Phonological theory and the Bayesian philosophy of science

Ollie Sayeed
EGG – Brno
July 2022

Day 1

Outline

Day 1: intro to phonological issues

Day 2: intro to Bayesian inference

Day 3: arguments for and against Bayesianism

Day 4-5: applications to phonology

Spoiler alert

This class will **not** have much to say about which theory of phonology is correct.

Our goals:

- introduce the basic disagreement between substance-free phonology and other theories
- introduce the Bayesian worldview
- look at issues in phonological theory through a Bayesian lens

Introduction

Today’s goals:

- introduce **substance-free phonology**
- introduce **evolutionary phonology**
- look at a concrete example

Asymmetries in phonetic substance

Not all surface phonetic patterns are equally common:

- final devoicing >> final voicing
- place assimilation >> place dissimilation
- voiced sonorants >> voiceless sonorants
- oral vowels >> nasal vowels
- *f* >> *θ*

The basic disagreement between **substanceful** and **substance-free** theories of phonology: how do we account for this?

‘Markedness’

A few different kinds of phonological asymmetry have been grouped together under the name **‘markedness’**. Marked segments...

- ...have low frequency within a language
- ...have low frequency across languages
- ...have a more restricted distribution within a language
- ...are acquired later in infancy
- ...are more likely to be impaired in cases of aphasia.

(Haspelmath 2006, Jakobson 1941)

‘Markedness’

Empirically, it seems to be **true** that these things correlate.

- Gordon (2014): within-language frequency correlates with across-language frequency.
- Romani et al. (2017): both age of acquisition and probability of impairment in aphasia correlate with both kinds of frequency.
    - And the correlation with across-language frequency holds **independent** of within-language frequency.

Formal markedness

Traditionally, markedness has been treated as a subject for theoretical phonologists:

- Prague School structuralism (Jakobson, Trubetzkoy)
- SPE markedness statements (Chomsky, Halle)
- OT markedness constraints (Prince, Smolensky)

In each of these cases, we're building **formal** markedness into our theory of mental representations.

But why?

Analytic bias v channel bias

Two kinds of causal explanation for a sound pattern:

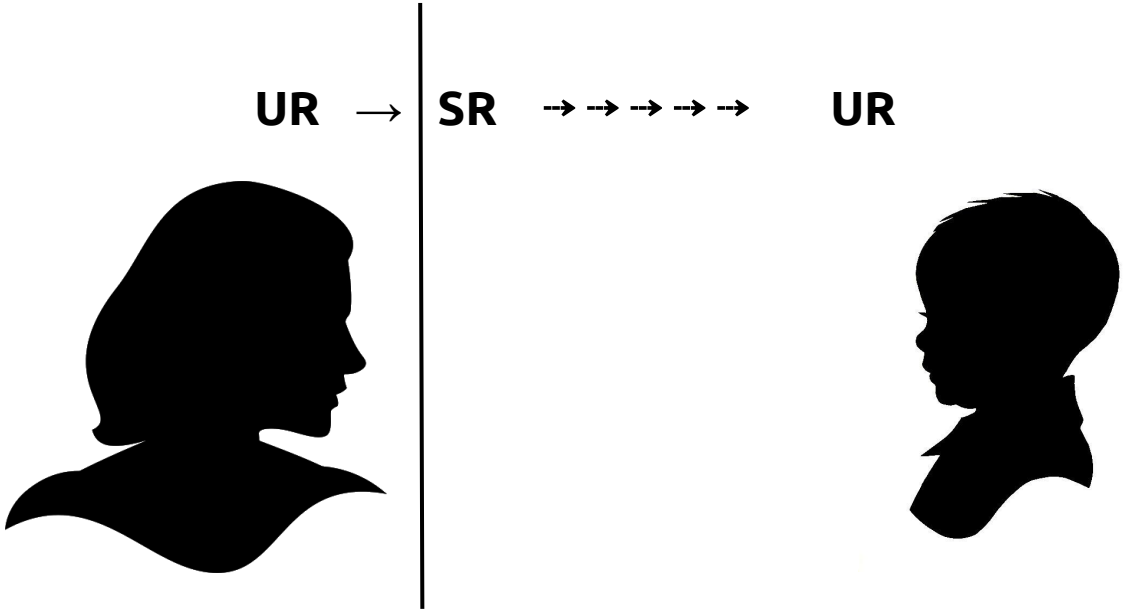
1. **Analytic bias** – facts about the language faculty;
   e.g. formal markedness, learning bias

2. **Channel bias** – facts about the transmission process.
   e.g. articulation, perception

In each individual case, either an analytic bias or a channel bias (or both) could be at work.

A constraint: the Gulf

All explanations in theoretical linguistics have to contend with **the Gulf**. Learners only have **indirect** access to the adult grammar.



SFP and EP

My view is that there's a explanatory division of labour between:

1. **Substance-free phonology** (SFP) as an account of representations

   and

2. **Evolutionary phonology** (EP) as an account of substance

This isn't normally put quite this explicitly in SFP circles.

Substance-free phonology

Hale and Reiss (2008) make two separate claims about what it means for phonology to be **“substance-free”**:

1. Mental symbols are not **the same** as muscle movements or sound waves.

2. The phonological component of the grammar isn’t **constrained to recapitulate** the properties of muscle movements and sound waves.
   (i.e. there is no formal markedness)

My view: 1 is trivial. 2 is interesting.

Evolutionary phonology

Blevins (2004): the typological frequency of a pattern needn't be built into the grammar. Among other things, it's also a function of:

- the frequency of sound changes that **create** it
- the frequency of sound changes that **destroy** it

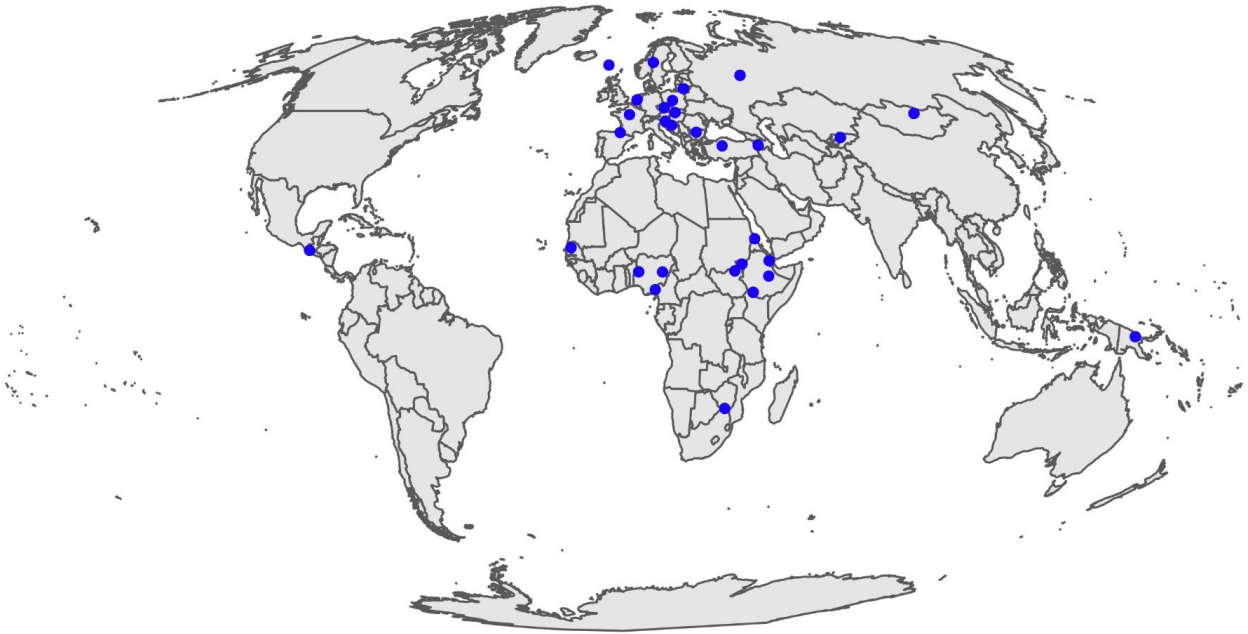
Explaining typological patterns becomes a job for phoneticians and historical phonologists, not (only) synchronic theoretical phonologists.

