CHAPTER 3

CONSTRAINT-BASED PRAGMATIC PROCESSING

JUDITH DEGEN AND MICHAEL K. TANENHAUS

3.1 The problem

How do listeners make sense of language? Bottom-up information from the linguistic signal (whether spoken or written) is never processed in isolation. Higher level (top-down) expectations combine with bottom-up information to determine how incoming information is interpreted. The role of top-down information has become increasingly important within an emerging perspective that the brain is a predictive machine with expectations about the upcoming signal that the actual input may match to a greater or lesser extent (for a review and proposal, see Clark, 2013).

While language processing in general addresses the question of how bottom-up information from the linguistic signal is integrated with top-down expectations generated from information about the discourse context (speaker identity, Question Under Discussion (QUD), common ground, utterance alternatives) and world knowledge, these questions take on a specific flavour in pragmatics, where the central questions focus on how context contributes to meaning.

Theories of information processing often distinguish between *automatic processes*, which are fast, dumb, and are not resource intensive; and *strategic processes*, which are slow, flexible, and resource intensive (Posner & Snyder, 1975; Neely, 1977; Shiffrin & Schneider, 1977; see also Kahneman, 2011, for a related framework.). Given the speed of online comprehension and dependence on literal meaning, it is often assumed that the context-dependent inferences, which are central to pragmatic processing, are strategic and, unless pre-compiled, must necessarily lag behind computation of literal meaning.

In contrast we have argued that whereas *unconstrained* inference or inference in situations of cue conflict might be slow and costly, inference that is *constrained* by rich conversational context and natural use of linguistic forms can be remarkably fast and easy (Degen & Tanenhaus, 2016). Moreover, because language unfolds over time, just as lexical and syntactic expectations are available to predict and explain the linguistic signal, so are expectations from constrained contexts. From this perspective, it is unexpected linguistic forms in unnatural contexts that make language processing slow and costly.

In this chapter, we argue for a *constraint-based* account of pragmatic processing. We begin by describing central features of constraint-based approaches to language processing and then consider how constraint-based accounts have furthered our understanding in other domains, using syntactic ambiguity resolution as an example. We then explore the logical space of possibilities for accounts of pragmatic processing. Finally, we review one area of pragmatic processing in depth—scalar implicature—before discussing the application of constraint-based approaches to other areas of pragmatics.

3.2 Constraint-based approaches to language

There are four defining characteristics of constraint-based approaches. First, as an utterance unfolds, listeners rapidly integrate multiple probabilistic sources of information in a weighted manner. Second, listeners generate expectations of multiple types about the future, including the acoustic/phonetic properties of utterances, syntactic structures, referential domains, and possible speaker meanings. Third, speakers and listeners can rapidly adjust expectations to different speakers and different situations, etc. Fourth, explanations that depend upon architectural constraints (e.g. information-encapsulated modules, discrete sequential processing stages) are only considered as a last resort.

Many early explanations of the speed and efficiency of language comprehension were inconsistent with a constraint-based approach. The speed of processing was attributed to encapsulated, autonomous subsystems or modules, which perform specialized computations, much like a worker on an assembly line (Forster, 1979; Fodor, 1983).

Syntactic ambiguity was an important test domain: is ambiguity resolved by domain-specific processes or by appeal to multiple sources of information, including the general context of an utterance? The influential Garden Path model of syntactic parsing proposed that syntactic heuristics were privileged in processing compared to other information sources. Initial results suggested that listeners do indeed make initial commitments without taking context into account (Ferreira & Clifton, 1986; for a review, see Frazier, 1987). However, as ideas about probabilistic constraints, including those provided by context, became more refined, evidence that listeners rapidly integrate multiple constraints emerged (for proposals and reviews, see MacDonald et al., 1994; Tanenhaus & Trueswell, 1995; Jurafsky, 1996; Altmann, 1998; Gibson & Pearlmutter, 1998; Levy, 2008).

Several factors were crucial for the emergence and convincing evaluation of successful constraint-based models. The relevant constraints needed to be identified and quantified. It was crucial to understand the central role of context and to develop methods that allowed language processing to be explored in well-defined contexts, which often required having clearly specified goal structures. Explicit testable linking hypotheses had to be articulated that mapped predictions of models onto dependent measures (e.g. reading times). Finally, it was important to appreciate the rich sources of information provided by the signal. In the following, we only implicitly address the state of pragmatic processing with respect to these five factors, but we include an explicit discussion at the end of the chapter.

3.3 Leading example: Scalar implicature, the drosophila of experimental pragmatics

The literal content of the sentence in (1) is given in (1a). However, an utterance of (1) is likely to give rise to the scalar inference in (1b). Since Grice (1975), the accepted explanation is this: upon observing (1), the listener reasons that, if the speaker was cooperative and knowledgeable, he should have produced the relevant stronger alternative in (1c). Because he did not, he can be assumed to be implicating its negation, thus arriving at the inference in (1b).

(1) Some of the people in the White House are nuts.
   a. Some, and possibly all, of the people in the White House are nuts.
   b. Some, but not all, of the people in the White House are nuts.
   c. All of the people in the White House are nuts.

Questions around scalar implicature involve the regularity, probability, and strength with which the inference arises, and whether listeners necessarily pass through an informationally privileged stage of literal (or conventional) processing before integrating contextual information that results in enrichment of the literal meaning to yield the inference (or the cancellation of the inference to yield the literal meaning). In contrast to these informational privilege accounts, constraint-based accounts propose that both the probability with which the inference arises and the speed with which it is processed are functions of the contextual support for the implicature (i.e. that the speaker intended to communicate the negation of the stronger alternative).

