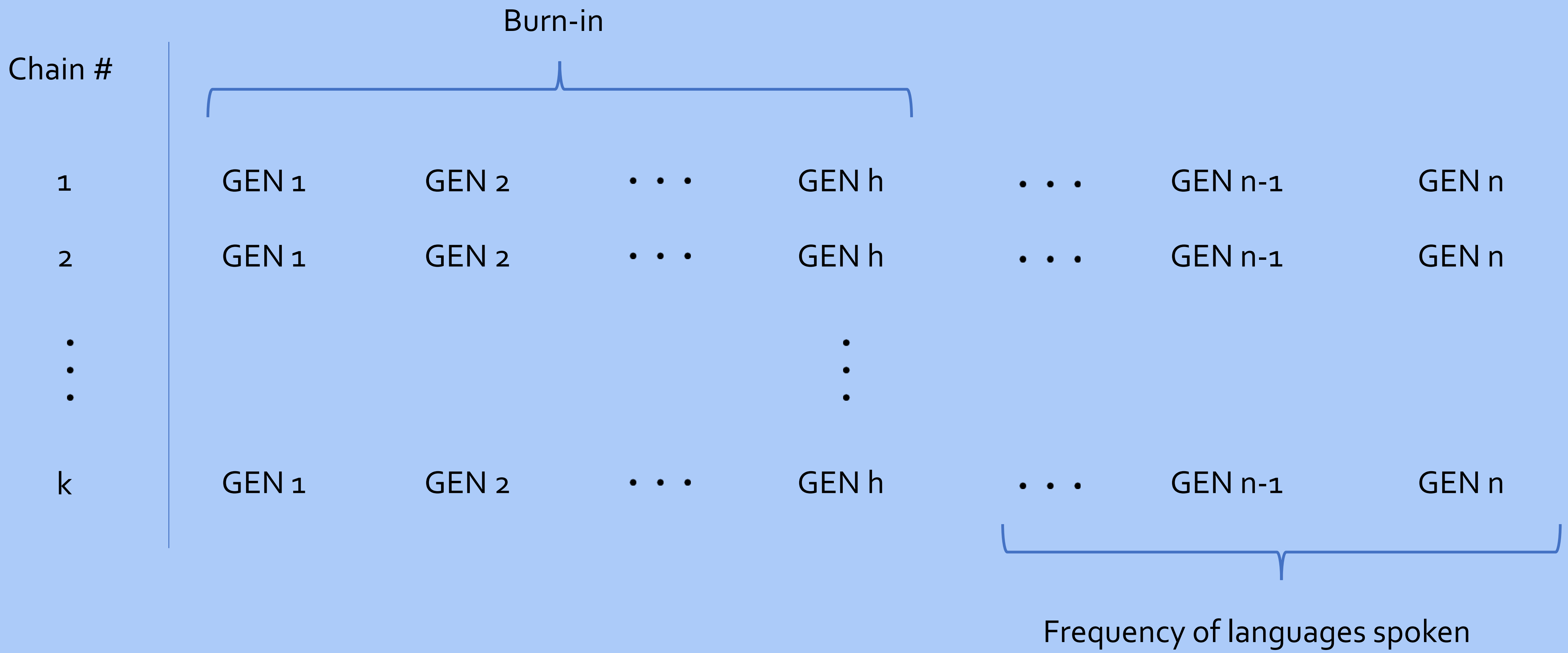


Cultural parent
(Gen n)

Cultural child
(Gen n+1)