Classic example: final devoicing

For example: German, Russian, Turkish, Wolof, Ojibwe and other languages have **final devoicing**.

The reverse, **final voicing**, is rare or unattested. Why the asymmetry?

Typological aside



OT on final devoicing

Classic Optimality Theory says there is a constraint *D]$_{\sigma}$, but no constraint *T]$_{\sigma}$.

So grammars with final voicing are **representationally impossible**.

This is an example of formal markedness: building typological patterns into the space of representationally possible grammars.

EP on final devoicing

Evolutionary phonology says there are phonetic pressures that diachronically lead to final devoicing, but not to final voicing:

- voicing is aerodynamically difficult without a following vowel (Westbury and Keating 1986)
- the glottis tends to close before a following pause (Cho et al. 2019)
- final stops tend to lengthen, which is a cue to voicelessness (Wightman et al. 1992).

All this is consistent with final voicing being **representationally possible** but **unlikely to appear** in a real language.

A response to EP

But Kiparsky (2006) points out there are more complicated routes to (something like) final devoicing under EP:

- final degemination followed by *tt* > *t* and *t* > *d*
  - (*batt* > *bat* > *bad*, but *batta* > *bata*)
- medial voicing followed by apocope
  - (*bata* > *bada* > *bad*)
- post-nasal voicing plus cluster simplification and deletion
  - (*banta* > *banda* > *bada* > *bad*, but *bata* > *bata*)

So what rules these out?

Probabilities and sound change

Kiparsky's point is a **quantitative** one.

Working out whether the data support an analytic bias (OT), or a channel bias (SFP/EP) is a complicated game of probabilities.

Beguš (2020) tries to **bootstrap** sound change probabilities from a database of sound changes, and finds support for the channel bias view of final devoicing.

Some lessons

1. Theoretical phonologists can’t escape **probabilities**.

Even without any randomness inside the grammar, there is uncertainty about the process that produces the grammars we see.

2. Theoretical phonology can’t take place in a **vacuum**.

Our evidence about the world’s languages is the product of a jumble of processes – we can’t do theoretical phonology without trying to disentangle them.

Probability in science

The Bayesian view is that **everything** in science is a game of probabilities.

(Except for 2 + 2 = 4?)

Probabilities are a way of capturing our **uncertainty** about the world beyond textbook settings like coin tosses and die rolls.

The relationship between a theory and the data is **very complicated**, but there are still lessons to learn from a Bayesian perspective.

Tomorrow

Day 2:

- the central idea: Bayes’ rule
- what Bayes’ rule means in practice

Day 2

A general problem

We have a set of **theories**.

(Logical Phonology, Optimality Theory, MaxEnt…)

We have **data**.

(written grammars, corpora, speaker judgements, experimental data…)

What do the data tell us about our theories?

Bayesianism

Bayesianism is a framework for **reasoning about evidence**. It has a **static** part and a **dynamic** part.

> Static: represent our uncertainty as a **probability distribution** over states of the world.
>
> Dynamic: **update** this distribution as new evidence comes in according to Bayes' rule.

Probability

P(A) means “the **probability** that A is true”.

P(B|C) means “the probability that B is true, **given** that C is true”.

P(D&E) means “the probability that D and E are **both** true”.

All probabilities are real numbers between 0 and 1, and probabilities of mutually exclusive events are **additive**.

Examples of probability

For example:

Examples of probability

For example:

- P(die rolls a 6) =

Examples of probability

For example:

- P(die rolls a 6) = ⅙

Examples of probability

For example:

- P(die rolls a 6) = ⅙
- P(die rolls an even number) =

Examples of probability

For example:

- P(die rolls a 6) = ⅙
- P(die rolls an even number) = ½

Examples of probability

For example:

- P(die rolls a 6) = ⅙
- P(die rolls an even number) = ½
- P(die rolls a number divisible by 3 | die rolls an even number) =

Examples of probability

For example:

- P(die rolls a 6) = ⅙
- P(die rolls an even number) = ½
- P(die rolls a number divisible by 3 | die rolls an even number) = ⅓

Examples of probability

For example:

- P(die rolls a 6) = ⅙
- P(die rolls an even number) = ½
- P(die rolls a number divisible by 3 | die rolls an even number) = ⅓
- P(die rolls a 3 | the sun is shining) =

Examples of probability

For example:

- P(die rolls a 6) = ⅙
- P(die rolls an even number) = ½
- P(die rolls a number divisible by 3 | die rolls an even number) = ⅓
- P(die rolls a 3 | the sun is shining) = ⅙

Examples of probability

For example:

- P(die rolls a 6) = ⅙
- P(die rolls an even number) = ½
- P(die rolls a number divisible by 3 | die rolls an even number) = ⅓
- P(die rolls a 3 | the sun is shining) = ⅙
- P(die rolls a 4 & die rolls a 5) =

Examples of probability

For example:

- P(die rolls a 6) = ⅙
- P(die rolls an even number) = ½
- P(die rolls a number divisible by 3 | die rolls an even number) = ⅓
- P(die rolls a 3 | the sun is shining) = ⅙
- P(die rolls a 4 & die rolls a 5) = 0

The chain rule

There is a relationship between conditional probabilities and joint probabilities called the **chain rule**:

$$P(A\&B) = P(A|B) \times P(B)$$

- P(die rolls an even number) = ½
- P(die rolls a number divisible by 3 | die rolls an even number) = ⅓
- P(die rolls a 6) = ⅙

Bayes' rule

The dynamic part of Bayesianism is encoded in an equation called **Bayes' rule**.

$$\mathbf{P(A|B) = P(B|A) \times P(A) / P(B)}$$

The probability of **A given B** is:

- the probability of **B given A**
- multiplied by the ratio of the probabilities of **A** and **B**

This follows from the chain rule.

Bayesian inference

Bayes' rule lies at the heart of **inference** problems.

