Pragmatic Inference

1 Pragmatic Inference: A Concept with Many Faces

Inference, broadly speaking, is the deriving of conclusions from facts or evidence. Inference can be thought of as either the conclusion itself, or the process of deriving that conclusion. Inferential processes can take different forms—such as deductive, inductive, or abductive—and different kinds of processes can take different kinds of information as their inputs.

This Element is about a specific kind of inference, namely pragmatic inference. Pragmatic inference is a heterogeneous concept, but, roughly speaking, it is the deriving of conclusions about meaning based on linguistic communication. From now on, if we drop the prefix ‘pragmatic’ in favour of simply referring to ‘inferences’ and ‘inferential processes’, bear in mind that these are in reference to the study of language and meaning, unless otherwise specified.

A prudent first question to ask is: why study pragmatic inference in the first place? Well, until the 1970s, it wasn’t widely studied in linguistics. It was assumed that if speakers communicate their messages directly and explicitly, hearers—assuming they ‘speak the same language’—should have no problem understanding what the speaker said, as all they have to do is ‘decode’ the message using their knowledge of the linguistic system of the language they speak. But the problem is that speakers do not always—or arguably ever—convey messages directly and explicitly. In fact, it wouldn’t be efficient to try: the physical (vocal or gestural) apparatus that speakers have for conveying thoughts is grossly limited in scope compared to the richness of human cognition, creating a bottleneck in communication; as Levinson (2000: 29) says, ‘inference is cheap, articulation is expensive’.

So while meanings can be conveyed more or less directly via the language system of whatever natural language (English, French, Japanese, etc.) they use, speakers can communicate much more information than that encoded by the words uttered, which hearers can then infer. This Element discusses various ways in which scholars have studied pragmatic inference, as well as various issues that have arisen alongside those studies.

1.1 Some Ways to Study Pragmatic Inference

The study of pragmatic inference came to the fore in Anglo-American pragmatic theory through Grice’s (1975, 1978) work on conversational implicatures: aspects of meaning that go beyond the explicit content of ‘what is said’. This work put focus on the role of speakers’ intentions in the recovery of meaning, challenging the traditional ‘code model’ that assumed language to ‘encode’ meanings to be ‘decoded’ by hearers. Rather, Grice’s theory of speaker meaning proposed that successful communication relies on speakers abiding by the Cooperative Principle, ‘Make your conversational contribution such as is required, at the stage at which it...
occurs, by the accepted purpose or direction of the talk exchange in which you are engaged’ (Grice 1989: 26), and the four categories of conversational maxims, of quantity (‘make your contribution as informative as required’ but ‘do not make your contribution more informative than is required’), quality (‘try to make your contribution one that is true’), relation (‘be relevant’) and manner (‘be perspicuous’). In case a speaker intends to communicate a conversational implicature, that is, something other than what was explicitly said, a hearer is licensed to infer that implicitly communicated meaning on the assumption that the speaker is a rational communicator who is following the Cooperative Principle and the maxims (see Grice 1975, ‘Logic and Conversation’ for his classic examples).

Implicatures are typically viewed as the archetype of pragmatic inference, and differ in important ways from logical inference, that is, conclusions derived from premises via valid arguments. This is not to say that logical inference does not also play an important role in the study of language and meaning: if I know some sentence to be true, I can construct inferences about the world based on the facts described by that sentence. For example, if I know that all sheep are mammals, and I know that Dolly is a sheep, then I can infer – via deductive inference – that Dolly is a mammal. But logical inference differs from pragmatic inference insofar as it is typically monotonic: the conclusion cannot be changed with the addition of new premises. Pragmatic inference, by contrast, is usually non-monotonic. This is because when hearers make inferences about what speakers say, they do so on the basis of partial evidence, and the conclusions they draw can be overridden (or ‘cancelled’) in the face of new, competing, information (see Section 4 on cancellation). In other words, pragmatic inferences are defeasible. This was, in fact, one of Grice’s tests for an aspect of meaning constituting an implicature, as opposed to, say, an entailment (a fact that logically follows from what is said): implicatures are defeasible, while entailments are not, and presuppositions (background information that is required for comprehending what has been said) lie somewhere in the middle (see Beaver et al. 2021 on presuppositional inferences).

