1 Introduction: Grammar, pragmatics, and what’s between them

Pragmatics has been notoriously hard to define. Or rather, it has proven quite impossible to reconcile between the patterning of phenomena assumed to be classical pragmatic topics (deixis and reference, speech acts, conversational and conventional implicatures, presuppositions, functional syntax) and the common set of definitions for pragmatics (most notably, context dependency, inferentiality, nontruth conditionality and others). In order to resolve the delimitation problem of the field we are forced to first abandon the expectation that all the definitional criteria converge on classifying some phenomenon as pragmatic (or as grammatical). In other words, we cannot expect that any given pragmatic phenomenon will simultaneously meet all the criterial definitions for pragmatics (and vice versa for grammatical phenomena). For example, while deixis is pragmatic in that it is context-dependent, it cannot meet the nontruth-conditionality criterion (for it contributes a truth-conditional meaning).

In addition, we must give up on what I have elsewhere called the topical approach to pragmatics, which assumes that all aspects of some phenomenon (e.g. of deixis, of presupposition, etc.) uniformly belong in pragmatics, or else, that all of them uniformly belong in grammar (see Ariel, forthcoming and chapter 2). Any specific instance of language use is neither wholly grammatical nor wholly pragmatic. To pick deixis again, it combines grammatical aspects (there is a grammatically specified difference between I and this) with pragmatic aspects (pinning down who the speaker is, what object this denotes). Hence, instead of struggling to find just the right set of definitions which would include all and only the canonical list of pragmatic topics (a mission impossible – see Levinson, 1983), we must choose one criterion to define pragmatics. We cannot even expect it to apply to all (aspects) of the topics on the classical pragmatic list (see Ariel, forthcoming). Perforce, a single criterion will offer a consistent division of labor between the grammatical and the pragmatic.

Now, if we can only choose one criterion for drawing a consistent and coherent grammar/pragmatics division of labor, which criterion should we opt for? This book, along with researchers such as Sperber and Wilson (1986/1995), adopts the code/inference distinction as the basis for the grammar/pragmatics division of labor. Why pick the code/inference distinction (as opposed to any other criterion, e.g. truth conditionality)? The simple answer is that we cannot afford not to adopt this distinction, given the nature of grammar. Whatever else grammar may be, there can be no controversy about it consisting of a set of codes. The essence of grammar is a set of conventional associations correlating specific forms with their
obligatory or optional, rule-governed positioning, meaning, and distributional patterns. Whether or not it is in addition devoted, for example, to truth-conditional interpretations remains to be seen.

At the same time, grammar obviously falls quite short of meeting our communicative needs. There is a consensus today about the underdeterminacy of grammar, i.e. the fact that our coded messages never exhaust the meaning we intend to convey (see especially Carston, 2002a, and see also Levinson, 2002a: 8). This is where pragmatics comes in, enriching our encoded messages with pragmatic, i.e. plausible inferred interpretations (as opposed to formal, logical deductions). Grammar and pragmatics always go together. You can’t have one without the other for effective communication.

This book is about the complex relationship between grammar and pragmatics, that is, between codes and inferences involved in human communication. The relationship is not unidimensional. It has a few facets, and each one of them needs to be examined. Naturally enough, researchers interested in the grammar/pragmatics division of labor have focused on the complementarity between the two competencies. Such research seeks to establish the precise borderline between codes and inferences. Indeed, this is one important facet of the relationship between grammar and pragmatics, and part I is devoted to this topic. We try to resolve a number of intriguing questions, all of which have the same format: given a certain correlation between some linguistic form and some interpretation or use conditions, should it be grammar or should it be pragmatics that accounts for it? In other words, is the correlation encoded or is it derived by plausible inferences? Now, up till recently, all inferences were automatically seen as conversational implicatures. Relevance theoreticians have proposed that not all pragmatic inferences are implicatures, however. If so, every interpretation we view as a pragmatic inference needs to be further classified as to which type of inference it is.

But there is much more to explore about the relationship between grammar and pragmatics. Codes and inferences must make contact in diachronic change. Research by functional and historical linguists (e.g. Bybee et al., 1994; Traugott and Dasher, 2002) has convincingly demonstrated that, historically, pragmatic inferences routinely become grammatical conventions (via the processes of grammaticization and semanticization). In fact, most, if not all, of our synchronic grammar is the product of our diachronic extralinguistic regularities, which have turned grammatical. If so, any analysis of the grammar/pragmatics divide must also meet the challenge of the penetrability of this divide: grammar is not only distinct from pragmatics, it is also its product. Part II addresses this facet of the relationship between grammar and pragmatics.