3.4 Constraint-based approaches to pragmatics

Constraint-based approaches to pragmatics share the assumptions of constraint-based approaches laid out in section 3.2. We focus here specifically on how the meaning of an utterance is processed. We do not commit ourselves to talking only about phenomena that fall solely within the domain of pragmatics; we, like the rest of the field, do not know where to draw the line between semantics and pragmatics (Szabó, 2005). Therefore, whenever we speak of 'pragmatic processing' we have in mind a listener processing information in service of inferring the speaker's intended meaning.

How are the multiple sources of information integrated in online pragmatic processing? Views of this process differ along at least two separable dimensions which span two extreme positions and many nuanced ones. One relates to when a piece of information is assumed to be processed compared to others; the other to how strongly it is weighted in the resulting interpretation. In the syntactic parsing example mentioned in section 3.2, much of the discussion focused on when contextual pragmatic information was used—that is, whether it was necessary to assume that there are distinct stages of processing. Researchers interested in the 'when' question typically use online measures of processing like eye movements in the visual world and, in reading, Event-Related Potentials (ERPs) and self-paced reading times. Some information about processing effort can also be gleaned from response times in truth-value judgement tasks. Researchers interested in how strongly a piece of information enters the resulting interpretation typically use offline measures like judgement data from truth-value judgement tasks, survey studies, and choices in interpretation selection tasks.

The two extreme positions are:

1. *Extreme informational privilege*: some types of information are privileged over others. In online processing, they are processed earlier and weighted most heavily in the resulting interpretation.
2. *Extreme parallelism*: all available information is processed in parallel. In online processing, all available information is processed simultaneously and weighted equally in the resulting interpretation.

Between these extremes lie other logical possibilities: for example, certain types of information may be privileged in processing but not in the resulting interpretation; or weighted more heavily in the resulting interpretation but nevertheless processed in parallel with other less heavily weighted types of information. Another intermediate position is that there is no principled default privilege of any type of information; how information is processed depends on how useful it is in the immediate context and how useful it has been in the listener's linguistic experience. Intermediate positions that cluster close to caricature 2 can be considered constraint-based theories. Positions that cluster close to caricature 1 have different names depending on the type of information that is claimed to be privileged and the phenomena to which they apply.

For scalar implicature, two extreme positions about time-course (close to caricature 1) are the Literal-First hypothesis (discussed by Huang & Snedeker, 2009a, 2011; Bott et al., 2012) and the Default hypothesis (Levinson, 2000a). The Literal-First hypothesis proposes that an expression's literal meaning is computed before pragmatic factors that lead to its enrichment. The Default hypothesis proposes that an implicature conventionally associated with an expression is automatically computed before additional pragmatic factors that lead to its cancellation to yield the literal meaning. In contrast, a constraint-based account claims that the speed and probability with which an implicature is computed is a function of the contextual support it receives (Degen & Tanenhaus, 2015, 2016).

The Literal-First and Default hypotheses share the assumption of informational privilege in the same way that the Garden Path model assumed that syntactic heuristics were privileged. Using the example of scalar implicature, we survey the evidence and conclude that the data are compatible with constraint-based theories but not with informational privilege theories, unless one makes ad hoc auxiliary assumptions.

We begin by introducing factors that we assume are generally at play in pragmatic processing. Each may receive a different setting in a particular experimental context.[1]

[1] We use 'setting' to refer to different values each factor can plausibly take on.

Researchers typically focus on only one factor at a time in their experiments. They rarely control for other factors, explore their effects, or speculate on their likely factor settings and how that might influence the results. Section 3.5 is intended to be both an overview of factors that affect pragmatic processing and a checklist for researchers to consider in designing experimental tasks and conditions, and in writing paper discussions.

3.5 Constraints involved in pragmatic processing

Constraint-based approaches are appealing in their broad coverage of phenomena yet are often criticized for being too permissive. How do we know which factors listeners are likely to be sensitive to in interpretation? How strongly should we expect these factors to matter? How can a coherent theory emerge from a framework whose main message appears to be 'context matters'? We believe that the situation is much less dire in that, as for syntactic processing, a constraint-based approach to pragmatic processing lends itself to explicitly quantitative investigations in ways that informational privilege approaches do not.

We now survey some factors that we propose are relevant for deriving any type of pragmatic inference and indeed, for processing any type of utterance, which include the QUD, world knowledge, properties of the utterance and its alternatives, properties of the speaker, and common ground. We distinguish these from more idiosyncratic inference-specific factors, some of which are discussed in section 3.6.

3.5.1 Question Under Discussion

An important observation going back at least to Grice (1975) is that language is interpreted with respect to a QUD (Roberts, 2012a) or goal that listeners expect the speaker to be addressing. The QUD establishes states of the world worth distinguishing for interlocutors. It need not be explicit, and in fact often isn't. For example, the precision with which speakers answer the overt question 'Where are you?' depends on the relevance of being maximally precise with respect to interlocutors' joint goal (Potts, 2012). There may also be uncertainty about the QUD, which is humorously illustrated in a dialogue between Harry and Jess in the film *When Harry Met Sally*:[2]

(2) Jess: If she's so great why aren't YOU taking her out?
Harry: How many times do I have to tell you, we're just friends.
Jess: So you're saying she's not that attractive.
Harry: No, I told you she IS attractive.
Jess: But you also said she has a good personality.
Harry: She HAS a good personality.

[2] This example was first made famous by Larry Horn (2004).

Jess: [Stops walking, turns around, throws up hands, as if to say 'Aha!']
Harry: What?
Jess: When someone's not that attractive they're ALWAYS described as having a good personality.
Harry: Look, if you were to ask me what does she look like and I said she has a good personality, that means she's not attractive. But just because I happen to mention that she has a good personality, she could be either. She could be attractive with a good personality or not attractive with a good personality.
Jess: So which one is she?
Harry: Attractive.
Jess: But not beautiful, right?