Assuming that ‘speakers implicate, hearers infer’ (Horn 2004: 6), pragmatic inference is very often equated with the hearer’s recovery of the speaker’s intended meaning. One direction in which scholars have since reframed Grice’s original philosophical account of speaker meaning is to address the question: how do hearers infer what speakers mean from what they say? One of the most notable developments in this regard is due to Relevance Theory (Sperber and Wilson 1986/1995), that aims to explain the cognitive processes in which a hearer engages in order to recover a speaker’s intended meaning. This explanation relies on the notion of ‘relevance’. An assumption is relevant in a context if it yields high cognitive effects (changes to one’s immediate and
manifest cognitive environment) alongside low processing effort. The crux of
two principles of relevance: first, ‘human cognition tends to be geared to the
maximisation of relevance’, and second, ‘every act of ostensive communication
communicates a presumption of its own optimal relevance’ (Sperber and
Wilson 1986/1995: 260). If a hearer presumes a speaker to be abiding by
these principles, hearers recover speakers’ intended meanings through a
process of least effort: a trade-off between maximising cognitive effects and
minimising processing effort. This is a subtle move in focus away from offering
generalisations of speakers’ language use in the spirit of Grice, and towards
psychological account of utterance processing that addresses the cognitive
question of how hearers recover speakers’ intentions.

In addition to the questions of what pragmatic inference is, as well as the
question of how pragmatic inference is derived in the mind, there is also the
question of why, under particular circumstances, would a hearer make
a particular inference over another? This question moves us away from
a cognitive account of what people actually do, and instead towards considering
the kinds of (potential) inferences that are licensed by a given sentence when it
is uttered in context. Such a question is typically addressed by normative
accounts that stem from commitment-based approaches to communication,
championed by Hamblin (1970), Brandom (1994), Geurts (2019), and many
others. There are many kinds of inferences that may be licensed by an utterance.
These include the conventional meanings of words and sentences as they are
produced, but can also extend to other kinds of meanings that are derivable from
the uttered sentence, including entailments, presuppositions, and even implicitly
communicated implicatures (see Section 3 for an overview of normative
commitment accounts).

Of course, there is a difference between what may be inferable from a given
utterance, and what is intended to be communicated by the speaker. Indeed, the
vast array of potential inferences that are available from a given utterance will
unlikely coincide one-to-one with the set of inferences that the speaker intended
the hearer to recover. Nor will the set of potential inferences perfectly align with
the set of inferences that the hearer actually recovers, whether or not the speaker
intended them to do so. Since as analysts we do not have direct access to
speakers’ actual mental states, one way to study the inferences that interlocutors
make is to observe recipients’ responses as evidence for the ways in which they
have understood themselves and others. This is the approach of work in talk-in-
interaction and Conversation Analysis that centralises the responses of others in
the ‘interactional achievement’ of understandings (Scheinoff 1981; see
Section 2 for further detail). On this approach, meanings are not tied to
individual utterances or to individual speakers’ intentions. Rather, meanings are viewed as intersubjective (shared between people) and emergent (can change over time) as interaction progresses.

Conversation Analysis tends to preclude theorisation about language; however, the insights that it offers by observing on-record interaction can be considered alongside both cognitive accounts and normative accounts to see whether the inferences that may be licensed by a given utterance are, indeed, the ones that participants appear to actually infer. As we progress through this Element, we will see the benefits of developing a hybrid account of pragmatic inference that encompasses speakers’ intentions in the attribution of meaning, alongside accounts of objectively available meanings, both by what is inferable by the uttered sentence, as well as by the on-record ways in which speakers and hearers orient themselves to different meanings in interaction.

1.2 The Scope of the Study of Pragmatic Inference – For Now

We can study inferences as products, namely inferences as meanings derived, or at least derivable, from what is said; or we can study inferences as processes, whether cognitive, logical, or normative, and, as summarised by Terkourafi (2021), different inferential processes can lead to the same product, while the same process can lead to different products. In this Element, we will look at the study of inferences-as-products under three different, albeit interrelated, guises. Inferences can be studied as, first, the intentions of the speaker that the hearer ought to, assuming communication to be functioning as it should, infer; second, the potential inferences that are licensed by a given sentence or utterance, for example entailments and presuppositions, as well as normative conventional meanings, that may or may not be entertained by speakers but are nevertheless justifiably inferable; and third, the inferences that hearers actually make regarding a speaker’s meaning, which may or may not align with the meaning that the speaker intended to communicate. We will also see how all of these inference products can be pursued via different inferential processes: the cognitive processes through which hearers entertain and ‘pick’ the inferences they make; the a priori logical inferential processes qua relations (e.g. entailment relations) that yield different inferential outputs; or the ways in which speakers make available inferences from what they say.

