The goal of today’s class is to begin to explore one fairly influential ‘competitive’ model of pragmatic reasoning: the Rational Speech Acts model (Goodman & Frank, 2016).

RSA is a computational model that aims to formalize pragmatic reasoning processes as probabilistic inference. RSA invites us to view pragmatic reasoning as rational, goal-driven probabilistic inference, of the sort used in many different situations. In this sense, it realizes some of the key insights/claims of Grice and Lewis: that speakers are rational, that listeners rely on this observation in drawing inferences about a whole host of implicated meaning beyond the literal meaning of a linguistic expression, and that pragmatic reasoning can be understood as one sort of goal-directed action using very broad inferential principles, some of which are probably not too specific to language per se.

RSA stands out among approaches, however, in a couple of ways. First, it leverages the probability calculus as a tool to model the process of reasoning under uncertainty. The use of the probability calculus allows for a principled model of how speakers navigate uncertainty at multiple levels of representation, as we will see. Second, and as a consequence, RSA departs from previous approaches (Grice’s cooperative principle; Sperber and Wilson’s Relevance Theory; and others) in eschewing a handful of hard principles that determine the distribution and character of generalized implicatures. In doing so, it sidesteps some of the difficulties that these theories have when different maxims come into conflict, and can yield principled, graded predictions about what listeners might do in situations where the principles come into conflict.

0.1 Probability theory: Rapid primer

It’ll be useful to put into our heads some basic ideas about probability theory to get our thinking off the ground on this. A probability distribution is a function that maps events to real numbers bounded by 0 and 1. Events are subsets of a sample space, $\Omega$ that is a set of all possible outcomes we might observe.

A probability distribution function respects the three fundamental axioms of probability theory:

1. Non-negative probabilities: $P(E) \in R, P(E) \geq 0 \forall E \in F$ (where $F$ is the event space, the power set of the sample space $\Omega$)
2. Unity: $P(\Omega) = 1$
3. Disjoint additivity: $P(\cup A_i) = \sum_i A_i$ for all mutually exclusive events $A_i$

So, for example, if I flipped two coins, I would have four basic outcomes in my sample space: $\text{HH, TH, HT, TT}$. If I assume each of the coinflips is fair and balanced, then each of these outcomes is equally likely (a uniform distribution). By unity, this yields a probability of .25 to each basic outcome. With the axiom of disjoint additivity, we can calculate the probability of any arbitrary event (subset of the sample space) that we might be interested in. For example, the event of ‘getting at least one head’ corresponds to the set $\text{TH, HT, HH}$; the probability of this event is 0.75, summing across all the mutually exclusive basic outcomes that could realize this event.

A very thorny issue in probability theory is how exactly these numbers are to be interpreted. Broadly speaking, there are two camps for us to keep in mind:

4. Frequentist interpretation: The probability of some event is the long-run (asymptotic) relative frequency of that event. So if I flipped my two coins an infinity of times, I would expect that .75 of that infinite series of flips would have at least one head.

5. Bayesian interpretation: The probability of some event is subjective degree of belief in the likelihood of that event.
I suspect most of us are naive Bayesians: one important dividing line is whether or not you feel you can assign probabilities to non-repeatable events. What’s the probability that I gave a good class today? Well, an orthodox frequentist would say that this is not a well-founded notion, or that it would simply be 0 or 1 (I did, or did not, give a good class, determined post-facto). A Bayesian would permit a more nuanced stance: you might have some sense, for example, of how often I give a good class, which might in turn allow you to form a subjective estimate of the likelihood I’ll give a good class (fifty-fifty!).

This intuitive notion of naive Bayesianism is the starting point for the RSA: we can use the calculus of the probability theory to represent subjective uncertainty about different events or, as we will see, states of affairs in the world.

There’s one other thing we need to get on our radar to understand the RSA model, and that is probabilistic inference: drawing conclusions from observations in the face of uncertainty. How does the calculus of probability theory allow us to do this?