Harry's explanation reveals that he and Jess were assuming that different QUDs were being addressed by 'She has a good personality', perhaps 'What does Sally look like?' or 'What is Sally's best feature?', as opposed to 'What are some general features of Sally?' or 'What is Sally's personality like?' Interpreting the utterance as the response to one of the former QUDs yields a very different interpretation than as a response to one of the latter.

Although there are to date no good empirical measures of the QUD, researchers have had some success in manipulating the presumed implicit QUD via cover stories (Zondervan, 2009, 2010; Degen, 2013; Degen & Goodman, 2014). In production tasks, manipulating the QUD has direct effects on the specificity of produced utterances (Potts, 2012). In visual world studies in particular, the QUD is often manipulated via the participants' task; for example, referent identification by clicking, pointing, or performing an action makes salient the QUD 'Which of the pictured images is the target?' (Sedivy et al., 1999). The QUD is also increasingly serving as an explanation for why the prejacents of presupposition triggers project less frequently than previously assumed (Beaver et al., 2017); for instance, the Projection Principle formulated by Simons et al. (2010) proposes that content only projects if it doesn't address the QUD. While a systematic empirical exploration of this principle is still outstanding, it exemplifies how widely useful the QUD is as an explanation for listeners' interpretation.

3.5.2 World knowledge

World knowledge (i.e. prior beliefs) strongly guides utterance interpretation. For instance, the pronoun 'they' is generally interpreted to refer to the city council in (3a) but to the demonstrators in (3b) (Kehler et al., 2008; adapted from Winograd, 1972b).

(3) The city council denied the demonstrators a permit because
a. they feared violence.
b. they advocated violence.

Similarly, listeners use world knowledge about object affordances in reference resolution (Chambers et al., 2004) and the interpretation of declarative sentences (Hagoort et al., 2004).

3.5.3 Properties of the observed utterance and its alternatives

Some aspects of the observed utterance related to the above factors are generally at play in pragmatic utterance interpretation: first and foremost, the cost and informativeness of an utterance compared to its alternatives (as estimated by the listener to be available to the speaker).

The speaker's utterance *cost* combines cognitive and production cost. The cost of an utterance is inherently difficult to determine and likely a function of non-independent factors (which may differ across the spoken and written modality): its frequency of occurrence, length, phonological and syntactic complexity, social value (e.g. taboo words likely have low cost in colloquial settings and high cost in formal settings), etc. How to best characterize the cost of an utterance is currently an open empirical question.

While a sentence's *informativeness* is often described in terms of its entailment relations compared to its alternatives, a more general notion of informativeness takes into account the QUD: what is often described as an utterance's relevance can be interpreted as its informativeness with respect to a partition induced by the QUD (e.g. Russell, 2012; Degen & Goodman, 2014; Kao et al., 2014).

In general, what an utterance's alternatives are is an open question. As with the QUD, there are currently no good empirical measures for estimating alternatives. Some have proposed constraints on alternatives based on complexity (Fox & Katzir, 2011); others have found that explicitly manipulating the set of utterances within experimental settings changes expectations about the available alternatives and their subsequent pragmatic inferences (Degen & Tanenhaus, 2015, 2016). The set of alternatives the listener assumes to be available for the speaker clearly affects the inferences drawn: Jess's response 'But not beautiful, right?' indicates he assumed that 'beautiful' was an available alternative to 'attractive', and thereby drew a scalar inference from 'attractive' to 'attractive, but not beautiful'.

3.5.4 Properties of the speaker

A listener's belief about the speaker's epistemic state is likely to affect interpretation of any utterance. A listener should only expect the speaker to have intended, for instance, the negation of a contextually more informative utterance, if he can rely on the speaker being in an epistemic state fit to assess the truth of the more informative utterance. This assumption of speaker knowledgeability has been discussed mainly in the context of scalar implicature as the 'competence assumption' (Geurts, 2010) or the 'epistemic step' (Sauerland, 2004), but is relevant more broadly and has also been a main focus in reference resolution (Keysar et al., 2000; Heller et al., 2008).

All of the above factors are directly related to the listener's beliefs about the speaker's cooperativity and reliability. Grice (1975) takes the assumption of speaker cooperativity for granted. However, there are many ways in which speakers are uncooperative. In politics, for instance, speakers are often less informative than required and lying is not unheard of. In advertising, QUD-relevance is often exploited and outright violated (Sedivy & Carlson, 2011). The meat section at a grocery store will invariably contain a label that says 'This meat is gluten-free!', suggesting that perhaps other meat is not (however, meat is generally gluten-free). Finally, children are less likely to learn from unreliable than from reliable speakers (Koenig & Harris, 2005; Gweon et al., 2014) and adults become less likely to draw contrastive inferences from utterances produced by unreliable compared to reliable speakers (Grodner & Sedivy, 2011).

3.5.5 Common ground

An additional complicating factor is that for each of the abovementioned factors, what is presumed to be in common ground between speaker and listener can influence interpretation. Consider the exchange in (4).

(4) A: Why didn't Jane get tenure?
B: Because she's a woman.
a. . . . so she's not good enough.
b. . . . so the field is biased against her.

Depending on the beliefs shared by A and B, either one of the inferences in (a) or (b) may be drawn from B's utterance.