Say we have data D and want to know what this tells us about a hypothesis H.

If we know how likely the **data** are **given the theory…**

…then Bayes' rule tells us how likely the **theory** is **given the data**.

Example

We see that the ground outside is wet (the **data**).

Did it rain overnight (the **theory**)?

P(it rained | ground is wet)

= P(ground is wet | it rained) $\times$ P(it rained) / P (ground is wet)

Some terminology

The probability P(H) of the hypothesis before looking at the data is called a **prior**.

The probability P(H|D) of the hypothesis given the data is called a **posterior**.

The probability P(D|H) of the data given a hypothesis is called a **likelihood**.

Bayes' rule says: **posterior ∝ prior** $\times$ **likelihood**

$$P(H|D) \qquad P(H) \qquad P(D|H)$$

Example

Thinking in terms of **ratios** of probabilities means we can ignore the prior P(D):

$$P(H|D) / P(\neg H|D)$$
$$=$$
$$P(D|H) / P(D|\neg H)$$
$$\times$$
$$P(H) / P(\neg H)$$

Bayes' rule converts the **prior odds** into the **posterior odds** through a **likelihood ratio**.

Forwards and backwards

We can think of Bayes' rule as converting **"forward"** (or "generative") probabilities like P(D|H) into **"backward"** (or "inverse") probabilities like P(H|D).

In general:

- we're uncertain about hypothesis H
- there is some kind of generative process in the world that, if H is true, produces data D with probability P(D|H)
- we observe data D

In situations like this, Bayes' rule tells us how to work "backwards" to reach P(H|D), telling us how likely H is given the data.

Two images

Two mental images that might help with the intuitions behind Bayesian inference:

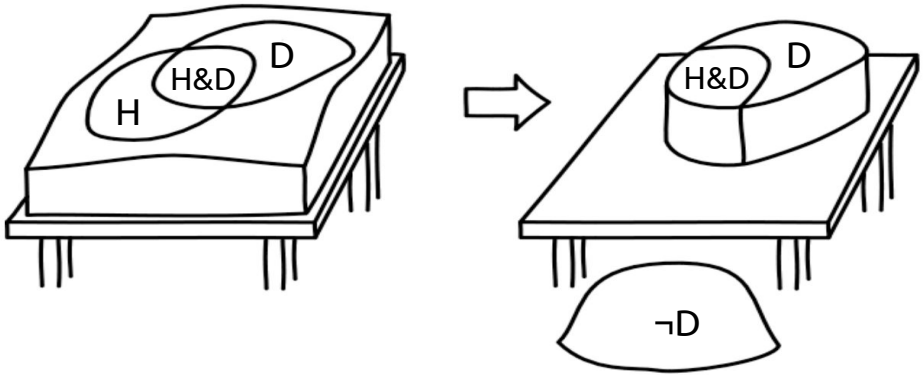
1. **Cutting the cake**
2. **Adjusting the scales**

Cutting the cake

Our space of hypotheses is like a cake.

When we learn data D, we perform Bayesian inference by **cutting out** all parts of the cake inconsistent with D.

Everything else is left in the same ratio as before.

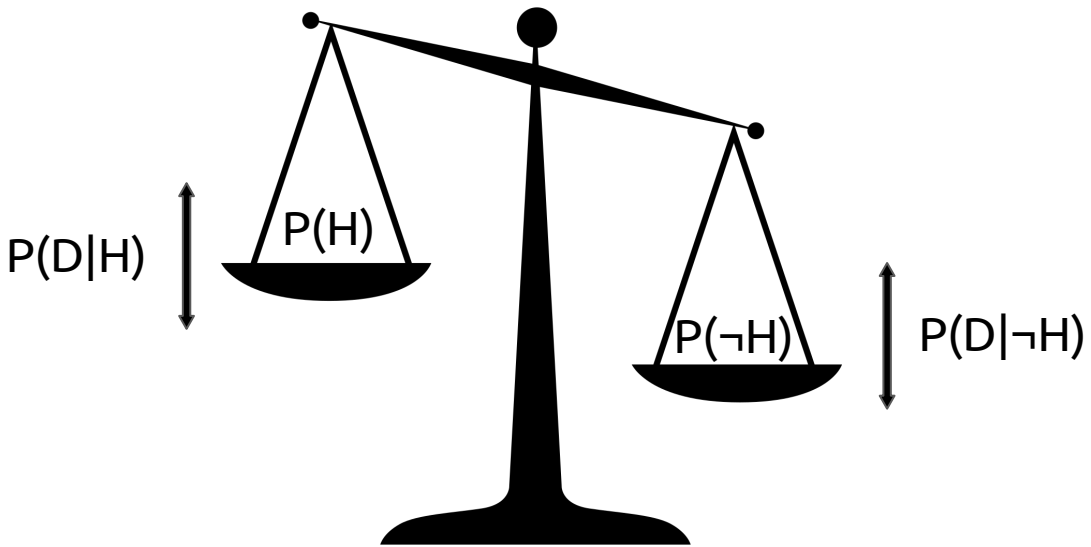


Norton (2011)

Adjusting the scales

Learning data D adjusts the balance of probability between H and ¬H.

If D adjusts H downward, it **automatically** adjusts ¬H upward and vice versa.



Adjusting the scales

All that matters is which of P(D|H) and P(D|¬H) is **bigger**.

P(D|H) is bigger:

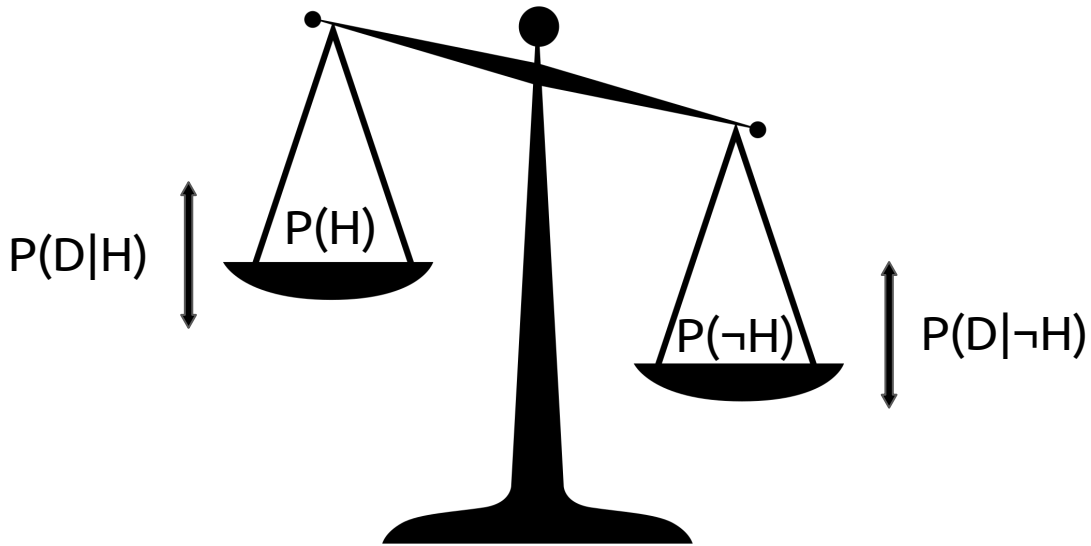
D is evidence **for** H

P(D|¬H) is bigger:

D is evidence **against** H

P(D|H) = P(D|¬H):

D **says nothing** about H



The likelihood principle

The **only** way that data D can give us information about a hypothesis H is through the likelihoods P(D|H) and P(D|¬H).