sample agents

Overall IL model

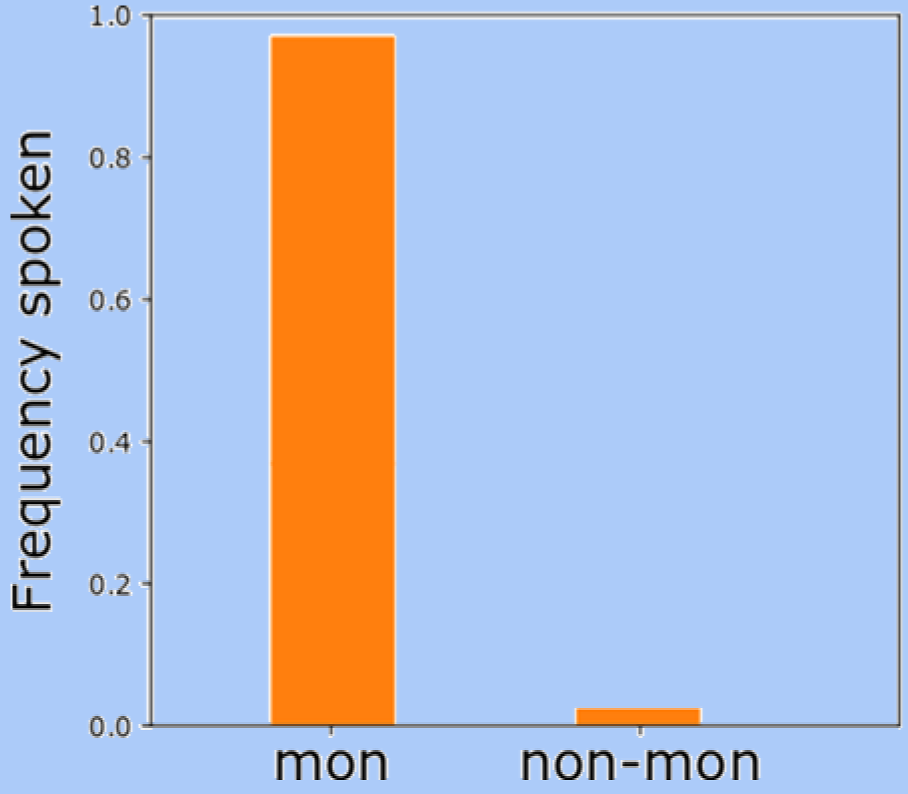


Burn-in

Chain #							
1	GEN 1	GEN 2	• • •	GEN h	• • •	GEN n-1	GEN n
2	GEN 1	GEN 2	• • •	GEN h	• • •	GEN n-1	GEN n
⋮				⋮			
k	GEN 1	GEN 2	• • •	GEN h	• • •	GEN n-1	GEN n

Frequency of languages spoken

First model: IL with gradable adjectives



- 3000 generations of IL
- 100 agents
- 100 runs of the simulation
- Burn-in: 500 generations