Mutual knowledge cannot be computed algorithmically without resort to heuristics, famously characterized by Clark & Marshall (1981) as falling into one of three broad categories: Community Membership, Linguistic Co-Presence, and Physical Co-Presence. Each factor discussed in this section can in principle be characterized as belonging to one of these three categories. However, listeners can have uncertainty about any of the abovementioned factors themselves (e.g. what is the QUD?) and about which knowledge can be presumed to be in common ground (e.g. does my interlocutor share my beliefs about the QUD?). These complexities might seem to pose insurmountable challenges to the study of pragmatic processing. However, substantial progress has been made and can be made in the future by systematically investigating these aspects of interpretation and formalizing how they interact within a testable framework.

Next, we survey the empirical facts about scalar implicature processing in detail from a constraint-based perspective and will show that all of the abovementioned factors play a role. These factors stand in contrast with inference-specific idiosyncrasies (e.g. lexical, syntactic), which we also discuss. Throughout, we contrast the predictions of the constraint-based and the informational privilege approaches.

3.6 Constraint-based approaches to scalar implicature processing

3.6.1 What's wrong with informational privilege accounts?

Early experiments in scalar implicature processing were aimed at testing the informational privilege Default Hypothesis (Levinson, 2000a)—the idea that, because certain scalar implicatures appear conventionally tied to certain scalar expressions (e.g. ‘some’) and arise seemingly with great regularity, they should also be automatically computed in processing. Only if the context fails to support the implicature should it be cancelled, in a second step of processing, thus yielding the literal interpretation. A truth-value judgement task in which participants were asked to evaluate the truth of under-informative utterances like ‘Some elephants are mammals’ suggested that, in contrast to the predictions of the Default Hypothesis, scalar implicatures incur a processing cost (Noveck & Posada, 2003; Bott & Noveck, 2004): response times were slower for pragmatic FALSE responses (indicating that participants drew the implicature that ‘Some, but not all, elephants are mammals’) than for literal TRUE responses (indicating that participants interpreted the utterance as indicating that ‘Some, and possibly all, elephants are mammals’).

These initial results were followed by a slew of studies replicating this ‘costly implicature’ effect in many different paradigms using many different measures, including reading times (Breheny et al., 2006; Bergen & Grodner, 2012), truth-value judgements (Degen & Tanenhaus, 2015), mouse-tracking (Tomlinson et al., 2013), speed-accuracy tradeoff paradigms (Bott et al., 2012), and eye movements in the visual world (Huang & Snedeker, 2009a, 2011; Degen & Tanenhaus, 2016). In fact, so numerous are the contexts in which processing the implicature is delayed compared to literal controls that some have concluded that it is this point ‘well-known’ that ‘scalar implicatures are derived with a delay’ (Chemla & Singh, 2014). In light of the early evidence, a new informational privilege proposal was formulated for how scalar implicatures are processed: the Literal-First hypothesis (Huang & Snedeker, 2009a). Under this account, a pragmatic processing stage follows a literal processing stage, thus explaining why scalar implicatures generally show increased processing times compared to literal content. This account is also linguistically appealing because it neatly divides semantics and pragmatics into two separate stages and treats semantics as prior to pragmatics in processing, mirroring the relative positions of semantics and pragmatics in the hierarchy of linguistic representation.

There is here a clear analogy to the case of syntactic processing: the Literal-First hypothesis assumes an initial stage of context-insensitive processing of literal information, after which contextual, pragmatic information is integrated. However, the 1990s saw the downfall of informational privilege accounts of syntactic parsing, as evidence increased that contextual information is integrated in syntactic processing as soon as it becomes available. In much the same way, evidence is now mounting that there are contexts in which scalar implicatures do not incur a processing cost. Most of this evidence has come from visual world eye-tracking (Grodner et al., 2010; Breheny et al., 2013; Degen & Tanenhaus, 2016). For instance, Degen & Tanenhaus showed that scalar implicature computation for ‘some’ was delayed when number terms like ‘two’ and ‘three’ were contextually available alternatives for the speaker, but not when the speaker could only use the quantifiers ‘some’ and ‘all’. There is also evidence from reading times showing that listeners exhibit no processing cost in reading quantifiers like ‘some’ in implicature-supporting compared to non-supporting contexts (Politzer-Ahles & Fiorentino, 2013), suggesting that, in contrast to predictions of informational privilege accounts, contextual information can rapidly enter scalar implicature computation.

Taken together, the empirical evidence suggests that the answer to the question ‘Do scalar implicatures incur a processing cost?’ is: ‘It depends’. That is, neither one of the two informational privilege accounts under consideration—the Default or the Literal-First hypothesis—can capture the full range of data. Sometimes scalar implicatures are processed quickly; and sometimes they are processed slowly. Just as in the case of syntactic processing, this is where constraint-based accounts become a useful way forward, both for providing explanatory value as well as for generating new predictions.

3.6.2 Beyond informational privilege accounts

How does a constraint-based account go beyond simply claiming that scalar implicature processing is context-dependent? First, it states explicitly that both the probability and the speed with which an implicature is derived is a function of the contextual support for the implicature. Thus, if one can identify the strength of different cues to interpretation, one should be able to predict, for a novel situation with different combinations of cues, what listeners' likely interpretations will be.

Identifying the set of cues that listeners use in interpretation is a daunting task and presumably why, despite the large amount of variability in implicature rates across experiments (from 0 to 100%; see Degen, 2013, and Dieussaert et al., 2011, for overviews), there have been relatively few investigations into the different contextual factors that determine the probability of drawing the inference. However, one step in this direction has recently been taken by Degen (2015): by combining corpus analyses and web-based experiments through Amazon's Mechanical Turk, she showed that for 1,363 naturally occurring instances of 'some' in the Switchboard corpus (Godfrey et al., 1992), not only was there a large amount of variation in the strength of scalar implicatures, but this variation was systematically predictable from contextual features, including syntactic features (e.g. whether the 'some'-Noun Phrase (NP) was realized in the partitive or non-partitive form; or whether it occurred in subject position), semantic features (e.g. whether 'some' was relatively weak/indefinite or strong/quantificational), and pragmatic features (e.g. whether the embedded NP referent had been previously mentioned). Note than none of these cues are the general ones mentioned in section 3.5—speaker epistemic state, QUD, world knowledge, properties of utterances and their alternatives, speaker cooperativity, and common ground. However, some of the investigated cues are plausibly mediators for these more general cues—for instance, speakers may be more likely to use the partitive when the stronger alternative is QUD-relevant. In general, the relation between bottom-up cues to speaker meaning available in the linguistic signal itself and top-down expectations that listeners bring to bear on utterance situations is still vastly under-explored.