This is called the **likelihood principle** (Berger and Wolpert 1988).

The key question for theoretical phonologists: would we be **more likely** to see the data if your theory is true than if your theory is false?

Tomorrow

Day 3:

- assumptions behind Bayesianism
- arguments that statements about uncertainty should obey those assumptions
- challenges for Bayesianism

Day 3

Why Bayes?

We’ve seen the basic machinery of Bayesian inference at work.

But why think this has anything to do with **scientific uncertainty**?

Probabilism

Bayes' rule itself is a trivial bit of algebra.

What's more at issue is **probabilism**: the philosophical view that statements about uncertainty should follow the laws of probability in the first place.

Probability as “extended logic”

We normally teach probability in a restricted way as a theory of **chance**, with examples from random processes like coin tosses and die rolls.

Bayesianism uses probability in a different and more general way: as a kind of **extended logic**.

- Propositional logic: normative reasoning under **certainty**
- Probability theory: normative reasoning under **uncertainty**

Interpretations of probability

There are lots of **different** systems that obey the laws of probability:

- long-run frequencies
- physical propensities/chances
- areas on a Venn diagram
- mod-squared amplitudes in quantum mechanics
- statements of uncertainty (according to probabilism)

No single one of these is “the interpretation” of probability.

(What is “the interpretation” of $y = x^2$?)

Arguments for probabilism

A general strategy we'll see today:

- give a set of conditions it would seem unreasonable for statements of our uncertainty about the world not to obey
- show that those conditions imply the laws of probability

A result like this is called a **representation theorem**: "if a system obeys conditions ABC, it's isomorphic to some other object D".

Arguments for probabilism

1. The **Fernandez argument**
2. The **Dutch book argument**
3. The **scoring rule argument**
4. **Cox's theorem**
5. **Savage's representation theorem**
6. The **complete class theorems**

1. The Fernandez argument

An intuitive argument that puts some metaphysics behind the cutting-the-cake procedure (Fernandez 2020):

1. There is a set of ways the world could be (“possible worlds”).
2. When we learn D, we rule out all the worlds where D is false.
3. The relative plausibility of all other worlds is unchanged, so they all become uniformly more plausible.

This is equivalent to **cutting** and **rescaling** the space of possible worlds: giving us Bayesian inference.

2. The Dutch book argument

A “**Dutch book”** is a combination of odds and bets that guarantees a win for the bookmakers whatever happens. For example:

- pay $1.50 now
- win $1 if it rains tomorrow
- win $1 if it doesn’t rain tomorrow

Ramsey (1926), de Finetti (1931): you are vulnerable to Dutch books if and only if your distribution of uncertainty violates the laws of probability.

3. The scoring rule argument

Forecasters are often judged by **scoring rules** that compare their forecasts over time to reality. For example:

> "Tomorrow it will rain with probability 70%"
>
> If rain: win $1-(1-0.7)^2 = 0.91$ points
>
> If no rain: win $1-0.7^2 = 0.51$ points

de Finetti (1974): if your forecasts violate the laws of probability, they are **dominated** by another set of forecasts that wins more points whatever happens.

4. Cox's theorem

Cox (1946) lays out three axioms about the **"plausibility"** of a claim:

1. plausibility is always a real number
2. **"and"** is compositional in a particular way: $p(A \wedge B) = g(p(A), p(B|A))$
3. **"not"** is compositional: $p(\neg A) = -f(p(A))$

If these axioms hold for some nondecreasing functions f and g, then we can convert each plausibility into a **probability** between 0 and 1.

5. Savage's representation theorem

Savage (1954): if our (revealed) **preferences** over states of the world:

1. … are complete and don't form cycles
2. … are independent of the outcomes of irrelevant events
3. … don't depend on the *current* state of the world
4. … don't change direction when all options get uniformly better

5. Savage's representation theorem

5. … are not totally indifferent between all outcomes

6. … let us accept any tiny amount of risk

… then we are acting as if we're maximizing expected utility according to an uncertainty distribution obeying the laws of probability.

6. Complete class theorems

A decision rule δ_1 is **dominated** by rule δ_2 if δ_2 is guaranteed to give a better outcome than δ_1 whatever happens. For example:

> If bringing a laptop is useful whether there is Wi-Fi or not, then “bring a laptop” dominates “don’t bring a laptop”.

A **complete class** is a set of rules that, between them, dominate any rule outside the set.

Wald (1947): the class of rules that obey the laws of probability is complete.

The Problem of the Priors

There are technical philosophical discussions about each of these arguments.

In my view, the big issue is how to use Bayesian inference in a principled way.

Specifically: how should we assign **priors**?

Indifference and entropy

Two intuitive answers:

1. the **principle of indifference**
   e.g. “assign $p = ½$ to each side of a coin”
   “assign $p = ⅙$ to each face of a die”

2. the **principle of maximum entropy**: maximize the “spread” of probability
   e.g. the normal distribution is the maximum-entropy distribution with mean μ and variance σ^2

Norton’s no-go theorem

Norton (2019): a system of **inductive inference** over a set of propositions can’t simultaneously:

1. … depend only on the deductive structure of the set
2. … be stable under finer grainings of propositions of the set
3. … assign more similar propositions closer levels of support
4. … distinguish between different propositions!

Norton calls this “the incompleteness of calculi of inductive inference”.

Bayesianism **lacks** a way of assigning priors that doesn’t violate condition 2.

Alternative formal systems

Standard probability theory is the orthodoxy in epistemology, but variations include:

- **imprecise probabilities**, e.g. letting P(A) be an interval (p_1, p_2)
- **ordinal probabilities**, where P(A) and P(B) are placed on a ranking but don't have numerical values
- **Dempster-Shafer theory**, relaxing the rule P(A or B) = P(A) + P(B) for disjoint A and B
- **infinitesimal probabilities**, where P(A) can be smaller than any real number but bigger than zero.

Tomorrow and Friday

Days 4+5:

applications to phonology:

- explanation and explanatory power
- the role of phonetics in phonological theorizing
- scientific realism
- picking between analyses
- science as a social enterprise

Days 4+5

So what?

Some issues that Bayesianism can contribute to:

- explanation and explanatory power
- the role of phonetics in phonological theorizing
- scientific realism
- picking between analyses
- science as a social enterprise

The likelihood principle again

The likelihood principle tells us:

- **The *only* thing that matters in empirically evaluating a theory is *how likely* it expected the data to be.**

This is a demanding criterion: **nothing** else is evidentially relevant.

