Probabilistic Semantics and Pragmatics: Uncertainty in Language and Thought

Noah D. Goodman and Daniel Lassiter

Stanford University
{ngoodman,danlassiter}@stanford.edu

Language is used to communicate ideas. Ideas are mental tools for coping with a complex and uncertain world. Thus human conceptual structures should be key to language meaning, and probability—the mathematics of uncertainty—should be indispensable for describing both language and thought. Indeed, probabilistic models are enormously useful in modeling human cognition (Tenenbaum *et al.*, 2011) and aspects of natural language (Bod *et al.*, 2003; Chater *et al.*, 2006). With a few early exceptions (e.g. Adams, 1975; Cohen, 1999b), probabilistic tools have only recently been used in natural language semantics and pragmatics. In this chapter we synthesize several of these modeling advances, exploring a formal model of interpretation grounded, via lexical semantics and pragmatic inference, in conceptual structure.

Flexible human cognition is derived in large part from our ability to imagine possibilities (or possible worlds). A rich set of concepts, intuitive theories, and other mental representations support imagining and reasoning about possible worlds—together we will call these the conceptual lexicon. We posit that this collection of concepts also forms the set of primitive elements available for lexical semantics: word meanings can be built from the pieces of conceptual structure. Larger semantic structures are then built from word meanings by composition, ultimately resulting in a sentence meaning which is a phrase in the "language of thought" provided by the conceptual lexicon. This expression is truth-functional in that it takes on a Boolean value for each imagined world, and it can thus be used as the basis for belief updating. However, the connection between cognition, semantics, and belief is not direct: because language must flexibly adapt to the context of communication, the connection between lexical representation and interpreted meaning is mediated by pragmatic inference.

A draft chapter for the Wiley-Blackwell *Handbook of Contemporary Semantics — second edition*, edited by Shalom Lappin and Chris Fox. This draft formatted on 25th June 2014.

There are a number of challenges to formalizing this view of language: How can we formalize the conceptual lexicon to describe generation of possible worlds? How can we appropriately connect lexical meaning to this conceptual lexicon? How, within this system, do sentence meanings act as constraints on possible worlds? How does composition within language relate to composition within world knowledge? How does context affect meanings? How is pragmatic interpretation related to literal meaning?

In this chapter we sketch an answer to these questions, illustrating the use of probabilistic techniques in natural language pragmatics and semantics with a concrete formal model. This model is not meant to exhaust the space of possible probabilistic models—indeed, many extensions are immediately apparent—but rather to show that a probabilistic framework for natural language is possible and productive. Our approach is similar in spirit to cognitive semantics (Jackendoff, 1983; Lakoff, 1987; Cruse, 2000; Taylor, 2003), in that we attempt to ground semantics in mental representation. However, we draw on the highly successful tools of Bayesian cognitive science to formalize these ideas. Similarly, our approach draws heavily on the progress made in formal model-theoretic semantics (Lewis, 1970; Montague, 1973; Gamut, 1991; Heim & Kratzer, 1998; Steedman, 2001), borrowing insights about how syntax drives semantic composition, but we compose elements of stochastic logics rather than deterministic ones. Finally, like game-theoretic approaches (Benz *et al.*, 2005; Franke, 2009), we place an emphasis on the the refinement of meaning through interactional, pragmatic reasoning.

In section 1 we provide background on probabilistic modeling and stochastic λ-calculus, and introduce a running example scenario: the game of tug-of-war. In section 2 we provide a model of literal interpretation of natural language utterances and describe a formal fragment of English suitable for our running scenario. Using this fragment we illustrate the emergence of non-monotonic effects in interpretation and the interaction of ambiguity with background knowledge. In section 3 we describe pragmatic interpretation of meaning as probabilistic reasoning about an informative speaker, who reasons about a literal listener. This extended notion of interpretation predicts a variety of implicatures and connects to recent quantitative experimental results. In section 4 we discuss the role of semantic indices in this framework and show that binding these indices at the pragmatic level allows us to deal with several issues in context-sensitivity of meaning, such as the interpretation of scalar adjectives. We conclude with general comments about the role of uncertainty in pragmatics and semantics.

1 Probabilistic models of commonsense reasoning

Uncertainty is a key property of the world we live in. Thus we should expect reasoning with uncertainty to be a key operation of our cognition. At the same time our world is built from a complex web of causal and other structures, so we expect structure within our representations of uncertainty. Structured knowledge of an uncertain world can be naturally captured by *generative* models, which make it possible to flexibly imagine (simulate) possible worlds in proportion to their likelihood. In this section, we first introduce the basic operations for dealing with uncertainty—degrees of belief and probabilistic conditioning. We then introduce formal tools for adding compositional structure to these models—the stochastic λ-calculus—and demonstrate how these tools let us build generative models of the world and capture commonsense reasoning. In later sections, we demonstrate how these tools can be used to provide new insights into issues in natural language semantics and pragmatics.

Probability is fundamentally a system for manipulating degrees of belief. The probability[1] of a proposition is simply a real number between 0 and 1 describing an agent's degree of belief in that proposition. More generally, a *probability distribution* over a *random variable* A is an assignment of a probability $P(A{=}a)$ to each of a set of exhaustive and mutually exclusive outcomes a, such that $\sum_a P(A{=}a) = 1$. The joint probability $P(A{=}a, B{=}b)$, of two random variable values is the degree of belief we assign to the proposition that both $A{=}a$ and $B{=}b$. From a joint probability distribution, $P(A{=}a, B{=}b)$, we can recover the *marginal* probability distribution on A: $P(A{=}a) = \sum_b P(A{=}a, B{=}b)$.

The fundamental operation for incorporating new information, or assumptions, into prior beliefs is *probabilistic conditioning*. This operation takes us from the *prior* probability of A, $P(A)$, to the *posterior* probability of A given proposition B, written $P(A|B)$. Conditional probability can be defined, following Kolmogorov (1933), by:

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{1}$$

This unassuming definition is the basis for much recent progress in modeling human reasoning (e.g. Oaksford & Chater, 2007; Griffiths *et al.*, 2008; Chater & Oaksford, 2008; Tenenbaum *et al.*, 2011). By modeling uncertain beliefs in probabilistic terms, we can understand reasoning as probabilistic conditioning. In particular, imagine a person who is trying to establish which hypothesis $H \in \{h_1, \ldots, h_m\}$ best explains a situation, and does so on the basis of a

[1] In describing the mathematics of probabilities we will presume that we are dealing with probabilities over discrete domains. Almost everything we say applies equally well to probability densities, and more generally probability measures, but the mathematics becomes more subtle in ways that would distract from our main objectives.

series of observations $\{o_i\}_{i=1}^N$. We can describe this inference as the conditional probability:

$$P(H|o_1,\ldots,o_N) = \frac{P(H)P(o_1,\ldots,o_N|H)}{P(o_1,\ldots,o_N)}. \tag{2}$$

This useful equality is called *Bayes' rule*; it follows immediately from the definition in equation 1. If we additionally assume that the observations provide no information about each other beyond what they provide about the hypothesis, that is they are *conditionally independent*, then $P(o_i|o_j,H) = P(o_i|H)$ for all $i \neq j$. It follows that:

$$P(H|o_1,\ldots,o_N) = \frac{P(H)P(o_1|H)\cdots P(o_N|H)}{P(o_1)\cdots P(o_N|o_1,\ldots,o_{N-1})} \tag{3}$$

$$= \frac{P(H)P(o_1|H)\cdots P(o_N|H)}{\sum_{H'} P(o_1|H')P(H')\cdots \sum_{H'} P(o_N|H')P(H'|o_1,\ldots,o_{N-1})}. \tag{4}$$

From this it is a simple calculation to verify that we can perform the conditioning operation sequentially rather than all at once: the *a posteriori* degree of belief given observations $o_1,\ldots,o_i$ becomes the *a priori* degree of belief for incorporating observation o_{i+1}. Thus, when we are justified in making this conditional independence assumption, understanding the impact of a sequence of observations reduces to understanding the impact of each one separately. Later we will make use of this idea to reduce the meaning of a stream of utterances to the meanings of the individual utterances.

1.1 Stochastic λ-Calculus and Church

Probability as described so far provides a notation for manipulating degrees of belief, but requires that the underlying probability distributions be specified separately. Frequently we wish to describe complex knowledge involving relations among many non-independent propositions or variables, and this requires describing complex joint distributions. We could write down a probability for each combination of variables directly, but this quickly becomes unmanageable—for instance, a model with n binary variables requires $2^n - 1$ probabilities. The situation is parallel to deductive reasoning in classical logic via truth tables (extensional models ascribing possibility to entire worlds), which requires a table with 2^n rows for a model with n atomic propositions; this is sound, but opaque and inefficient. Propositional logic provides structured means to construct and reason about knowledge, but is still too coarse to capture many patterns of interest. First- and higher-order logics, such as λ-calculus, provide a fine-grained language for describing and reasoning about (deterministic) knowledge. The *stochastic* λ-calculus (SLC) provides a formal, compositional language for describing probabilities about complex sets of interrelated beliefs.

At its core SLC simply extends the (deterministic) λ-calculus (Barendregt, 1985; Hindley & Seldin, 1986) with an expression type $(L \oplus R)$, indicating random choice between the sub-expressions L and R, and an additional reduction

rule that reduces such a choice expression to its left or right sub-expression with equal probability. A sequence of standard and random-choice reductions results in a new expression and some such expressions are in normal form (i.e. irreducible in the same sense as in λ-calculus); unlike λ-calculus, the normal form is not unique. The reduction process can be viewed as a distribution over reduction sequences, and the subset which terminate in a normal-form expression induces a (sub-)distribution over normal-form expressions: SLC expressions denote (sub-)distributions over completely reduced SLC expressions. It can be shown that this system can represent any computable distribution (see for example Ramsey & Pfeffer, 2002; Freer & Roy, 2012).

The SLC thus provides a fine-grained compositional system for specifying probability distributions. We will use it as the core representational system for conceptual structure, for natural language meanings, and (at a meta-level) for specifying the architecture of language understanding. However, while SLC is simple and universal, it can be cumbersome to work with directly. Goodman *et al.* (2008a) introduce Church, an enriched SLC that can be realized as a *probabilistic programming language*—parallel to the way that the programming language LISP is an enriched λ-calculus. In later sections we will use Church to actually specify our models of language and thought. Church starts with the pure subset of Scheme (which is itself essentially λ-calculus enriched with primitive data types, operators, and useful syntax) and extends it with elementary random primitives (ERPs), the inference function `query`, and the memoization function `mem`. We must take some time to describe these key, but somewhat technical, pieces of Church before turning back to model construction. Further details and examples of using Church for cognitive modeling can be found at `http://probmods.org`. In what follows we will assume passing familiarity with the Polish notation used in LISP-family languages (fully parenthesized and operator initial), and will occasionally build on ideas from programming languages—Abelson & Sussman (1983) is an excellent background on these ideas.

Rather than restricting to the $\oplus$ operation of uniform random choice (which is sufficient, but results in extremely cumbersome representations), Church includes an interface for adding elementary random primitives (ERPs). These are procedures that return random values; a sequence of evaluations of such an ERP procedure is assumed to result in independent identically distributed (i.i.d.) values. Common ERPs include `flip` (i.e. Bernoulli), `uniform`, and `gaussian`. While the ERPs themselves yield i.i.d. sequences, it is straightforward to construct Church procedures using ERPs that do not. For instance `((λ (bias) (λ () (flip bias))) (uniform 0 1))` creates a function that "flips a coin" of a specific but unknown `bias`. Multiple calls to the function will result in a sequence of values which are not i.i.d., because they jointly depend on the unknown `bias`. This illustrates how more complex distributions can be built by combining simple ones.