There is a problem…

Degeneracy

L	d_1	d_2	d_3
s_1	1	0	1
s_2	0	1	0
s_3	0	1	1

L	d_1	d_2	d_3
s_1	0	1	0
s_2	0	1	1
s_3	1	0	0

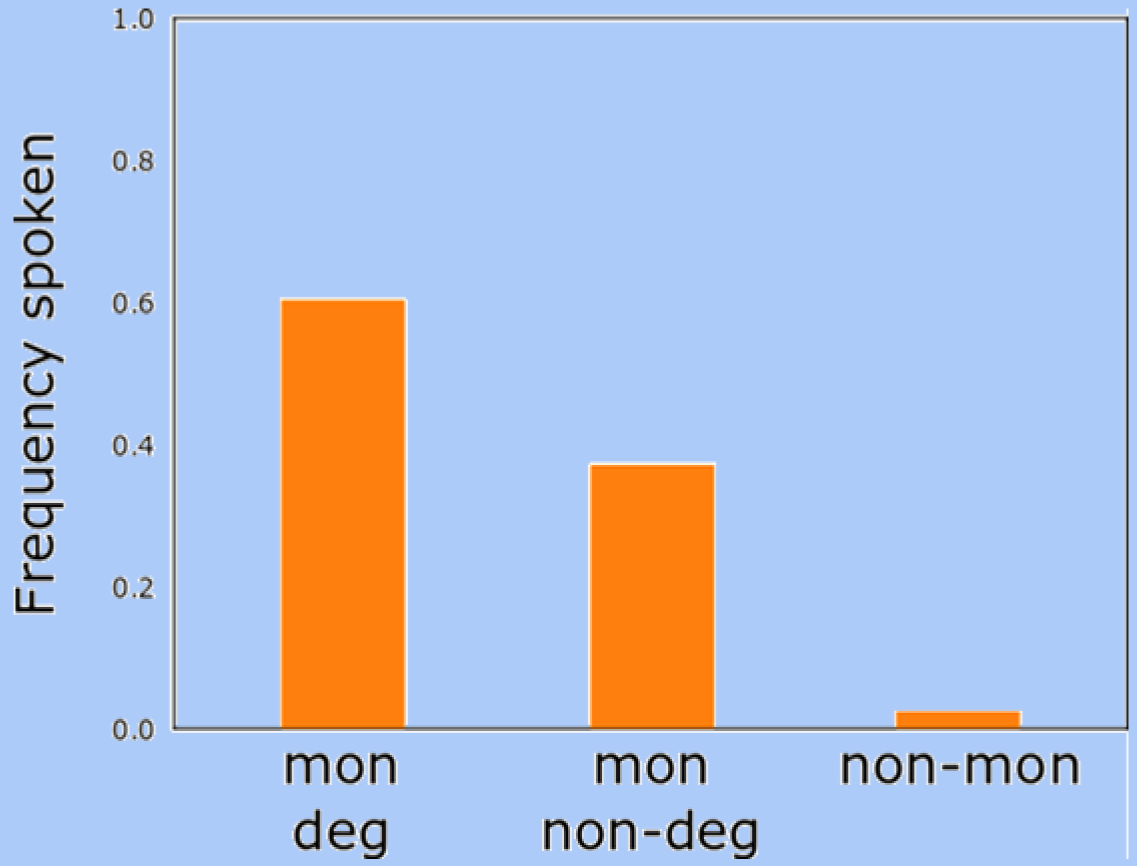
Non-degenerate languages

L	d_1	d_2	d_3
s_1	1	1	1
s_2	0	0	0
s_3	1	1	1

L	d_1	d_2	d_3
s_1	0	0	0
s_2	1	1	1
s_3	0	0	0

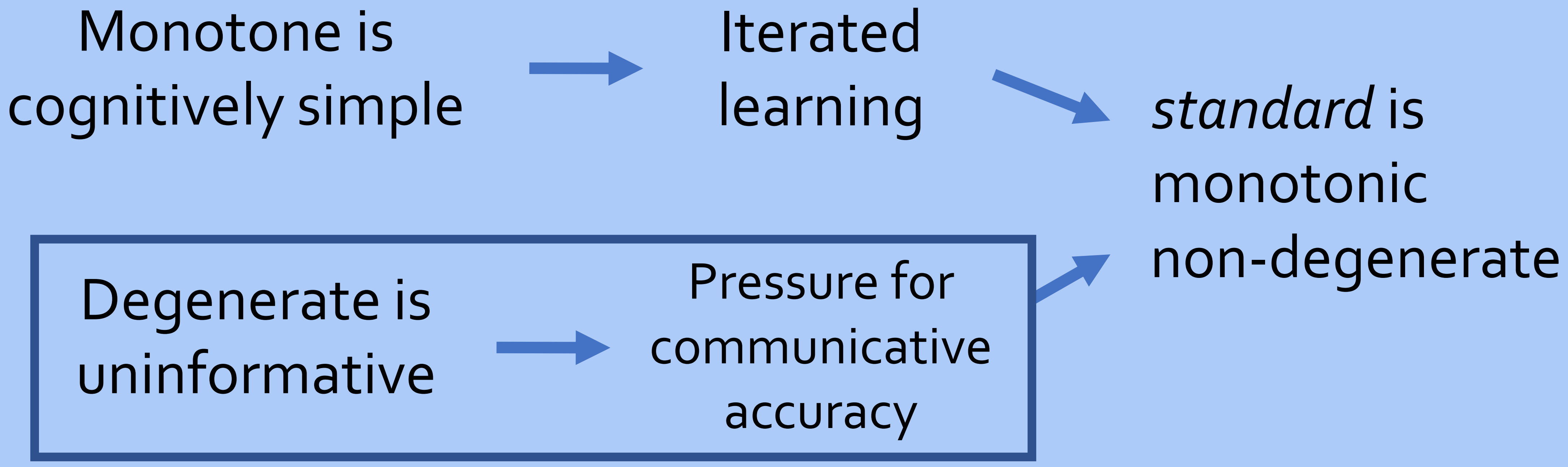
Degenerate languages

First model: IL with gradable adjectives



Not quite right!

A functional explanation



Monotone is cognitively simple

Iterated learning

Degenerate is uninformative

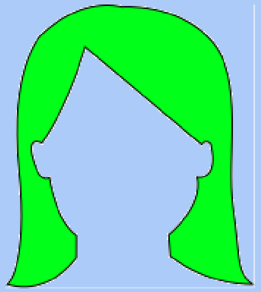
Pressure for communicative accuracy

standard is monotonic non-degenerate

Second model: communication

d_2

d_2

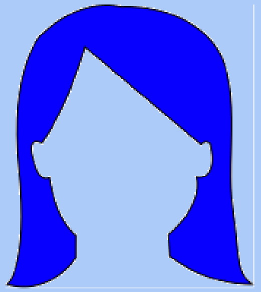


Speaker

	d_1	d_2	d_3
s_1	1	0	0
s_2	1	0.5	0
s_3	0	0.5	1

$$\xrightarrow{s_2}$$

	d_1	d_2	d_3
s_1	1	0	0
s_2	0	1	0
s_3	0	1	1

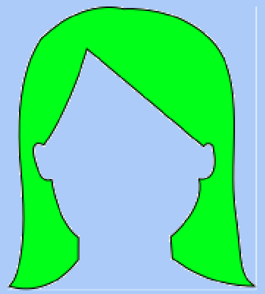


Hearer

Second model: communication

$$d_2 \qquad = \qquad d_2$$

Success!

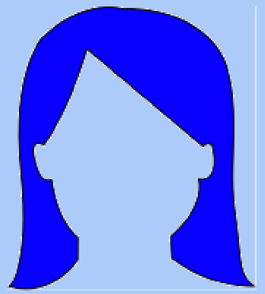


Speaker

	d_1	d_2	d_3
s_1	1	0	0
s_2	1	0.5	0
s_3	0	0.5	1

$$\xrightarrow{\quad s_2 \quad}$$

	d_1	d_2	d_3
s_1	1	0	0
s_2	0	1	0
s_3	0	1	1



Hearer

Second model: communicative pressure

(Uniform probability of observing degrees)