First, consider some definitions:

6 Joint probability: \( P(A, B) = P(A|B)P(B) \)

7 Joint probability, with conditional independence:
\( P(A, B) = P(A)P(B) \) (if A, B are conditionally independent)

8 Conditional probability: \( P(A|B) = \frac{P(A, B)}{P(B)} \)

Conditional probability can be rewritten using these identities to yield Bayes’ rule:

9 Bayes’ rule: \( P(A|B) = \frac{P(B|A)P(A)}{P(B)} \)

10 Bayes’ rule (w/ marginalization): \( P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)} \)

Bayes’ rule in itself is an unremarkable rearranging of some of the expressions above. It gets its ‘oomph’ from the power is gives us to draw inferences under uncertainty, and this is why you probably hear this rule hyped a bit. To see this, imagine we rewrite A and B with more evocative terms: D for data, or some set of observations, and H for hypothesis, or some conclusion we seek to evaluate in light of our observed data. Now we could write:

11 Bayes’ rule (cool version): \( P(H|D) = \frac{P(D|H)P(H)}{P(D)} \)

This expression formalizes the degree of subjective belief we should have in some hypothesis H after observing some data D. It is the normatively correct expression for this: this expression yields the probability that should be assigned to H upon seeing some data, assuming you know all of the other parts of the expression. This says that the probability of a hypothesis given some data is the probability that some data would arise under that hypothesis, multiplied with the prior probability of that hypothesis, divided by the probability of the data.

In the Bayesian framework, there are canonical terms given to each of the subexpressions in the right-hand side of Bayes’ rule:

12 Posterior probability: \( P(H|D) \)

13 Likelihood: \( P(D|H) \)

14 Prior: \( P(H) \)

15 Evidence: \( P(D) \)

If we are interested only in evaluating the relative amount of evidence in favor of a hypothesis (for example, to select the best hypothesis from a set of competing hypotheses), then we may leave out the denominator of this expression, noting that it is invariant in H. For this reason, you will often see Bayes’ rule simply abbreviated to:

16 Bayes’ rule (coolest version): \( P(H|D) \propto P(D|H)P(H) \)

For our purposes, it is interesting to note that the calculus of probability theory is automatically ‘competitive’ in a specific sense: probabilities are zero sum, and so evidence for one hypothesis is also evidence against some other hypothesis/hypotheses, because raising the probability of some hypothesis will necessarily lower the probability on other hypotheses (because of Unity).

So that’s it: Bayes rule offers the normatively rational way of using the calculus of probability to reason to conclusions (hypotheses) on the basis of data in the presence of uncertainty. From here on out, it’s all variations on a theme: drawing graded inferences from observed data using Bayes’ theorem, as applied to particular instances.
0.2 Rational Speech Acts model

Imagine you see the following display and hear: my friend has glasses. Which picture do you suppose is the friend?

You likely chose the person who only has glasses, i.e., a strengthened version of what the speaker actually said. This is an example of a type of generalized conversational implicature (Goodman and colleagues call this an ad hoc implicature because it leverages features of the visual context in an ad hoc way, rather than e.g. relying on Horn scales or the like). This kind of ad hoc implicature arises from a Gricean perspective because you as a listener are assuming I the speaker am following something like the cooperative principle. So you might expect me to follow e.g. the Maxim of Quantity. If you’re assuming I’m being informative in this way, you can recover the implication that I am talking about the person who only has glasses, not the hat and glasses.

The Rational Speech Acts takes a slightly different perspective on this process, although there is a close correspondence between this and the Gricean reasoning sketched above. The central hypothesis in the RSA perspective is that contextualized language interpretation is a recursive process wherein the listener simulates a (rational) speaker, who in turn, simulates a listener. Something like:

Critically, this involves inference under uncertainty at multiple levels, uncertainty at the listener level, at the simulated speaker level, and at the simulated listener level. Bayesian inference allows us link this uncertain chain of reasoning in a principled way.

The Rational Speech Acts model models this process as follows. Let $m$ be the message, a discrete random variable that consists of the set of all possible messages. For our purposes, let’s just suppose that these are undecomposed sentences. Let $w$ be the world, a discrete random variable that consists of the state of affairs in the world.