While a systematic investigation of the extent to which these cues matter is still lacking, their role in scalar implicature has been demonstrated in various paradigms via individual manipulations in controlled experimental settings. For instance, the speaker's epistemic state modulates the speed with which listeners draw scalar inferences (Bergen & Grodner, 2012; Breheny et al., 2013) and the probability with which listeners draw the inference (Goodman & Stuhlmüller, 2013). The QUD modulates the probability with which listeners draw scalar inferences (Zondervan, 2010; Degen, 2013; Degen et al., 2014). World knowledge in the form of listeners' prior beliefs about the effect of certain actions on objects modulates the probability with which listeners draw scalar inferences (Degen et al., 2015). Properties of utterances themselves—both the observed utterance and its alternatives—modulate the probability of a scalar inference. For example, when listeners believe the speaker could have used number terms instead of 'some', implicatures are drawn less often (Degen & Tanenhaus, 2016; see also Skordos & Barner, Chapter 2 in this volume, for the importance of reasoning about alternatives in children's scalar inferences). The costlier the stronger alternative, the less likely the implicature is drawn (Rohde et al., 2012; Degen et al., 2013). Contrastive stress on the scalar increases the probability of the implicature (Cummins & Rohde, 2015; De Marneffe & Tonhauser, to appear). And while the investigation of speaker reliability effects on implicatures is even more sparse, there is evidence that in contexts in which the stronger alternative is face-threatening for the listener, implicatures are drawn less frequently (Bonnefon et al., 2009), suggesting that using the weaker alternative is also a device for speakers to avoid being impolite.

In addition, resource limitations affect the probability with which scalar implicatures are drawn. An increasing number of studies have shown that listeners are less likely to respond pragmatically when under cognitive load (De Neys & Schaeken, 2007; Dieussaert et al., 2011; Marty & Chemla, 2013). While some see this as conclusive evidence that scalar implicatures incur a cost, the constraint-based perspective suggests an alternative explanation: each of these studies used the Bott & Noveck (2004) style 'Some elephants are mammals' stimuli. For these stimuli, world knowledge does not support the implicature (i.e. we already know that all elephants are mammals), while the need to retain the assumption that the speaker is being informative does. This constitutes a situation of *cue conflict*. Under constraint-based accounts, the resolution of cue conflicts is one of the contributors to processing difficulty (Elman et al., 2004). It may thus be the resolution of this cue conflict in favour of speaker cooperativity and against world knowledge that is incurring the cost. The constraint-based approach predicts that if one were to set up situations where the implicature is strongly supported by the context, then increasing cognitive load should lead to *more* implicatures.

The variation in probability and speed with which scalar inferences are drawn has various implications, both methodological and theoretical. Methodologically, it suggests that experimenters, when investigating a particular factor in a particular paradigm, should be vigilant of the settings the other factors are likely to take on in their experiment. Not manipulating a factor explicitly does not mean that participants don't assign it a setting. For instance, when investigating the effect of the QUD on scalar inferences it doesn't matter how strong the manipulation is; if participants believe that the speaker is unlikely to know about the truth of the stronger alternative (or if they have uncertainty about whether she is), then it will appear as though there is no effect of the QUD because participants will at most draw ignorance inferences. We believe that for the factors we have listed above, a researcher should have a good estimate of that factor's setting in his experiment (or figure out a way to measure it). Otherwise, rather than evaluating hypotheses about naturalistic language interpretation, the results might reflect effects of complicated meta-linguistic reasoning.

The above review suggests that informational privilege accounts like the Default or the Literal-First hypothesis are on a weak footing. Constraint-based accounts offer promise as a unifying framework within which to explore scalar implicature processing (see Breheny, Chapter 4 in this volume, for a similar review and conclusion, though couched within a slightly different framework).

3.7 Constraint-based approaches in other areas of pragmatics

Discussing constraint-based approaches to all areas of pragmatics is far outside the scope of this chapter. Other chapters in this Handbook highlight the ways in which multiple constraints affect referential choice (Davies & Arnold in Chapter 28); focus (Kim, Chapter 25; Tonhauser, Chapter 29); exhaustivity inferences associated with *it*-clefts (Onea, Chapter 24); and negation (Tian & Breheny, Chapter 12). Here we touch on the application of constraint-based approaches to a few additional domains.

3.7.1 Perspective-taking

Physical co-presence, operationalized as information that interlocutors have visual access to, has served as a testing ground for asking when interlocutors can take into account differences in each other's perspectives. Experimenters manipulate which information is common (shared) and privileged (available only to one interlocutor) (Keysar et al., 2000). Physical co-presence has been used to address questions about production and comprehension. Production studies under the rubric 'audience design' (Clark & Murphy, 1982) ask to what extent speakers tailor their utterances for their interlocutors. Comprehension studies, under the rubric of 'perspective-taking' (Keysar et al., 2000; Heller et al., 2008), ask to what extent listeners consider differences between their knowledge and that of the speaker in online processing. We focus here on comprehension, noting that the literature has followed the same trajectory as we have discussed for syntactic ambiguity and scalar implicature.