To represent conditional probabilities in SLC and Church we introduce the `query` function. Unlike simpler representations (such as Bayes nets) where

conditioning is an operation that happens to a model from the outside, `query` can be defined within the SLC itself as an ordinary function. One way to do this is via *rejection sampling*. Imagine we have a distribution represented by the function with no arguments `thunk`, and a predicate on return values `condition`. We can represent the conditional distribution of return values from `thunk` that satisfy `condition` by:

```
(define conditional
  (λ ()
    (define val (thunk))
    (if (condition val) val (conditional))))
```

where we have used a stochastic recursion (conveniently specified by the named define) to build a conditional. Conceptually this recursion samples from `thunk` until a value is returned that satisfies `condition`; it is straightforward to show that the distribution over return values from this procedure is exactly the ratio used to define conditional probability in equation 1 (when both are defined). That is, the `conditional` procedure samples from the conditional distribution that could be notated $P(\texttt{(thunk)}=\texttt{val}|\texttt{(condition val)}=\texttt{True})$. For parsimony, Church uses a special syntax, `query`, to specify such conditionals:

```
(query
  ...definitions...
  qexpr
  condition)
```

where `...definitions...` is a list of definitions, `qexpr` is the expression of interest whose value we want, and `condition` is a condition expression that must return `true`. This syntax is internally transformed into a thunk and predicate that can be used in the rejection sampling procedure:

```
(define thunk (λ () ...definitions... (list condition qexpr)))
(define predicate (λ (val) (equal? true (first val))))
```

Rejection sampling can be taken as the definition of the `query` interface, but it is very important to note that other implementations that approximate the same distribution can be used and will often be more efficient. For instance, see Wingate *et al.* (2011) for alternative implementations of `query`. In this chapter we are concerned with the computational (or competence) level of description and so need not worry about the implementation of `query` in any detail.

Memoization is a higher-order function that upgrades a stochastic function to have persistent randomness—a memoized function is evaluated fully the first time it is called with given arguments, but thereafter returns this "stored" value. For instance `(equal? (flip) (flip))` will be `true` with probability 0.5, but if we define a memoized flip, `(define memflip (mem flip))`, then `(equal? (memflip) (memflip))` will always be `true`. This property is convenient for representing probabilistic dependencies between beliefs that rely on common properties, for instance the strengths and genders of people in a game (as illustrated below). For instance, memoizing a function `gender` which maps individuals to their gender will ensure that gender is a stable property, even if it is not known

in advance what a given individual's gender is (or, in effect, which possible world is actual).[2]

In Church, as in most LISP-like languages, source code is a first-class data type: it is represented by lists. The quote operator tells the evaluation process to treat a list as a literal list of symbols, rather than evaluating it: `(flip)` results in a random value `true` or `false`, while `'(flip)` results in the list `(flip)` as a *value*. For us this will be important because we can "reverse" the process by calling the `eval` function on a piece of reified code. For instance, `(eval '(flip))` results in a random value `true` or `false` again. Usefully for us, evaluation triggered by `eval` happens in the local context with any bound variables in scope. For instance:

```
(define expression '(flip bias))
(define foo ((λ (bias) (λ (e) (eval e))) (uniform 0 1)))
(foo expression)
```

In this snippet the variable `bias` is not in scope at the top level where `expression` is defined, but it is in scope where `expression` is evaluated, inside the function bound to `foo`. For the natural language architecture described below this allows utterances to be evaluated in the local context of comprehension. For powerful applications of these ideas in natural language semantics see Shan (2010).

Church is a dynamically typed language: values have types, but expressions don't have fixed types that can be determined *a priori*. One consequence of dynamic typing for a probabilistic language is that expressions may take on a distribution of different types. For instance, the expression `(if (flip) 1 true)` will be an integer half the time and Boolean the other half. This has interesting implications for natural language, where we require consistent dynamic types but have no particular reason to require deterministically assigned static types. For simplicity (and utility below) we assume that when an operator is applied to values outside of its domain, for instance `(+ 1 'a)`, it returns a special value *error* which is itself outside the domain of all operators, except the equality operator `eq?`. By allowing `eq?` to test for error we permit very simple error handling, and allow `query` (which relies on a simple equality test to decide whether to "keep going") to filter out mis-typed sub-computations.

1.2 Commonsense knowledge

In this chapter we use sets of stochastic functions in Church to specify the intuitive knowledge—or theory—that a person has about the world. To illustrate this idea we now describe an example, the tug-of-war game, which we will use later in the chapter as the non-linguistic conceptual basis of a semantics

[2] A technical, but important, subtlety concerns the "location" where a memoized random choice is created: should it be at the first use, the second, ...? In order to avoid an artificial symmetry breaking (and for technical reasons), the semantics of memoization is defined so that all random values that may be returned by a memoized function are created when the memoized function is *created*, not where it is called.

and pragmatics for a small fragment of English. Tug-of-war is a simple game in which two teams pull on either side of a rope; the team that pulls hardest will win. Our intuitive knowledge of this domain (and indeed most similar team games) rests on a set of interrelated concepts: players, teams, strength, matches, winners, etc. We now sketch a simple realization of these concepts in Church. To start, each player has some traits, strength and gender, that may influence each other and his or her contribution to the game.

```
(define gender (mem (λ (p) (if (flip) 'male 'female))))
(define gender-mean-strength (mem (λ (g) (gaussian 0 2))))
(define strength
  (mem (λ (p) (gaussian (gender-mean-strength (gender p)) 1))))
```

We have defined the strength of a person as a *mixture model*: strength depends on a latent class, gender, through the (a priori unknown) gender means. Note that we are able to describe the properties of people (`strength`, `gender`) without needing to specify the people—instead we assume that each person is represented by a unique symbol, using memoized functions from these symbols to properties to create the properties of a person only when needed (but then hold those properties persistently). In particular, the person argument, `p`, is never used in the function `gender`, but it matters because the function is memoized—a gender will be persistently associated to each person even though the distribution of genders doesn't depend on the person. We will exploit this pattern often below. We are now already in a position to make useful inferences. We could, for instance observe the strengths and genders of several players, and then Pat's strength but not gender, and ask for the latter:

```
(query
  (define gender (mem (λ (p) (if (flip) 'male 'female))))
  (define gender-mean-strength (mem (λ (g) (gaussian 0 2))))
  (define strength
    (mem (λ (p) (gaussian (gender-mean-strength (gender p)) 1))))

  (gender 'Pat)

  (and (equal? (gender 'Bob) 'male) (= (strength 'Bob) -1.1)
       (equal? (gender 'Jane) 'female) (= (strength 'Jane) 0.5)
       (equal? (gender 'Jim) 'male) (= (strength 'Jim) -0.3)
       (= (strength 'Pat) 0.7)))
```

The result of this query is that Pat is more likely to be female than male (probability .63). This is because the observed males are weaker than Jane, the observed female, and so a strong player such as Pat is likely to be female as well.

In the game of tug-of-war players are on teams:

```
(define players '(Bob Jim Mary Sue Bill Evan Sally Tim Pat Jane Dan Kate))
(define teams '(team1 team2 ... team10))

(define team-size (uniform-draw '(1 2 3 4 5 6)))
(define players-on-team (mem (λ (team) (draw-n team-size players))))
```

Here the `draw-n` ERP draws uniformly but without replacement from a list. (For simplicity we draw players on each team independently, allowing players

to potentially be on multiple teams.) In addition to players and teams, we have matches: events that have two teams and a winner. The winner depends on how hard each team is pulling, which depends on how hard each team member is pulling.

```
(define teams-in-match (mem (λ (match) (draw-n 2 teams))))
(define players-in-match (λ (match) (apply append (map players-on-team
    (teams-in-match match)))))
(define pulling (mem (λ (player match)
                       (+ (strength player) (gaussian 0 0.5)))))
(define team-pulling (mem (λ (team match)
    (sum (map (λ (p) (pulling p match)) (players-on-team team))))))
(define (winner match)
  (define teamA (first (teams-in-match match)))
  (define teamB (second (teams-in-match match)))
  (if (> (team-pulling teamA) (team-pulling teamB)) teamA teamB))
```

Notice that the team pulling is simply the sum of how hard each member is pulling; each player pulls with their intrinsic strength, plus or minus a random amount that indicates their effort on this match.

```
(define players '(Bob Jim Mary Sue Bill Evan Sally Tim Pat Jane Dan Kate))
(define teams '(team1 team2 ... team10))
(define matches '(match1 match2 match3 match4))
(define individuals (append players teams matches))

(define gender (mem (λ (p) (if (flip) 'male 'female))))
(define gender-mean-strength (mem (λ (g) (gaussian 0 2))))
(define strength (mem (λ (p) (gaussian (gender-mean-strength (gender p))
    1))))

(define team-size (uniform-draw '(1 2 3 4 5 6)))
(define players-on-team (mem (λ (team) (draw-n team-size players))))

(define teams-in-match (mem (λ (match) (draw-n 2 teams))))
(define players-in-match (λ (match) (apply append (map players-on-team
    (teams-in-match match)))))
(define pulling (mem (λ (player match) (+ (strength player) (gaussian 0
    0.5)))))
(define team-pulling (mem (λ (team match)
    (sum (map (λ (p) (pulling p match)) (players-on-team team))))))
(define (winner match)
  (let ([teamA (first (teams-in-match match))]
        [teamB (second (teams-in-match match))])
    (if (> (team-pulling teamA match) (team-pulling teamB match))
        teamA
        teamB)))
```

Figure 1. The collected Church definitions forming our simple intuitive theory (or conceptual lexicon) for the tug-of-war domain.

The intuitive theory, or conceptual lexicon of functions, for the tug-of-war domain is given altogether in Figure 1. A conceptual lexicon like this one describes generative knowledge about the world—interrelated concepts that can be used to describe the causal story of how various observations come

to be. We can use this knowledge to reason from observations to predictions or latent states by conditioning (i.e. `query`). Let us illustrate how a generative model is used to capture key patterns of reasoning. Imagine that Jane is playing Bob in match 1; we can infer Jane's strength before observing the outcome of this match:

```
(query
  ...ToW theory...
  (strength 'Jane) ;; variable of interest
  (and ;; conditioning expression
    (equal? (players-on-team 'team1) '(Jane))
    (equal? (players-on-team 'team2) '(Bob))
    (equal? (teams-in-match 'match1) '(team1 team2))))
```

In this and all that follows `...ToW theory...` is an abbreviation for the definitions in Figure 1. The result of this inference is simply the prior belief about Jane's strength: a distribution with mean 0 (Figure 2). Now imagine that Jane wins this match:

```
(query
  ...ToW theory...
  (strength 'Jane) ;; variable of interest
  (and ;; conditioning expression
    (equal? (players-on-team 'team1) '(Jane))
    (equal? (players-on-team 'team2) '(Bob))
    (equal? (teams-in-match 'match1) '(team1 team2))
    (equal? (winner 'match1) 'team1)))
```

If we evaluate this query we find that Jane is inferred to be relatively strong: her mean strength after observing this match is around 0.7, higher than her *a priori* mean strength of 0.0.

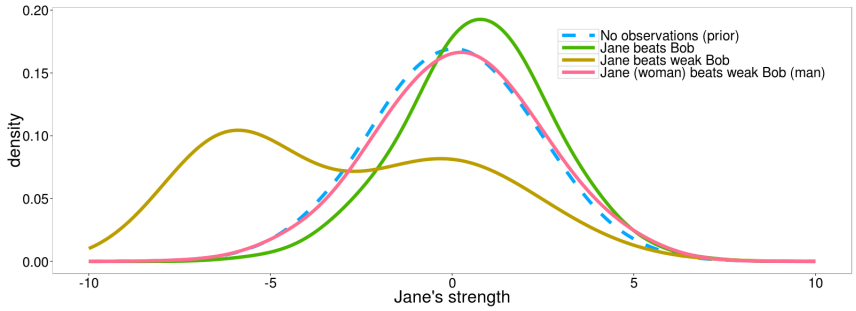

Figure 2. An example of explaining away. Lines show the distribution on Jane's inferred strength after (a) no observations; (b) observing that Jane beat Bob, whose strength is unknown; (c) learning that Bob is very weak, with strength -8. (d) learning that Jane and Bob are different genders

However, imagine that we then learned that Bob is a weak player:

```
(query
```

```
  ...ToW theory...
  (strength 'Jane) ;; variable of interest
  (and ;; conditioning expression
    (equal? (players-on-team 'team1) '(Jane))
    (equal? (players-on-team 'team2) '(Bob))
    (equal? (teams-in-match 'match1) '(team1 team2))
    (equal? (winner 'match1) 'team1)
    (= (strength 'Bob) -8.0)))
```

This additional evidence has a complex effect: we know that Bob is weak, and this provides evidence that the mean strength of his gender is low; if Jane is the same gender, she is also likely weak, though stronger than Bob, who she beat; if Jane is of the other gender, then we gain little information about her. The distribution over Jane's strength is bimodal because of the uncertainty about whether she has the same gender as Bob. If we knew that Jane and Bob were of different genders then information about the strength of Bob's gender would not affect our estimate about Jane:

```
(query
  ...ToW theory...
  (strength 'Jane) ;; variable of interest
  (and ;; conditioning expression
    (equal? (players-on-team 'team1) '(Jane))
    (equal? (players-on-team 'team2) '(Bob))
    (equal? (teams-in-match 'match1) '(team1 team2))
    (equal? (winner 'match1) 'team1)
    (= (strength 'Bob) -8.0)
    (equal? (gender 'Bob) 'male)
    (equal? (gender 'Jane) 'female)))
```

Now we have very little evidence about Jane's strength: the inferred mean strength from this query goes back to (almost) 0, because we gain no information via gender mean strengths, and Jane beating Bob provides little information given that Bob is very weak. This is an example of *explaining away* (Pearl, 1988): the assumption that Bob is weak has explained the observation that Jane beat Bob, which otherwise would have provided evidence that Jane is strong. Explaining away is characterized by *a priori* independent variables (such as Jane and Bob's strengths) becoming coupled together by an observation (such as the outcome of match 1). Another way of saying this is that our knowledge of the world, the generative model, can have a significant amount of modularity; our inferences after making observations will generally not be modular in this way. Instead, complex patterns of influence can couple together disparate pieces of the model. In the above example we also have an example of *screening off*: the observation that Bob and Jane are of different genders renders information about Bob's (gender's) strength uninformative about Jane's. Screening off describes the situation when two variables that were *a priori* dependent become independent after an observation (in some sense the opposite of explaining away). Notice that in this example we have gone through a non-monotonic reasoning sequence: Our degree of belief that Jane is strong went up from the first piece of evidence, down below the prior from the second, and then back up from the third.