(17) **Messages**: $m = \{’my friend has glasses’, ’my friend has nothing’, ’my friend has a hat and glasses’\}

(18) **Worlds**: $w = \{bare guy is my friend, glasses guy is my friend, hat and glasses guy is my friend\}$

Given these two random variables (and their associated basic outcomes / sample spaces), RSA models the interpretation process as proceeding in three stages:

(19) **Pragmatic listener**: $P_{L_1}(w|m) \propto P_S(m|w)P(w)$

(20) **Pragmatic speaker**: $P_S(m|w) \propto \exp(\alpha U_S(m; w))$
(21) **Literal speaker:** \( P_{L_0}(w|m) \propto [m]^{w} P(w) \)

Our model of the listener is what is labeled above as the **Pragmatic Listener**. We will recognize this model of the listener as a simple application of Bayes rule: she estimates the probability of some world \( w \) given a message \( m \) by inverting this conditional probability according to Bayes rule. Here, then, it decomposes into the prior probability of that world (prior probability), multiplied by the probability of the message given some world. This latter probability, our Likelihood term, is understood to be the listener’s internal representation of the speaker. In this way, the listener incorporates an internal model of the speaker in her reasoning processes: she is simulating what a speaker would do if a speaker intended to select a message to communicate \( w \).

This internal simulation is what is known as the **Pragmatic Speaker**, which is the model of speaker that is rational / cooperative (and a little selfish, as we’ll see). Knowledge of how such a speaker would act allows the listener to estimate the probability of a message given some world. And although it’s a little harder to see, this is also closely related to an expression of Bayes’ rule as well:

(22) **Pragmatic speaker:** \( P_{S}(m|w) \propto \exp(\alpha U_{S}(m; w)) \)

(23) **Optimality parameter:** \( \alpha \) a free parameter that dictates ‘how rational’ a speaker will be

(24) **Utterance utility function:** \( U_{S}(m; w) = \log P_{L_0}(w|m) \)

The exponentiation in the pragmatic speaker level is what is known as a **softmax** function: it takes a vector of real numbers and normalizes them to sum to 1, such that they can be interpreted as probabilities. And the values that this function takes in are the relative likelihoods of different messages given some world. This part is crucial, and we will come back to it!

First, though, let’s wrap this up with the Literal listener. I find this name counterintuitive: the literal listener is not literally a listener: it is the internal listener model of the internal speaker model. It is the literal listener in that it does not draw any implicatures, it merely computes the semantic value of an utterance. According to the literal listener level, the probability of some world given a message is product of the prior probability of that world \( P(w) \) times the truth value of that message/world pair. If the message evaluates to true in that world, this is 1; if it evaluates to false, it is 0.

### 0.3 Example

Let’s work an example to make sure that we understand how this goes. First, you can surf over to the course webpage and retrieve `signalinggame.txt`. This is code that implements this model, and can be run at WebPPL (http://webppl.org), an online implementation of a Javascript-based probabilistic programming language. The code for this model is from Greg Scrontas’ excellent tutorial course on the Rational Speech Acts model (https://www.problang.org/).

Consider again the world and message pair in this image:

```
My friend has glasses.
```

According to this model, the probability of each world when the listener hears ‘my friend has glasses’ is: bareguy: 0, glassesguy: 0.75, hatglassesguy: 0.25. The model gives 3:1 odds that it’s the image in the middle. When the listener hears ‘my friend has nothing’, bareguy has probability 1 and everything else 0; When the listener hears ‘my friend has hat and glasses’, hatglassesguy has probability 1 and everything else 0. These latter two are straightforward to calculate by hand: when a word is incompatible with the truth-conditional semantics of an utterance, its probability is zero at the level of the literal listener. This null probability will propagate through all the levels of the pragmatic listener’s reasoning.

So let’s focus on the interesting case: ‘my friend has glasses’. Where, in this model, does the relevant ad hoc implicature come from?

First, note that in our little mini context, ‘my friend has a hat and glasses’ is a stronger statement in a couple of senses. Logically it is so: ‘my friend has a hat and glasses’ entails ‘my friend has glasses.’ It is also ‘stronger’ in the sense that it is more informative, if it is true: the set of worlds compatible with this utterance is smaller (one versus two), and so
it is more informative in that it would allow the listener to narrow down the set of possible worlds more than a weaker statement.

