Vagueness & Polysemy 4

Why are languages vague and polysemous?

Peter Sutton
peter.sutton@hhu.de
Heinrich Heine University Düsseldorf

Winter School on Semantics in Barcelona

January, 2018

Main questions for the course

Main Questions

- Why is vagueness a challenge for a formal semantics for natural languages?
  - Day 1: Background
    - Properties of Classical First Order Logic (CFOL) & connection to modern semantic theory
    - Diagnosing Vagueness
- What are the main reactions to this challenge?
  - Days 1-3: Responses to vagueness
    - fuzziness (day 1)
    - S'valuationism & polysemy, contextualism, degrees
    - Probabilistic approaches
- How does vagueness connect to polysemy?
  - Part of Day 2: Is vagueness a kind of polysemy?
  - Day 4: The common origins of vagueness and polysemy?

Plan for Day 4

1. Two questions about vagueness
2. Introduction to Information Theory
3. Information Theoretic Models for the origins of Ambiguity and Polysemy
   - Ferrer i Cancho and Solé (2003): Polysemy arises from a competition between speaker effort and hearer effort
   - Piantadosi et al. (2011)- No polysemy would be redundant
4. An Information Theoretic Model for the origin of Vagueness
   - The probabilistic iterated learning model
   - Results
   - Discussion and Conclusions

So far...

Semantic (or Epistemic) theories of vagueness

- What is the representation of vagueness?
  - Degrees of truth
  - Underspecification: filling in forming gaps or gluts
  - Context shifts
  - Uncertainty/Ignorance
- What logical/semantic properties do these approaches have?

A different question:

- What is the origin of vagueness?/Why do natural languages have vague expressions?

Eliminating Vagueness and Polysemy

- Early days of analytic philosophy: Vagueness, ambiguity and polysemy are imperfections that should be eradicated in the development of a language for science and the investigation of concepts

  "Science is perpetually trying to substitute more precise beliefs for vague ones; this makes it harder for a scientific proposition to be true than for the vague beliefs of uneducated persons to be true, but it makes scientific truth better worth having if it can be obtained." (Russell 1923, 91)

  "So long as the reference remains the same, such variations of sense may be tolerated, although they are to be avoided in the theoretical structure of a demonstrative science and ought not to occur in a perfect language." (Frege 1948/1892, p. 210n)

Eliminating Vagueness and Polysemy

- Even now: Vagueness most often framed as a problem for (classical and non-classical) logic, semantics etc.
  - As a failure e.g. to fully grasp the complexities of language usage (Williamson 1994)
  - As underspecification of/uncertainty about precise predicates (S'valuationism)
  - As a failure to fully grasp the contextual standards at play (Degree semantics, PLK)

Embracing Vagueness and Polysemy

- Today: Connection to a viewpoint that has roots in Zipf's work (Zipf 1935, 1949)
  - We find NL phenomena such as vagueness, ambiguity, polysemy, crosslinguistic variation in syntax and semantics
  - These phenomena can be seen as the byproduct of communication and learning
    - ★ E.g. balancing ease of learning with ease of (successful) communication
- Vagueness and polysemy should not be eliminated
  - They are crucial feature of communication systems that balance the goals and needs of:
    - ★ Speakers and hearers
    - ★ Competent agents and learners
  - We learn and communicate in conditions of uncertainty.

Zipf and the principle of least effort

- Zipf (1935, 1949): Ambiguity arises from two competing pressures on learning and communication
  - Ideal language for a speaker: a single lexical item for every sense (a 'Ba' language)
  - Ideal language for hearer: for every sense, a single lexical item
- Main idea:
  - Languages balance these two pressures: *The principle of least effort*
- This principle was held to explain Zipf's Law:
  - The frequency of a lexical item is proportional its frequency rank
  - List of words: $w_1, w_2, ..., w_n$
  - Frequency ranking r of words $R(w_i) = r$
  - Frequency/Probability of w_i is proportional to i
  - $P(w_i) \propto R(w_i)^{-\alpha}$ (where $\alpha \approx 1$)

Zipf's Law: Example

Rank	*word*	$P(word)$	
1	*a*	$P(a) \propto 1^{-1}$	$P(a) \propto 1$
2	*b*	$P(b) \propto 2^{-1}$	$P(b) \propto 0.5$
3	*c*	$P(c) \propto 3^{-1}$	$P(c) \propto 0.\overline{3}$
4	*d*	$P(d) \propto 4^{-1}$	$P(d) \propto 0.25$
5	*e*	$P(e) \propto 5^{-1}$	$P(e) \propto 0.2$
6	*f*	$P(f) \propto 6^{-1}$	$P(f) \propto 0.1\overline{6}$

- Informal explanation: The principle of least effort can explain these patterns
  - More frequently used (accessible forms) will be more heavily weighted towards use

Zipf onwards

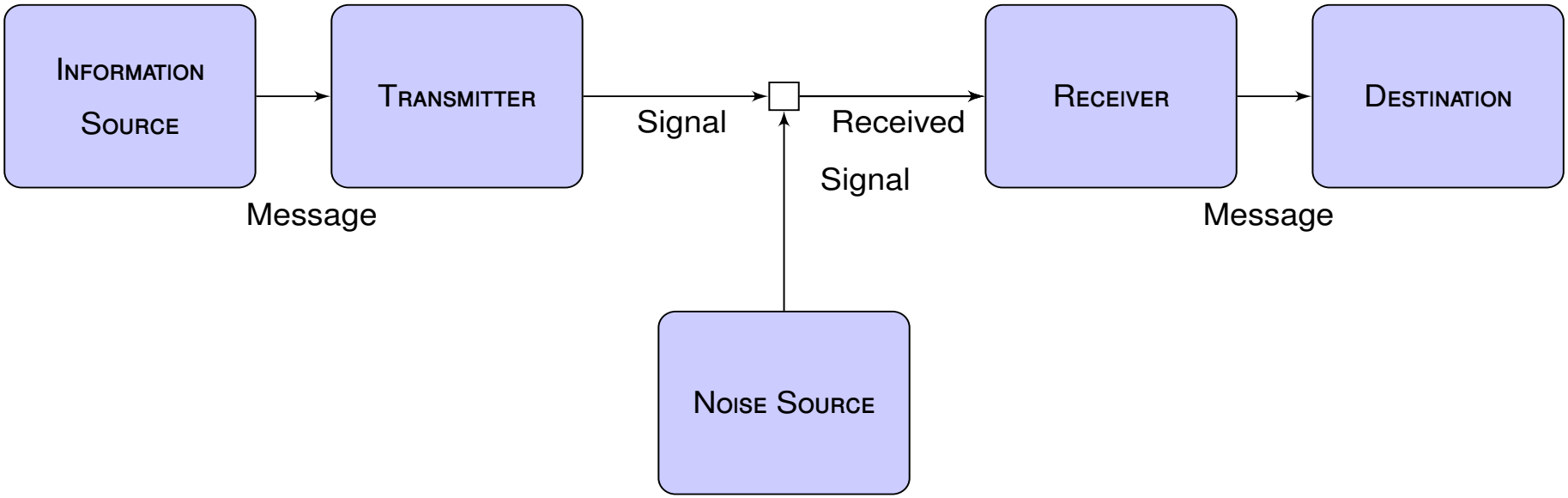
- Developments of Zipfian ideas formalised in terms of information theoretic constraints (Ferrer i Cancho and Solé 2003; Piantadosi et al. 2011)
  - Ferrer i Cancho and Solé (2003): Zipf's law can be explained in terms of balancing hearer and speaker effort (measured information theoretically)
  - Piantadosi et al. (2011): Speaker/hearer effort can be understood information theoretically in terms of processing costs
  - Ambiguity and polysemy are features of a non-redundant system that makes use of contextual information.

Zipf onwards II

- Other linguistic phenomena also investigated in terms of competing information theoretic pressures:
  - vagueness evolves in language when boundedly rational agents repeatedly engage in cooperative signalling (Franke et al. 2010).
  - languages optimize information density (Levy and Jaeger 2007; Jaeger 2010)
  - general efficiency principles can explain a wide range of crosslinguistic syntactic patterns (Hawkins 2014)
  - conflict between information-theoretic pressures can explain crosslinguistic count/mass variation (Sutton and Filip 2017);

Information Theory

- Developed by Claude Shannon (1948)
- Originally for telecommunications
  - How can messages be efficiently transmitted over a noisy channel?



"We have to be clear about the rather strange way in which, in this theory, the word 'information' is used; for it has a special sense which, among other things, must not be confused with meaning." (Weaver 1979, p.30)

"We are concerned not with the meaning of individual messages but with the whole statistical nature of the information source" (Weaver 1979, p. 31)

Quantifying Information

Some questions:

- How much information is contained at a source?
- How much is encoded by the signal?
- How much information is decoded?

One common unit for measuring information is the *binary unit* (bit).

$U = \{$  ,  $\}$

This situation contains 1 bit of information.

One binary decision to determine single outcome (e.g. heads or tails)

$U = \{\langle$  $\rangle, \langle$  $\rangle, \langle$  $\rangle, \langle$  $\rangle, \}$

This situation contains 2 bits of information.