Such complex, non-monotonic patterns of reasoning are extremely common in probabilistic inference over structured models.

There are a number of other patterns of reasoning that are common results of probabilistic inference over structured models, including Occam's razor (complexity of hypotheses is automatically penalized), transfer learning (an inductive bias learned from one domain constrains interpretation of evidence in a new domain), and the blessing of abstraction (abstract knowledge can be learned faster than concrete knowledge). These will be less important in what follows, but we note that they are potentially important for the question of language learning—when we view learning as an inference, the dynamics of probabilistic inference come to bear on the learning problem. For detailed examples of these patterns, using Church representation, see `http://probmods.org`.

1.3 Possible worlds

We have illustrated how a collection of Church functions—an intuitive theory—describes knowledge about the world. In fact, an intuitive theory can be interpreted as describing a probability distribution over possible worlds. To see this, first assume that all the (stochastic) functions of the intuitive theory are memoized.[3] Then the value of any expression is determined by the values of those functions called (on corresponding inputs) while evaluating the expression; any expression is assigned a value if we have the values of *all* the functions on *all* possible inputs. A possible world then, can be represented by a complete assignment of values to function-argument pairs, and a distribution over worlds is defined by the return-value probabilities of the functions, as specified by the intuitive theory.

We do not need to actually *compute* the values of all function-argument pairs in order to evaluate a specific expression, though. Most evaluations will involve just a fraction of the potentially infinite number of assignments needed to make a complete world. Instead, Church evaluation constructs only a partial representation of a possible world containing the minimal information needed to evaluate a given expression: the values of function applications that are actually reached during evaluation. Such a "partial world" can be interpreted as a set of possible worlds, and its probability is the sum of the probabilities of the worlds in this set. Fortunately this intractable sum is equal to the product of the probabilities of the choices made to determine the partial world: the partial world is independent of any function values not reached during evaluation, hence marginalizing these values is the same as ignoring them.

In this way, we can represent a distribution over all possible worlds *implicitly*, while explicitly constructing only partial worlds large enough to be relevant to a given query, ignoring irrelevant random values. The fact that

[3] If not all stochastic functions are memoized, very similar reasoning goes through: now each function is associated with an infinite number of return values, individuated by call order or position.

infinite sets of possible worlds are involved in a possible worlds semantics has sometimes been considered a barrier to the psychological plausibility of this approach. Implementing a possible worlds semantics via a probabilistic programming language may help defuse this concern: a small, finite subset of random choices will be constructed to reason about most queries; the remaining infinitude, while mathematically present, can be ignored because the query is statistically independent of them.

2 Meaning as condition

Following a productive tradition in semantics (Stalnaker, 1978; Lewis, 1979; Heim, 1982, etc.), we view the basic function of language understanding as belief update: moving from a prior belief distribution over worlds (or situations) to a posterior belief distribution given the *literal meaning* of a sentence. Probabilistic conditioning (or `query`) is a very general way to describe updating of degrees of belief. Any transition from distribution P_{before} to distribution P_{after} can be written as multiplying by a non-negative, real-valued function and then renormalizing, provided P_{before} is non-zero whenever P_{after} is.[4] From this observation it is easy to show that any belief update which preserves impossibility can be written as the result of conditioning on some (stochastic) predicate. Note that conditioning in this way is the natural analogue of the conception of belief update as intersection familiar from dynamic semantics.

Assume for now that each sentence provides information which is logically independent of other sentences given the state of the world (which may include discourse properties). From this it follows, parallel to the discussion of multiple observations as sequential conditioning above, that a sequence of sentences can be treated as sequentially updating beliefs by conditioning—so we can focus on the literal meaning of a single sentence. This independence assumption can be seen as the most basic and important *compositionality* assumption, which allows language understanding to proceed incrementally by utterance. (When we add pragmatic inference, in section 3, this independence assumption will be weakened, but it remains essential to the basic semantic function of utterances.)

How does an utterance specify which belief update to perform? We formalize the literal listener as:

```
(define (literal-listener utterance QUD)
  (query
    ...theory...
    (eval QUD)
    (eval (meaning utterance))))
```

This function specifies the posterior distribution over answers to the Question Under Discussion (`QUD`) given that the literal meaning of the utterance is true.[5] Notice that the prior distribution for the literal listener is specified by a conceptual lexicon—the `...theory...`—and the `QUD` will be evaluated in the local environment where all functions defined by this theory are in scope. That is,

[4] For infinite spaces we would need a more general condition on the measurability of the belief update.

[5] QUD theories have considerable motivation in semantics and pragmatics: see Ginzburg 1995; Van Kuppevelt 1995; Roberts 2012; Beaver & Clark 2008 among many others. For us, the key feature of the `QUD` is that it denotes a partition of W that is naturally interpreted as the random variable of immediate interest in the conversation.

the question of interest is determined by the *expression* `QUD` while its answer is determined by the *value* of this expression in the local context of reasoning by the literal listener: the value of `(eval QUD)`. (For a description of the `eval` operator see section 1.1 above.) Hence the semantic effect of an utterance is a function from `QUD`s to posteriors, rather than directly a posterior over worlds. Using the `QUD` in this way has two beneficial consequences. First, it limits the holism of belief update, triggering representation of only the information that is needed to capture the information conveyed by a sentence about the question of current interest. Second, when we construct a speaker model the `QUD` will be used to capture a pressure to be informative about the topic of current interest, as opposed to global informativity about potentially irrelevant topics.

2.1 Composition

The `meaning` function is a stochastic mapping from strings (surface forms) to Church expressions (logical forms, which may include functions defined in `...theory...`). Many theories of syntactic and semantic composition could be used to provide this mapping. For concreteness, we consider a simple system in which a string is recursively split into left and right portions, and the meanings of these portions are combined with a random combinator. The first step is to check whether the utterance is syntactically atomic, and if so look it up in the lexicon:

```
(define (meaning utterance)
  (if (lexical-item? utterance)
      (lexicon utterance)
      (compose utterance)))
```

Here the predicate `lexical-item?` determines if the (remaining) utterance is a single lexical item (entry in the lexicon), if so it is looked up with the `lexicon` function. This provides the base case for the recursion in the `compose` function, which randomly splits non-atomic strings, computes their meanings, and combines them into a list:

```
(define (compose utterance)
  (define subs (random-split utterance))
  (list (meaning (first subs)) (meaning (second subs))))
```

The function `random-split` takes a string and returns the list of two substrings that result from splitting at a random position in the length of the string.[6]

Overall, the `meaning` function is a stochastic mapping from strings to Church expressions. In `literal-listener` we `eval` the representation constructed by `meaning`

[6] While it is beyond the scope of this chapter, a sufficient syntactic system would require language-specific biases that favor certain splits or compositions on non-semantic grounds. For instance, lexical items and type shifters could be augmented with word-order restrictions, and conditioning on sentence meaning could be extended to enforce syntactic well-formedness as well (along the lines of Steedman 2001). Here we will assume that such a system is in place and proceed to compute sample derivations.

in the same environment as the `QUD`. Because we have formed a list of the sub-meanings, evaluation will result in forward application of the left sub-meaning to the right. Many different meanings can get constructed and evaluated in this way, and many of them will be mis-typed. Critically, if type errors are interpreted as the non-`true` value *error* (as described in section 1.1), then mis-typed compositions will not satisfy the condition of the `query` in the `literal-listener` function—though many ill-typed compositions can be generated by `meaning`, they will be eliminated from the posterior, leaving only well-typed interpretations.

To understand what the `literal-listener` does overall, consider rejection sampling: we evaluate both the `QUD` and meaning expressions, constructing whatever intermediate expressions are required; if the meaning expression has value `true`, then we return the value of `QUD`, otherwise we try again. Random choices made to construct and evaluate the meaning will be reasoned about jointly with world states while interpreting the utterance; the complexity of interpretation is thus an interaction between the domain theory, the meaning function, and the lexicon.

2.2 Random type shifting

The above definition for `meaning` always results in composition by forward application. This is too limited to generate potential meanings for many sentences. For instance "Bob runs" requires a backward application to apply the meaning of "runs" to that of "Bob". We extend the possible composition methods by allowing the insertion of type-shifting operators.

```
(define (meaning utterance)
  (if (lexical-item? utterance)
      (lexicon utterance)
      (shift (compose utterance))))

(define (shift m)
  (if (flip)
      m
      (list (uniform-draw type-shifters) (shift m))))

(define type-shifters '(L G AR1 AR2 ...))
```

Each intermediate meaning will be shifted zero or more times by a randomly chosen type-shifter; because the number of shifts is determined by a stochastic recursion, fewer shifts are *a priori* more likely. Each lexical item thus has the potential to be interpreted in any of an infinite number of (static) types, but the probability of associating an item with an interpretation in some type declines exponentially with the the number of type-raising operations required to construct this interpretation. The use of a stochastic recursion to generate type ambiguities thus automatically enforces the preference for interpretation in lower types, a feature which is often stipulated in discussions of type-shifting (Partee & Rooth, 1983; Partee, 1987).

We choose a small set of type shifters which is sufficient for the examples of this chapter:

- **L**: `(λ (x) (λ (y) (y x)))`
- **G**: `(λ (x) (λ (y) (λ (z) (x (y z)))))`
- **AR1**: `(λ (f) (λ (x) (λ (y) (x (λ (z) ((f z) y))))))`
- **AR2**: `(λ (f) (λ (x) (λ (y) (y (λ (z) ((f x) z))))))`

Among other ways they can be used, the shifter **L** enables backward application and **G** enables forward composition. For instance, *Bob runs* has an additional possible meaning `((L 'Bob) runs)` which applies the meanings of *runs* to that of *Bob*, as required.

Type shifters **AR1** and **AR2** allow flexible quantifier scope as described in Hendriks (1993); Barker (2005). (The specific formulation here follows Barker, 2005, pp.453ff.) We explore the ramifications of the different possible scopes in section 2.5. This treatment of quantifier scope is convenient, but others could be implemented by complicating the syntactic or semantic mechanisms in various ways: see e.g. May (1977); Steedman (2012).

2.3 Interpreting English in Church: the Lexicon

Natural language utterances are interpreted as Church expressions by the `meaning` function. The stochastic λ-calculus (implemented in Church) thus functions as our intermediate language, just as the ordinary, simply-typed λ-calculus functions as an intermediate translation language in the fragment of English given by Montague (1973). A key difference, however, is that the intermediate level is not merely a convenience as in Montague's approach. Conceptual representations and world knowledge are also represented in this language as Church function definitions. The use of a common language to represent linguistic and non-linguistic information allows lexical semantics to be grounded in conceptual structure, leading to intricate interactions between these two types of knowledge. In this section we continue our running tug-of-war example, now specifying a lexicon mapping english words to Church expressions for communicating about this domain.