(e.g. how beautiful an analysis is, how neatly or satisfyingly it captures the data, how many generalizations it makes)

Forwards and backwards again

Bayesian inference converts forward probabilities P(D|H) into backward probabilities P(H|D).

In phonology, the process generating the data is the centuries-long cycle of language acquisition and change. Put another way:

> **Historical phonology** and **phonological acquisition** are the study of the forward problem.
>
> **Theoretical phonology** is the study of the backward problem.

Explanatory power and the role of phonetics

Platonic vs embedded explanations

Platonic:

e.g. abstract constraints that express a language's preference to have a symmetrical, dispersed vowel system (Flemming 2004)

Language as an abstract object

Embedded:

e.g. a model of the learner that predicts asymmetrical, undispersed systems to be less easily learnable (Vaux and Samuels 2015, Roberts and Clark 2020)

Language as a phenomenon in the physical universe

Related ideas

Platonic explanation ≈ Gorrie's (2015) **ideal-typology**, a paradigm that says linguistic theories refer to 'ideal-types' that need not exist in the real world.

Graf (2019) distinguishes two kinds of theoretical linguist:

- **formalists** believe the formalism is literally real (≈ embedded explanation);
- **analysts** believe the formalism is just a notation for an abstract "analysis" (≈ Platonic explanation).

Against Platonic explanations

My view is that Bayesianism requires **embedded** explanations.

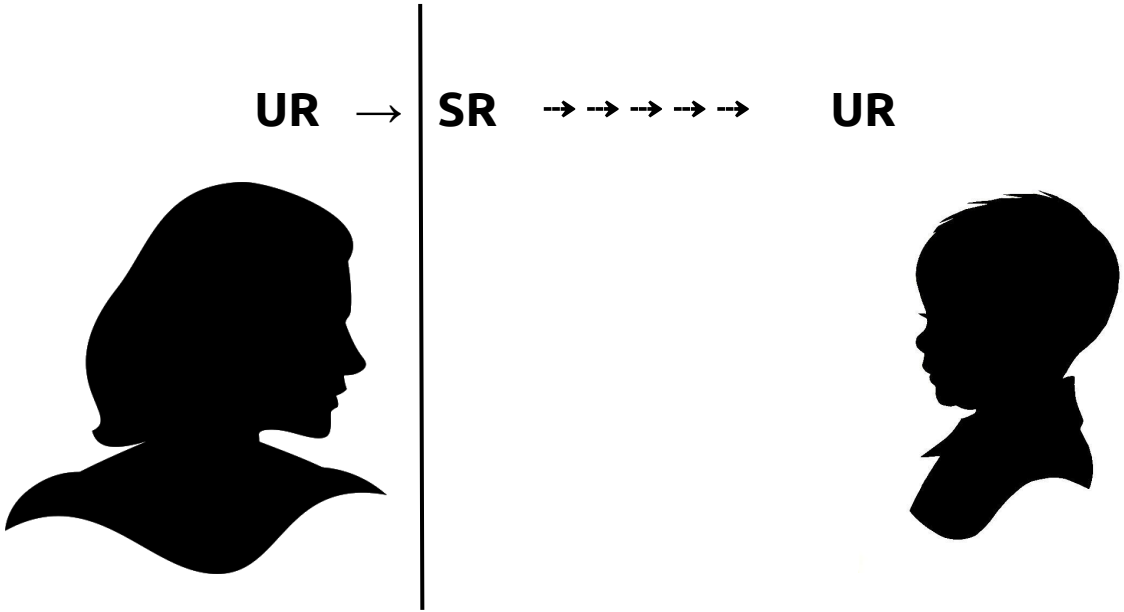
- We need a **causal story** that explains why the data are evidentially relevant to the theory we're evaluating.

- Platonic explanations don't have this property, until we cash them out in terms of claims about the real world.

We could annotate a Platonic explanation with likelihoods, but they wouldn't be likelihoods **of** anything.

The Gulf again

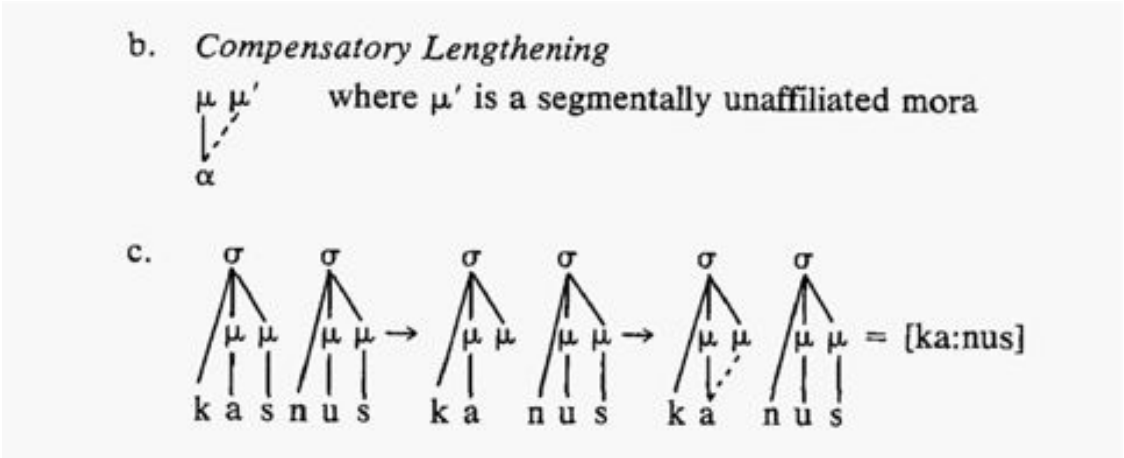
Learners only have **indirect** access to the adult grammar.



The Gulf

This rules out explanations that refer to structure the child can't see.

e.g. the moraic analysis of compensatory lengthening (Hayes 1989):



A child who fails to hear the first **s** doesn't **know** there's a mora that needs to be reattached.

Phonetics vs phonology

In general, representational naturalness is not causally relevant when it conflicts with phonetics:

- Lip rounding in American English **ɹ**
  - rhoticity and roundedness share low F3, despite the lack of [+labial] (Mielke et al. 2016);
- Rhinoglottophilia, e.g. **h** > **ŋ** in Avestan
  - has both articulatory and acoustic reasons, despite the lack of [+nasal] (Matisoff 1975);
- Epenthesis in lateral-fricative clusters, e.g. **ls** > **lts**,
  - has a gestural timing explanation, despite the lack of [–cont] (Ohala 1997).

Moral: representational theories don’t make predictions on their own. We need to consider the Gulf.

What makes an explanation?

What are we looking for in an explanation anyway? Some popular candidates:

- The **deductive-nomological** model (Hempel 1965): the theory explains the data as the (ideally deductive) consequence of a **law**.
  e.g. “apples fall to the ground because of the law of gravity”

- **Statistical relevance** (Peirce 1931): the theory makes the data **more likely** to be true than if the theory was false.
  e.g. “adaptive complexity would be unlikely by chance, but very likely given a creator”

D-N is a special case of SR.

What makes an explanation?