There was always consensus that listeners *can* take the speaker's perspective. The question was how heavily they weigh differences in perspective and whether there is an informationally privileged stage of egocentric-first processing. The rationale for an initial egocentric stage is that (a) taking into account an interlocutor's perspective is resource-demanding; (b) the listener's own perspective is a good proxy for the speaker's perspective; and (c) errors can be corrected when miscommunication occurs. Competing constraint-based accounts assume that multiple probabilistic constraints, which include beliefs about the reliability of physical co-presence, the amount of information provided by the specific context, and goodness of fit between a linguistic expression and potential referents, determine whether there will be strong and immediate or weaker and delayed effects of physical co-presence.

Keysar et al. (2000) provided striking evidence for initial egocentrism. A confederate speaker and a naïve listener sat on opposite sides of a box with cubbyholes: Some contained objects both participants could see (shared objects); others could only be seen by the listener (privileged objects). The speaker directed the listener to move objects. When a privileged object was a better fit for the referential expression than a shared object (e.g. 'tape' is the referring expression: a roll of sticky tape is privileged, and a cassette tape is shared), listeners typically looked first to the privileged object and often began to reach for it.

Subsequent research by Nadig & Sedivy (2002) strongly qualified the interpretation of these results. When potential referents are equally good fits for the referring expression and the description is felicitous (Hanna et al., 2003; Heller et al., 2008), listeners neither look at nor reach for the privileged object. Hawkins & Goodman (2016), using a production task, demonstrated that the referring expressions used by Keysar et al. (2000) are less informative than those generated by naïve speakers. Ironically, then, apparent failures of perspective-taking might be attributable to sophisticated expectations about speaker behaviour—that is, to perspective-taking. An explicitly constraint-based approach to resolving the apparent conflict between the Keysar et al. (2000) and Heller et al. (2008) studies, couched in Bayesian terms (Heller et al., 2016), predicts the differences by assuming that listeners consider and apply appropriate weights to privileged and common ground. Listeners need to monitor both objects in privileged and shared ground for independent reasons: for example, a speaker asking an information-seeking question is most likely referring to an unseen object. Indeed, when a confederate asks a question, listeners rapidly direct their attention to cubbyholes with privileged objects (Brown-Schmidt et al., 2008; Brown-Schmidt, 2009).

3.7.2 Lexical precedents

A question that is ripe for a constraint-based approach is whether there are immediate partner-specific effects on lexical precedents, as argued by Metzing & Brennan (2003). Consider a situation where an ambiguous object was previously labelled a vase or a funnel and then subsequently referred to by the same name, *maintaining* the precedent, or by a different name, *breaking* the precedent by either the same speaker or a different speaker. Based on a meta-analysis of results, Kronmüller & Barr (2007) documented three different effects with different time-courses. They proposed that each arises from distinct processes: (a) an early automatic priming effect, which leads to a small, fast same-speaker advantage for maintaining a precedent; (b) large, speaker-independent costs when a precedent is broken; and (c) a late recovery stage where costs for broken precedents are reduced when the precedent established by the first speaker is broken by a new speaker.

A constraint-based analysis would attribute effects to different stages or types of processing only a last resort after taking into account factors such as: (1) how confident the speaker was in using a name (confidence might be conveyed by different constructions and variations in prosody); (2) listeners' probabilistic beliefs about whether or not a lexical precedent used by an interlocutor was a good fit for the object; and (3) whether the listener would expect a new speaker to use the same form as the previous speaker. For example, weaker effects of broken precedents should arise when the speaker expressed less confidence.

3.7.3 Learning: adaptation, generalization, and speaker-specificity

A growing body of research in language processing motivated by how listeners cope with speaker-variability has focused on questions of speaker-specificity, adaptation, and generalization. For example, in speech perception there is a general absence of simple, reliable acoustic cues to phonemes and word boundaries. Moreover, there is variability within a talker and across talkers. Multiple factors contribute to variability, including variants of a language (e.g. American versus Australian English), smaller regional dialects (accents), socioeconomic class, and even different groups within the same high school (Eckert, 1989). From a constraint-based perspective, this leads to questions such as: (1) how do listeners adapt on the fly to specific speakers, modifying existing hypotheses based on prior knowledge? (2) How and when do listeners generalize to new speakers (see, for example, Kleinschmidt & Jaeger, 2015)?

Similar questions are beginning to be addressed in pragmatic processing. An eye-tracking study by Grodner & Sedivy (2011) suggested that when a speaker uses modification unreliably and listeners believe the speaker might be pragmatically challenged, they no longer anticipate that a prenominal scalar adjective is likely to signal contrast. This finding has recently been replicated and extended by Ryskin et al. (2016), who demonstrate similar effects from exposure to a range of utterances, without prior knowledge about the speaker's pragmatic competence. Similarly, when a talker uses contrastive focus associated with the pitch accent L+H* unreliably with pre-nominal adjectives, listeners down-weight it as a cue when it occurs in a prosodic contour (tune) in a different construction (Kurumada et al., n.d.).

Focusing on *informativity*, Pogue et al. (2016) demonstrated that when listeners are exposed to two talkers, one of whom is under-informative, they judge new under-informative utterances as more likely to have been produced by that speaker. The same pattern did not hold for over-informative talkers. This is not surprising given that an utterance that technically gives more information that is strictly needed to identify a referent might be efficient because the additional information is useful, especially when there is uncertainty in the referential domain (Graf et al., 2016; Degen et al., 2017).

In addition, Yildirim et al. (2016) showed that listeners' expectations about how quantifiers map onto specific quantities (e.g. judgements about how likely a speaker is to use 'some' to refer to a given quantity of green candies) shift with exposure to different distributions in the input.