We abbreviate the denotations of expressions `(meaning` α`)` as $[\![\alpha]\!]$. The simplest case is the interpretation of a name as a Church symbol, which serves as the unique mental token for some object or individual (the name-bearer).

- $[\![Bob]\!]$: `'Bob`
- $[\![Team\ 1]\!]$: `'team1`
- $[\![Match\ 1]\!]$: `'match1`
- ...

Interpreted in this way names are directly referential since they are interpreted using the same symbol in every situation, regardless of inferences made during interpretation.

A one-place predicate such as *player* or *man* is interpreted as a function from individuals to truth-values. Note that these denotations are grounded in aspects of the non-linguistic conceptual model, such as `players`, `matches`, and `gender`.

- $[\![player]\!]$: `(λ (x) (element? x players))`
- $[\![team]\!]$: `(λ (x) (element? x teams))`
- $[\![match]\!]$: `(λ (x) (element? x matches))`
- $[\![man]\!]$: `(λ (x) (equal? (gender x) 'male))`
- $[\![woman]\!]$: `(λ (x) (equal? (gender x) 'female))`

Similarly, transitive verbs such as *won* denote two-place predicates. (We simplify throughout by ignoring tense.)

- $[\![won]\!]$: `(λ (match) (λ (x) (equal? x (winner match))))`
- $[\![played\ in]\!]$: `(λ (match) (λ (x) (or (element? x (teams-in-match match)) (element? x (players-in-match match)))))`
- $[\![is\ on]\!]$: `(λ (team) (λ (x) (element? x (players-on-team team))))`

Intensionality is implicit in these definitions because the denotations of English expressions can refer to stochastic functions in the intuitive theory. Thus predicates pick out functions from individuals to truth-values in any world, but the specific function that they pick out in a world can depend on random choices (e.g., values of `flip`) that are made in the process of constructing the world. For instance, *player* is true of the same individuals in every world, because `players` is a fixed list (see Figure 1) and `element?` is the deterministic membership function. On the other hand, *man* denotes a predicate which will be *a priori* true of a given individual (say, `'Bob`) in 50% of worlds—because the memoized stochastic function `gender` returns `'male` 50% of the time when it is called with a new argument.

For simplicity, in the few places in our examples where plurals are required, we treat them as denoting lists of individuals. In particular, in a phrase like *Team 1 and Team 2*, the conjunction of NPs forms a list:

- $[\![and]\!] =$ `(λ (x) (λ (y) (list x y)))`

Compare this to the set-based account of plurals described in Scha & Winter 2014 (this volume). To allow distributive properties (those which require atomic individuals as arguments) to apply to such collections we include a type-shifting operator (in `type-shifters`, see section 2.2) that universally quantifies the property over the list:

- **DIST**: `(λ (V) (λ (s) (all (map V s))))`

For instance, *Bob and Jim played in Match 1* can be interpreted by shifting the property $[\![played\ in\ Match\ 1]\!]$ to a predicate on lists (though the order of elements in the list will not matter).

We can generally adopt standard meanings for functional vocabulary, such as quantifiers.

- $[\![every]\!]$: `(λ (P) (λ (Q) (= (size P) (size (intersect P Q)))))`
- $[\![some]\!]$: `(λ (P) (λ (Q) (< 0 (size (intersect P Q)))))`
- $[\![no]\!]$: `(λ (P) (λ (Q) (= 0 (size (intersect P Q)))))`
- $[\![most]\!]$: `(λ (P) (λ (Q) (< (size P) (* 2 (size (intersect P Q))))))`

For simplicity we have written the quantifiers in terms of set size; the `size` function can be defined in terms of the domain of `individuals` as `(λ (S) (length (filter S individuals)))`.[7]

We treat gradable adjectives as denoting functions from individuals to degrees (Bartsch & Vennemann, 1973; Kennedy, 1997, 2007). Antonym pairs such as *weak/strong* are related by scale reversal.

- ⟦*strong*⟧: `(λ (x) (strength x))`
- ⟦*weak*⟧: `(λ (x) (- 0 (strength x)))`

This denotation will require an operator to bind the degree in any sentence interpretation. In the case of the relative and superlative forms this operator will be indicated by the corresponding morpheme. For instance, the superlative morpheme *-est* is defined so that *strongest player* will denote a property that is true of an individual when that individual's strength is equal to the maximum strength of all players:[8]

- ⟦*-est*⟧: `(λ (A) (λ (N) (λ (x) (= (A x) (max-prop A N)))))`

For positive form sentences, such as *Bob is strong*, we will employ a type shifting operator which introduces a degree threshold to bind the degree—see section 4.

2.4 Example interpretations

To illustrate how a (literal) listener interprets a sequence of utterances, we consider a variant of our explaining-away example from the previous section. For each of the following utterances we give one expression that could be returned from `meaning` (usually the simplest well-typed one); we also show each meaning after simplifying the compositions.

- Utterance 1: *Jane is on Team 1.*
  `meaning`: `((L 'Jane) (λ (team) (λ (x) (element? x (players-on-team team))) 'team1))`
  simplified: `(element? 'Jane (players-on-team 'team1))`
- Utterance 2: *Bob is on Team 2.*
  `meaning`: `((L 'Bob) (λ (team) (λ (x) (element? x (players-on-team team))) 'team2))`
  simplified: `(element? 'Bob (players-on-team 'team2))`
- Utterance 3: *Team 1 and Team 2 played in Match 1.*
  `meaning`: `((L ((L 'team 1) ((λ (x) (λ (y) (list x y))) 'team2))) (DIST ((λ (match) (λ (x) (element? x (teams-in-match match)))) 'match1)))`
  simplified: `(all (map (λ (x) (element? x (teams-in-match 'match1)))) '(team1 team2))`

[7] In the examples below, we assume for simplicity that many function words, for example *is* and *the*, are semantically vacuous, i.e., that they denote identity functions.

[8] The set operator `max-prop` implicitly quantifies over the domain of discourse, similarly to `size`. It can be defined as `(lambda (A N) (max (map A (filter N individuals))))`.

- Utterance 4: *Team 1 won Match 1.*
  `meaning`: `((L 'team1) ((λ (match) (λ (x) (equal? x (winner match)))) 'match1))`
  simplified: `(equal? 'team1 (winner 'match1))`

The literal listener conditions on each of these meanings in turn, updating her posterior belief distribution. In the absence of pragmatic reasoning (see below), this is equivalent to conditioning on the conjunction of the meanings of each utterance—essentially as in dynamic semantics (Heim, 1992; Veltman, 1996). Jane's inferred strength (i.e. the posterior on `(strength 'Jane)`) increases substantially relative to the uninformed prior (see Figure 3).

Suppose, however, the speaker continues with the utterance:

- Utterance 5: *Bob is the weakest player.*
  `meaning`: `((L 'Bob) (((L (λ (x) (- (strength x)))) (λ (A) (λ (N) (λ (x) (= (A x) (max-prop A N)))))) (λ (x) (element? x players))))`
  simplified: `(= (- (strength 'Bob)) (max (λ (x) (- (strength x))) (λ (x) (element? x players))))`

This expression will be true if and only if Bob's strength is the smallest of any player. Conditioning on this proposition about Bob, we find that the inferred distribution of Jane's strength decreases toward the prior (see Figure 3)—Jane's performance is explained away. Note, however, that this non-monotonic effect comes about not by directly observing a low value for the strength of Bob and information about his gender, as in our earlier example, but by conditioning on the truth of an utterance which does not entail *any* precise value of Bob's strength. That is, because there is uncertainty about the strengths of all players, in principle Bob could be the weakest player even if he is quite strong, as long as all the other players are strong as well. However, the other players are most likely to be about average strength, and hence Bob is particularly weak; conditioning on Utterance 5 thus lowers Bob's expected strength and adjusts Jane's strength accordingly.

2.5 Ambiguity

The `meaning` function is stochastic, and will often associate utterances with several well-typed meanings. Ambiguities can arise due to any of the following:

- Syntactic: `random-split` can generate different syntactic structures for an utterance. If more than one of these structures is interpretable (using the type-shifting operators available), the literal listener will entertain interpretations with different syntactic structures.
- Compositional: Holding the syntactic structure fixed, insertion of different (and different numbers of) type-shifting operators by `shift` may lead to well-typed outputs. This can lead, for example, to ambiguities of quantifier scope and in whether a pronoun is bound or free.

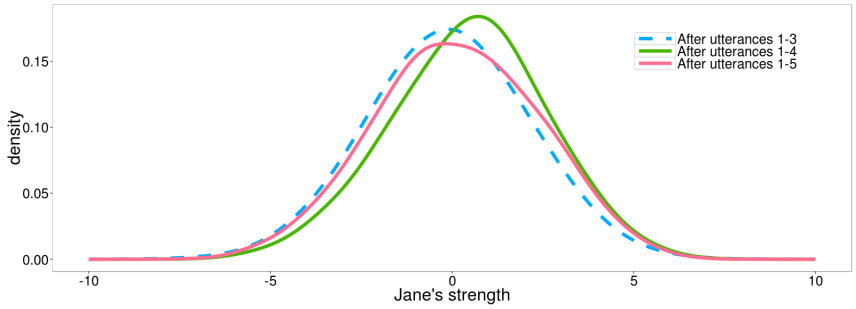

Figure 3. A linguistic example of explaining away, demonstrating that the literal listener makes non-monotonic inferences about the answer to the QUD "How strong is Jane?" given the utterances described in the main text. Lines show the probability density of answers to this QUD after (a) utterances 1-3; (b) utterances 1-4; (c) utterances 1-5.

- Lexical: the `lexicon` function may be stochastic, returning different options for a single item, or words may have intrinsically stochastic meanings. (The former can always be converted to the latter.)

In the literal interpretation model we have given above, `literal-listener`, these sources of linguistic ambiguity will interact with the interpreter's beliefs about the world. That is, the `query` implies a *joint inference* of sentence meaning and world, given that the meaning is true of the world. When a sentence is ambiguous in any of the above ways, the listener will favor plausible interpretations over implausible ones, because the interpreter's model of the world is more likely to generate scenarios which make the sentence true.

For example, consider the utterance "Most players played in some match". Two (simplest, well-typed) interpretations are possible. We give an intuitive paraphrase and the meanings for each (leaving the leaving lexical items in place to expose the compositional structure):

- Subject wide scope:
  "For most players x, there was a match y such that x played in y."
  ((L ([[*Most*]] [[*players*]])) ((AR2 (AR1 [[*played in*]])) ([[*some*]] [[*match*]])))
- Object wide scope:
  "For some match y, most players played in y."
  ((L ([[*Most*]] [[*players*]])) ((AR1 (AR2 [[*played in*]])) ([[*some*]] [[*match*]])))

Both readings equally *a priori* probable, since the `meaning` function draws type-shifters uniformly at random. However, if one reading is more likely to be true, given background knowledge, it will be preferred. This means that we can influence the meaning used, and the degree to which each meaning influences the listener's posterior beliefs, by manipulating relevant world knowledge.

To illustrate the effect of background knowledge on choice of meaning, imagine varying the number of matches played in our tug-of-war example.

Recall (see Figure 1) that all teams are of size `team-size`, which varies across worlds and can be anywhere from 1 to 6 players, with equal probability. If the number of matches is large (say we `(define matches '(match1 ... match10))`), then the subject-wide scope reading can be true even if `team-size` is small: it could easily happen that most players played in one or another of ten matches even if each team has only one or two players. In contrast, the object-wide scope reading, which requires most players on a *single* match, can be true only if teams are large enough (i.e. `team-size` is ≥ 4, so that more than half of the players are in each match). The `literal-listener` jointly infers `team-size` and the reading of the utterance, assuming the utterance is true; because of the asymmetry in when the two readings will be true, there will be a preference for the subject-wide reading if the number of matches is large—it is more often true. If the number of matches is small, however, the asymmetry between readings will be decreased. Suppose that only one match was played (i.e. `(define matches '(match1))`), then both readings can be true only if the team size is large. The listener will thus infer that `team-size`≥ 4 and the two readings of the utterance are equally probable. Figure 4, left panel, shows the strength of each reading as the number of matches varies from 1 to 10, with the number of teams fixed to 10. The right panel shows the mean inferred team size as the number of matches varies, for each reading and for the marginal. Our model of language understanding as joint inference thus predicts that the resolution of quantifier scope ambiguities will be highly sensitive to background information.