This intuition falls out from the probability calculus. To see this, let’s first consider the probability of different worlds for a message under the literal listener:

\[
P(w = \text{bareguy}|m = \text{‘my friend has glasses’}) = \frac{1}{3}
\]

\[
P(w = \text{glassesguy}|m = \text{‘my friend has glasses’}) = \frac{1}{3}
\]

\[
P(w = \text{hatglassesguy}|m = \text{‘my friend has glasses’}) = \frac{1}{3}
\]

This yields an improper probability distribution, because the probabilities do not sum to one. The true probability distribution is achieved by renormalizing to yield a proper probability distribution. Thus, the literal listener’s probability distribution over worlds given the utterance ‘my friend has glasses’ is the following:

\[
P(w = \text{bareguy}|m = \text{‘my friend has glasses’}) = 0
\]

\[
P(w = \text{glassesguy}|m = \text{‘my friend has glasses’}) = 0.5
\]

\[
P(w = \text{hatglassesguy}|m = \text{‘my friend has glasses’}) = 0.5
\]

Let’s stop and reflect on how this relates to more familiar Gricean principles. The process of conditioning the probability of a world given a message has an effect similar to the Gricean Maxim of Quality: say things that are true in the world. Here, if a message is not true for some world in our world set, it is simply zero.

These probabilities are then input to the pragmatic speaker level. The speaker level is thought of as a rational agent trying to choose the best out of a number of different actions; in our case, the actions are different utterances. Any given action has a utility associated with it, and the assumption of rationality of the speaker boils down to assuming that the speaker will choose the action (utterance) with the highest utility. Utility, as we will see, comes out to be related to the negative surprisal of an utterance: how much information the listener will not know about the state of affairs in the world after an utterance. This rationality assumption is the correlate of the Cooperative Principle in the RSA model: utterances with higher utility for the speaker are those that──all else being equal──communicate more information to the listener. We will revise this shortly, but it’s what we’ve got now.

Let’s walk through this. The pragmatic speaker level needs to calculate a probability distribution over messages given worlds. So now let’s suppose that the world we want to choose a message for is glassesguy. The speaker has three message options to choose from:

\[
P(m = \text{‘my friend has nothing’}|w = \text{glassesguy})
\]

\[
P(m = \text{‘my friend has glasses’}|w = \text{glassesguy})
\]

\[
P(m = \text{‘my friend has a hat and glasses’}|w = \text{glassesguy})
\]

We can calculate the utility for each of these. Let’s suppose that \( \alpha \) is 1 for now to simplify things.

\[
P(m = \text{‘my friend has nothing’}|w = \text{glassesguy}) = \exp(\log P(w = \text{glassesguy}|m = \text{‘my friend has nothing’})) = 0
\]

\[
P(m = \text{‘my friend has glasses’}|w = \text{glassesguy}) = \exp(\log P(w = \text{glassesguy}|m = \text{‘my friend has glasses’})) = 0.5
\]

\[
P(m = \text{‘my friend has a hat and glasses’}|w = \text{glassesguy}) = \exp(\log P(w = \text{glassesguy}|m = \text{‘my friend has a hat and glasses’})) = 0.5
\]

Renormalizing to respect Unity, we get:

\[
P(m = \text{‘my friend has nothing’}|w = \text{glassesguy}) = \exp(\log P(w = \text{glassesguy}|m = \text{‘my friend has nothing’})) = 0
\]

\[
P(m = \text{‘my friend has glasses’}|w = \text{glassesguy}) = \exp(\log P(w = \text{glassesguy}|m = \text{‘my friend has glasses’})) = 1
\]

\[
P(m = \text{‘my friend has a hat and glasses’}|w = \text{glassesguy}) = \exp(\log P(w = \text{glassesguy}|m = \text{‘my friend has a hat and glasses’})) = 0.5
\]

The negative log probability of an event is its surprisal. When the logarithm is in base 2, this is interpreted as the amount of bits of information (in the Shannon Information Theory sense) that are communicated by an event. Unlikely events have higher surprisal; likely events have lower surprisal. We can see that the utility is related to the negative surprisal of the world under an utterance: the negative surprisal value is higher the higher the probability of a world given an utterance, and utility tracks negative surprisal.