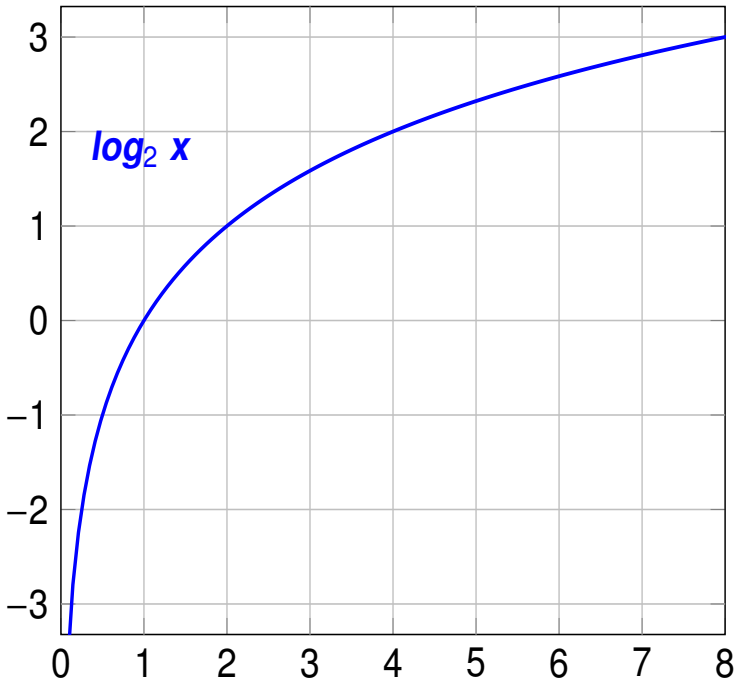
Two binary decisions must be made to determine a single outcome. E.g.

- Decision 1: $\{\langle H, H\rangle, \langle H, T\rangle\}$ or $\{\langle T, H\rangle, \langle T, T\rangle\}$
- Decision 2: $\langle H, H\rangle$ or $\langle H, T\rangle$

How much information does a situation with only one possible outcome contain?

Bits and Logs

This relationship is captured by a base 2 logarithm (for bits)



y := *Information contained in a situation.*
x := *Number of possible alternative outcomes in s.*

$$y = \log_2 x$$
$$2^y = x$$

However, other alternative bases can be used. E.g.:

- Euler's number $e \approx 2.718$
- $\log_e = ln$ (natural logarithm)

Informational content of messages

On a noise- and equivocation-free channel, a signal *s* can convey as a message *m* at most the amount of information in the source situation.

Information of the received message, given the signal $\mathbb{I}_s(r)$ can be less than at the source situation σ
$\mathbb{I}_s(m) = \mathbb{I}(\sigma)$ – *equivocation*

equivocation: information for the number of possibilities left, given the signal *r*
or, roughly, information lost in transmission

Example: Informational content of messages

Source situation: Tossing three coins

$$\{\langle H, H, H\rangle, \langle H, H, T\rangle, \langle H, T, H\rangle, \langle H, T, T\rangle, \langle T, H, H\rangle, \langle T, T, H\rangle, \langle T, H, T\rangle, \langle T, T, T\rangle\}$$

Maximum information contained at source:

$$\mathbb{I}(\sigma) = \log_2(8) = 3 \text{ bits}$$

Signal	No. of possibilities left, given signal	Information conveyed
$\langle ?, ?, ? \rangle$	8	$\log_2(8) - \log_2(8) = 0$
$\langle H, ?, ? \rangle$	4	$\log_2(8) - \log_2(4) = 1$
$\langle H, T, ? \rangle$	2	$\log_2(8) - \log_2(2) = 2$
$\langle H, T, H \rangle$	1	$\log_2(8) - \log_2(1) = 3$

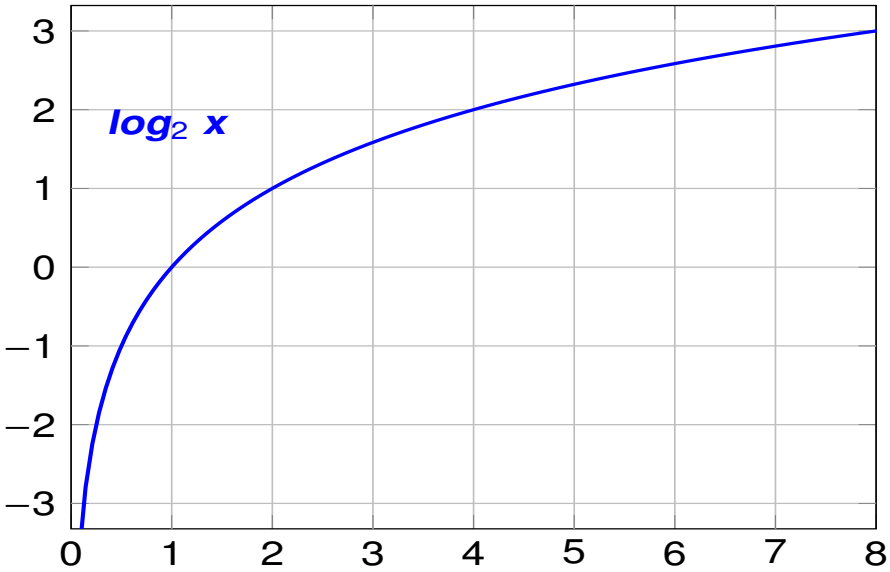
More complex situations

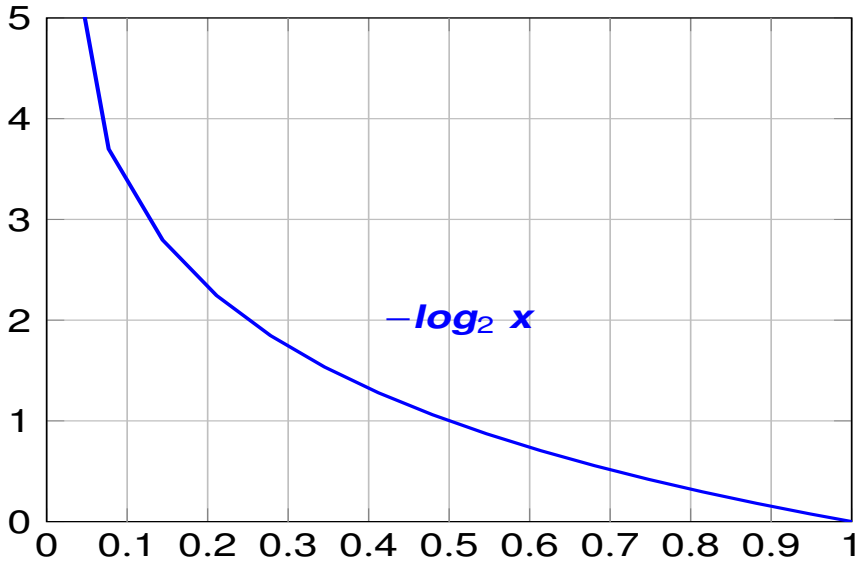
Up-to-now: A special case

- All the possibilities had an equal chance (coin tosses etc.)
- We can extend method to cases with different prior probabilities.

Surprisal: Intuitive idea

- More information conveyed by learning that something improbable is the case than learning that something probable is the case.
- An outcome with a .5 probability = 1 bit, with .25 probability = 2 bits





$$\mathbb{I}(s_n) = -\log_2 P(s_n)$$

Note on Logs and Probabilities

To generalize from earlier. Logs express a relationship in terms of powers of bases:

$$\log_z x = y$$

$$z^y = x$$

For us:

- x is a probability value.
- y is a negative surprisal value ($-y$ is the surprisal value)

This means that we can move backwards and forwards between probabilities and surprisal values:

$$\log_2 0.25 = -2$$

$$2^{-2} = \frac{1}{2^2} = 0.25$$

Example: Surprisal

Suppose the following situation w:

Weather (i):	Rain	Sun	Snow	Hail
$P(w_i)$:	0.5	0.25	0.125	0.125

We can calculate the surprisal for each outcome ($-log(P(w))$):

$\mathbb{I}(w_i)$	1	2	3	3

- In other words, were we to find out it was snowing/hailing, we would have received more information than to find out it is sunny/raining

Entropy

Entropy is a measure of:

- The average amount of information in a situation
- Or: the average number of binary decisions needed to find out what the situation is like

Entropy for a situation $H(\sigma)$ is calculated as the sum of the surprisals of each outcome proportional to their probability:

$$\mathcal{H}(\sigma) = \sum_i P(\sigma_i) \times \log P(\sigma_i)$$

Entropy: Example

Weather (i):	Rain	Sun	Snow	Hail
$P(w_i)$:	0.5	0.25	0.125	0.125

What is the entropy for w?

$$\begin{aligned}\mathcal{H}(w) &= \sum_i P(w_i) \times \mathbb{I}(w_i) \\ &= (0.5 \times 1) + (0.25 \times 2) + (0.125 \times 3) + (0.125 \times 3) \\ &= 0.5 + 0.5 + +0.375 + 0.375 \\ &= 1.75\end{aligned}$$

What does this mean?

- The average number of binary decisions it would take to establish what the actual weather is is 1.75

Surprisal values for signals in conditions of uncertainty

- So far, we have mostly talked only about surprisal of outcomes in a situation/entropy of a situation
- For communication, we have a system of signals and some possible messages.