The final chapter of this book (part III) brings together these two very different approaches and research traditions. We discuss grammar/pragmatics interfaces, i.e. representational levels where the two combine. Synchronically, we have at least two such levels, conveyed meanings and basic-level meanings. We hardly touch on the first type of representation, despite the fact that it consists of the linguistic meaning plus all the conversational implicatures generated by the speaker (but see
section 1.2 below). The reason is that here, although semantics and pragmatics combine, they remain interactionally independent of each other. There is also no controversy regarding this interface level. We focus on the basic, minimal meaning level which the speaker is necessarily seen as committed to. This interactionally significant (synchronic) meaning is an integrative level, where codes and inferences are intimately woven together to form one proposition. Here codes and inferences are in fact so well integrated that it’s not clear that they are separable (for cognitive and discoursal purposes). We consider proposals as to how to define this meaning representation, namely, minimally (i.e. with as few inferences as possible), or maximally (incorporating a substantial inferential contribution). The conclusion we reach is that this hybrid-level meaning is minimalist (à la Grice) on some occasions and maximalist (à la Sperber and Wilson) on other occasions.

Now, although this basic-level interpretation has been proposed as a synchronic, real-time representation, we end the book by entertaining the possibility that it also serves as the input for semanticization and grammaticization. In other words, it is possible that the synchronic basic-level grammar/pragmatics hybrid representation is not only the significant meaning level in terms of the ongoing interaction, it is also the diachronic arena where pragmatics may turn grammar. After all, linguistic change must occur in real-time discourse. If this is true, the most significant grammar/pragmatics interface level both synchronically and diachronically is the basic interactional interpretation, where codes and inferences co-mingle, regardless of the fact that they are arrived at via quite distinct cognitive processes.

Let’s begin our introduction now, which will prepare the ground for the whole book. We set out from the now well-accepted assumption that we always communicate by combining codes (grammar) with inferences (pragmatics). We introduce the basic facts about inferencing in section 1.1. We briefly outline the Gricean mechanism for generating inferences (conversational implicatures) in section 1.2, and we define and distinguish between codes and inferences in section 1.3. Section 1.4 introduces pragmatic inferences other than conversational implicatures. We end with section 1.5, where we introduce the two challenges facing research into the grammar/pragmatics relationship: drawing the code/inference distinction on the one hand, and accounting for the process whereby inferences cross over and become codes on the other hand.

### 1.1 On inferring

We get more for our words when they are embedded in natural discourse context. Here are two examples where it is quite clear that we need to read between the lines:

1. a. **LEWINSKY**: . . . See my mom’s big fear is that he’s ((President Clinton – MA)) going to send somebody out to kill me. ((PART OMITTED))
TRIPP: Oh, my God. Don’t even say such an asinine thing. **He’s not that stupid. He’s an arrogant . . . but he’s not that stupid.**

LEWINSKY2: Well, you know, **accidents happen.**


LEWINSKY2: Well, you know, **accidents happen.**

b. REBECCA1: So you can testify to two of [em].

RICKIE1: [Yeah].

REBECCA2: That’s why I had you come up, **beau[se],**

RICKIE2: [Yeah],

REBECCA3: . . . um,

. . . that’s great. (SBC: 008)

While the above exchanges are natural enough, and easy to interpret (for their participants, at least), there is something strange about each of them. **(1a)** seems to involve irrelevant responses: why is Clinton’s intelligence (Tripp’s response) relevant to the fear of Lewinsky’s mother that he will have her killed (a topic raised by Lewinsky)? And what is the relevance of the occurrence of accidents (Lewinsky’s response) to Clinton’s intelligence (discussed by Tripp)? It seems that while there is a connection between Lewinsky’s and Tripp’s contributions above, it is not explicitly stated. It is inferred. Tripp intends to convey that since Clinton is intelligent he won’t have Lewinsky killed, because killing her would be a stupid act (probably since he would be caught and get into even deeper trouble). Lewinsky then retorts that he may indeed have her killed, except it would be made to look like an accident (so he may not be caught). We similarly need to enrich Rebecca’s contribution in (1b), since her sentence is cut off in the middle. What is the reason that Rebecca (a prosecuting attorney) had Rickie (a sexual abuse victim preparing to testify against her sexual abuser) come up to her office? Given the specific context, we can assume that it must be the fact that Rickie can testify to two cases of sexual abuse by the defendant, as opposed to other witnesses who “only” experienced one sexual abuse from him. In the light of what has already been said in the exchange, this reason is quite obvious, and hence need not be explicitly mentioned. The addressees can easily infer it. The exchanges in (1) are quite typical of natural interaction. Communication begins with the coded message, but it never ends there. Inferences are an inherent part of it.