In future work, it will be crucial for the pragmatic processing community to examine the kind of variability that speakers naturally produce and, therefore, listeners encounter, using corpus studies and production studies to determine both the utterance choices speakers make in a given situation and how those choices might vary. Here again there is a lesson to be learned from constraint-based approaches to syntactic ambiguity, where architecture-based hypotheses were tested with stimuli for which the relevant constraints had not been identified or quantified. The importance of understanding what speakers typically say and the extent to which that influences listeners' expectations is further highlighted by the trajectory of work on scalar implicature and perspective-taking.

3.8 Information integration

Across the domains of experimental pragmatics reviewed in sections 3.6 and 3.7, there does not appear to be a case where a certain type of information is systematically preferred over another in processing. Instead, which type of information is processed when, and which type of information is weighted more heavily in the resulting interpretation, varies as a function of the experimental context. This is the signature of a domain-general information processing mechanism that weighs and prioritizes information differentially.

What is the principle according to which information is integrated (weighted in the resulting interpretation; prioritized during online processing)? Under the constraint-based approach, the probability with which an interpretation arises and the speed with which it is processed is a function of the contextual support for that interpretation. But what function? And how is contextual support computed?

Because they are probabilistic, combine multiple sources of information, and typically appeal to architectural constraints as a last resort, constraint-based models share many of the same principles as rational (ideal observer) approaches to perception and cognition (Ernst & Banks, 2002; Alais & Burr, 2004; Knill, 2007). Over the past ten years, explicit computational models in the rational tradition have been developed within the emerging sub-field of probabilistic pragmatics (Franke & Jäger, 2016; Goodman & Frank, 2016). These formalize the Gricean programme by assuming that speakers try to produce utterances that strike a balance between being true and informative while minimizing utterance cost, implicitly encoding that speakers are cooperative.

Interpretation is considered to be probabilistic inference. Listeners are hypothesized to combine information about their prior beliefs and the speaker's likely utterances via Bayes' Rule, while making minimal assumptions about resource limitations, thereby accommodating the fact that rationality is bounded. The speaker's likelihood function provides a principled way to incorporate the factors discussed in sections 3.5—3.7. These models have been successful at explaining (variation in) a variety of pragmatic inferences, including scalar implicature (Russell, 2012; Goodman & Stuhlmüller, 2013; Degen et al., 2015); embedded implicatures (Potts et al., 2015; Bergen et al., 2016); hyperbole (Kao et al., 2014); interpretation of under-informative referring expressions (Frank & Goodman, 2012; Franke & Degen, 2016); and phenomena at the semantics-pragmatics interface such as the interpretation of gradable adjectives (Lassiter & Goodman, 2013) and vague quantifiers (Schöller & Franke, 2016), as well as the production and interpretation of pronouns (Rohde, Chapter 27 in this volume).

While rational computational models of pragmatics have achieved remarkable success in explaining the quantitative patterns in the *outcome* of the inference process at the sentence level, they have not yet been applied to online processing. We see this as a crucial step in developing constraint-based theories of pragmatic processing. The general question is: how is one to link computational model predictions to measures of processing effort?

One a priori possibility is that the probability of an interpretation as predicted by a rational model directly predicts the processing effort associated with arriving at that interpretation. For instance, when the probability of a scalar implicature is high, it will be computed quickly.

However, this simple idea cannot be right: in many of the Bott & Noveck (2004) style 'Some elephants are mammals' experiments, the probability of the implicature appears to be high, as evidenced by large implicature response proportions, and yet the implicature is slow to process compared to the sentence's literal interpretation.

In addition, focusing only on the resulting implicature probability obscures how that probability arises: there are many circumstances in which implicatures are highly probable, only some of which rely on cooperative speaker behaviour. For example, listeners may have uncertain prior beliefs about the world and then hear an utterance from a competent speaker that clearly addresses the QUD and that strongly biases towards the inference. In this case, one would expect the inference to arise quickly and seamlessly. By contrast, in other situations, listeners may have strong priors about what the world is like (and assume that these priors are in common ground, for example by assuming that all elephants are mammals and that this is a commonly known fact). Assuming that speakers are cooperative and trying to produce utterances that strike a balance between being true and informative, certain speaker utterances will be unexpected because they are either very unlikely to describe a true world or because they are highly under-informative, given the prior—'Some elephants are mammals' is an example of the latter.

The reading time literature on syntactic processing suggests that reading times—a proxy for processing difficulty—are logarithmically related to *surprisal* (Levy, 2008; Smith & Levy, 2013): The more unexpected the utterance, the greater the processing effort incurred by the listener/reader. We believe that this approach can be straightforwardly extended to pragmatic processing, using richer contextual utterance probabilities as the basis for the computation of surprisal. Under the rational models, the surprisal of an utterance like 'Some elephants are mammals' will be high. If listeners are drawing on their model of the speaker in comprehension, then after Bayesian inference, the most likely explanation of the utterance is that the speaker intended the implicature. This would thus constitute a case of a high probability of the implicature being intended, while nevertheless being a difficult utterance to process because of its large surprisal value.

Preliminary evidence for the promise of this 'pragmatic surprisal' view comes from Augurzky & Franke (2016), who showed using ERPs that listeners' N400 amplitude scales with the contextual surprisal of utterances like 'Some of the dots are black'. How does this relate to response times in a truth-value judgement task? While reading times and the N400 component tend to scale with utterance surprisal, truth-value judgements (and eye movements in the visual world) are likely to reflect a combination of both utterance surprisal and the resulting evidence for a particular interpretation. This requires a more complex linking function. The precise interplay of production and comprehension is an exciting avenue for future research, and one that is only made possible by building testable and incrementally refinable computational models that make quantitative predictions about the complex interplay between the various factors we have discussed in this chapter.