Surprisal for a message m given a signal s:

$$\mathbb{I}(m|s) = -\log_2 P(m|s)$$

- A signal can underspecify the message
  - Some messages would contain more information than others (if they were the right message)

Entropy for conditional values

The average amount of information conveyed by a specific signal s_j:

$$\mathcal{H}(m|s_j) = -\sum_i P(m_i|s_j) \times \log_2 P(m_i|s_j)$$

- A signal can underspecify the message
  - $\mathcal{H}(m|s_j)$ is a measure on how much information we need to determine a single message
  - Higher entropy = more indeterminate signal

Example: Surprisal values and entropy for signals and messages

Example we had before:

Weather (w_i):	Rain	Sun	Snow	Hail
$P(w)$:	0.5	0.25	0.125	0.125
$\mathcal{I}(w_i)$	1	2	3	3

Signal s_1: "There will be s#?#?# today"

- Given the possibilities, s_1 tells us something, but not enough to discern the full message
  - Weather begins with 's', but we can't tell if it's sun or snow
  - Priors for sun and snow can help weight the conditional probability
  - Can also calculate the surprisal values, given the signal

$P(w_i\|s_1)$	0	2/3	1/3	0
$\log_2 P(w_i\|s_1)$	$-\infty$	0.58	-1.58	$-\infty$

Example: Surprisal values and entropy for signals and messages

$P(w_i\|s_1)$	0	2/3	1/3	0
$\log_2 P(w_i\|s_1)$	$-\infty$	-0.58	-1.58	$-\infty$

The entropy for the message, give the signal is:

$$\begin{aligned}\mathcal{H}(w_i|s_1) &= -\sum_i P((w_i|s_1)) \times \log_2 P(w_i|s_1) \\ &= (0 \times -\infty) + (2/3 \times -0.58) + (1/3 \times -1.58) + (0 \times -\infty) \\ &= 0.92\end{aligned}$$

- So s_1 equivocates between messages, and some more information is needed to determine the full message (on average 0.92 bits)

Adding further explanation to Zipf

Ferrer i Cancho and Solé (2003) (simplified)

- Ambiguity/polysemy can arise as a result of balancing speaker and hearer effort
- Speaker effort = signal entropy
- Hearer effort = entropy of message, given signal

Signal entropy as a measure of Speaker Effort

Two different 'languages':

1. $l_1 : \mathcal{I}(s_1) = \{m_1, m_2\}$
2. $l_2 : \mathcal{I}(s_1) = m_1, \mathcal{I}(s_2) = m_2$

Speaker effort = effort in selecting a signal s_i from the vocabulary given an message m_j

- l_1 requires minimum effort. No matter what is being referred to, the same word is used
- l_2 requires more effort involved. Speaker must select the right message for the intended referent.

Entropy as a measure of Speaker Effort II

Entropy can be used as a measure on this:

- Assume $P(m_1) = P(m_2) = 0.5$
- Then $P(s_i) = \sum_j \mathcal{I}(s_i) = m_j$
- Speaker entropy: $\mathcal{H}(s) = \sum_i p(s_i) \times \log_2 p(s_i)$

$$
\begin{array}{lcccc}
 & I_1 & & & \\
P(s_1) & = & 0.5 + 0.5 & = & 1 \\
H(s) & = & 1 \times \log_2(1) & = & 0
\end{array}
$$

$$
\begin{array}{lcccc}
I_2 & & & & \\
P(s_1) & = & 0.5 & & \\
P(s_2) & = & 0.5 & & \\
H(s) & = & 0.5 \times \log_2(0.5) + 0.5 \times \log_2(0.5) & = & 1
\end{array}
$$

Signal entropy as a measure of Speaker Effort III

Far more signals (even with synonymy) creates more effort for the speaker

- 100 signals each with probability 0.01
  - Speaker effort = $100 \times (0.01 \times \log 0.01) = 6.64$
- 100 signals each with probability 0.001
- Speaker effort = 9.97

Conditioned Entropy as a measure of Hearer Effort

- Intuitive idea: The effort for the hearer is increased when the signal under-specifies the message

$$\mathcal{H}(M|s_i) = \sum_j P(m_j|s_1) \log P(m_j|s_1)$$

- One extreme: Maximum equivocation
  - Every message is (potentially) expressed by a single single
- The other extreme: No equivocation
  - Every signal determines only one message

- $I_1 : \mathcal{I}(s_1) = \{m_1, m_2\}$
- $I_2 : \mathcal{I}(s_1) = m_1, \mathcal{I}(s_2) = m_2$

Conditioned Entropy as a measure of Hearer Effort II

- $l_1 : \mathcal{I}(s_1) = \{m_1, m_2\}$
- $l_2 : \mathcal{I}(s_1) = m_1, \mathcal{I}(s_2) = m_2$

$$
\begin{array}{lcl}
l_1 & & \\
P(m_1|s_1) & = & 0.5 \\
P(m_2|s_1) & = & 0.5 \\
H(M|s_1) & = & 0.5 \times \log_2(0.5) + 0.5 \times \log_2(0.5) \quad = \quad 1
\end{array}
$$

$$
\begin{array}{lcl}
l_2 & & \\
P(m_1|s_1) & = & 1 \\
P(m_2|s_1) & = & 0 \\
P(m_1|s_2) & = & 0 \\
P(m_2|s_2) & = & 1 \\
H(M|s_1) & = & 1 \times \log_2(1) + 0 \times \log_2(0) \quad = \quad 0 \\
H(M|s_2) & = & 1 \times \log_2(1) + 0 \times \log_2(0) \quad = \quad 0 \\
H(M|S) & = & (0.5 \times 0) + (0.5 \times 0) \quad = \quad 0
\end{array}
$$

Hearer Effort versus Speaker Effort

- The two pressures work in opposition

Ferrer i Cancho and Solé (2003) propose a function with a weighting variable $\lambda \in [0, 1]$:

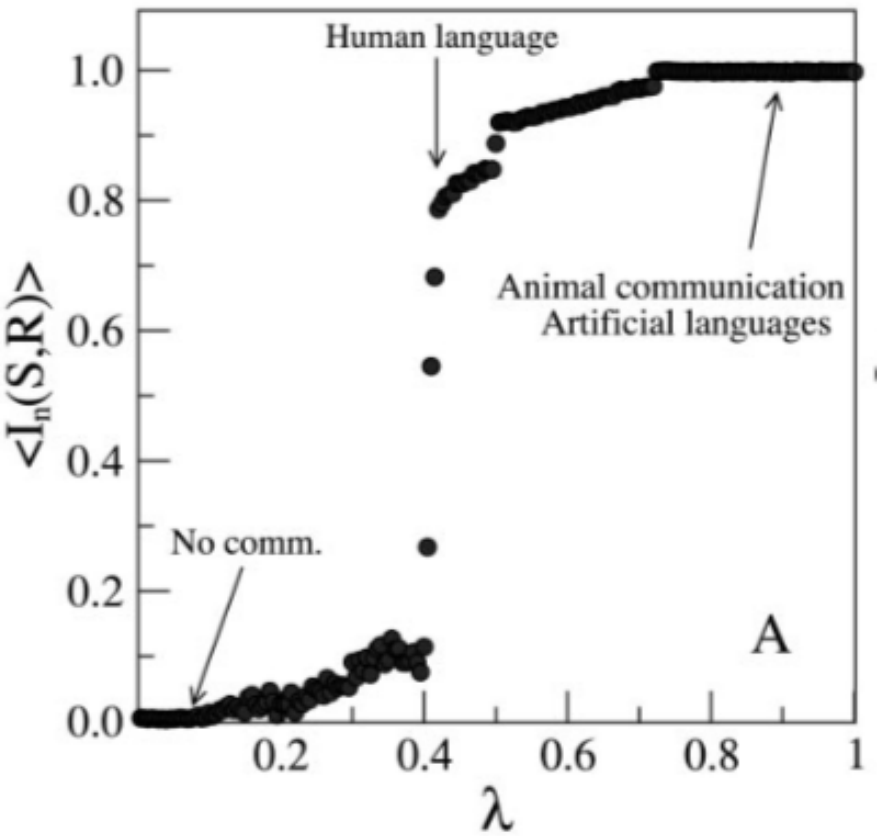
$$\Omega(\lambda) = \lambda(\mathcal{H}(M|S)) + (1 - \lambda)(\mathcal{H}(S))$$

Depending on the setting for λ, this favours either reducing speaker effort or reducing hearer effort

Some of the results:

An evolutionary model of language adjustments, given a setting for speaker/hearer effort weighting:

- x-axis: weighting variable
- y-axis: measure of mutual information



- When speaker effort is minimised, there is no communication
- When hearer effort is minimised there is perfect communication, but speaker cost that is too high
- At a mid point of around 0.4, there is a sharp transmission to a polysemous and synonymous language that still has high levels of mutual information

Clarity and Ease

- Zipf's intuition: ambiguity and polysemy arise from a rational process of communication is correct
  - Speaker effort not really the right conceptualisation though

Instead: *clarity* and *ease*

- Clear signal: intended meaning can be recovered with high probability
- Easy signal: efficiently produced, communicated and processed
  - E.g., easy to process includes: word that "are likely short, frequent, and phonotactically well- formed."

"Clarity and ease are opposed because there are a limited number of 'easy' signals which can be used. This means that in order to assign meanings unambiguously or clearly, one must also use words which are more difficult." Piantadosi et al. (2011, p. 281)

Ambiguity and Polysemy reduce redundancy

Like Ferrer i Cancho and Solé (2003), Piantadosi et al. (2011) use entropy measures, e.g, for meaning space entropy

$$\mathcal{H}(M) = -\sum_i P(m_i) \log P(m_i)$$

However, in context, the probability distribution over meanings may change. I.e.:

$P(m_i)$ is not always equal to $P(m_i|c_j)$

So, the entropy for the meaning space should be *conditional entropy*

$$\mathcal{H}(M|C) = -\sum_j P(c_j) \sum_i P(m_i|c_j) \log P(m_i|c_j)$$

Furthermore, provably:

$$\mathcal{H}(M) > \mathcal{H}(M|C)$$

Ambiguity and Polysemy reduce redundancy II

$$\mathcal{H}(M) > \mathcal{H}(M|C)$$

What does this mean?

- The average amount of information to discern a meaning is decreased by context.