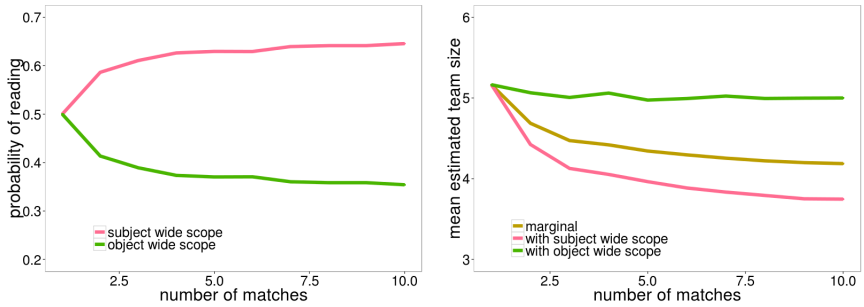

Figure 4. The probability of the listener interpreting the utterance *Most players played in some match* according to the two possible quantifier scope configurations depends in intricate ways on the interpreter's beliefs and observations about the number of matches and the number of players on each team (left). This, in turn, influences the total information conveyed by the utterance (right). For this simulation there were 10 teams.

More generally, an ambiguous utterance may be resolved differently, and lead to rather different belief update effects, depending on the plausibility of the various interpretations given background knowledge. Psycholinguistic re-

search suggests that background information has exactly this kind of graded effect on ambiguity resolution (see, for example, Crain & Steedman, 1985; Altmann & Steedman, 1988; Spivey *et al.*, 2002). In a probabilistic framework, preferences over alternative interpretations vary continuously between the extremes of assigning equal probability to multiple interpretations and assigning probability 1 to a single interpretation. This is true whether the ambiguity is syntactic, compositional, or lexical in origin.

2.6 Compositionality

It should be clear that compositionality has played a key role in our model of language interpretation thus far. It has in fact played several key roles: Church expressions are built from simpler expressions, sequences of utterances are interpreted by sequential conditioning, the `meaning` function composes Church expressions to form sentence meanings. There are thus several, interlocking "directions" of compositionality at work, and they result in interactions that could appear non-compositional if only one direction was considered. Let us focus on two: compositionality of world knowledge and compositionality of linguistic meaning.

Compositionality of world knowledge refers to the way that we use SLC to build distributions over possible worlds, not by directly assigning probabilities to all possible expressions, but by an evaluation process that recursively samples values for sub-expressions. That is, we have a compositional language for specifying generative models of the world. Compositionality of linguistic meaning refers to the way that *conditions* on worlds are built up from simpler pieces (via the `meaning` function and evaluation of the meaning). This is the standard approach to meaning composition in truth-conditional semantics. Interpreted meaning—the posterior distribution arrived at by `literal-listener`—is not immediately compositional along either world knowledge or linguistic structure. Instead it arises from the interaction of these two factors. The glue between these two structures is the intuitive theory; it defines the conceptual language for imagining particular situations, and the primitive vocabulary for semantic meaning.

An alternative approach to compositional probabilistic semantics would be to let each linguistic expression denote a distribution or probability directly, and build the linguistic interpretation by composing them. This appears attractive: it is more direct and simpler (and does not rely on complex generative knowledge of the world). How would we compose these distributions? For instance take "Jack is strong and Bob is strong". If "Jack is strong" has probability 0.2 and "Bob is strong" has probability 0.3, what is the probability of the whole sentence? A natural approach would be to multiply the two probabilities. However this implies that their strengths are independent—which is intuitively unlikely: for instance, if Jack and Bob are both men, then learning that Jack is strong suggests than men are strong, which suggests that Bill is strong. A more productive strategy is the one we have taken: world knowledge

specifies a joint distribution on the strength of Bob and Jack (by first sampling the prototypical strength of men, then sampling the strength of each), and the sentence imposes a *constraint* on this distribution (that each man's strength exceeds a threshold). The sentence denotes not a world probability simpliciter, but a constraint on worlds which *is* built compositionally.

2.7 Extensions and related work

The central elements of probabilistic language understanding as described above are: grounding lexical meaning into a probabilistic generative model of the world, taking sentence meanings as conditions on worlds (built by composing lexical meanings), and treating interpretation as joint probabilistic inference of the world state and the sentence meaning conditioned on the truth of the sentence. It should be clear that this leaves open many extensions and alternative formulations. For instance, varying the method of linguistic composition, adding static types that influence interpretation, and including other sources of uncertainty such as a noisy acoustic channel are all straightforward avenues to explore.

There are several related approaches that have been discussed in previous work. Much previous work in probabilistic semantics has a strong focus on vagueness and degree semantics: see e.g. Edgington 1997; Frazee & Beaver 2010; Lassiter 2011, discussed further in section 4 below and in Lassiter 2014 (this volume). There are also well-known probabilistic semantic theories of isolated phenomena such as conditionals (Adams, 1975; Edgington, 1995, and many more) and generics (Cohen, 1999a,b). We have taken inspiration from these approaches, but we take the strong view that probabilities belong at the foundation of an architecture for language understanding, rather than treating it as a special-purpose tool for the analysis of specific phenomena.

In Fuzzy Semantics (Zadeh, 1971; Lakoff, 1973; Hersh & Caramazza, 1976, etc.) propositions are mapped to real values that represent degrees of truth, similar to probabilities. Classical fuzzy semantics relies on strong independence assumptions to enable direct composition of fuzzy truth values. This amounts to a separation of uncertainty from language and non-linguistic sources. In contrast, we have emphasized the interplay of linguistic interpretation and world knowledge: the probability of a sentence is not defined separate from the joint-inference interpretation, removing the need to define composition directly on probabilities.

A somewhat different approach, based on type theory with records, is described by Cooper *et al.* (2014). Cooper et al.'s project revises numerous basic assumptions of model-theoretic semantics, with the goals of better explaining semantic learning and "pervasive gradience of semantic properties." The work described here takes a more conservative approach, by enriching the standard framework while preserving most basic principles. As we have shown, this gives rise to gradience; we have not addressed learning, but there is an extensive literature on probabilistic learning of structured representations similar to

those required by our architecture: see e.g. Goodman *et al.* 2008b; Piantadosi *et al.* 2008, 2012; Tenenbaum *et al.* 2011. It may be, however, that stronger types than we have employed will be necessary to capture subtleties of syntax and facilitate learning. Future work will hopefully clarify the relationship between the two approaches, revealing which differences are notational and which are empirically and theoretically significant.

3 Pragmatic interpretation

The `literal-listener` described above treats utterances as true information about the world, updating her beliefs accordingly. In real language understanding, however, utterances are taken as speech acts that inform the listener indirectly by conveying a speaker's intention. In this section we describe a version of the *Rational Speech Acts* model (Goodman & Stuhlmüller, 2013; Frank & Goodman, 2012), in which a sophisticated listener reasons about the intention of an informative speaker.

First, imagine a speaker who wishes to convey that the question under discussion (`QUD`) has a particular answer (i.e. value). This can be viewed as an inference: what utterance is most likely to lead the (literal) listener to the correct interpretation?

```
(define (speaker val QUD)
  (query
    (define utterance (language-prior))
    utterance
    (equal? val (literal-listener utterance QUD))))
```

The `language-prior` forms the *a priori* (non-contextual and non-semantic) distribution over linguistic forms, which may be modeled with a probabilistic context free grammar or similar model. This prior inserts a *cost* for each utterance: using a less likely utterance will be dispreferred *a priori*. Notice that this speaker conditions on a single sample from `literal-listener` having the correct `val` for the `QUD`—that is, he conditions on the `literal-listener` "guessing" the right value. Since the listener may sometimes accidentally guess the right value, even when the utterance is not the most informative one, the speaker will sometimes choose sub-optimal utterances. We can moderate this behavior by adjusting the tendency of the listener to guess the most likely value:

```
(define (speaker val QUD)
  (query
    (define utterance (language-prior))
    utterance
    (equal? val ((power literal-listener alpha) utterance QUD) )))
```

Here we have used a higher-order function `power` that raises the return distribution of the input function to a power (and renormalizes). When the power `alpha` is large the resulting distribution will mostly sample the maximum of the underlying distribution—in our case the listener that `speaker` imagines will mostly sample the *most* likely `val`.

Writing the distribution implied by the `speaker` function explicitly can be clarifying:

$$P(\texttt{ut}|\texttt{val},\texttt{QUD}) \propto P(\texttt{ut})P_{\texttt{listener}}(\texttt{val}|\texttt{ut},\texttt{QUD})^{\alpha} \tag{5}$$

$$\propto e^{\alpha \ln(P_{\texttt{listener}}(\texttt{val}|\texttt{ut},\texttt{QUD}))+\ln(P(\texttt{ut}))} \tag{6}$$

Thus, the `speaker` function describes a speaker who chooses utterances using a soft-max rule $P(\texttt{utt}) \propto e^{\alpha U(\texttt{utt})}$ (Luce, 1959; Sutton & Barto, 1998). Here the utility $U(\texttt{utt})$ is given by the sum of

- the informativity of `utt` about the `QUD`, formalized as negative surprisal of the intended value: $\ln(P_{\texttt{listener}}(\texttt{val}|\texttt{ut}, \texttt{QUD}))$,
- a cost term $\ln(P(\texttt{utt}))$, which depends on the language prior.

Utterance cost plausibly depends on factors such as length, frequency, and articulatory effort, but the formulation here is noncommittal about precisely which linguistic and non-linguistic factors are relevant.

A more sophisticated, pragmatic, listener can now be modeled as a Bayesian agent updating her belief about the value of the question under discussion given the observation that the speaker has bothered to make a particular speech act:

```
(define (listener utterance QUD)
  (query
    ...theory...
    (define val (eval QUD))
    val
    (equal? utterance (speaker val QUD))))
```

Notice that the prior over `val` comes from evaluating the `QUD` expression given the `theory`, and the posterior comes from updating this prior given that the speaker has chosen `utterance` to convey `val`.

The force of this model comes from the ability to call the `query` function within itself (Stuhlmueller & Goodman, 2013)—each query models the inference made by one (imagined) communicator, and together they capture sophisticated pragmatic reasoning. Several observations are worth making: First, alternative utterances will enter into the computation in sampling (or determining the probability of) the actual utterance from `speaker`. Similarly, alternative values are considered in the listener functions. Second, the notion of informativity captured in the `speaker` model is not simply information transmitted by `utterance`, but is *new* information conveyed to the listener *about* the `QUD`. Information which is not new to the listener or which is not relevant to the `QUD` will not contribute to the speaker's utility.

3.1 Quantity implicatures

We illustrate by considering quantity implicatures: take as an example the sentence "Jane played in some match". This entails that Jane did not play in zero matches. In many contexts, it would also be taken to suggest that Jane did not play in all of the matches. However, there are many good reasons for thinking that the latter inference is not part of the basic, literal meaning of the sentence (Grice, 1989; Geurts, 2010). Why then does it arise? Quantity implicatures follow in our model due to the pragmatic listener's use of "counterfactual" reasoning to help reconstruct the speaker's intended message from

his observed utterance choice. Suppose that the `QUD` is "How many matches did Jane play in?" (interpreted as $[\![\textit{the number of matches Jane played in}]\!]$). The listener considers different answers to this question by simulating partial worlds that vary in how many matches Jane played in and considering what the speaker would have said for each case. If Jane played in every match, then "Jane played in every match" would be used by the speaker more often than "Jane played in some match". This is because the speaker model favors more informative utterances, and the former is more informative: a literal speaker will guess the correct answer more often after hearing "Jane played in every match". Since the speaker in fact chose the less informative utterance in this case, the listener infers that some precondition for the stronger utterance's use—e.g., its truth—is probably not fulfilled.

For example, suppose that it is common knowledge that teams have four players, and that three matches were played. The speaker knows exactly who played and how many times, and utters "Jane played in some match". How many matches did she play in? The speaker distribution is shown in Figure 5. If Jane played in zero matches, the probability that the speaker will use either utterance is zero (instead the speaker will utter "Jane played in no match"). If she played in one or two matches, the probability that the speaker will utter "Jane played in some match" is non-zero, but the probability that the speaker will utter "Jane played in every match" is still zero. However, the situation changes dramatically if Jane in fact played in all the matches: now the speaker prefers the more informative utterance "Jane played in every match".