It is clear from the outset that the kinds of inferences that fall in the remit of one’s area of study will depend on the perspective that the analyst takes. While we can follow a broadly Gricean approach and focus on inferences about the speaker’s intended meaning, we may also acknowledge that hearers can draw all kinds of inferences from speakers’ utterances that the speaker did not intend to
communicate. This of course occurs in straightforward cases of misunderstanding, such as if a hearer misheard or misinterpreted what the speaker said, or if the speaker miscommunicated their intended meaning, for example through ‘slips of the tongue’ or mispronunciations. As these kinds of cases evidence ‘faulty’ communication, they may appear to be of little interest to a theory of communication that aims to describe and explain how communication works when things go as they should. However, as we will explore throughout this Element, inferences lie on a cline from clearly intended to clearly unintended, and disputes over what was said or what was meant can be very important to a theory of communication: they provide evidence for the kinds of meanings that are inferable from different kinds of utterances, and they demonstrate where potential problems in communication can lie. They can also be important to speakers themselves as they work with their conversational partners to figure out their joint attitudes towards relevant issues, or to avoid responsibility for inferences attributed to them but that they didn’t mean to communicate.

So, although different scholarly accounts may ask different questions and take different kinds of evidence as input, looking at the different theoretical options on offer – and note that we will only highlight a handful here – hints at how combining approaches can be fruitful for gaining a full picture of the nature of pragmatic inference as their insights inform one another. For example, as more progress is made on understanding the cognitive processing of language, more accurate developments can be made towards generalisations of normative language use that are faithful to how people actually use language, while observations about how speakers orient to certain aspects of meaning in real time can provide evidence for how speakers prioritise and structure different kinds of inferences in the mind.

Now, while pragmatic inference is typically equated with the meaning of a given utterance, it has to be recognised that people do not just communicate one single message when they speak. We already know that they can simultaneously convey explicit and implicit messages. But as we also know, it’s not just what the speaker said that has communicative importance, it’s how they said it, or, as we will discuss, in some cases what they didn’t say. In addition to implicatures, presuppositions, and entailments that we have touched upon so far, there is a wealth of other types of inference that can arise in the course of linguistic communication, including inferences relating to social dynamics, attitudes, and emotions. For example, use of indirect or mitigating language can offer insights into how the speaker perceives their relationship with their interlocutor. Use of irony, sarcasm and other figurative language might indicate something about the speaker’s mood or attitude. Use of slang or taboo language may invoke inferences about the speaker’s emotional state. And paralinguistic cues such as tone of voice, accent,
volume of speech, gestures, and facial expressions can accompany language use to help draw such inferences. All of these inferences will be ‘pragmatic’ in the sense that they are defeasible, although they will range in the extent to which they are intended or unintended to be communicated, or indeed consciously or subconsciously communicated.

Viewing, as Haugh (2007: 90–1) does, implication as ‘anything that is the consequence of something else’, and ‘anything that can be anticipated or inferred by the addressee from what is said’ (my emphasis), what is to say that the study of pragmatic inference should not extend to consider this wider array of inferences? To do so would not only go beyond the scope of pragmatic inference as the recovery of speakers’ intentions in the spirit of post-Gricean pragmatics, but also beyond the joint process of co-constructing propositional meanings, or negotiating commitment for communicating propositions. It would take us further into the realm of attitudes, emotions, social propriety, and other ways of expressing oneself that can, in some situations, have a greater impact on communicative outcomes and interpersonal relations than the content of what is expressed. Scholars have started to consider this vast array of inferences that one can obtain from utterances that go beyond what was said or what was meant, and full consideration of how such inferences have been and can be studied will certainly take us beyond the scope of this Element. But what will hopefully become clear through this brief precis of theoretical options is that the question of where the locus of meaning lies is not only relevant to pragmatic theory, but is also of paramount importance for real-life communicative issues, including managing everyday instances of interpersonal communication conflict where ‘what is said’, ‘what is meant’, and ‘what is communicated’ are at issue.

2 ‘Faulty Inferences’: Speaker Intentions, Indeterminate Meanings, and Misunderstandings

Since Grice’s (1975) seminal work on the relationship between what is said and what is meant, the term ‘pragmatic inference’ typically refers to a hearer’s understanding of a speaker’s intended meaning. On this view, as long as a hearer infers a speaker’s intention from their utterance, communication can proceed unhindered. But we also know that hearers can draw all kinds of inferences from a speaker’s utterance that the speaker did not intend to communicate, which are usually seen to lie outside Grice’s account. This section begins with an overview of Grice’s distinction between ‘what is said’ and ‘what is implicated’ and the debates that followed in distinguishing these two aspects of meaning, before moving beyond Grice’s account to consider different kinds of inferences that
Pragmatic Inference