Shannon's coding theorem:

- $\mathcal{H}(X)$ bits of uncertainty information cannot be disambiguated with fewer than $\mathcal{H}(X)$ bits of information without error

The consequence:

- If languages are efficient, they will disambiguate in context.
- Due to the inequality $\mathcal{H}(M) > \mathcal{H}(M|C)$, this will never disambiguate meaning out of context.
- Efficient languages will display ambiguity and polysemy

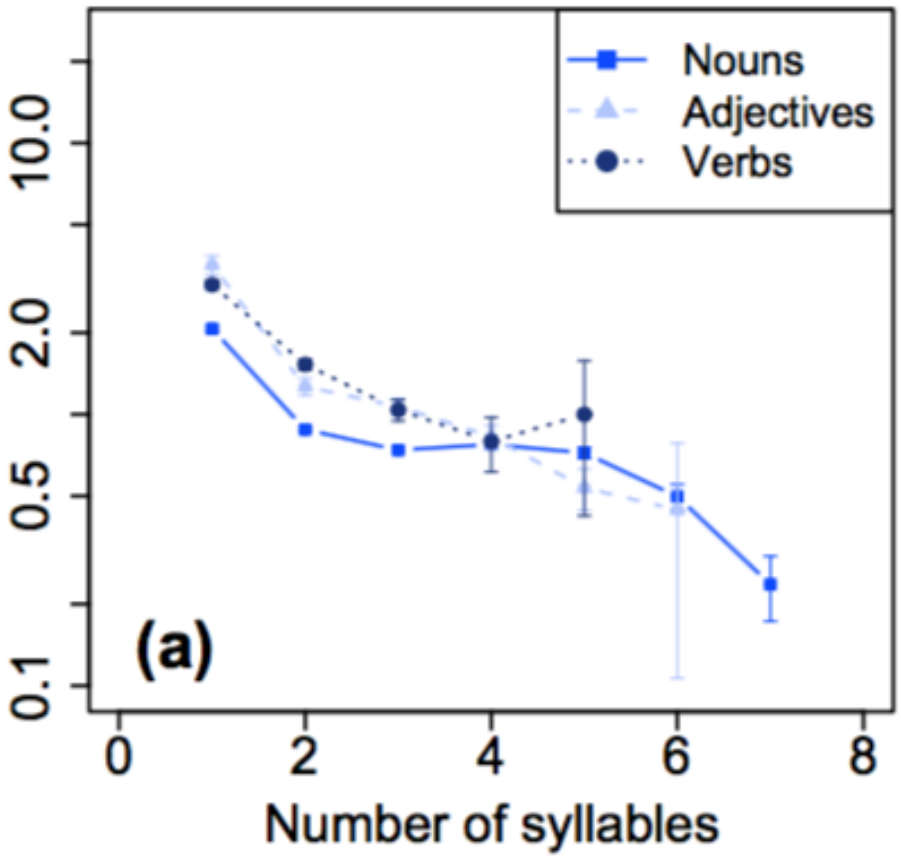
Ease and polysemy

Suppose s_1 is easier than s_2 (e.g., it is shorter, more frequent, or phonotactically simpler etc.)

- If s_1 conveys m_1 and s_2 conveys m_2, an easier language (for the speaker) is one in which s_1 conveys both m_1 and m_2
  - However, this comes at more cost for the hearer ....
  - ... unless m_1 and m_2 are not probable in the same contexts
- So, easier languages have polysemous/ambiguous items, but this need not increase hearer costs, provided context aids sense disambiguation.

Corpus Study Results on Polysemy

- y axis: Raw number of additional senses a word has



- There is a correlation between ease (in terms of fewer syllables) and number of senses
  - Similar results for other metrics of ease

Interim summary

- Polysemy arguably arises as a result of conflicts between speaker effort and hearer effort (Ferrer i Cancho and Solé 2003)
- Polysemy arguably arises because ease for the hearer (polysemy), needn't create unclarity for the hearer if meaning can be disambiguated in context (Piantadosi et al. 2011)

Parallels with polysemy

- Some amount of vagueness can be the sign of a balance between speaker and hearer goals
- Vagueness is not a negative feature of language if messages can be sufficiently disambiguated in context.

Problem:

- Why would this not just generate polysemy?
- What brings about vagueness?

Hypothesis 1: Learning

- Learner-hearers only get a sample of the full language. The learning bottle neck leads to vagueness

Hypothesis 2: Noise and equivocation

- Models for polysemy assume that noise and equivocation can be overcome e.g., by context
- Some noise/equivocation is not removable by context, but forced upon us by, e.g., the granularity of our sense perception

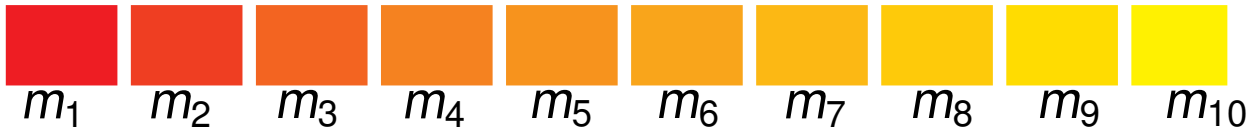
Iterated Learning Models (ILMs)

- Pioneered by Kirby and Hurford (Brighton and Kirby 2001; Kirby and Hurford 2002; Kirby 2007, amongst many others)
- Basic pattern: Generations of agents $A_1, ..., A_n$. A_1 has a language, A_i learns from A_{i-1}
- Kirby and Hurford et al. looking at the SEMANTIC BOTTLENECK and how properties of language emerge from pressures of LEARNABILITY and STABILITY
  - the semantic bottleneck: can the language be learnt without experiencing every expression-denotation pairing
  - learnability: can the language be learnt from sparse data
  - stability: not too much meaning change across generations
- Some very interesting results. E.g.:
  - compositionality improves language stability
  - grammatical expressions are highly stable
  - less common peripheral predicates vary
- Some limitations
  - learning was pretty trivial (total learning from one instance)
  - no noise

The plan

- Follow up on some informal ideas in Sutton 2013
  - Vagueness arises from competing pressures from learning and communication
  - (focus there on context-sensitivity)
  - Today: Formalize an abstract version of this in terms of communication over a noisy channel
- Adopt the ILM paradigm (to model the bottleneck), but enhance it with a learning model that incorporates noisy channels
- Follow in the Zipfian paradigm to identify vagueness as arising from:
  - Learning pressures: the semantic bottleneck
  - Communication pressures: signalling over a noisy channel

- Simulation coded in Matlab
- A set of messages: $\mathbb{M} = \{m_1, m_2, m_3, m_4, m_5, m_6, m_7, m_8, m_9, m_{10}\}$
- A total order
  - Intuitively, something like a set of colours that grade from one to another



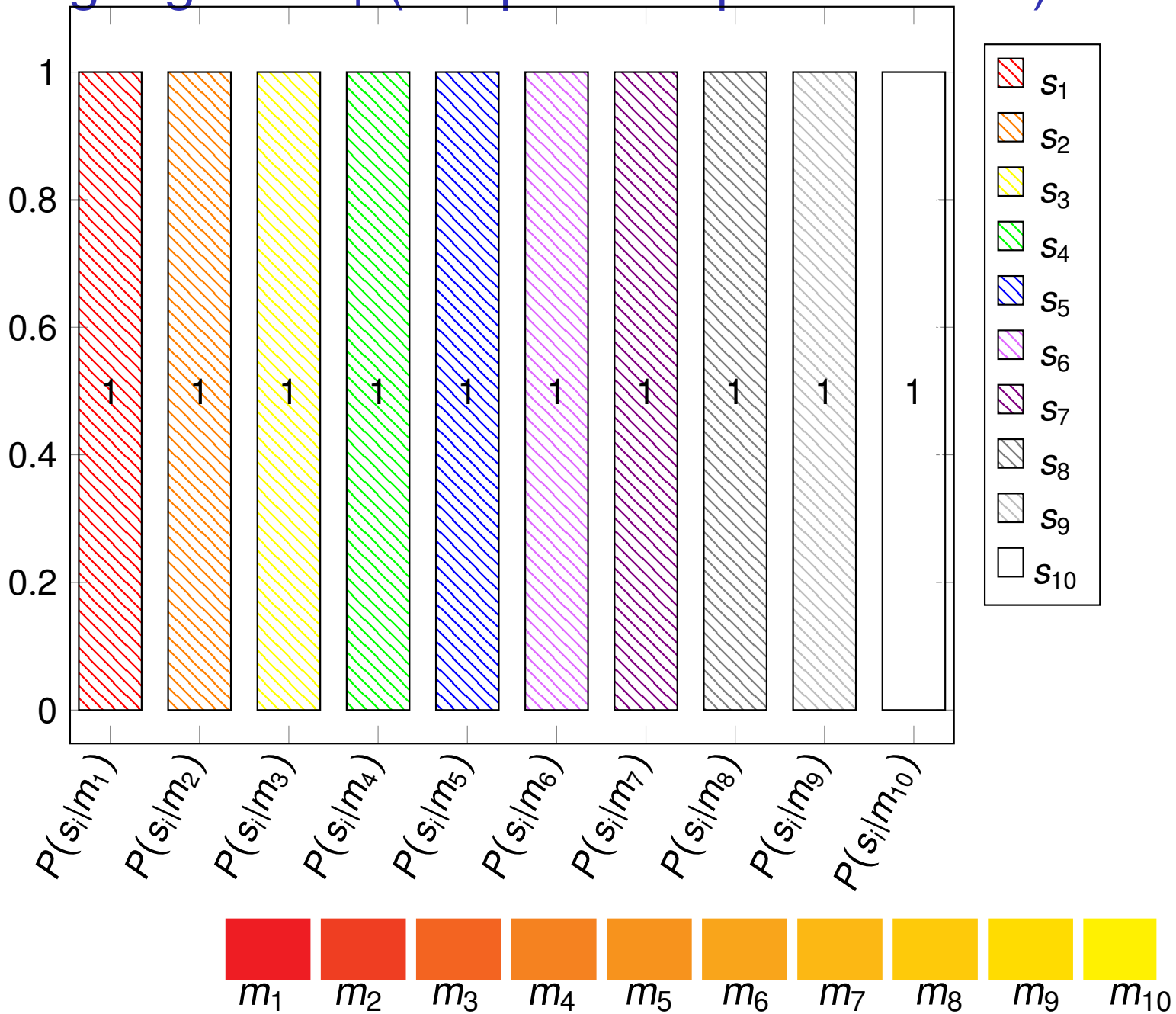
- A distance function e.g., for some real number δ:
  $\mathcal{D}(m_i, m_j) = \delta \times abs(i - j)$
- A set of signals: $\mathbb{S} = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}$
- A set of n generations agents $\mathbb{A} = \{A_1, ..., A_n\}$
- A set of languages (one for each agent) $\mathbb{L} = \{L_1, ..., L_n\}$
  - Languages are characterised as probability distributions $P(s_i|m_j)$