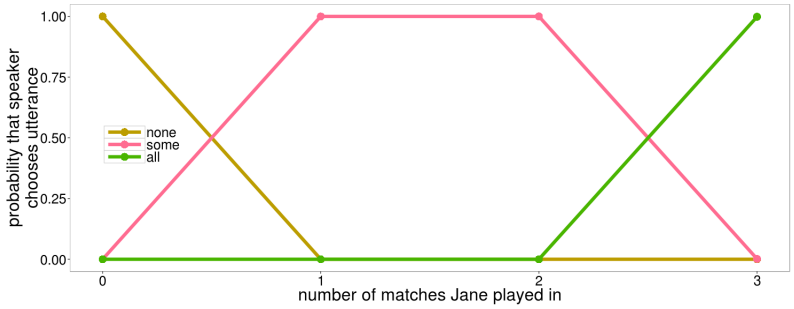

Figure 5. Normalized probability that the speaker will utter "Jane played in no/-some/every match" in each situation, generated by reasoning about which utterance will most effectively bring the literal listener to select the correct answer to the QUD "How many matches did Jane play in?". (The parameter `alpha` is set to 5.)

The pragmatic listener still does not know how many matches Jane played in but can reason about the speaker's utterance choice. If the correct answer were 3 the speaker would probably not have chosen "some", because the literal listener is much less likely to choose the answer 3 if the utterance is "some"

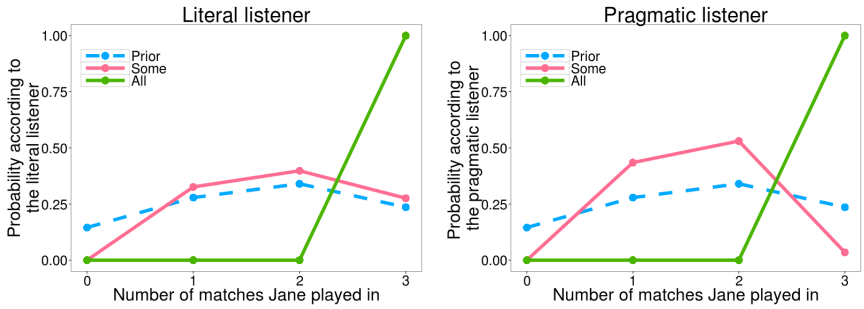

Figure 6. Interpretation of "Jane played in some match" by the literal and pragmatic listeners, assuming that the only relevant alternatives are "Jane played in no/every match". While the literal listener (left pane) assigns a moderate probability to the "all" situation given this utterance, the pragmatic listener (right pane) assigns this situation a very low probability. The difference is due to the fact that the pragmatic listener reasons about the utterance choices of the speaker (Figure 5 above), taking into account that the speaker is more likely to say "every" than "some" if "every" is true.

as opposed to "every". The listener can thus conclude that the correct answer probably is not 3. Figure 6 shows the predictions for both the literal and pragmatic listener; notice that the interpretation of "some" differs only minimally from the prior for the literal listener, but is strengthened for the pragmatic listener. Thus, our model yields a broadly Gricean explanation of quantity implicature. Instead of stipulating rules of conversation, the content of Grice's Maxim of Quantity falls out of the recursive pragmatic reasoning process whenever it is reasonable to assume that the speakers is making an effort to be informative. (For related formal reconstructions of Gricean reasoning about quantity implicature, see Franke 2009; Vogel *et al.* 2013.)

3.2 Extensions and related work

The simple Rational Speech Acts (RSA) framework sketched above has been fruitfully extended and applied to a number of phenomena in pragmatic understanding; many other extensions suggest themselves, but have not yet been explored. In Frank & Goodman 2012 the RSA model was applied to explain the results of simple reference games in which a speaker attempted to communicate one of a set of objects to a listener by using a simple property to describe it (e.g. *blue* or *square*). Here the intuitive theory can be seen as simply a prior distribution, `(define ref (ref-prior objects))` over which object is the referent in the current trial, the `QUD` is simply `ref`, and the properties have their standard extensions. By measuring the `ref-prior` empirically Frank & Goodman (2012) were able to predict the speaker and listener judgements with high quantitative accuracy (correlation around 0.99).

In Goodman & Stuhlmüller 2013 the RSA framework was extended to take into account the speaker's belief state. In this case the speaker should choose an utterance based on its *expected* informativity under the speaker's belief distribution. (Or, equivalently, the speaker's utility is the negative Kullback-Leibler divergence of the listener's posterior beliefs from the speaker's.) This extended model makes the interesting prediction that listeners should not draw strong quantity implicatures from utterances by speakers who are not known to be informed about the question of interest (cf. Sauerland, 2004; Russell, 2006). The experiments in Goodman & Stuhlmüller (2013) show that this is the case, and the quantitative predictions of the model are borne out.

As a final example of extensions to the RSA framework, the `QUD` itself can be an object of inference. If the pragmatic listener is unsure what topic the speaker is addressing, as must often be the case, then she should jointly infer the `QUD` and its `val` under the assumption that the speaker chose an utterance to be informative about the topic (whatever that happens to be). This simple extension can lead to striking predictions. In Kao *et al.* (2014); Kao *et al.* such `QUD` inference was shown to give rise to non-literal interpretations: hyperbolic and metaphoric usage. While the literal listener will draw an incorrect inference about the state of the world from an utterance such as "I waited a million hours", the speaker only cares if this results in correct information about the `QUD`; the pragmatic listener knows this, and hence interprets the utterance as only conveying information about the `QUD`. If the `QUD` is inferred to be a non-standard aspect of the world, such as whether the speaker is irritated, then the utterance will convey only information about this aspect and not the (false) literal meaning of the utterance: the speaker waited longer than expected and is irritated about it.

The RSA approach shares elements with a number of other formal approaches to pragmatics. It is most similar to game theoretic approaches to pragmatics. In particular to approaches that treat pragmatic inference as iterated reasoning, such as the Iterated Best Response (IBR) model (Franke, 2009; Benz *et al.*, 2005). The IBR model represents speakers and listeners recursively reasoning about each other, as in the RSA model. The two main differences are that IBR specifies unbounded recursion between speaker and listener, while RSA as presented here specifies one level, and the IBR specifies that optimal actions are chosen, rather than soft-max decisions. Neither of these differences is critical to either framework. We view it as an empirical question whether speakers maximize or soft-maximize and what level of recursive reasoning people actually display in language understanding.

4 Semantic indices

In formal semantics sentence meanings are often treated as intensions: functions from semantic indices to truth functions (Lewis, 1970, 1980; Montague, 1973). The semantic theory has little or nothing to say about how these indices are set, except that they matter and usually depend in some way on context. We have already seen that a probabilistic theory of pragmatic interpretation can be used to describe and predict certain effects of context and background knowledge on interpretation. Can we similarly use probabilistic tools to describe the ways that semantic indices are set based on context? We must first decide how semantic indices should enter into the probabilistic framework presented above (where we have so far treated meanings simply as truth functions). The simplest assumption is that they are random variables that occur (unbound) in the meaning expression and are reasoned about by the literal listener:

```
(define (literal-listener utterance QUD)
  (query
    ...theory...
    (define index (index-prior))
    (define val (eval QUD))
    val
    (eval (meaning utterance))))
```

Here we assume that the meaning may contain an unbound occurrence of `index` which is then bound during interpretation by the `(define index ...)` definition. Because there is now a joint inference over `val` and `index`, the index will tend to be set such that the utterance is most likely to be true.

Consider the case of gradable adjectives like *strong*. In section 2.3 we have defined $[\![strong]\!]$ = `(λ (x) (strength x))`; to form a property from the adjective in a positive form sentence like *Bob is strong*, we must bind the degree returned from `strength` in some way. A simple way to do this is to add a type-shifter that introduces a free threshold variable θ—see, for example, Kennedy 2007 and Lassiter 2014 (this volume). We extend the set of type shifters that can be inserted by `shift` (see section 2.2) with:

- **POS**: `(λ (A) (λ (x) (>= (A x) θ)))`

In this denotation the variable θ is a free index that will be bound during interpretation as above. Now consider possible denotations that can be generated by `meaning`.

- $[\![Bob\ is\ strong]\!]$=`('Bob (λ (x) (strength x)))`
- $[\![Bob\ is\ strong]\!]$=`((L 'Bob) (λ (x) (strength x)))`
- $[\![Bob\ is\ strong]\!]$=`((L 'Bob) (POS (λ (x) (strength x))))`

The first of these returns *error* because `'Bob` is not a function; the second applies `strength` to `'Bob` and returns a degree. Both of these meanings will be removed in the `query` of `literal-listener` because their values will never equal true. The third meaning tests whether Bob is stronger than a threshold variable and

returns a Boolean—it is the simplest well-typed meaning. With this meaning the utterance "Bob is strong" (with QUD "How strong is Bob?") would be interpreted by the literal listener (after simplification, and assuming for simplicity a domain of -100 to 100 for the threshold) via:

```
(query
  ...theory...
  (define θ (uniform -100 100))
  (define val (strength 'Bob))
  val
  (>= (strength 'Bob) θ))
```

Figure 7 shows the prior (marginal) distributions over θ and Bob's strength, and the corresponding posterior distributions after hearing "Bob is strong". The free threshold variable has been influenced by the utterance: it changes from a uniform prior to a posterior that is maximum at the bottom of its domain and gradually falls form there—this makes the utterance likely to be true. However, this gives the wrong interpretation of *Bob is strong*. Intuitively, the listener ought to adjust her estimate of Bob's strength to a fairly high value, relative to the prior. Because the threshold is likely very low, the listener instead learns very little about the variable of interest from the utterance: the posterior distribution on Bob's strength is almost the same as the prior.

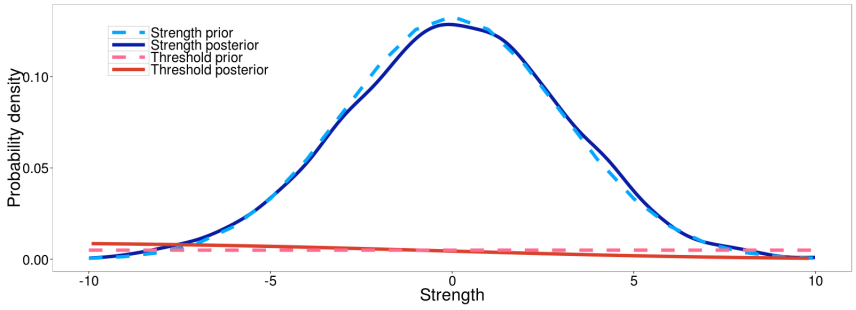

Figure 7. The literal listener's interpretation of an utterance containing a free threshold variable θ, assuming an uninformative prior on this variable. This listener's exclusive preference for true interpretations leads to a tendency to select extremely low values of θ ("degree posterior"). As a result the utterance conveys little information about the variable of interest: the strength posterior is barely different from the prior.

What is missing is the pressure to adjust θ so that the sentence is not only true, but also *informative*. Simply including the informative speaker and pragmatic listener models as defined above is not enough: without additional changes the index variables will be fixed by the literal listener with no pragmatic pressures. Instead, we *lift* the index variables to the pragmatic level. Imagine a pragmatic listener who believes that the index variable has a value

that she happens not to know, but which is otherwise common knowledge (i.e. known by the speaker, who assumes it is known by the listener):

```
(define (listener utterance QUD)
  (query
    ...theory...
    (define index (index-prior))
    (define val (eval QUD))
    val
    (equal? utterance (speaker val QUD index))))

(define (speaker val QUD index)
  (query
    (define utterance (language-prior))
    utterance
    (equal? val (literal-listener utterance QUD index))))

(define (literal-listener utterance QUD index)
  (query
    ...theory...
    (define val (eval QUD))
    val
    (eval (meaning utterance))))
```

In most ways this is a very small change to the model, but it has important consequences. At a high level, index variables will now be set in such a way that they both make the utterance likely to be true *and* likely to be pragmatically useful (informative, relevant, etc); the tradeoff between these two factors results in significant contextual flexibility of the interpreted meaning.