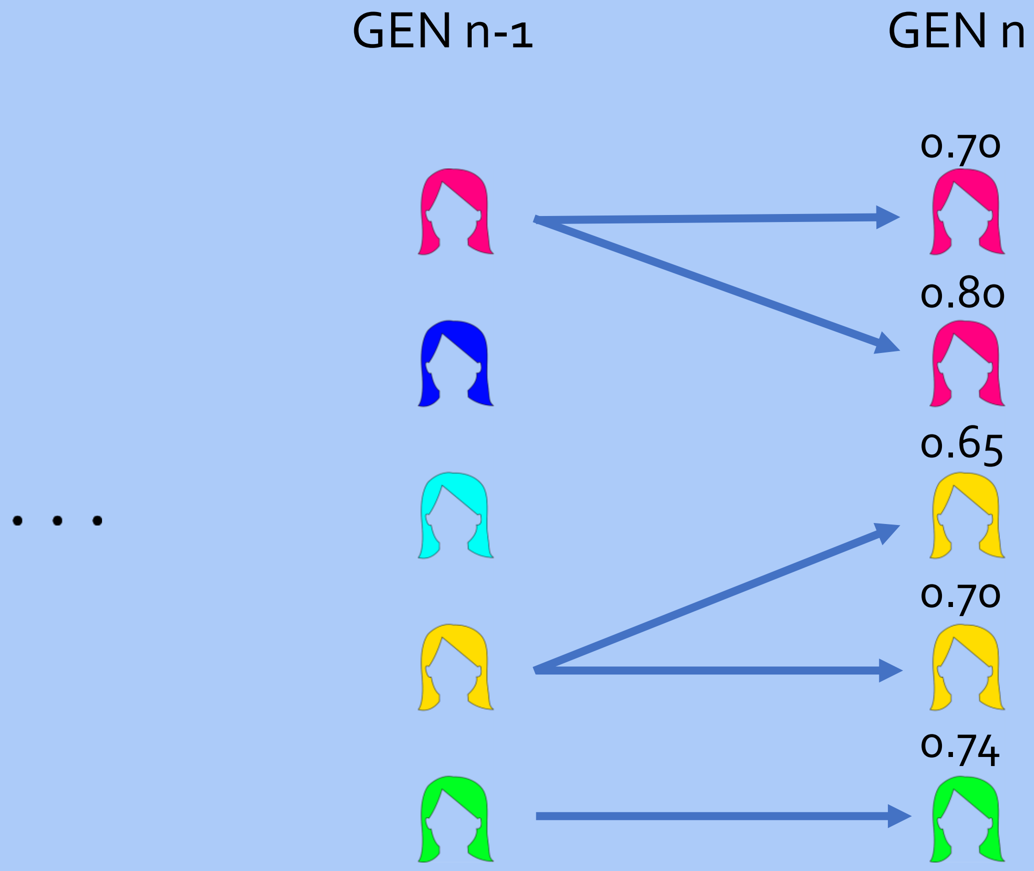
Expected communicative accuracy:

$$c(L_h, L_s) = \sum_{d_i \in D} \sum_{s_j \in S} p(s_j \mid d_i, L_s) p(d_i \mid s_j, L_h) p(d_i)$$

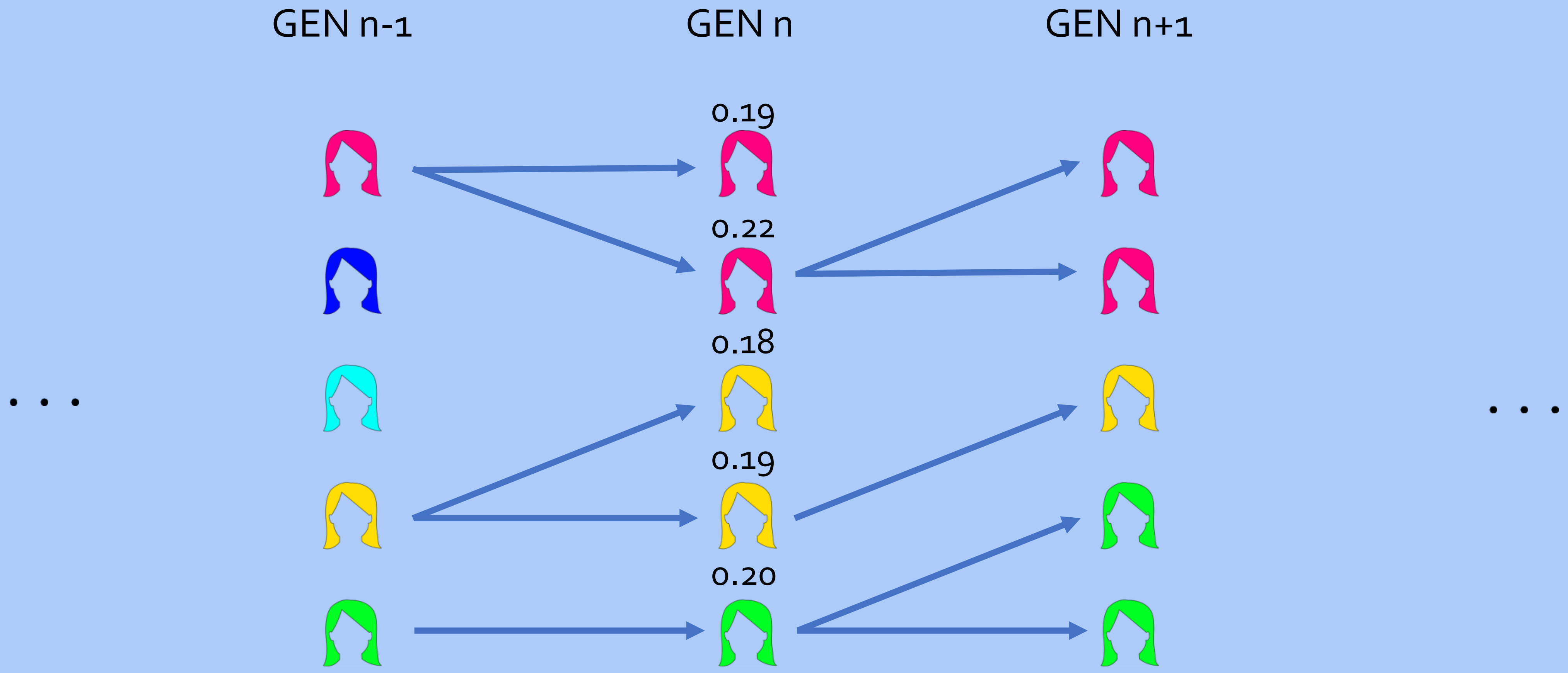
Probability that the speaker uses that signal for that degree