5
Here, the utility of the utterance ‘my friend has glasses’ is 1. But in general, the negative surprisal should decrease when an utterance is compatible with more worlds. So what would this look like if the to-be-communicated world was hatglassesguy?

\[P(m = \text{‘my friend has nothing’}|w = \text{hatglassesguy}) = \exp(\log P(w = \text{hatglassesguy}|m = \text{‘my friend has nothing’})) = 0\]

\[P(m = \text{‘my friend has glasses’}|w = \text{hatglassesguy}) = \exp(\log P(w = \text{hatglassesguy}|m = \text{‘my friend has glasses’})) = 0.5\]

\[P(m = \text{‘my friend has a hat and glasses’}|w = \text{hatglassesguy}) = \exp(\log P(w = \text{hatglassesguy}|m = \text{‘my friend has a hat and glasses’})) = 1\]

Renormalizing, we get:

\[P(m = \text{‘my friend has nothing’}|w = \text{hatglassesguy}) = \exp(\log P(w = \text{hatglassesguy}|m = \text{‘my friend has nothing’})) = 0\]

\[P(m = \text{‘my friend has glasses’}|w = \text{hatglassesguy}) = \exp(\log P(w = \text{hatglassesguy}|m = \text{‘my friend has glasses’})) = 1/3\]

\[P(m = \text{‘my friend has a hat and glasses’}|w = \text{hatglassesguy}) = \exp(\log P(w = \text{hatglassesguy}|m = \text{‘my friend has a hat and glasses’})) = 2/3\]

These are our speaker utility for these utterances given this world. From this we can see the utility of the message ‘my friend has glasses’ is higher for the glassesguy world than is the hatglassesguy world. This is because it has higher negative surprisal for the glassesguy world than the hatglassesguy world; it is a better (more informative) utterance in the former world. In this way, negative surprisal implements the Gricean Maxim of Quality.

Here’s another perspective. When a given utterance is compatible with fewer worlds, the probability mass is distributed among fewer options; if the probability mass is distributed among fewer options, all of those options have overall higher probability. So logical strength directly corresponds to higher utility (higher negative surprisal) in this model.

Finally, we arrive at the pragmatic listener level:

\[\text{Pragmatic listener: } P_L(w|m) \propto P_S(m|w) P(w)\]

If we hear the message ‘my friend has glasses’, what should we infer?

\[\text{Pragmatic listener: } P_L(w = \text{glassesguy}|\text{‘my friend has glasses’}) \propto P_S(\text{‘my friend has glasses’}|\text{glassesguy}) P(\text{glassesguy}) = 1 \times 1/3 = 0.333\]

\[\text{Pragmatic listener: } P_L(w = \text{hatglassesguy}|\text{‘my friend has glasses’}) \propto P_S(\text{‘my friend has glasses’}|\text{hatglassesguy}) P(\text{hatglassesguy}) = 1/3 \times 1/3 = 0.111\]

\[\text{Pragmatic listener: } P_L(w = \text{bareguy}|\text{‘my friend has glasses’}) \propto P_S(\text{‘my friend has glasses’}|\text{bareguy}) P(\text{bareguy}) = 0 \times 1/3 = 0\]

Renormalizing, we get:

\[\text{Pragmatic listener: } P_L(w = \text{glassesguy}|\text{‘my friend has glasses’}) \propto P_S(\text{‘my friend has glasses’}|\text{glassesguy}) P(\text{glassesguy}) = 0.75\]

\[\text{Pragmatic listener: } P_L(w = \text{hatglassesguy}|\text{‘my friend has glasses’}) \propto P_S(\text{‘my friend has glasses’}|\text{hatglassesguy}) P(\text{hatglassesguy}) = 0.25\]

\[\text{Pragmatic listener: } P_L(w = \text{bareguy}|\text{‘my friend has glasses’}) \propto P_S(\text{‘my friend has glasses’}|\text{bareguy}) P(\text{bareguy}) = 0\]

This way, we get the ‘ad hoc scalar implicature’ for this reference game.