Language of A_1

Assume that the first agent has a categorical language (not sneaking vagueness in)

$P(s_i \mid m_j)$	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
m_1	1	0	0	0	0	0	0	0	0	0
m_2	0	1	0	0	0	0	0	0	0	0
m_3	0	0	1	0	0	0	0	0	0	0
m_4	0	0	0	1	0	0	0	0	0	0
m_5	0	0	0	0	1	0	0	0	0	0
m_6	0	0	0	0	0	1	0	0	0	0
m_7	0	0	0	0	0	0	1	0	0	0
m_8	0	0	0	0	0	0	0	1	0	0
m_9	0	0	0	0	0	0	0	0	1	0
m_{10}	0	0	0	0	0	0	0	0	0	1

Language of A_1 (Graphical representation)



A_1 provides learning data for A_2

- σ is a parameter governing the number of messages (governed by situations to be described)
- Then draw a random sample of σ messages.
- A_1 then produces a signal for each message in the sample in line with her language.
  - In a noise-free model, this is the learning data for A_2 (D_{A_2})
- For example (for a space of three messages):

$$
\begin{array}{lcl}
L_{A_1} & = & \begin{array}{c|ccc} P(s_i|m_j) & s_1 & s_2 & s_3 \\ m_1 & 1 & 0 & 0 \\ m_2 & 0 & 1 & 0 \\ m_3 & 0 & 0 & 1 \end{array} \\
M & = & \{m_1, m_3, m_1, m_2, m_1, m_3\} \\
D_{A_2} & = & \{\langle m_1, s_1\rangle, \langle m_3, s_3\rangle, \langle m_1, s_1\rangle, \langle m_2, s_2\rangle, \langle m_1, s_2\rangle, \langle m_3, s_3\rangle\}
\end{array}
$$

Here, no bottleneck: A_2 has witnessed a signal for every message
Also no noise or equivocation: The signals from A_1 do not add or lose any information

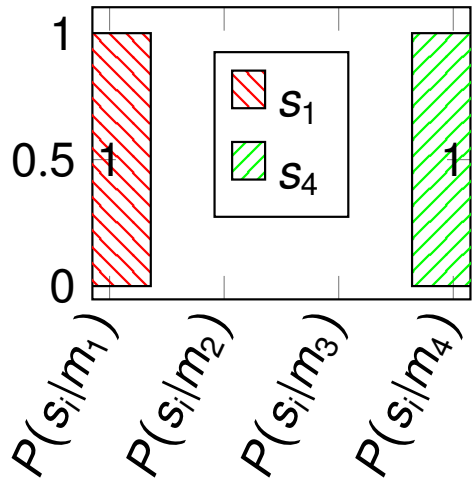
- Result: A_2 learns the same language as A_1

The semantic learning bottleneck

- A_{n+1} does not get a description for every situation from A_n
  - For example (restricted to four messages):

$$
\begin{array}{lcl}
 & & \begin{array}{c|cccc} P(s_i|m_j) & s_1 & s_2 & s_3 & s_4 \\ m_1 & 1 & 0 & 0 & 0 \\ m_2 & 0 & 1 & 0 & 0 \\ m_3 & 0 & 0 & 1 & 0 \\ m_4 & 0 & 0 & 0 & 1 \end{array} \\
L_{A_1} & = & \\
M & = & \{m_1, m_1, m_4, m_4\} \\
D_{A_2} & = & \{\langle m_1, s_1\rangle, \langle m_1, s_1\rangle, \langle m_4, s_4\rangle, \langle m_4, s_4\rangle\}
\end{array}
$$

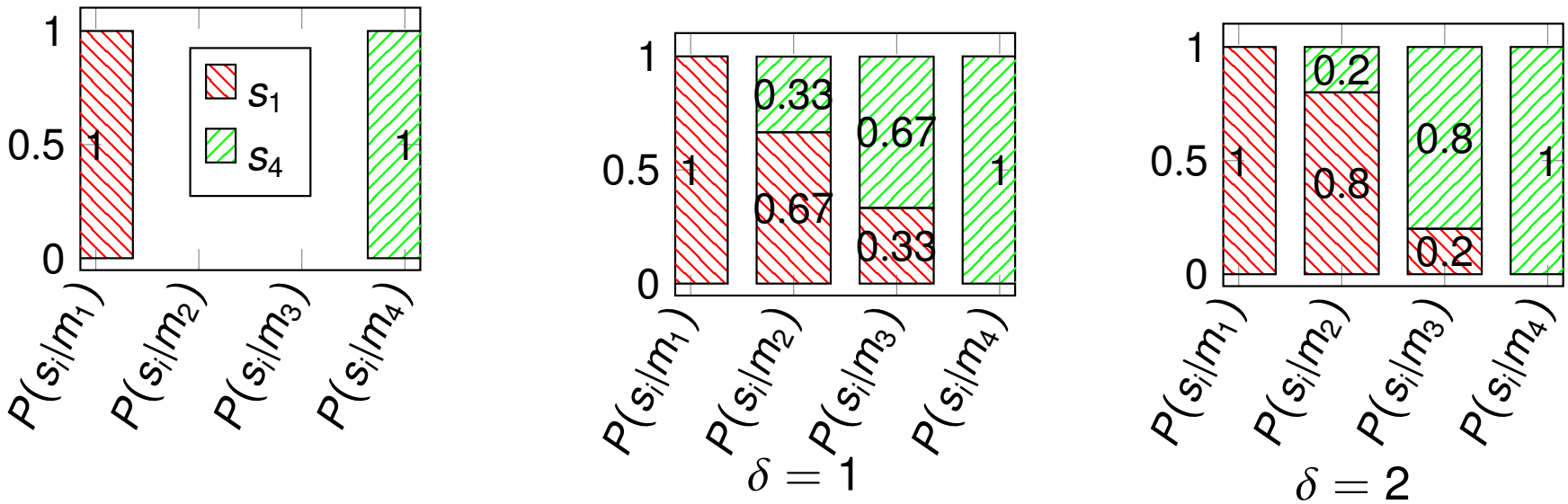
A_2 has insufficient data to know how to communicate m_2 and m_3.
Can only directly infer a partial language:



Reasoning given the learning bottleneck

A_{n+1} may have to infer how to communicate an unwitnessed message:

- Informally: e.g. A_2 must look 'up' and 'down' along the series of witnessed messages and infer how certain she should be about using any particular predicate
  - This should be a function of 'distance' between witnessed and unwitnessed messages ($D(m_i, m_j) = \delta \times abs(m_i, m_j)$);
  - Predicates should be convex e.g. not $\{\langle m_1, s_1\rangle, \langle m_2, s_2\rangle, \langle m_3, s_1\rangle\}$ (Gärdenfors 2000)
- E.g. $P(s_1|m_2) \propto \log_{-1}(\log(P(s_1|m_1)) - D(m_2, m_1))$



Example of encoding a noisy channel

A_n tries to communicate $\langle m_1, s_1 \rangle$ and $\langle m_2, s_2 \rangle$.
A_{n+1} receives $D = \{\langle m_1, s_1 \rangle, \langle m_2, s_2 \rangle\}$, but also $D' = \{\langle m_2, s_1 \rangle, \langle m_1, s_2 \rangle\}$
Parameter $\mathcal{U} \in [0, 1]$ determines how well A_{n+1} can determine the 'true' message.

$$P_{A_{n+1}}(s_j|m_j) = \frac{P_D(m_j, s_j) + \mathcal{U} \times P_{D'}(m_j, s_j)}{P_D(m_j) + \mathcal{U} \times P_{D'}(m_j)}$$

$U = 0$			$U = 0.5$			$U = 1.0$		
$P(s_i\|m_j)$	s_1	s_2	$P(s_i\|m_j)$	s_1	s_2	$P(s_i\|m_j)$	s_1	s_2
m_1	1	0	m_1	$0.\overline{6}$	$0.\overline{3}$	m_1	0.5	0.5
m_2	0	1	m_2	$0.\overline{3}$	$0.\overline{6}$	m_2	0.5	0.5