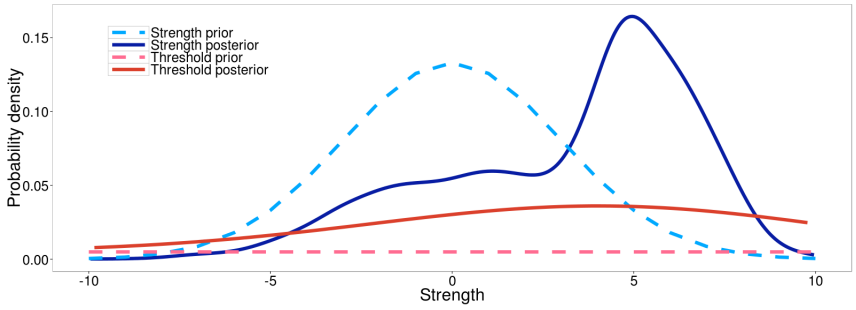

Figure 8. The pragmatic listener's interpretation of an utterance such as "Bob is strong," containing a free threshold variable θ that has been lifted to the pragmatic level. Joint inference of the degree and the threshold leads to a "significantly greater than expected" meaning. (We assume that the possible utterances are to say nothing (cost 0) and "Bob is strong/weak" (cost 6), and `alpha`= 5, as before.)

In the case of the adjective *strong*, Figure 8, the listener's posterior estimate of strength is shifted significantly upward from the prior, with mean at roughly one standard deviation above the prior mean (though the exact distribution depends on parameter choices). Hence *strong* is interpreted as

meaning "significantly stronger than average", but does not require maximal strength (most informative) or permit any strength (most often true). This model of gradable adjective interpretation (which was introduced in Lassiter & Goodman 2013) has a number of appealing properties. For instance, the precise interpretation is sensitive to the prior probability distribution on answers to the QUD. We thus predict that gradable adjective interpretation should display considerable sensitivity to background knowledge. This is indeed the case, as for example in the different interpretations of "strong boy", "strong football player", "strong wall", and so forth. Prior expectations about the degree to which objects in a reference class have some property frequently plays a considerable role in determining the interpretation of adjectives. This account also predicts that vagueness should be a pervasive feature of adjective interpretation, as discussed below. See Lassiter & Goodman 2013 for detailed discussion of these features.

We can motivate from this example a general treatment of semantic indices: lift each index into the pragmatic inference of `listener`, passing them down to `speaker` and on to `literal-listener`, allowing them to bind free variables in the literal meaning. As above all indices will be reasoned over jointly with world states. Any index that occurs in a potential meaning of an alternative utterance must be lifted in this way, to be available to the `literal-listener`. If we wish to avoid listing each index individually, we can modify the above treatment with an additional indirection: For instance by introducing a memoized function `index` that maps variable names to (random) values appropriate for their types.

4.1 Vagueness and indeterminate boundaries

Probabilistic models of the type described here make it possible to maintain the attractive formal precision of model-theoretic semantics while also making room for vagueness and indeterminate boundaries in both word meanings and psychological categories. There is considerable evidence from both psychological (e.g. Rosch, 1978; Murphy, 2002; Hampton, 2007) and linguistic (Taylor, 2003) research that a lack of sharp boundaries is a pervasive features of concept and word usage. Linguistic indeterminacy and vagueness can be understood as uncertainty about the precise interpretation of expressions in context. As discussed in section 2.5, uncertainty can enter from a number of sources in constructing meaning from an utterance; to those we can now add uncertainty that comes from a free index variable in the meaning, which is resolved at either the literal or pragmatic listener levels. Each source of uncertainty about the meaning leads to an opportunity for context-sensitivity in interpretation. These sources of context-sensitivity predict a number of important features of vagueness. We illustrate this by discussing how key features of vagueness in adjective interpretation are predicted by our treatment of gradable adjectives, above. For more discussion of vagueness and an overview of theories see Lassiter 2014 (this volume).

Borderline cases. While the underlying semantics of *Bill is strong* yields a definite boundary, introduced to the meaning by **POS**, there is posterior uncertainty over the value of this threshold. Hence, an individual whose degree of strength falls in the middle of the posterior distribution (see Figure 8) will be a borderline case of *strong*. In the example above, an individual with strength 3 will have a roughly equal chance of counting as strong and as not strong.

Tolerance principles. Suppose Bill has strength 4.5 and Mary has strength 4.4. It would be odd for someone to confidently agree to the claim that Bill is strong, but to deny confidently that Mary is strong. Our model explains this intuition: when two individuals' strength are separated by a small gap, the posterior probability that the threshold falls in this gap is very small—hence it is very rarely the case that one counts as strong and the other does not. Indeed, this could happen only if the posterior distribution over strength had a sharp discontinuity, which in turn would imply that the prior had an abrupt boundary (Lassiter & Goodman, 2013).

The sorites paradox. The following is an instance of a famous puzzle:

- Bill is strong.
- A person who is slightly less strong than a strong person is also strong.
- Therefore, everyone is strong, no matter how weak.

People generally find the premises plausible, but the conclusion (which follows logically by induction) not at all plausible. Evidently something is wrong with the second premise, but what?

Our probabilistic approach, built as it is upon a bivalent logic, requires that the conclusion is true in a given world if the premises are true. However, if the second premise is interpreted as universally quantified it will rarely be true: if there are enough individuals, there will be two separated by a small amount, but on either side of the threshold. Yet this answer—that the second premise is in fact false in most relevant situations—does not explain the psychological aspect of the puzzle (Graff, 2000): people express high confidence in the second premise.

Lassiter & Goodman (2013) argue that the second premise is not interpreted in a simple universally quantified way, but is evaluated probabilistically as a conditional: given that person x (of *a priori* unknown strength) is strong, form the posterior distribution over θ as above; under this distribution what is the probability that a person with strength slightly smaller is strong, i.e. the probability that `(- (strength x) ϵ)>` θ.[9] This probability depends on the prior distribution, but for reasonably gradual priors and fairly small gaps ϵ it will be quite high. Figure 4.1 shows the probability of the inductive premise as a function of the gap for the setup used before. This account builds on previous probabilistic approaches to the vagueness and the sorites (Borel, 1907;

[9] An extension to the linguistic fragment described above would be necessary to derive this interpretation formally. One approach would be to treat the relative clause an embedded `query`.

Black, 1937; Edgington, 1997; Lawry, 2008; Frazee & Beaver, 2010; Égré, 2011; Lassiter, 2011; Sutton, 2013), but is the first to offer a specific account of why vague adjectives should have context-sensitive probabilistic interpretations, and of how the distribution is determined in a particular context of utterance.

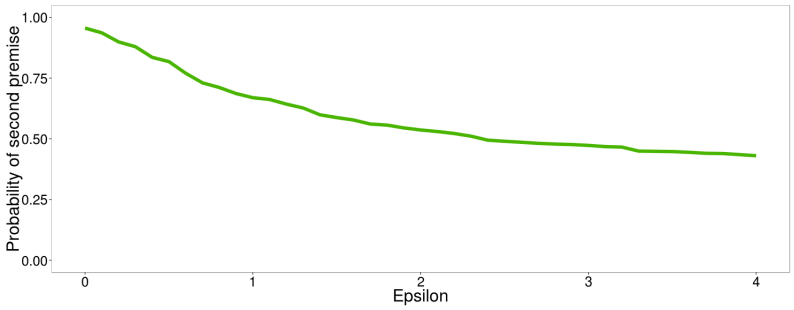

Figure 9. With prior distributions and parameters as above, the probability of the second premise of the sorites paradox is close to 1 when the inductive gap is small, but decreases as the size of the gap increases.

4.2 Extensions and related work

Another interpretation of the above modeling approach (indeed, the original interpretation, introduced in Bergen *et al.* (2012)) is as the result of *lexical uncertainty*: each index represents a lingering uncertainty about word meaning in context which the listener must incorporate in the interpretation process.[10] This interpretation is appealing in that it connects naturally to language acquisition and change (Smith *et al.*, 2013). For instance, upon hearing a new word a learner would initially treat its meaning as underdetermined—in effect, as an index variable ranging over all expressions of the appropriate type—and infer its meaning on each usage from contextual cues. Over time the prior over this 'index' would tighten until only the correct meaning remained, and no contextual flexibility was left. A difficulty with the lexical uncertainty interpretation is explaining why certain aspects of a word's meaning are so much more flexible than others and why this appears to be regular across words of a given type. The free-index interpretation accounts for this naturally because the dimensions of flexibility are explicitly represented as unbound variables in lexical entries or in type shifters used in the compositional construction of meaning. A more structured (e.g. hierarchical) notion of lexical uncertainty may be able to reconcile these interpretations, which are essentially equivalent.

[10] Note that lexical uncertainty is a form of lexical ambiguity, but is the special form in which the choice of ambiguous form is lifted to the pragmatic listener.

The use of lifted semantic indices, or lexical uncertainty, can account for a number of puzzling facts about language use beyond those considered above. The original motivation for introducing these ideas (Bergen *et al.*, 2012) was to explain the Division of Pragmatic Labor (Horn, 1984) : why are (un)marked meanings assigned to (un)marked utterances, even when the utterances have the same literal semantics? The basic RSA framework cannot explain this phenomena. If however we assume that the meanings can each be refined to more precise meanings, the correct alignment between utterances and interpretations is achieved.

An important question is raised by this section: which, if any, ambiguities or under-specifications in meaning are resolved at the literal listener level, and which are lifted to the pragmatic listener? This choice has subtle but important consequences for interpretation, as illustrated above for scalar adjectives, but it is empirical question that must be examined for many more cases before we are in a position to generalize.

5 Conclusion

In this chapter we have illustrated the use of probabilistic modeling to study natural language semantics and pragmatics. We have described how stochastic λ-calculus, as implemented in Church, provides compositional tools for probabilistic modeling. These tools helped us to explicate the relationship between linguistic meaning, background knowledge, and interpretation.

On the one hand we have argued that uncertainty, formalized via probability, is a key organizing principle throughout language and cognition. On the other hand we have argued, by example, that we must still build detailed models of natural language architecture and structure. The system we have described here provides important new formalizations of how context and background knowledge affect language interpretation—an area in which formal semantics has been largely silent. Yet the enterprise of formal semantics has been tremendously successful, providing insightful analyses of many phenomena of sentence meaning. Because compositional semantics plays approximately its traditional role within our architecture, many of the theoretical structures and specific analyses will be maintained. Indeed, seen one way, our probabilistic approach merely augments traditional formalizations with a theory of interpretation in context—one that makes good on many promissory notes from the traditional approaches.

There are several types of uncertainty and several roles for uncertainty in the architecture we have described. While the fundamental mechanisms for representing and updating beliefs are the same for discrete variables (such as those that lead to scope ambiguity for quantifiers) and continuous variables (such as the threshold variable we used to interpret scalar adjectives in the positive form), there are likely to be phenomenological differences as well as similarities. For instance, continuous variables lend themselves to borderline cases in a way that discrete variables don't, while both support graded judgements. Similarly, the point at which a random variable is resolved—within the literal listener, in the pragmatic listener, or both—can have profound effects on its role in language understanding. Variables restricted to the literal listener show plausibility but not informativity effects; variables in the pragmatic listener that are not indices show informativity but limited context sensitivity; etc. Overall then, uniform mechanisms of uncertainty can lead to heterogeneous phenomenology of language understanding, depending on the structure of the language understanding model.

In the architecture we have described, uncertainty is pervasive through all aspects of language understanding. Pervasive uncertainty leads to complex interactions that can be described by joint inference of the many random choices involved in understanding. Joint inference in turn leads to a great deal of flexibility, from non-monotonic effects such as explaining away (section 2), through ambiguous compositional structure (section 2.5) and pragmatic strengthening (section 3), to vagueness and context-specificity of indices (section 4). It is particularly important to note that even when the archi-

tecture specification is relatively modular, for instance separate specification of world knowledge (the `...theory...`) and meaning interpretation (the `meaning` function), the inferential effects in sentence interpretation will have complex, bi-directional interactions (as in the interaction of background knowledge and quantifier scope ambiguity in section 2). That is, language understanding is analyzable but not modular.

6 Acknowledgements

We thank Erin Bennett for assistance preparing this chapter, including simulations and editting. We thank Henk Zeevat, Shalom Lappin, Scott Martin, and Adrian Brasoveanu for helpful comments on early versions of this chapter or related presentations.

This work was supported in part by a John S. McDonnell Foundation Scholar Award (NDG), and Office of Naval Research grants N000141310788 and N000141310287 (NDG).

References

Abelson, Harold & Gerald Jay Sussman (1983), Structure and interpretation of computer programs .

Adams, Ernest W. (1975), *The logic of conditionals: An application of probability to deductive logic*, Springer.

Altmann, Gerry & Mark Steedman (1988), Interaction with context during human sentence processing, *Cognition* 30(3):191–238.