Probability that the listener guesses that degree for that signal

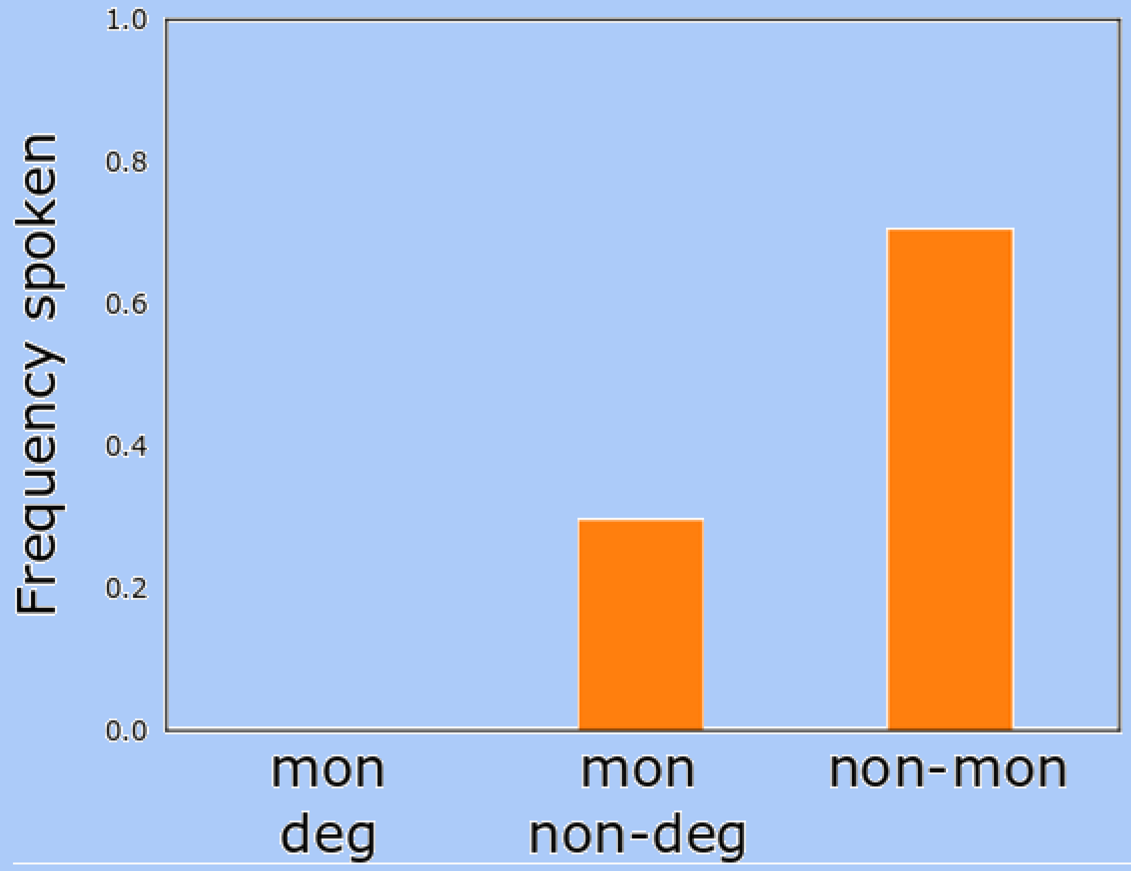
Implementing communicative pressure



Implementing communicative pressure



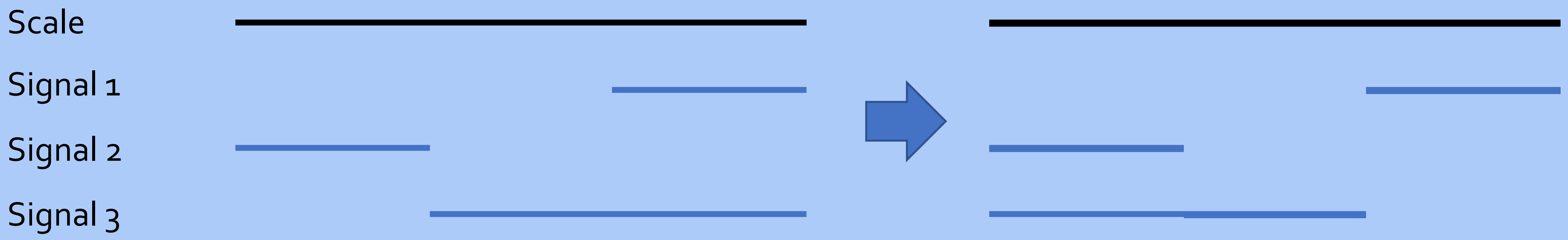
Results of second model



- 3000 generations of IL
- 100 agents
- 100 runs of the simulation
- Burn-in: 500 generations

Incorrect prediction!

Why non-monotonic?



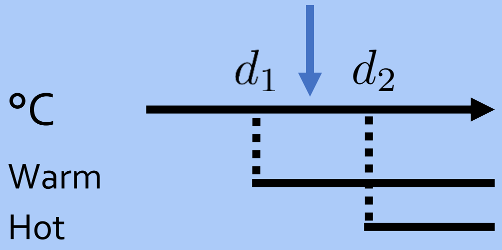
Scale

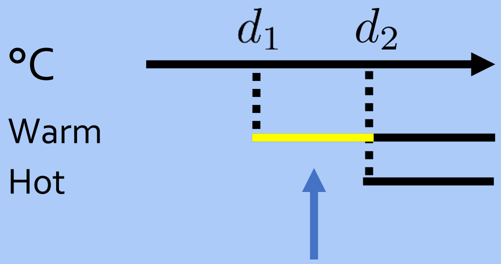
Signal 1

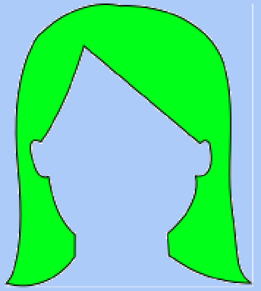
Signal 2

Signal 3

A bit of pragmatics: scalar implicatures



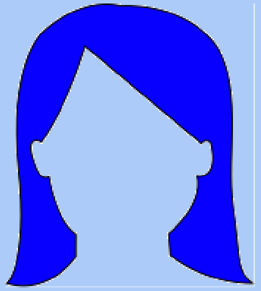




"Warm"