If $\mathcal{U} = 0$: No perplexity
If $\mathcal{U} = 1$: Maximum perplexity

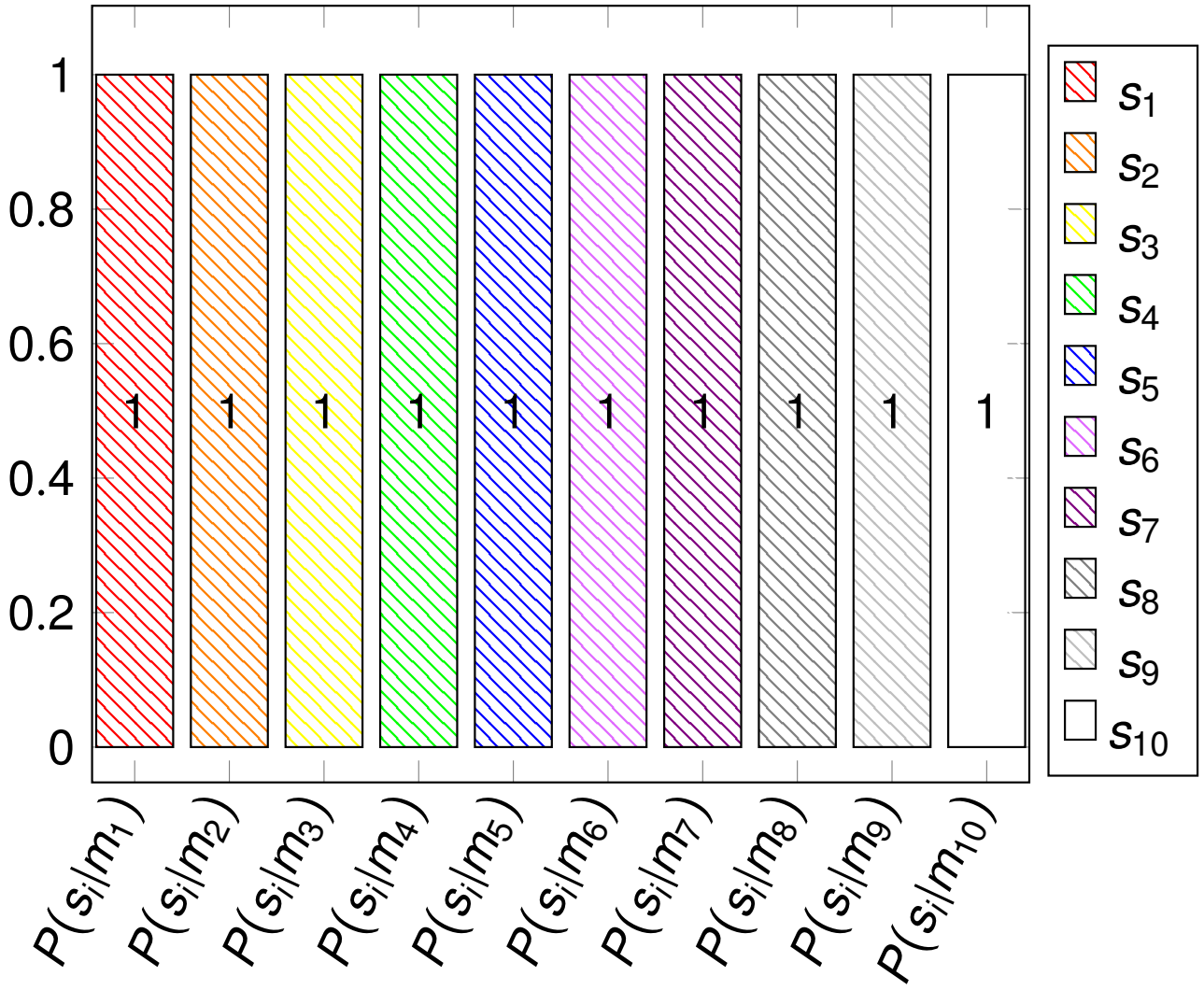
Another parameter, $\mathcal{N}$, the NOISE PARAMETER determines how probable it is that A_{n+1} receives noisy information (on top of the 'correct' information)

Summary

- Parameters kept fixed:
  - Number of messages (=10)
  - Starting language for A_1 (categorical, one signal/predicate for each message)
  - Number of generations = 100
- Parameters controlling learnability and the bottleneck:
  - Number of messages described by A_i to A_{i+1} (controls size of bottleneck)
  - Distance function parameter δ if there are bottleneck gaps between two signals, how sharply does the learner's representation approach 0.5-0.5 in the gap.
- Parameters controlling communication (information transmission):
  - $\mathcal{N} \in [0, 1]$: average proportion of noisy signals (from 0 – 100%)
  - $\mathcal{U} \in [0, 1]$: ability of A_{i+1} to discriminate the 'true' message from the noise (0= minimum perplexity, 1=maximum perplexity).