- The **causal-mechanical** model (Halpern and Pearl 2005): the theory gives a **causal mechanism** that produces the data.
    e.g. “finch beaks are produced by the process of natural selection”

- **Unification** (Kitcher 1981): the theory **unifies** multiple kinds of data into consequences of one phenomenon.
    e.g. “electricity and magnetism are consequences of the same field”

Or we could be **pluralist**: let “explanation” refer to any of the above depending on context.

A Bayesian perspective

My view:

- We should want **true** theories.
- Considerations of explanatory power are only relevant if there's a Bayesian reason that "explanatory" theories have higher posterior probability.

The D-N, statistical relevance, causal-mechanical, and unification models are arguably special cases of Bayes.

A Gricean account of explanation

But not all true statements make good explanations.

Explanations should also obey **Gricean** principles (Hao 2022):

- contain the right amount of information (Quantity)
- avoid irrelevant details (Relation)
- be easy to understand (Manner)

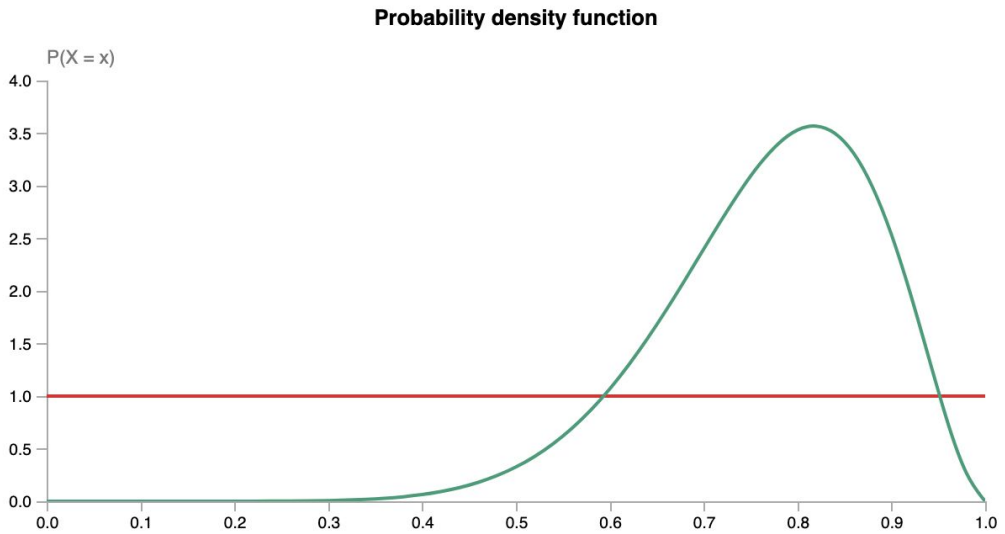
A list of the positions and momenta of 10^{50} atoms might technically be an accurate account of the result of the last UK general election, but it’s un-Gricean.

Predictiveness

For a Bayesian, a theory's predictions are a **finite resource**.

You can spread your expectations equally over all data sets…

…or put all your eggs in one basket with one strong prediction.



Predictiveness

Predictiveness/falsifiability ~ “scientific risk-taking”:

> A **more predictive** theory has more to gain from a correct prediction and more to lose from an incorrect one.
>
> A **less predictive** theory will not shift much from its prior probability, whatever the data are.

(But a less predictive theory could still be true!)

Predictiveness in phonology

In the phonology context, people often call theories like SFP “unrestrictive” (= “unpredictive”). But:

1. Predictive and unpredictive theories have the same likelihood given the data **in expectation**. Unpredictive theories are not *a priori* less likely.

2. SFP makes clearer predictions about probable languages **in tandem** with a theory of diachrony, like EP.

Falsifiability

The textbook scientific method is focused on **falsifying** hypotheses by experiment. In the Bayesian worldview:

- Rather than picking one hypothesis and trying to falsify it, in general we are trying to **distinguish between** sets of hypotheses.

- Disconfirmation is **gradient**, not discrete. A theory that continues to make false predictions get less and less likely given the data.

- New experimental data are not the only relevant evidence. Theories are also judged based on how well they **accommodate** existing data.

Ockham’s razor

The standard Bayesian approach to **simplicity/“Ockham’s razor”** is to assign simpler theories higher priors.

But why? A couple of responses:

1. Priors are arbitrary, so we’re free to build in Ockham’s razor if we like.

2. If simplicity is cashed out as “minimum description length” in a formal language, it’s impossible **not** to assign simpler theories higher priors beyond some cut-off (Carlsmith 2021).

Linguists vs learners

Distinguish between the mind of the **linguist** and the mind of the **learner**.

- If you accept the Bayesian arguments from days 2 and 3, linguists should be Bayesians.

- Child learners are probably **not** Bayesians.

When we write down an analysis, we are studying the **child's** reasoning: the task is to reason about a non-Bayesian reasoner in a Bayesian way.

Scientific realism

Scientific realism

Scientific theories often refer to new entities that help explain the data.

e.g. germs, atoms, electromagnetic fields, phonemes, syntactic structures…

Are these entities **real**?

“Circularity”

Sometimes theories of invisible entities are accused of being **“circular”**.

> “We propose floating features because of surface palatalization, but then say palatalization happens because of floating features.”

This is underlyingly a confusion between forward and backward probabilities:

- “palatalization ⇒ floating features” is a statement of **evidential support**
- “floating features ⇒ palatalization” is a statement of **causation**

Kinds of realism

We can distinguish three kinds of realism in science (Horwich 1982):

1. **Metaphysical realism**: there is a real world out there
2. **Semantic realism**: we should take scientific claims “at face value” as descriptions of the real world
3. **Epistemological realism**: our current best theories are *correct* descriptions of the real world

Under Bayesianism, 3 comes in degrees: we can have more or less confidence about our current best theories.

The no miracles argument

The classic argument for scientific realism: the "**no miracles" argument**.

> Putnam (1975): realism "is the only philosophy that doesn't make the success of science a miracle".

e.g. if quantum electrodynamics were false, it would be very unlikely for the magnetic moment of the electron to match the theory to ten significant figures

This is a straightforward Bayesian argument:

> "P(D|H) is very high and P(D|¬H) is very low, so D is strong evidence for H"

A pessimistic response

The classic anti-realist response: even if P(D|H) is high, P(H) is very low because most scientific theories have historically been false:

- phlogiston
- luminiferous aether
- spontaneous generation
- miasma
- …

This is sometimes called the **pessimistic meta-induction** (Laudan 1981).

A response to the response

One realist response to pessimistic meta-induction is to argue that superseded theories are not **false**: they're **special cases** of new theories under particular conditions.

- Newtonian mechanics = special relativity at low speeds
- classical mechanics = quantum mechanics at large scales

Phonological examples?

- Neogrammarian change = lexical diffusion after completion
- deterministic OT = stochastic OT with no noise

Verisimilitude

A different realist response: past theories are indeed false, but we're steadily getting **closer** to the truth.

> Compare: "all models are false, but some models are useful"

This relies on an idea of **verisimilitude**, or closeness-to-truth.