The most basic goal of pragmatic theories is to provide an account for how we go about interpreting such everyday exchanges. This is the goal informing Grice (1989), as well as Horn (1984 and onwards) and Sperber and Wilson (1986/1995 and onwards), all following Grice. We here briefly present only Grice’s pragmatic theory, because our purpose in this chapter, indeed in this book, is served by demonstrating that there is at least one pragmatic theory which can account for additional interpretations we get people to infer when we speak. In fact, all pragmatic theories are in this sense Gricean. They all assume that every act of communication is actually **inferential,** because the addressee is required to **infer** the speaker’s intention from whatever evidence is available to him (the linguistic
code constituting only one important source of information). At the same time, this chapter also serves as an introduction for later chapters, where disagreements between theorists will be discussed. These revolve around the role of particularized and generalized conversational implicatures, and the Relevance-theoretic concept of explication, all here introduced.

Natural discourse is a cooperative activity. It’s not a random collection of independent utterances (see Mann and Thompson, 1986). What makes discourse coherent? What expectations do we have from interlocutors’ utterances? The idea is that whatever those expectations may be, they cannot be fulfilled by our coded messages exclusively. It is only when we (also) consider inferred interpretations that we can ultimately explain how discourse works. In order to produce and interpret appropriate pieces of discourse, speakers must rely on the many inferences which accompany those codes. This is why codes and inferences are so intimately connected with each other. And this is why discourse coherence and inferencing are co-dependent on each other as well. Let’s examine briefly Grice’s (1975, 1989) proposals on how to account for both the nature of discourse and the mechanism responsible for the derivation of pragmatic inferences.

No doubt, one of the most important features of human discourse is that we assume that speakers’ utterances are somehow relevant to us. This is why we are led to interpret the following two statements as related, in fact, as constituting a contrast:

(2) The father was appointed chief of staff, the daughter refused to enlist (in the army).  
(Originally Hebrew, Hair headline, April 25, 2002)

Each statement by itself is quite irrelevant to the addressees. The readers of the newspaper need not be informed that someone’s father has been appointed chief of staff. They know the man by name, and they know he was appointed. The fact that some daughter refused to serve in the army is equally irrelevant. However, with the two pieces of information taken together, the headline becomes highly relevant: it turns out that the Israeli chief of staff himself (who had spoken about the importance of serving in the army) has a daughter who refused to serve. To see that we indeed expect discourse to be relevant, note what happens when Darryl can’t see Pamela’s utterance as relevant:

(3) PAMELA: I guess it’s j– looking at my mother, 
   too, 
   I n– — 
   (Hx)
DARRYL: . . . What does that have to do with why you’re reading a book 
on death?  (SBC: 005)

1 Indeed Tomasello (1999) argues forcefully that children’s ability to view others as intentional agents (an ability which emerges at nine to twelve months) is a prerequisite for language, and something that animals lack.
Darryl’s question shows that the default assumption is that interlocutors produce relevant utterances. Since he fails to find relevance, he asks about it. So, an important requirement on speakers is that they be relevant. This is what Grice’s maxim of Relation (“Be relevant!”) is about.

Next, let’s examine Grice’s maxim of Quantity. We expect utterances to be informative. Here’s a case where Joanne finds one of Lenore’s utterances (the one marked bold) not informative enough. She then asks for an elaboration:

(4) **JOANNE:** ... (H) He’s got iron, with his multiples.

**LENORE:** [Well, I] have iron too, but th– some of it isn’t absorbable.

**JOANNE:** What do you mean absorbable.

**LENORE:** (H) it’s good for your anemia. (SBC: 015)

Of course, Joanne knows what absorbable means (semantically). But she can’t get enough relevant information from Lenore, presumably because she is missing some background information that Lenore is taking for granted, something to the effect that absorbability is an important issue for the effectiveness of iron pills. In this sense, Lenore’s contribution was not informative enough, and hence is not quite appropriate.

A third maxim proposed by Grice is the maxim of Quality. We naturally assume that our interlocutors tell us the truth. This is why when Pete in (5) realizes he’s about to say something he is not sure is true (that salad spinners did not exist when they were growing up), he stops short, and asks his friends whether it’s true or not:

(5) **PETE:** I don’t think they ever —

. . . did they actually exist back then? (SBC: 003)

Grice’s maxim of Quality instructs speakers not only to tell only the truth, but also to avoid saying that which they lack evidence for. Pete is abiding by this maxim.