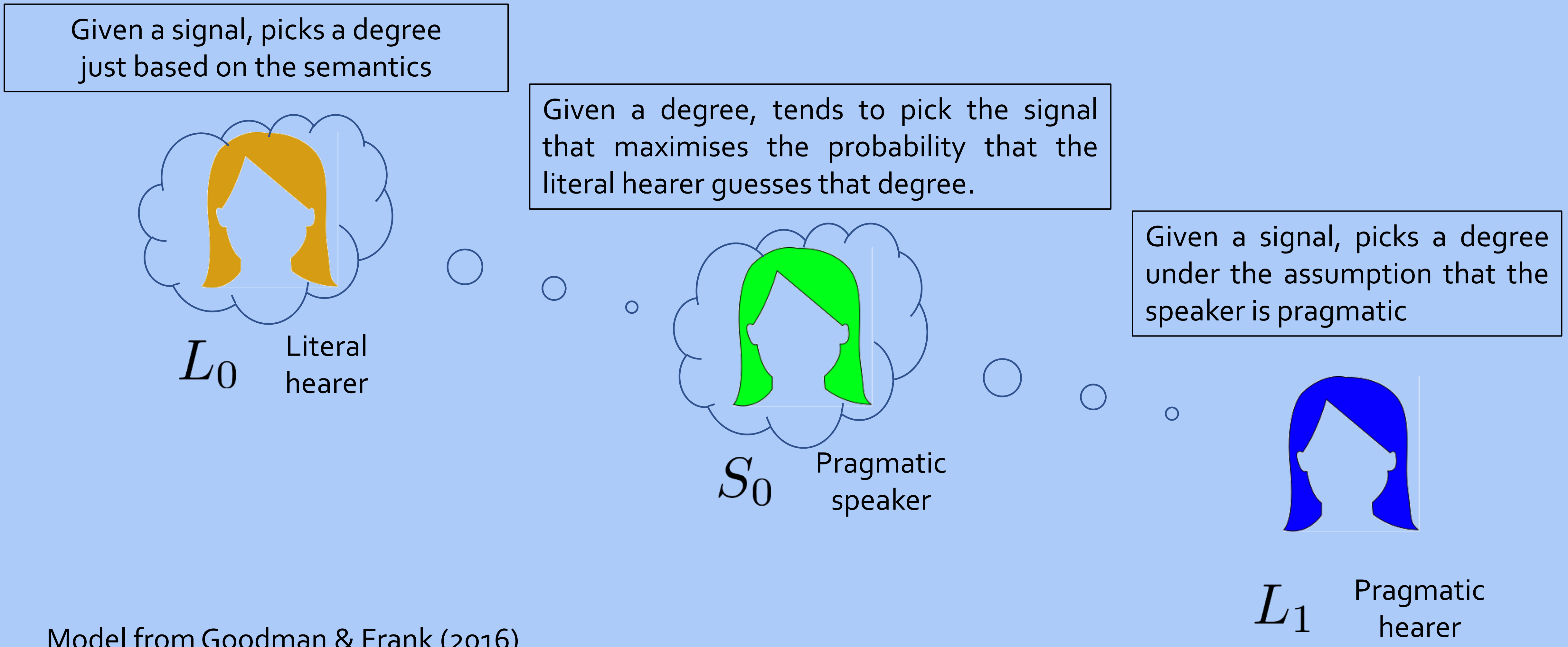
Speaker

If it had been greater than d_2 she would have said "hot".



Hearer

Modelling scalar implicatures



Given a signal, picks a degree just based on the semantics

L_0 Literal hearer

Given a degree, tends to pick the signal that maximises the probability that the literal hearer guesses that degree.

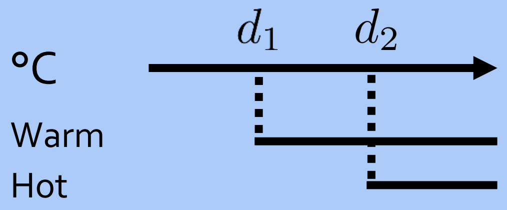
S_0 Pragmatic speaker

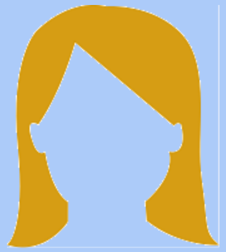