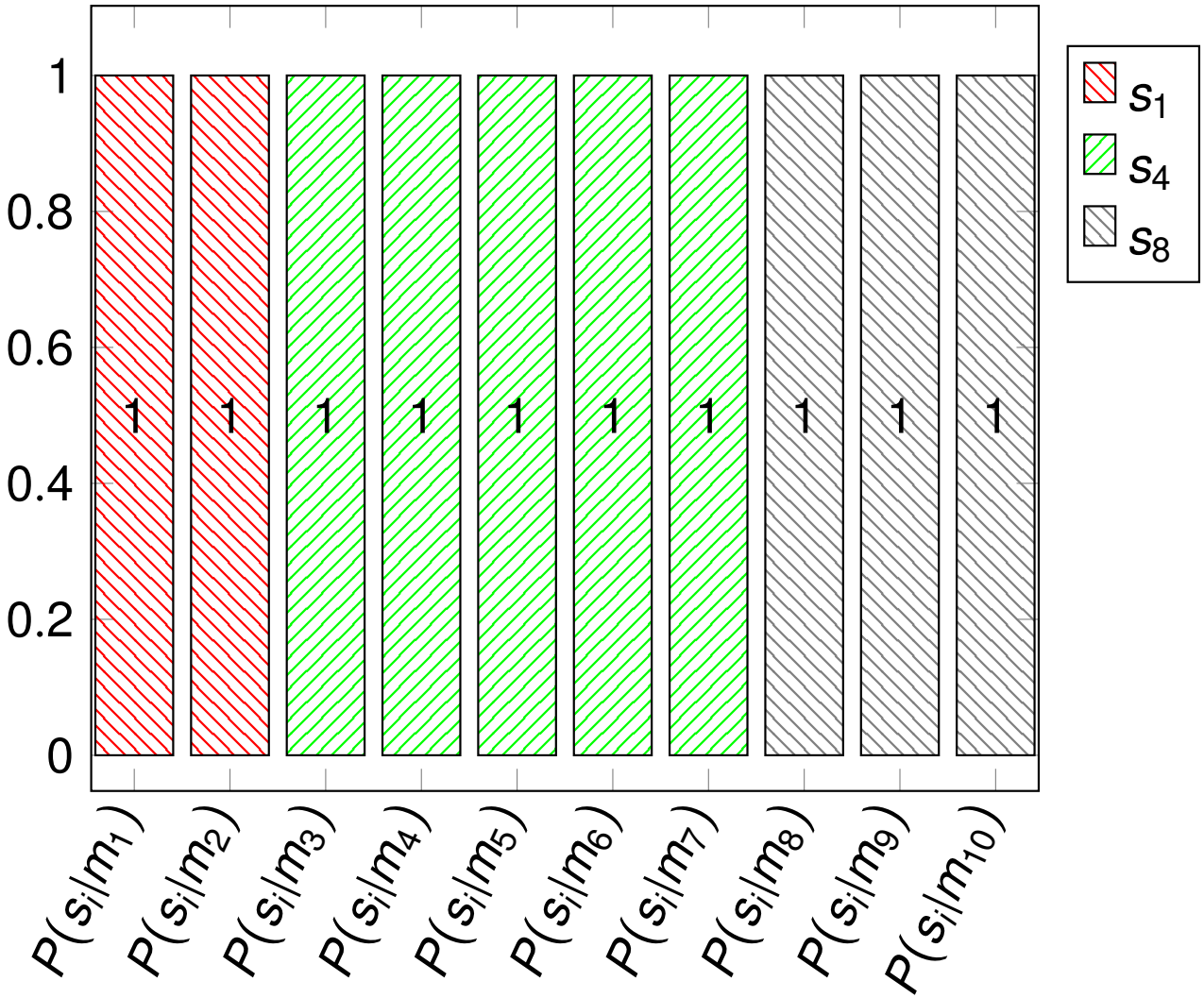
No bottleneck, no noise

- Sample size = 500, Noise level = 0
- Unsurprisingly, $L_{A_1} = L_{A_{100}}$



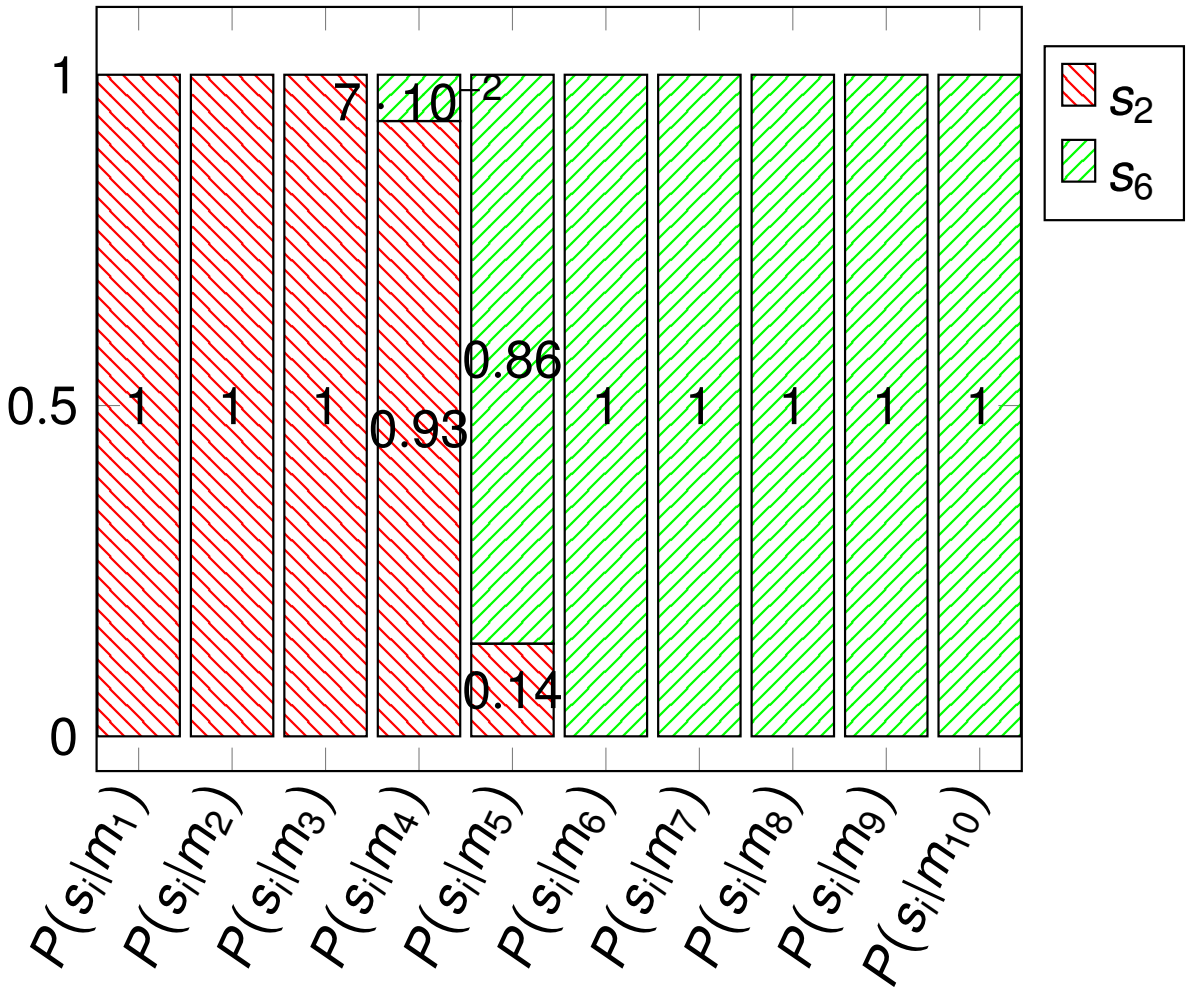
Bottleneck, but no noise

- Sample size = 30, Noise level = 0
- Reduction of number of predicates (typically 2-3). No vagueness.



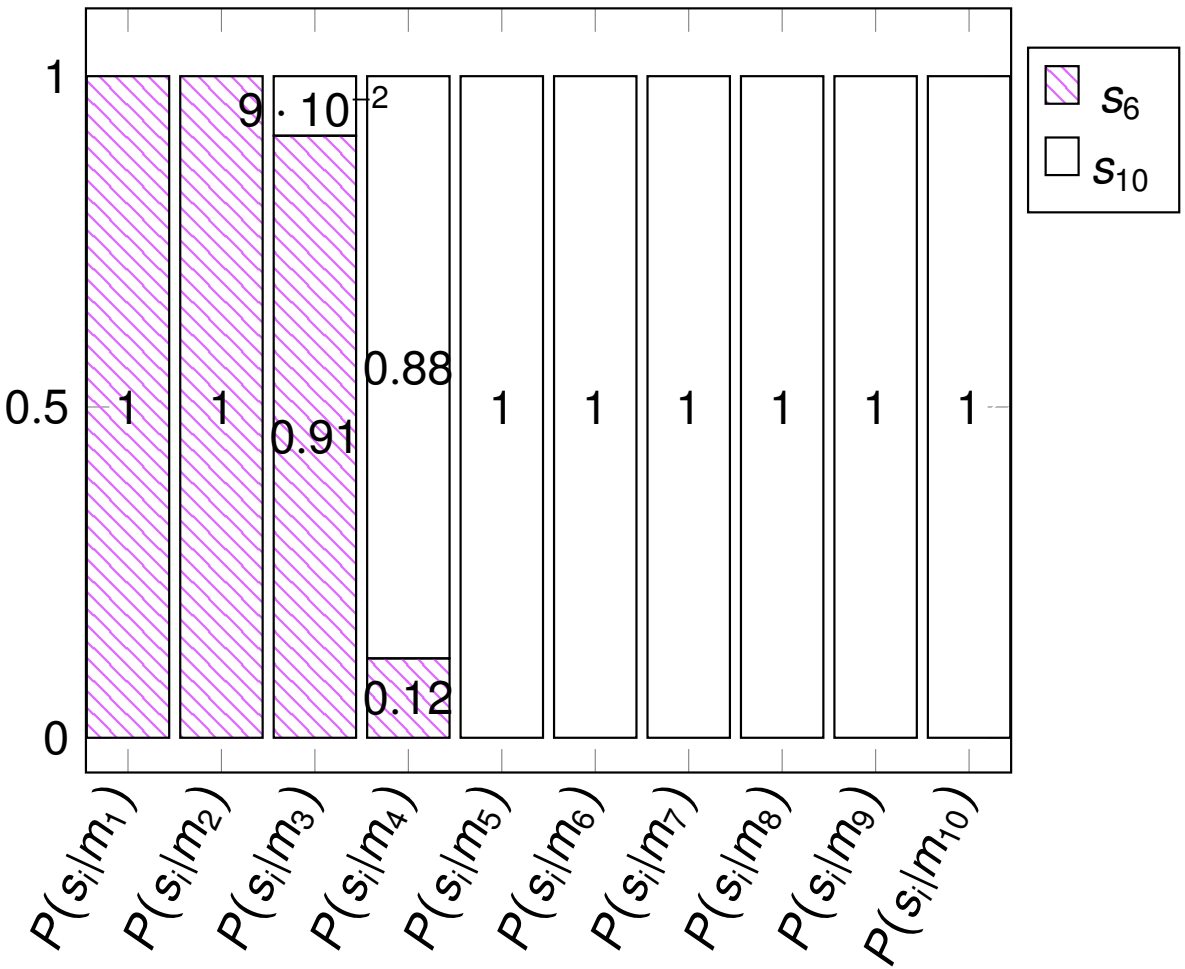
No bottle neck, with noise

- Sample size = 500, Noise level = 0.33 (1/3 of signals were noisy), perplexity set to 0.5
- Reduction of number of predicates (typically 2, sometimes three). Marginally graded boundaries.



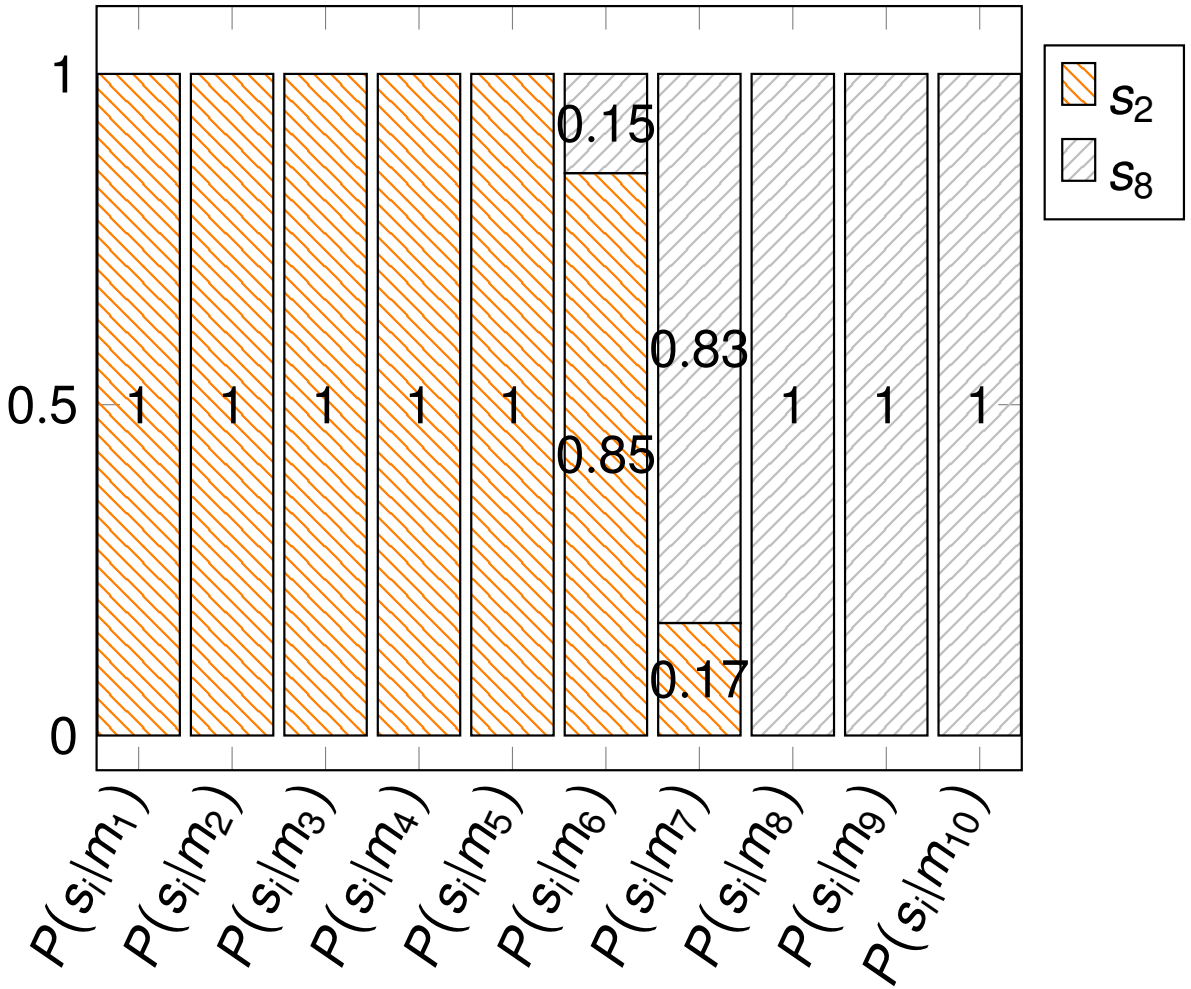
No bottle neck, with noise (cont.)