2.1 Pragmatic Inferences and ‘What Is Said’

We start here with Grice’s (1957) work on non-natural meaning, ‘meaning$_{\text{NN}}$’, that turned its attention away from how meanings are solely derived from sentences and their component parts, and towards speaker meaning as intentional meaning. For him, for a speaker to mean$_{\text{NN}}$ something by an utterance is for the speaker to produce an utterance with the intention of inducing a belief in the hearer by having them recognise this intention (see Grice 1989: 220). A rational, cooperative speaker – abiding by the Cooperative Principle and conversational maxims – is thus expected to formulate their utterance in such a way that it would be understood by the hearer in the way the speaker intended to be understood (see Jaszczolt 2023, chapter 7 for a recent – detailed and critical – overview of Grice’s work on speaker meaning).

Now, in framing meaning$_{\text{NN}}$ in terms of speakers’ intentions, Grice acknowledged that ‘speaker meaning’ can depart from sentence meaning and what is explicitly ‘said’, and instead constitute a conversational implicature. For him, meaning$_{\text{NN}}$ is the composition of ‘what is said’ and ‘what is implicated’, and the study of the two have subsequently been viewed as separate enterprises. ‘What is said’ is typically viewed as a product of the language system as ‘sentence meaning’, and is considered the bearer of truth conditions. Meanwhile, ‘what is implicated’ is often related to ‘speaker meaning’, involves contextual information for its recovery, and is traditionally not considered truth-conditional. However, later theorists in the neo- and post-Gricean traditions have long noted that the dividing line is not this clear-cut, with debates abounding as to how to distinguish ‘what is said’ and ‘what is implicated’: what Horn (2006) labelled the ‘border wars’.

Grice himself could be credited with paving the way for this debate in his observation that sentences underdetermine truth-conditional meaning in cases of ambiguity and reference assignment (see Grice 1978, ‘Further Notes on Logic and Conversation’), and that context is necessarily required for their resolution. This observation has been extended by various theorists to other syntactically complete but semantically underdetermined sentences for which additional contextual information is required to communicate something meaningful. On such a ‘contextualist’ view, sentences like (1) only bear truth conditions on the supply of information from context that indicates what the speaker
is not ready for. That is, (1) will mean something very different in a situation where the speaker is not ready to start writing, to the situation in which the speaker is not ready to eat lunch.

(1) I’m not ready.

This process of adding contextual information has been described as ‘filling in’ (Bach 1994), or ‘saturation’ (Recanati 2004), as it involves saturating the logical form by filling in the sentence from the ‘bottom up’.

There are also cases in which lexical items may need ‘fleshing out’ (Bach 1994) to make sense in a given context. Assuming that the referent of ‘he’ in (2) is a human, its utterance invites the hearer to create an ‘ad hoc concept’ (Carston 2002) of ‘snake’ by drawing on the relevant features of snakes that the referent shares, leading the hearer of the utterance to infer the likely intended meaning that the person in question is devious in some way.

(2) He’s a snake.

This kind of ‘top down’ ‘free enrichment’ (Carston 2002) or ‘modulation’ (Recanati 2010) of the logical form is not linguistically mandated insofar as it can be possible to obtain a fully fledged truth-conditional proposition from the uttered sentence. However, in seeking truth conditions that reflect the ways in which speakers use and understand their utterances, scholars of a ‘contextualist’ orientation take the view that such pragmatic inferences should be used to enrich the logical form of the sentence to generate a unit of ‘what is said’ that outputs truth conditions in line with speakers’ intuitions about them. Recanati (2004, 2010) goes so far as to argue that – while his pragmatic operation of modulation is an optional, context-dependent process – no truth-evaluable unit is free from pragmatic inferencing.

There is ample debate over which kinds of meanings are generated by bottom up versus top down processes, including but not limited to the domain of quantifier expressions (e.g. restricting the domain of ‘every’ in ‘every bottle is empty’ to those on the table or at the party; see Stanley and Szabó 2000), narrowed readings of logical connectives (e.g. taking ‘and’ to mean ‘and then’ or ‘and as a result’, see Carston 1988, 2002), and strengthened concepts encoded by lexical items (e.g. our ‘snake’ example above, see again Carston 2002). Likewise, how far we go with allowing context to intrude on the logical form of the uttered sentence depends on one’s theoretical commitments. While as we’ve seen Relevance Theory and Recanati equate ‘what is said’ with their own versions of enriched logical forms, other theorists (e.g. Jaszczolt 2005, 2010) go right to the end of the spectrum in prioritising the study of semantics with the main, ‘primary’, meaning that a speaker intended, which may adhere to