Given a signal, picks a degree under the assumption that the speaker is pragmatic

L_1 Pragmatic hearer

Model from Goodman & Frank (2016)

Modelling scalar implicatures

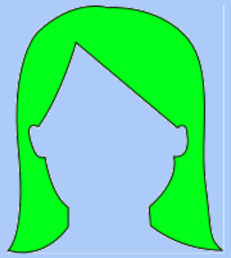




L_0 Literal hearer

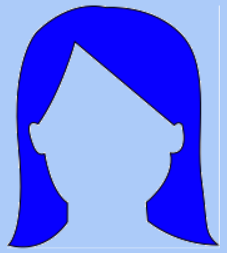
$$\mathcal{U}_{S_1}(s;d) = \log(p_{L_0}(d \mid s))$$

$$p_{S_1}(s \mid d) \propto \exp(\alpha \mathcal{U}_{S_1}(s;d))$$



S_1 Pragmatic speaker

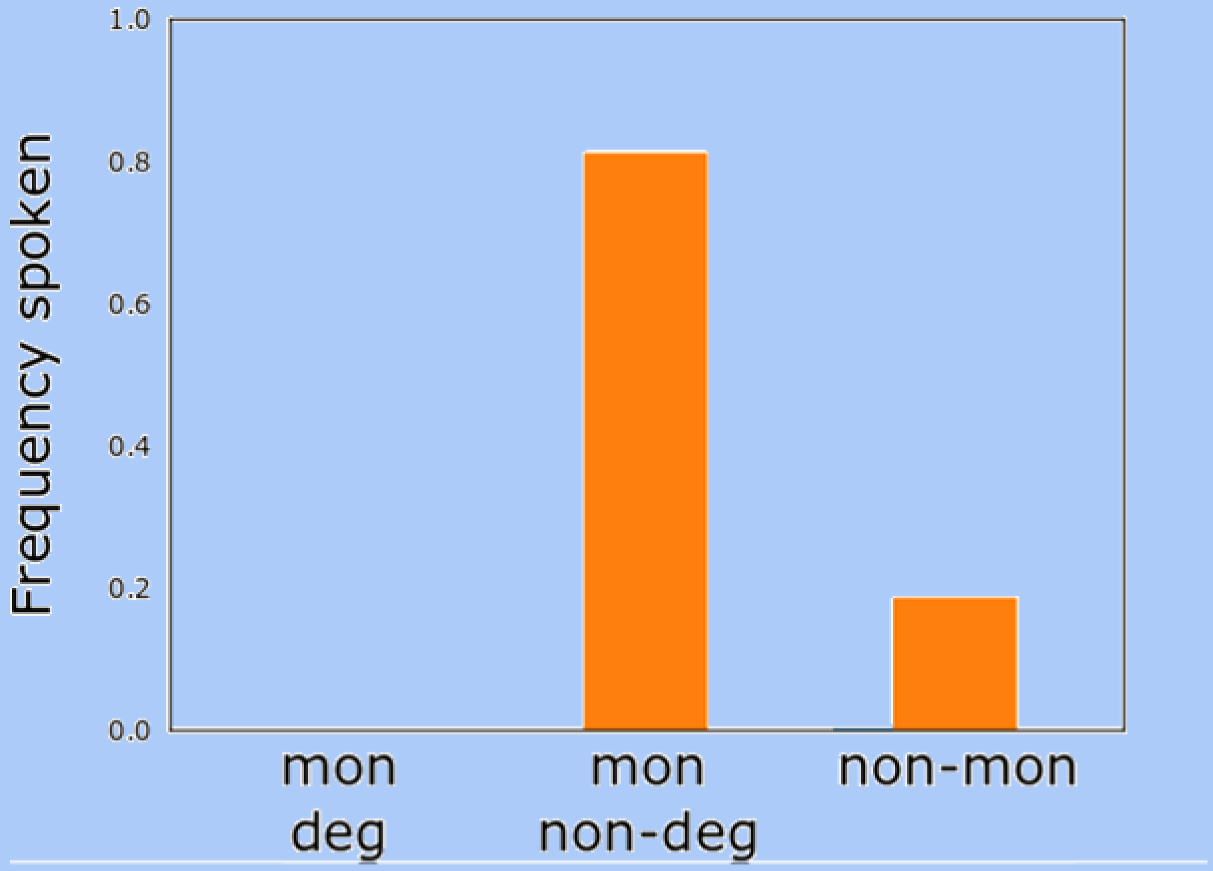
$$p_{L_1}(d \mid s) \propto p_{S_1}(s \mid d) p_{L_1}(d)$$