Barendregt, Hendrik Pieter (1985), *The lambda calculus: Its syntax and semantics*, volume 103, North Holland.

Barker, Chris (2005), Remark on jacobson 1999: Crossover as a local constraint, *Linguistics and Philosophy* 28(4):447–472.

Bartsch, Renate & Theo Vennemann (1973), *Semantic structures: a study in the relation between semantics and syntax*, Athenäum.

Beaver, David & Brady Clark (2008), *Sense and sensitivity: How focus determines meaning*, Wiley-Blackwell, ISBN 1405112646.

Benz, Anton, Gerhard Jäger, & Robert van Rooij (2005), *Game theory and Pragmatics*, Palgrave Macmillan.

Bergen, L., N.D. Goodman, & R. Levy (2012), That's what she (could have) said: How alternative utterances affect language use, in *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*.

Black, Max (1937), Vagueness. an exercise in logical analysis, *Philosophy of science* 4(4):427–455.

Bod, R., J. Hay, & S. Jannedy (2003), *Probabilistic linguistics*, The MIT Press.

Borel, Émile (1907), Sur un paradoxe économique: Le sophisme du tas de blé et les vérités statistiques, *Revue du Mois* 4:688–699.

Chater, Nick, Christopher D Manning, *et al.* (2006), Probabilistic models of language processing and acquisition, *Trends in cognitive sciences* 10(7):335–344.

Chater, Nick & Mike Oaksford (2008), *The probabilistic mind: Prospects for Bayesian cognitive science*, Oxford University Press.

Cohen, Ariel (1999a), Generics, frequency adverbs and probability, *Linguistics and Philosophy* 22:221–253.

Cohen, Ariel (1999b), *Think generic! The Meaning and Use of Generic Sentences*, CSLI.

Cooper, Robin, Simon Dobnik, Shalom Lappin, & Staffan Larsson (2014), A probabilistic rich type theory for semantic interpretation, in *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, (72–79).

Crain, Stephen & Mark Steedman (1985), On not being led up the garden path: The use of context by the psychological parser :320–358.

Cruse, D Alan (2000), *Meaning in language*, volume 2, Oxford University Press Oxford.

Edgington, Dorothy (1995), On conditionals, *Mind* 104(414):235, doi:10.1093/mind/104.414.235.

Edgington, Dorothy (1997), Vagueness by degrees, in R. Keefe & P. Smith (eds.), *Vagueness: A Reader*, MIT Press, (294–316).

Égré, Paul (2011), Perceptual ambiguity and the sorites, in *Vagueness in Communication*, Springer, (64–90).

Frank, M.C. & N.D. Goodman (2012), Predicting pragmatic reasoning in language games, *Science* 336(6084):998–998.

Franke, M. (2009), *Signal to act: Game theory in pragmatics*, Ph.D. thesis, Institute for Logic, Language and Computation, University of Amsterdam.

Frazee, Joey & David Beaver (2010), Vagueness is rational under uncertainty, *Proceedings of the 17th Amsterdam Colloquium* .

Freer, Cameron E & Daniel M Roy (2012), Computable de finetti measures, *Annals of Pure and Applied Logic* 163(5):530–546.

Gamut, L.T.F. (1991), *Logic, Language, and Meaning, volume 1: Introduction to Logic*, volume 1, University of Chicago Press.

Geurts, Bart (2010), *Quantity implicatures*, Cambridge University Press.

Ginzburg, J. (1995), Resolving questions, I, *Linguistics and Philosophy* 18(5):459–527.

Goodman, Noah D., Vikash K. Mansinghka, Daniel Roy, Keith Bonawitz, & Joshua B. Tenenbaum (2008a), Church: A language for generative models, in *Uncertainty in Artificial Intelligence 2008*.

Goodman, Noah D & Andreas Stuhlmüller (2013), Knowledge and implicature: Modeling language understanding as social cognition, *Topics in cognitive science* 5(1):173–184.

Goodman, Noah D., Joshua B. Tenenbaum, Jacob Feldman, & Thomas L. Griffiths (2008b), A rational analysis of rule-based concept learning, *Cognitive Science* 32(1):108–154.

Graff, Delia (2000), Shifting sands: An interest-relative theory of vagueness, *Philosophical Topics* 20:45–81.

Grice, H. Paul (1989), *Studies in the Way of Words*, Harvard University Press.

Griffiths, Thomas L., Charles Kemp, & Joshua B. Tenenbaum (2008), Bayesian models of cognition, in R. Sun (ed.), *Cambridge Handbook of Computational Psychology*, Cambridge University Press, (59–100).

Hampton, J.A. (2007), Typicality, graded membership, and vagueness, *Cognitive Science* 31(3):355–384.

Heim, Irene (1982), *The semantics of definite and indefinite noun phrases*, Ph.D. thesis.

Heim, Irene (1992), Presupposition projection and the semantics of attitude verbs, *Journal of Semantics* 9(3):183, ISSN 0167-5133.

Heim, Irene & Angelika Kratzer (1998), *Semantics in Generative Grammar*, Blackwell.

Hendriks, H.L.W. (1993), *Studied flexibility: Categories and types in syntax and semantics*, Institute for Logic, Language and Computation.

Hersh, Harry M & Alfonso Caramazza (1976), A fuzzy set approach to modifiers and vagueness in natural language., *Journal of Experimental Psychology: General* 105(3):254.

Hindley, James Roger & Jonathan Paul Seldin (1986), *Introduction to Combinators and (Lambda) Calculus*, volume 1, Cambridge [Cambridgeshire]; New York: Cambridge University Press.

Horn, Laurence (1984), Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature, in Deborah Schiffrin (ed.), *Meaning, Form, and Use in Context: Linguistic Applications*, Georgetown University Press, (11–42).

Jackendoff, Ray (1983), *Semantics and cognition*, volume 8, The MIT Press.

Kao, Justine T, Leon Bergen, & Noah D Goodman (2014), Formalizing the pragmatics of metaphor understanding, in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Kao, Justine T, Jean Y Wu, Leon Bergen, & Noah D Goodman (????), Nonliteral language understanding for number words, under review.

Kennedy, C. (2007), Vagueness and grammar: The semantics of relative and absolute gradable adjectives, *Linguistics and Philosophy* 30(1):1–45.

Kennedy, Chris (1997), *Projecting the adjective: The syntax and semantics of gradability and comparison*, Ph.D. thesis, U.C., Santa Cruz.

Kolmogorov, Andrey (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Julius Springer.

Lakoff, George (1973), Hedges: A study in meaning criteria and the logic of fuzzy concepts, *Journal of philosophical logic* 2(4):458–508.

Lakoff, George (1987), Women, fire, and dangerous things: What categories reveal about the mind .

Lassiter, D. (2011), Vagueness as probabilistic linguistic knowledge, in R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (eds.), *Vagueness in Communication*, Springer, (127–150).

Lassiter, Daniel (2014), Adjectival modification and gradation, in Shalom Lappin & Chris Fox (eds.), *Handbook of Contemporary Semantic Theory, 2^{nd} edition*, Blackwell.

Lassiter, Daniel & Noah D. Goodman (2013), Context, scale structure, and statistics in the interpretation of positive-form adjectives, to appear in *Semantics & Linguistic Theory (SALT) 23*.

Lawry, Jonathan (2008), Appropriateness measures: an uncertainty model for vague concepts, *Synthese* 161(2):255–269.

Lewis, David (1970), General semantics, *Synthese* 22(1):18–67.

Lewis, David (1979), Scorekeeping in a language game, *Journal of Philosophical Logic* 8(1):339–359, ISSN 0022-3611, doi:10.1007/BF00258436.

Lewis, Davis (1980), Index, context, and content, in Stig Kanger & Sven Öhman (eds.), *Philosophy and Grammar*, Reidel, (79–100).

Luce, R.D. (1959), *Individual choice behavior: A theoretical analysis*, John Wiley.

May, Robert (1977), *The grammar of quantification*, Ph.D. thesis, Massachusetts Institute of Technology.

Montague, Richard (1973), The proper treatment of quantification in ordinary English, in J. Hintikka, J. Moravcsik, & P. Suppes (eds.), *Approaches to Natural Language*, Reidel, volume 49, (221–242).

Murphy, Gregory (2002), *The Big Book of Concepts*, MIT Press.

Oaksford, Mike & Nick Chater (2007), *Bayesian rationality: The probabilistic approach to human reasoning*, Oxford University Press.

Partee, Barbara (1987), Noun phrase interpretation and type-shifting principles, *Studies in discourse representation theory and the theory of generalized quantifiers* 8:115–143.

Partee, Barbara & Mats Rooth (1983), Generalized conjunction and type ambiguity, *Formal Semantics: The Essential Readings* :334–356.

Pearl, Judea (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, ISBN 1558604790.

Piantadosi, Steven T., Noah D. Goodman, Benjamin A. Ellis, & Joshua B. Tenenbaum (2008), A bayesian model of the acquisition of compositional semantics, in

Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society, (1620–1625).

Piantadosi, Steven T, Joshua B Tenenbaum, & Noah D Goodman (2012), Bootstrapping in a language of thought: A formal model of numerical concept learning, *Cognition* 123(2):199–217.

Ramsey, Norman & Avi Pfeffer (2002), Stochastic lambda calculus and monads of probability distributions, in *ACM SIGPLAN Notices*, ACM, volume 37, (154–165).

Roberts, Craige (2012), Information structure in discourse: Towards an integrated formal theory of pragmatics, *Semantics & Pragmatics* 5:1–69.

Rosch, Eleanor (1978), Principles of categorization, in Eleanor Rosch & Barbara B. Lloyd (eds.), *Cognition and categorization*, Lawrence Erlbaum, (27–48).

Russell, Benjamin (2006), Against grammatical computation of scalar implicatures, *Journal of semantics* 23(4):361–382.

Sauerland, Uli (2004), Scalar implicatures in complex sentences, *Linguistics and philosophy* 27(3):367–391.

Scha, Remko & Yoad Winter (2014), The formal semantics of plurality, in Shalom Lappin & Chris Fox (eds.), *Handbook of Contemporary Semantic Theory, 2^{nd} edition*, Blackwell.

Shan, Chung-chieh (2010), The character of quotation, *Linguistics and Philosophy* 33(5):417–443.

Smith, N. J., N. D. Goodman, & M. C. Frank (2013), Learning and using language via recursive pragmatic reasoning about other agents, in *NIPS 2013*.

Spivey, Michael J, Michael K Tanenhaus, Kathleen M Eberhard, & Julie C Sedivy (2002), Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution, *Cognitive Psychology* 45(4):447–481.

Stalnaker, R. (1978), Assertion, in P. Cole (ed.), *Syntax and Semantics 9: Pragmatics*, Academic Press.

Steedman, Mark (2001), *The Syntactic Process*, MIT press.

Steedman, Mark (2012), *Taking Scope: The Natural Semantics of Quantifiers*, MIT Press.

Stuhlmueller, A. & N. D. Goodman (2013), Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs, *Journal of Cognitive Systems Research* .

Sutton, Peter (2013), *Vagueness, Communication and Semantic Information*, Ph.D. thesis.

Sutton, R.S. & A.G. Barto (1998), *Reinforcement learning: An introduction*, MIT Press.

Taylor, John R (2003), *Linguistic Categorization*, Oxford University Press.

Tenenbaum, J.B., C. Kemp, T.L. Griffiths, & N.D. Goodman (2011), How to grow a mind: Statistics, structure, and abstraction, *Science* 331(6022):1279.

Van Kuppevelt, Jan (1995), Discourse structure, topicality and questioning, *Journal of linguistics* 31(1):109–147.

Veltman, Frank (1996), Defaults in update semantics, *Journal of Philosophical Logic* 25(3):221–261, ISSN 0022-3611, doi:10.1007/BF00248150.

Vogel, Adam, Max Bodoia, Christopher Potts, & Dan Jurafsky (2013), Emergence of Gricean maxims from multi-agent decision theory, in *Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the*

Association for Computational Linguistics, Association for Computational Linguistics, Atlanta, Georgia, (1072–1081).

Wingate, David, Andreas Stuhlmueller, & Noah D Goodman (2011), Lightweight implementations of probabilistic programming languages via transformational compilation, in *Proceedings of the 14th international conference on Artificial Intelligence and Statistics*, (131).

Zadeh, Lotfi Asker (1971), Quantitative fuzzy semantics, *Information sciences* 3(2):159–176.