3.9 Quo vadis, constraint-based pragmatics?

In the introduction, we laid out five developments that were crucial for the emergence of successful constraint-based models in syntactic processing. We use these as a guide in assessing the current state of constraint-based approaches in experimental pragmatics. As a reminder, the developments were (a) identifying and quantifying relevant constraints; (b) understanding the central role of context; (c) clearly specifying goal structures;

(d) developing explicit testable linking hypotheses; and (e) appreciating the richness of the signal. Along the way we will try to identify further promising questions for future work that we think will advance the field.

Experimental pragmatics is in the early stages of identifying and quantifying the constraints involved in pragmatic processing. The initial focus on questions about how quickly pragmatic inferences are computed for a limited set of phenomena played a catalytic role in helping the field develop. But it also detracted both from the complexity of pragmatic inference and from efforts to understand the signal, that is, what speakers actually produce, which is a crucial part of understanding both how to prepare stimuli and how listeners process the necessarily artificial input provided to them by experimenters. One way of identifying and quantifying the constraints involved in pragmatic processing is to combine corpus studies with web-based experiments (e.g. Degen, 2015) to collect large numbers of judgements about, for instance, the role of each of the factors discussed in section 3.5 in the interpretation of naturally occurring language. Combining this distributional information with computational modelling and controlled psycholinguistic experiments will further allow for incrementally testing and refining explicit constraint-based theories of pragmatic processing that embrace the complexity of the phenomenon. While this might seem daunting, it is worth noting that studying vision in more complex natural tasks provided insights that ultimately simplified some problems (see Salverda et al., 2011). For instance, models of fixations in searching a scene on a screen account for less than 30 per cent of the variance, suggesting that there are numerous unexplained factors contributing to fixations. In more complex goal-based tasks (e.g. making tea or making a sandwich), more than 90 per cent of fixations are accounted for by the goal structure.

Part of this process is acknowledging the central role of context, a development that is clearly underway. Awareness of clearly specified goal structures in experimental investigations of pragmatic phenomena is still somewhat lacking in certain areas of experimental pragmatics (e.g. scalar implicature processing), while in other areas the experiment-level QUD or goal of the experimental task forms the basis of paradigms (e.g. visual world eye-tracking for investigating contrastive inferences from prenominal adjectives, where the goal, at least when coupled with a task (for discussion, see Salverda et al., 2011; and Salverda & Tanenhaus, 2017), involves referent identification and the QUD is therefore 'Which of these images does the speaker intend me to select?').

Formulating explicit, testable linking hypotheses is often ignored in experimental studies in language processing, and it has not been a strong suit of experimental pragmatics. The emergence of computational models in probabilistic pragmatics has brought with it the need to clearly specify the link between model predictions and various dependent measures of utterance interpretation (Degen & Goodman, 2014), which is only exacerbated once one moves to online processing measures. Formulating and evaluating linking hypotheses will play an essential role in developing and testing scalable models of pragmatic processing.

Finally, as touched upon earlier, experimental pragmatics quite generally would benefit from researchers appreciating the richness of the signal more than has hitherto been the case, via systematic investigations of naturally occurring language that are not subject to potential researcher bias in sampling relevant examples. Moreover, it will be increasingly important to take into account information provided by the speech signal, such as prosody, which provides important cues to a speaker's intentions and the myriad other cues in the speech stream that are partially linked to information.

Making progress on all of these fronts will facilitate the building of explicit models of the rich interplay between bottom-up information (from the signal and context) and top-down information (conversational expectations and world knowledge). There are many interesting avenues to pursue in modelling both the outcome of inference processes as a function of context, and incremental belief update as utterances unfold over time.

Another area in which this approach will have far-reaching consequences is in language acquisition and learning. Understanding the ways in which child-directed speech differs from adult-directed speech across different contexts will allow for extensions of pragmatic processing models to models of language acquisition.

Finally, a rich under-explored area of pragmatics that merits much further investigation is social meaning. The focus of pragmatics is typically on information that speakers communicate about the world rather than about themselves or about the relation between speaker and listener. Yet the long history of sociolinguistics has documented the many ways in which speakers constantly communicate about these dimensions, whether voluntarily or not (Eckert, 2012). Recent initial attempts at bridging the gap between probabilistic pragmatics and variationist sociolinguistics are very promising: for instance, Burnett (n.d.) demonstrated that the use of the in/ing variants ('runnin'' vs. 'running') can be explained by a game-theoretic model of speakers attempting to convey different personae (combinations of features like 'friendly' and 'competent').

Earlier we treated the speaker's epistemic state as one abstract factor that is likely to always matter for pragmatic interpretation. However, listeners' estimates of speakers' epistemic states are themselves subject to prior beliefs about the world. For instance, if a listener holds a prior belief that one gender is on average more competent than another, this would predict that listener should be less likely to derive scalar inferences and more likely to derive ignorance inferences from utterances produced by someone of the gender perceived to be less competent. In a similar vein, the phenomenon of testimonial injustice, whereby for instance the same utterance is interpreted as a command when produced by a man but as a request when produced by a woman (Fricker, 2007), highlights the importance of integrating prior beliefs about conversational goals based on perceived social category.

3.10 Conclusion

In this chapter, we have surveyed the developments in constraint-based pragmatic processing and contrasted this approach with various types of informational privilege approaches in some example domains. Constraint-based approaches to pragmatic processing are still in their infancy. Nonetheless, in the domains where they have been applied, they emerge as the account that is most strongly supported by the data. Moreover, constraint-based approaches are well-suited for important future directions, both in their merging with probabilistic models of pragmatics and in the extension to social meaning.