L_1 Pragmatic hearer

Model from Goodman & Frank (2016)

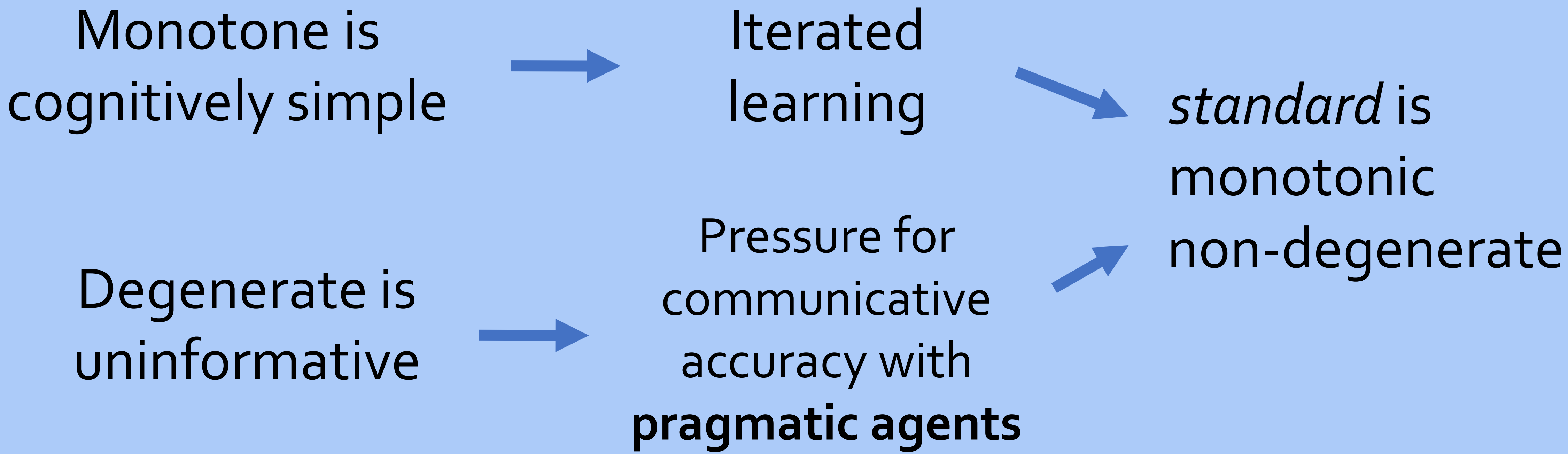
Results of third model



- 3000 generations of IL
- 100 agents
- 100 runs of the simulation
- Burn-in: 500 generations

Monotonicity evolves!

Conclusion of modelling



Monotone is cognitively simple

Iterated learning

Degenerate is uninformative

Pressure for communicative accuracy with **pragmatic agents**

standard is monotonic non-degenerate

A few final words

- There is other work on tradeoff analyses
  - Quantifiers (Steinert-Threlkeld, 2021)
  - Color naming, folks biology, number systems (Kemp, Xu, Regier 2018)
  - Modal semantics (Imel & Steinert-Threlkeld, 2022)
- Considering multiple pressures at once most promising explanation
- Many open questions in how to explain universals. Methods we've seen:
  - Learnability (ANNs, Bayesian pLoT, MDL)
  - Complexity (logical, Kolmogorov)
  - Cultural evolution (iterated learning)
  - Combination of pressures (tradeoff analysis, IL+communication)
- Many exciting avenues for future work!