Now it's not only *probability* that is gradient; (*closeness-to-*)*truth* itself is gradient.

Verisimilitude

A couple of ideas:

1. If we have a **metric** over the space of possible worlds, we can treat verisimilitude as “closeness to the actual world” under that metric.
   (Compare Lewis and Stalnaker’s “closeness” among possible worlds.)

2. We could model truth directly with a **fuzzy logic**, like the infinite-valued Łukasiewicz logic $Ł_{\aleph}$, making the truth of a statement a real number between 0 and 1.

If scientific theories are false but becoming less false, the pessimistic meta-induction is less concerning.

Picking between analyses

Analysis under a theory

Theoretical linguists are often doing a mixture of two things at once:

1. picking between **theories** based on the data
2. picking between **analyses** of the data under a given background theory

We should be careful not to confuse these two goals.

Linguists vs learners again

When we pick between analyses of a data set, we are modelling the reasoning of the **child** who constructed that analysis given their input.

Principles like Ockham's razor or the laws of Bayesianism are only applicable to linguistic analysis if we have reason to think the child follows them.

In particular: children's minds could be such that they come up with analyses that look silly to us as analysts.

Taking the language faculty seriously

The language faculty did not evolve to:

- be easily analysable by linguists
- produce grammars that look satisfyingly neat to linguists
- produce grammars that are uniquely identifiable to a linguist who doesn't know the space of possible grammars

If a theory or an analysis falls foul of these conditions – that's unlucky for us, but doesn't make it less likely to be **true**.

Example: “mirror-image” rules

In Lithuanian, both /ʃ+s/ and /s+ʃ/ surface as [ʃ] (Kenstowicz and Kisseberth 1979):

/neʃ-si/	→	[neʃi]	“you will carry”
/saus-ʃala/	→	[sauʃala]	“bitter cold”

How should we analyse this?

Example: “mirror-image” rules

Analysis 1: separate rules

s → ∅ / _ ʃ
s → ∅ / ʃ _

Analysis 2: allow disjunction of environments

s → ∅ / {_ ʃ, ʃ _}

Analysis 3: modify the syntax of phonological rules with an **“adjacency”** symbol |

s → ∅ | ʃ “delete s adjacent to ʃ”

Example: “mirror-image” rules

Which of analyses 1, 2 and 3 is more satisfying to us as linguists is **irrelevant** to the question of which analysis the child picks.

> If the phonological component doesn’t contain a | symbol, the child **can’t** pick option 3.
>
> If the phonological component doesn’t contain disjunction over environments, the child **can’t** pick option 2.

Before considering which analysis the child will pick, we need to consider which analyses are available to the child for picking.

Neat analyses vs possible analyses

Two separate claims:

1. “learners pick the neatest/simplest analysis possible given the options available”
2. “the space of possible analyses must be such that the neatest/simplest analysis that linguists can think of is available to the learner”

1 is plausibly true, but 2 is almost certainly false. The language faculty doesn’t care about us!

Science as a social enterprise

Science as a social enterprise

So far, we've been thinking of "all scientists" as a single agent – as if we all make up one giant brain.

But science in the real world takes place in a **community** of scientists.

Social epistemology is the field that studies reasoning by communities of reasoners.

Bayes vs Kuhn

Thomas Kuhn (1962) argued for a distinction between:

- **“normal science”**, where researchers within a community forbid questioning the assumptions of that community
- **“paradigm shifts”**, where a community suddenly shifts to a new set of assumptions

Bayesianism is permanently “open-minded”, with no distinction between normal science and paradigm shifts: everything is permanently up for re-evaluating.

Theoretical linguistics in the 21st century feels more Kuhnian to me.

“Bubbliness” in science

Zollmann (2022): under some conditions, a network of individually Bayesian scientists will reject “unlucky” hypotheses too quickly (the **Zollman effect**).

But if we **restrict** information flow between scientists, unlucky hypotheses can survive long enough to become locally popular.

An argument for separate “bubbles” of scientists, semi-insulated from criticism by other bubbles?

Social Bayesianism

There are interpretations of Bayesianism as a description of multiple agents.

Critch and Russell (2017): a single Bayesian agent is isomorphic to a sort of gambling-based democracy:

- hypotheses ~ voters
- likelihoods ~ voters' predictive abilities (success at gambling)
- an agent's decision ~ results of an election

(The formal analogy is fun, but I do not endorse replacing either science or democracy with a money-weighted election.)

An example of embedded explanation

Embedded explanation: markedness

If phonology is substance-free, what explains the “markedness” trends identified by Jakobson?

A formal model of sound change shows how EP can take some of the explanatory load off SFP and predict Jakobson’s correlations even if SFP is true (Sayeed and Ceolin 2019).

Modelling sound change

In a simple model of sound change, there are two possible events: **splits** and **mergers**.

- A split takes an existing category and cuts it in two; some of its tokens are assigned to a different (possibly new) category.
- A merger takes two existing categories and merges them; all their tokens now belong to one category.

We can think of an unconditioned shift **X** > **Y** as equivalent to a ‘merger’ with a new category.

Dynamics

Start with a vector of segment frequencies $\mathbf{p_i}$ at time **t**. At each time step **t+1**, apply a split or merger at random:

Split

$$p_i^{t+1} := \frac{p_i^t}{2}$$

$$p_j^{t+1} := \frac{p_i^t}{2} + p_j^t$$

$$p_k^{t+1} := p_k^t$$

Merger

$$p_i^{t+1} := 0$$

$$p_j^{t+1} := p_i^t + p_j^t$$

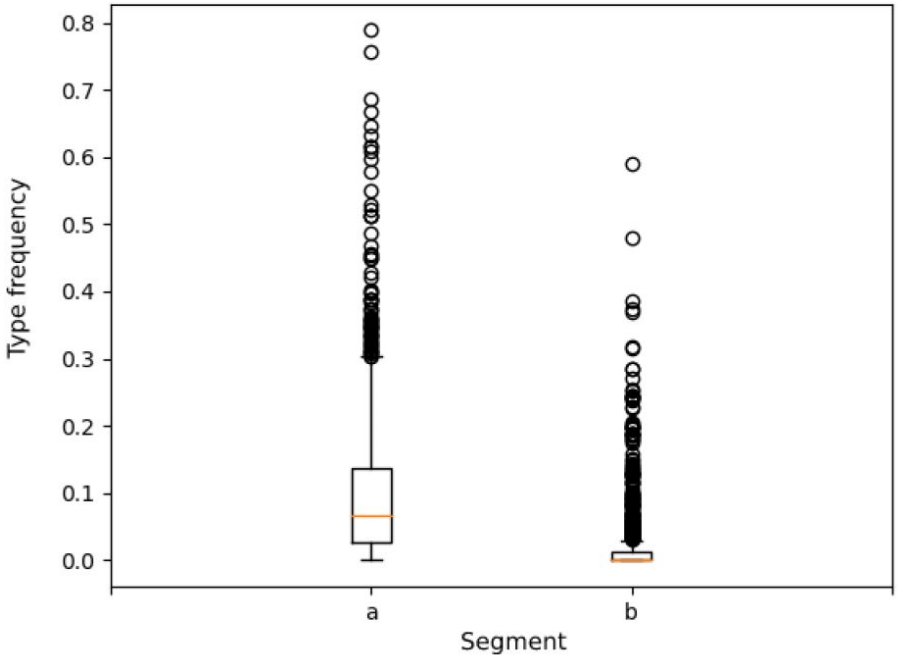
$$p_k^{t+1} := p_k^t$$

Splitwise and mergerwise bias

Each segment has a probability of being **created** or **destroyed** by sound change, which we encode as its **splitwise** and **mergerwise** bias respectively:

- Each bias is defined as a probability distribution over the segments in the language.
- When the algorithm calls for a random segment to be created in a split (destroyed in a merger), weight the choice of segment by its splitwise (mergerwise) bias.