0.4 Consequences, extensions, isssuds

OK: the model has to now simply derived aspects of the Gricean Maxims of Quality and Quantity from the probability calculus and assumptions about how the listener (and speaker) operates.

This basic model offers a framework, more than a hypothesis or theory: it is a calculus for estimating the probabilities of different interpretations given assumptions about the utterances a speaker will consider (the alternatives), about the worlds under consideration, and about what speakers take into account when speaking. The real theory, from this perspective, comes from substantive claims we might make at all levels of this model: what the worlds consist in, how truth value are determined, the alternative set of utterances, and the costs associated with different utterances.
One aspect of the model that changes quite a bit across implementations is the Pragmatic Speaker level.

(53) **Pragmatic speaker**: \( P_S(m|w) \propto \exp(\alpha U_S(m;w)) \)

(54) **Optimality parameter**: \( \alpha \) a free parameter that dictates ‘how rational’ a speaker will be

(55) **Utterance utility function**: \( U_S(m;w) = \log P_L(w|m) \)

In the basic form, there is a free parameter that dictates how rational a speaker will be. As this approaches positive infinity, a listener will tend to categorically choose the most likely interpretation as the only one; as it approaches zero, a listener will tend to choose randomly among possible meanings.

Another way in which this basic speaker model can be elaborated is by introducing a cost function:

(56) **Pragmatic speaker**: \( P_S(m|w) \propto \exp(\alpha U_S(m;w)) \)

(57) **Optimality parameter**: \( \alpha \) a free parameter that dictates ‘how rational’ a speaker will be

(58) **Utterance utility function**: \( U_S(m;w) = \log P_L(w|m) - \text{cost}(m) \)

This cost function penalizes utterances according to whatever metric you would like: in some implementations it is taken to be length, or frequency, which are supposed to correlate with ease of production. This model signalinggame.osty.txt implements this version of the speaker utility function by penalizing the number of predicates used. It can be seen that this blunts the effect of the Maxim of Quantity, making the advantage of the glasses interpretation less sharp. If there is a penalty for saying more, then this means that speakers will be somewhat less likely to draw the intended ad hoc implicature in this context.

The cost function can implement Grice’s Maxim of Manner; together with the negative surprisal of the world given some utterance it also can be seen as implementing something very close in spirit to Sperber and Wilson’s Relevance Theory, which proposes that informativity and cognitive cost trade off against each other to determine an utterance’s ‘relevance.’

Other extensions are possible. For example, the speaker utility function can be weighted to reflect what the listener assumes about the speaker’s epistemic access:

(59) **Pragmatic speaker w/ epistemic access**: \( U_S(m;k) = E_P(w|k)(U_S(m;w)) \)

This version derives speaker utility not with respect to the actual world \( w \) but with respect to the speaker’s knowledge state \( k \), which is the average utility across all worlds compatible with the speaker’s knowledge state. This interesting extension can capture the fact that a pragmatic listener will take the knowledge state of the speaker into considering when drawing scalar implicatures. Goodman and Stuhlmüller report an experiment on the rate of drawing scalar implicatures with some in which they experimentally manipulated the speaker’s epistemic state by having the speaker report that they either saw the contents of 1, 2, or 3 out of 3 envelopes. That same speaker then uttered a sentence with some like ‘Some of the envelopes have checks in them.’ An experimental participant was then asked how many of the envelopes they would bet had checks in them, based on what the speaker said:

An open issue, one we will pay close attention to next class, concerns the alternative set that is available. The predictions of the model are highly dependent on the alternatives that are considered as part of the calculation, a finding that will not surprise you if you are familiar with Gricean approaches to implicature more broadly. To see this, we can look at a model that implements scalar implicatures in WebPPL, and see what happens when we add different utterances to the alternative set. Can we understand why the alternatives matter the way they do, in the context of this model?

0.5 Back to pronouns

Schulz and colleagues (submitted) offer one example of how the RSA framework can be applied to the interpretation of pronominal expressions. The starting point for their investigation is a cross-linguistic comparison of the preferred interpretation of ambiguous pronouns in English, French
and German. Consider the following three sentences, all translations of each other:

(60) **English**: The postman met the streetsweeper before he went home.

(61) **French**: Le facteur a rencontré le balayeur avant qu’il rentre à la maison.