- Sample size = 500, Noise level = 0.33 (1/3 of signals were noisy), **perplexity set to** 1
- Increasing perplexity has no major impact on gradedness or number of predicates



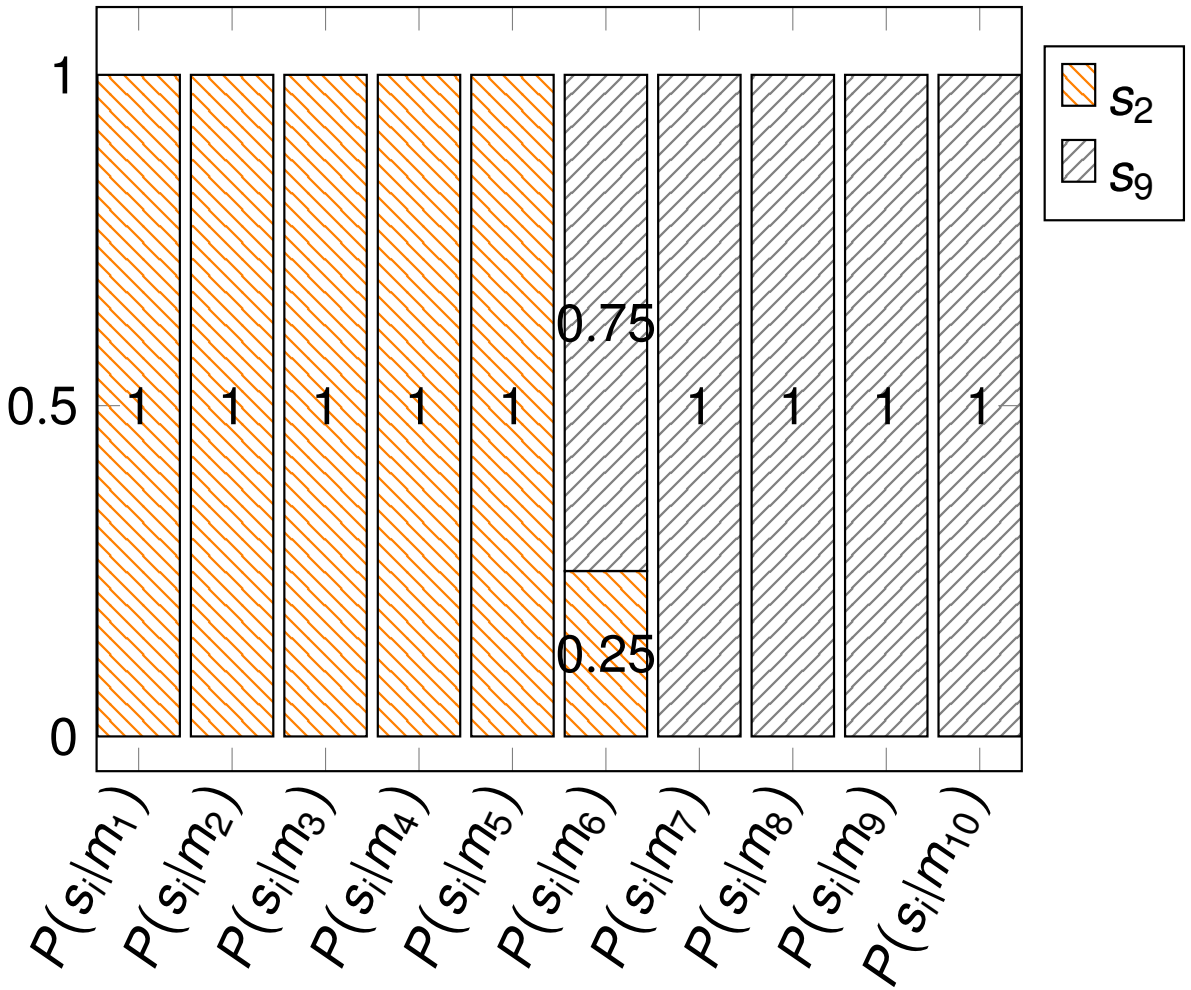
No bottle neck, with increased noise

- Sample size = 500, **Noise level = 0.75 (3/4 of signals were noisy)**, perplexity set to 0.5
- Slightly increased amount of gradedness (0.8-0.85 vs 0.85-0.96)



Bottleneck and noise

- Sample size = 30, Noise level = 0.2 (1/5 of signals were noisy), perplexity set to 0.66
- I.e. bottleneck, lower noise level, but higher perplexity
- Still more gradedness (0.66-0.85), but a bit less stability



Summary

Bottleneck	Noise	Predicate reduction	Vagueness
0	0	0	0
1	0	1	0
0	1	1	1 (some)
1	1	1	1 (more)

Discussion and Conclusions

- Some support for the hypothesis: Vagueness is a product of learning and communication over a noisy channel.
- Vagueness without precisifications or propositions? (correlation between uses of signals and specific messages)
  - But, messages here are just discriminable shades/heights
- Surprising result: No real impact of distance
  - Possible explanation: Small message space size (=10) makes gaps of > 1 and stability unlikely
  - Further work: run models with much larger numbers of messages/situations (and starting predicates/signals)
- Slight bias towards s_1 and s_{10}: need to fine-tune the model to reduce this.
- Further developments: incremental learning (via a generative model)

Gemma Boleda's Worry

The message/situation space assumed is discrete, but shouldn't this space be:

(a) Continuous

(b) Vague

Some answers

(a) It should be continuous, this was an idealisation

(b) It depends what 'vague' means here:

- On my model, noise means that agents are not sure what the world is like, given a signal
- There is a probability distribution over discrete e.g. shades
- This could be a distribution over portions of a continuum (a)
- If 'vague' means this kind of uncertainty: yes
- If vague means indeterminacy of object identity: no
- (Not in the model): For any particular object, might an agent be unsure what shade it is? – Yes
  - ★ E.g., an extra level of noise from perception itself
  - ★ Then again, this could be the source of why signals are noisy in the first place

Take-home messages for day 4

- The ingredients for polysemy
  - Some kind of conflict between speaker effort and hearer effort
  - Or between speaker ease/clarity for the hearer
- An alternative ingredient for polysemy?
  - Learning bottleneck
- The added ingredients for vagueness
  - Noise such that the hearer/learner is uncertain exactly how the world is, given the signal/utterance
  - Reasoning in such conditions results in:
    - ★ blurred boundaries
    - ★ borderline cases
    - ★ sorites susceptibility?

Take-home messages for the course

The problem of vagueness

- Blurriness and CL are in conflict
  - There is either a bullet to bite on the first order or massage the first order problem away
- So there may be a bullet to bite on the second/higher orders
- (I think) the price to pay is truth conditional semantics
  - Replaced with a correlational view of meaning

Vagueness and Polysemy

- Contrary to first appearances, both phenomena may be required for systems of communication that are suitable for speakers and hearers alike

An intriguing possibility

- Maybe there is a common mechanism to explain a wide array of linguistic phenomena
  - Entropy reduction

Thanks

Thanks to participants of TbiLLC 2017 for many helpful comments

Selected References I

Henry Brighton and Simon Kirby. Meaning space structure determines the stability of culturally evolved compositional language". *Technical report, Language Evolution and Computation Research Unit, Department of Theoretical and Applied Linguistics, The University of Edinburgh.*, 2001.

R. Ferrer i Cancho and R. Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):788–791, 2003.

Michael Franke, Gerhard Jäger, and Robert van Rooij. Vagueness, signaling & bounded rationality. proceedings of LENLS2010, 2010.

Gottlob Frege. Sense and reference. *Philosophical Review*, 57:209–230, 1948/1892.

Peter Gärdenfors. *Conceptual Space: The Geometry of Thought*. The MIT Press, Cambridge, MA, 2000.

Selected References II

John A. Hawkins. *Cross-linguistic Variation and Efficiency*. OUP, Oxford, 2014.

Florian T. Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62, 2010.

Simon Kirby. The evolution of meaning-space structure through iterated learning. In C. Lyon, C. Nehaniv, and A. Cangelosi, editors, *Emergence of Communication and Language*, pages 253–268. Springer, Verlag, London, 2007.

Simon Kirby and James Hurford. The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi and D. Parisi, editors, *Simulating the Evolution of Language*, pages 121–148. Springer, Verlag, London, 2002.

Selected References III

Roger Levy and Florian T. Jaeger. Speakers optimize information density through syntactic reduction. *Proceedings of the twentieth annual conference on neural information processing systems*, 19:849–856, 2007.

S. Piantadosi, H. Tily, and E. Gibson. The communicative function of ambiguity in language. *PNAS*, 108(9):3526–3529, 2011.

Bertrand Russell. Vagueness. *Australasian Journal of Psychology and Philosophy*, 1(2):84–92, 1923.

Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

Peter. R. Sutton. *Vagueness, Communication, and Semantic Information*. PhD thesis, King's College London, 2013.

Peter R. Sutton and Hana Filip. Probabilistic mereological type theory and the mass/count distinction. Forthcoming in *JLM*, 2017.

Selected References IV

Warren Weaver. The mathematics of communcation. In C. D. Mortensen, editor, *Basic Readings in Communication Theory*, pages 27–38. Harper and Row, 1979. Originally published in *Scientific American* 181 (1).

Timothy Williamson. *Vagueness*. Routledge, Abingdon, 1994.

G. Zipf. *The psychobiology of language*. Houghton Mifflin, 1935.

G. Zipf. *Human behavior and the principle of least effort*. Addison-Wesley, 1949.