Results

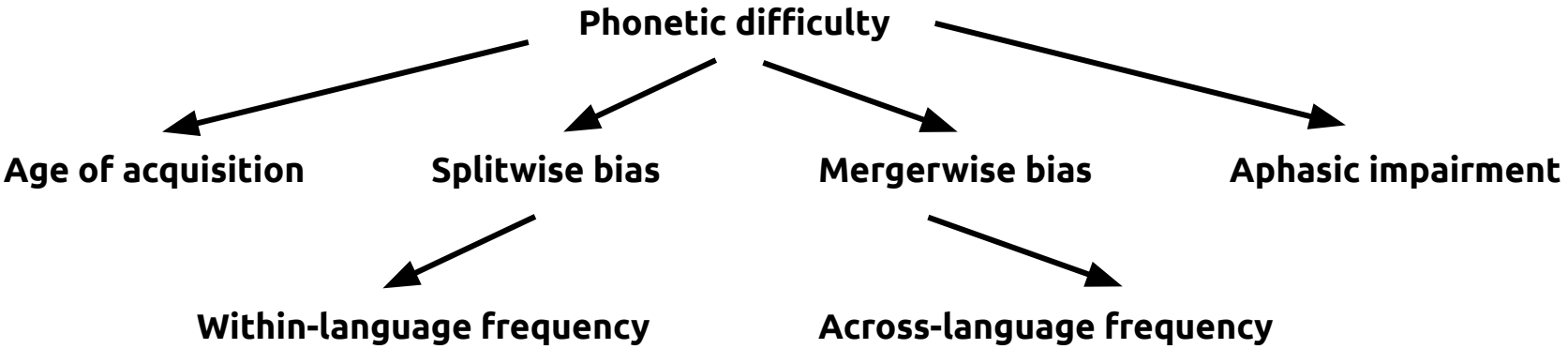


Within-language frequencies of +biased **a** and –biased **b** over 10,000 runs

Embedded explanation: markedness

A model of languages undergoing random sound change **derives** Jakobson's correlation between across-language frequency and within-language frequency.

If "acoustic/articulatory difficulty" drives sound change, age of acquisition, and impairment in aphasia, we have a potential causal diagram:



Miscellanea

Theories as notational variants

Maybe different phonological theories are expressing the same claims about language in a different **notation**?

Responses:

- This is **false** for e.g. SFP and classic OT, which allow different sets of possible languages.

- Even for theories that allow the same set of possible languages, we might be **atomists** in Graf's sense: realists about the individual atoms of our description.

Theories as notational variants

- If it were true that linguistic theories don’t make different predictions, there would be **no point** in doing linguistics! The data would give us precisely zero information about each theory.

But to the extent that two superficially different theories really are describing the same possible world, it would be a mistake to think of them as competitors.

Theories and truth

Are theories even the kind of thing that can be true or false?

For Graf's **analysts**, theories can't be "true" or "false" any more than Hausa or Tamil could be "true" or "false". An analyst's theory is just a formal language for describing **analyses**.

But what are the truth conditions of an analysis? If it's a claim about mental representations, we can ask what makes a possible representation – and we're back at formalism.

Generalizations of Bayes' rule

Bayesian inference in practice assumes we learn D with certainty. What if our certainty in D is only 0.7, and we're uncertain about the *data*? Two responses:

1. If we know **something** for certain, count that as "the data".

2. **Shrink** the probability of D by a factor of 0.7, leaving all other probabilities unchanged. (Bayes' rule is the special case where D has probability 1.)

2 is itself a special case of minimizing the **relative entropy** between the posterior and prior subject to the constraint that D has posterior probability 0.7 (Diaconis and Zabell 1982).

Probability vs explanatory power

David Deutsch (2014) gives three arguments against Bayesianism:

1. Say an explanatory theory (e.g. “the sun is powered by fusion”) has “explanatoriness” q. Its negation (“the sun is not powered by fusion”) has explanatoriness 0, not $1 - q$. So q can’t be a probability.
2. We have inconsistent explanatory theories (e.g. general relativity and quantum field theory) whose joint probability is zero, but they are still explanatory.
3. Pessimistic meta-induction: all of our current theories are false, but they are still explanatory, and their true negations are not explanatory.

Probability vs explanatory power

We've talked about responses to the pessimistic meta-induction and how false theories might differ in their gradient closeness-to-truth.

On the first point, it would be a mistake to **equate** probability with explanatory power. "2 + 2 = 4" is not maximally explanatory!

Sprenger and Hartmann (2019) list a few proposed Bayesian metrics of explanatory power.

> Example: the **Good-McGrew measure** log(P(D|H)/P(D)), which says H's explanatory power is the log of the fraction by which it increases the probability of the data D.

The problem of old evidence

Intuitively, if we discover that a theory H accounts for **existing** data D, this should be Bayesian evidence for H.

> Example: Einstein's general relativity explained an already known shift in the orbit of Mercury.

But if P(D) = 1, then P(D|H) = P(D|¬H), so D is apparently not evidence for H!

This is the **problem of old evidence**. Is it really true that theories get no points for explaining known facts?

The problem of old evidence

It's been suggested this is a problem with the assumption of **logical omniscience** (e.g. Garber 1983).

> A logically omniscient Bayesian already knows that H entails D – call this "X".
>
> A non-logically omniscient Einstein **learned** X when he worked out the implications of relativity, and P(H|X,D) > P(H|¬X,D). So it's really X that is evidence for H, not D.

Logical omniscience

We often express uncertainty about logical facts:

- “I’m not sure whether 4836 × 505 is divisible by 12”
- “The Riemann hypothesis is probably true”
- “All even numbers up to 4×10^{18} have been found to obey Goldbach’s conjecture, so the conjecture keeps looking more likely”

Is it **irrational** to think things like this?

Logical omniscience

Pettigrew (2021): Bayesianism can dispense with the assumption of logical omniscience, given computational or memory constraints.

Good's theorem says "information is valuable": you should expect yourself to make better decisions on Wednesday if you learn new information on Tuesday.

> So it's instrumentally irrational not to know logical truths if you are logically omniscient – information is "free" to you, so you're wasting potential value.
>
> But if you aren't logically omniscient, it **can** be rational to be ignorant about logical truths.