(62) **German**: Der Briefträger hat den Straßenfeger getroffen bevor er nach Hause ging.

English and German speakers show a systematic preference to resolve the ambiguous pronoun to the subject of the matrix clause in contexts like this one, above and beyond the particulars of any lexical content. Interestingly, French speakers show a different pattern: they show a systematic object bias. The starting point for Schulz et al’s investigation it the question of why this should be so.

French seems to be the odd one out here: English and German behave like what you would expect cross-linguistically: subject pronouns show a strong preference for first-mentioned, subject, topic antecedents, so we might have expected the German or English pattern based simply on these robust cross-linguistic patterns. So what is up with French?

The idea that this group has pursued in earlier work is that this preference results from the application of the Gricean Maxim of Manner/Quantity. They note that there is a more minimal structure in French that unambiguously codes the subject reference:

(63) **French, finite**: Le facteur a rencontré le balayeur avant qu’il rentre à la maison.

(64) **French, nonfinite**: Le facteur a rencontré le balayeur avant de rentrer à la maison.

The non-finite variant has a phonologically null PRO element that is obligatorily bound by the subject; German does not have this non-finite variant. French speakers might reason pragmatically that the object preference is better because if the speaker had meant the subject construal of the pronoun, she would have used the more minimal, less ambiguous variant.

Cool idea, but it really quickly runs into a problem. English has the same structural alternative as French, but it patterns like German in having a subject pronoun preference.

Why oh why doesn’t the same blocking effect arise in English as in French? The hypothesis that Schulz and colleagues make is that the alternatives are weighted by their frequency, and that this is where French and English diverge. French and English are the same in terms of what alternatives are technically possible; however, frequency of occurrence creates a cline of availability across these different alternatives, and this is ultimately responsible for the different preferences.

Indeed, this seems to be borne out in a corpus study they carried out. The English study was done using COCA, which has upwards of 570 million part-of-speech tagged words spanning 1990 to (at the time of this study) 2017. Importantly, this corpus can be subdivided into different genres: Fiction, Newspaper, and Spoken are the relevant sections for this study. The French corpora were more diverse, as there is not single resource comparable to COCA for French. They used three corpora: Frantext for Fiction-like genre, Est Républicain for Journalistic-like genre, and the Enquête sociolinguistique à Orléans for Spoken-like genre.

Here are the results:

<table>
<thead>
<tr>
<th></th>
<th>Spoken</th>
<th>Fiction</th>
<th>Newspaper</th>
</tr>
</thead>
<tbody>
<tr>
<td>finite construction</td>
<td>12,930</td>
<td>15,058</td>
<td>12,244</td>
</tr>
<tr>
<td>alternative construction</td>
<td>8,523</td>
<td>7,846</td>
<td>27,021</td>
</tr>
<tr>
<td>Ratio non-finite/finite</td>
<td>0.66</td>
<td>0.52</td>
<td>2.21</td>
</tr>
<tr>
<td>Proportion finite</td>
<td>0.60</td>
<td>0.66</td>
<td>0.31</td>
</tr>
</tbody>
</table>

*Table 3: Corpus counts of sentence-final after in English.*
The results of the corpus search reveal that English speakers use the finite version more often than the non-finite version, and this preference is more pitched for before than it is for after. This trend is similar across genres, though it is reversed in the Newspaper genre. French exhibits the reverse pattern: all genres show a preference for a non-finite construction, but especially so in the Newspaper genres.

Frequency based explanations can be a bit unsatisfying: in the end, we do need to say something about where the corpus frequencies come from! The different frequencies across languages Schulz et al. attribute to competition indicative and subjunctive mood in the finite subordinate clauses, where there is a tension between the prescriptively correct mood (subjonctif for avant, indicatif for après), and what speakers actually tend to use (the reverse). This uncertainty creates, where possible, an avoidance strategy on the part of the French speaker which yields the frequency counts observed.

Schulz et al. then develop an RSA model of this process. You can play with it by retrieving rsapronouns.txt from the course website and putting it into WebPPL.

Here’s how their model is specified:

(67) **Messages**: \( m = \{ \text{‘finite subordinate clause’, ‘nonfinite subordinate clause’} \} \)

(68) **Worlds**: \( w = \{ \text{subject is antecedent, object is antecedent} \} \)

Note that the two alternatives differ in their informativity (negative surprisal). Finite subordinate clause is compatible with either world; non-finite subordinate clause is only compatible with the first. Negative sur-
prisal should, all else being equal, create an object bias for pronouns in both French and English. In other words, English is strange in not patterning like French, although I presented it the other way around! (See also Levinson, 1987, for similar puzzling examples).

All else is not equal. Schulz et al. propose two different ways in which the competition is biased: in the costs associated with different messages / utterances, and in the prior probabilities of different worlds. The costs are taken directly from the corpus search, and are the logged ratio of the two structures for each language and connector. The priors come from previous work on English, which suggests that subject interpretations are more common (discovering this independent of the interpretive preferences for pronouns is tricky; on Kehler and Rohde’s 2013 analysis this next mention bias might be ultimately driven by an implicit QUD). Here is how each model is parameterized.

<table>
<thead>
<tr>
<th></th>
<th>English before</th>
<th>English after</th>
</tr>
</thead>
<tbody>
<tr>
<td>Messages</td>
<td>Finite (before he went...)</td>
<td>Finite (after he went...)</td>
</tr>
<tr>
<td></td>
<td>Non-finite (before going...)</td>
<td>Non-finite (after going...)</td>
</tr>
<tr>
<td>Costs</td>
<td>C(finite) = -0.117</td>
<td>C(finite) = -0.511</td>
</tr>
<tr>
<td></td>
<td>C(non-finite) = -2.207</td>
<td>C(non-finite) = -0.916</td>
</tr>
<tr>
<td>Prior beliefs</td>
<td>Pr(w₁) = 0.8</td>
<td>Pr(w₁) = 0.8</td>
</tr>
<tr>
<td></td>
<td>Pr(w₂) = 0.2</td>
<td>Pr(w₂) = 0.2</td>
</tr>
<tr>
<td>α</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Predictions for pronoun resolution in finite</td>
<td>P₁(w₁</td>
<td>finite) = 0.79 (subject)</td>
</tr>
<tr>
<td></td>
<td>P₁(w₂</td>
<td>finite) = 0.21 (object)</td>
</tr>
<tr>
<td></td>
<td>P₂(w₁</td>
<td>finite) = 0.36 (subject)</td>
</tr>
<tr>
<td></td>
<td>P₂(w₂</td>
<td>finite) = 0.64 (object)</td>
</tr>
</tbody>
</table>

Table 10: Comparison of models for French avant and après

Here are the resulting model predictions and data:

<table>
<thead>
<tr>
<th></th>
<th>French avant</th>
<th>French après</th>
</tr>
</thead>
<tbody>
<tr>
<td>Messages</td>
<td>Finite (avant que...)</td>
<td>Finite (après que...)</td>
</tr>
<tr>
<td></td>
<td>Non-finite (avant de...)</td>
<td>Non-finite (VOIR/ÊTRE)</td>
</tr>
<tr>
<td>Costs</td>
<td>C(finite) = -1.309</td>
<td>C(finite) = -1.470</td>
</tr>
<tr>
<td></td>
<td>C(non-finite) = -0.315</td>
<td>C(non-finite) = -0.261</td>
</tr>
<tr>
<td>Prior beliefs</td>
<td>Pr(w₁) = 0.8</td>
<td>Pr(w₁) = 0.8</td>
</tr>
<tr>
<td></td>
<td>Pr(w₂) = 0.2</td>
<td>Pr(w₂) = 0.2</td>
</tr>
<tr>
<td>α</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Predictions for pronoun resolution in finite</td>
<td>P₁(w₁</td>
<td>finite) = 0.36 (subject)</td>
</tr>
<tr>
<td></td>
<td>P₁(w₂</td>
<td>finite) = 0.64 (object)</td>
</tr>
</tbody>
</table>

Table 11: Comparison of models for English before and after
The model captures the major cross-linguistic difference: more object choices in French than in English (although note that in French they did not find an object preference per se). There are some unexplained features, such as the cross-over interaction between connector and language.