Finally, according to Grice’s maxim of Manner, utterances should be constructed in an optimal style or manner. They should be brief and clear. Here’s a case where the speaker was not clear enough (in the turn marked bold), and caused a problem for his addressee. She has to ask him about his intended message therefore:

(6) **W:** M!

**M:** I’m coming!

**W:** Watch out along the way.

**M:** Why?

**W:** Sandy!

**M:** ((Looks around, doesn’t see Sandy, the dog))

What?

She vomited? (Aug. 5, 2006)
While on other occasions it may be enough to use an NP utterance and trust the addressee to complete the missing propositional information (see *Iraq* standing for ‘the American war and occupation of Iraq’ in (13) below), this was not a successful act in (6). This is then a violation of Grice’s maxim of Manner.

The problems we detected in the discourses in (3)–(6) provide evidence for how proper discourse usually proceeds. The four Gricean maxims mentioned above characterize what Grice (1975, 1989) suggested are cooperative communicative principles. According to Grice, natural discourse reflects the application of the Cooperative Principle, which instructs speakers to “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, 1975: 45). This cooperative (some prefer rational – see Kasher, 1976) behavior translates into abiding by the four maxims we’ve mentioned: Quantity (informativity), Relation (relevance), Quality (truthfulness), and Manner. Speakers are expected to provide just the right amount of information (neither too much nor too little – Quantity), the information should be relevant (Relation) and true (Quality), and it should be optimally phrased, namely, the utterance should be as brief and clear as possible (Manner). As we saw in the above examples, whenever any one of these maxims was violated (or about to be violated), there was some breakdown in the communication, and speakers hastened to query about it, so as to somehow remedy the problem. I should add, however, that examples (3)–(6) were not at all easy to find. Most of the time, discourse does proceed rather smoothly, according to the Gricean maxims above.²

Now, what’s all this got to do with the question of codes and inferences? A lot, according to pragmatists. The argument is that we simply cannot comply with the cooperative principle, nor with the four maxims, by subjecting just our coded, explicit messages to these principles. As already exemplified in (1), we must take into account inferences in order to see how speakers go about cooperating with each other. If we restrict ourselves to examining only codes, we will find quite a few violations of the Gricean proposal, ones which are not accidental lapses (performance errors), like the exceptional examples in (3)–(6). The violations we are about to examine below (section 1.2) do not cause any communication problems. This is so because they are intended, and they give rise to inferences which then show that speakers were cooperative after all. Inferences serve discourse when codes fail. Still, why is it that we don’t communicate by codes alone? Why don’t we make sure that our codes fully abide by the four Gricean maxims? Wouldn’t it be clearer and more efficient to communicate this way? Why divide up our communicative task between two modes? For example, why didn’t the newspaper choose as its headline the Hebrew counterpart of “The father was appointed chief of staff but his daughter refused to enlist (in the army)” (instead of (2))?

² See Grice (1975, 1989) for a detailed outline of the theory.
First, explicating everything we wish to convey to the addressee would considerably slow us down. Following Levelt (1993), Levinson (2000a) has convincingly argued that our production mechanisms are far too slow to allow for an efficient, fully explicit mode of communication. Many pieces of information which constitute an integral part of our messages are better left to inferencing, because human beings are very apt at drawing inferences. Actually pronouncing out loud these assumptions would take a long time and effort on the speaker’s part, and at the same time would waste the addressee’s time in superfluous decoding. Many inferences are faster for the addressee to compute than for the speaker to form a verbalizing plan and to articulate with words. Here’s a relevant example (and see again (1)):

(7) **HAROLD:** that . . . really hot tap dancer.
**JAMIE:** [Oh] that kid.
**MILES:** . . . **He** was actually here two weeks ago, and I missed him.  
(SBC: 002)

In fact, what Miles is conveying to his interlocutors is that ‘He, i.e. the tap dancer, was actually here, i.e. in this town, two weeks ago for a public performance, and I missed that performance with him.’ But Miles can skip stating explicitly that the dancer gave a performance (see how much longer the explicit version is), because we can very easily infer it. The same applies to the deictic (here) and the anaphoric expressions (He, him, and see again (1b)). Leaving part of the message to inference can be an efficient step, then.

In addition to effort saving as in (7), speakers have a variety of reasons for preferring an implicit over an explicit mode for some interpretations. There is an interactional difference between explicit and implicit messages. Consider the following:

(8) a. The Americans know Netanyahu, who is actually Benjamin Nitai.
    (Originally Hebrew, Reshet Bet radio, Jan. 11, 2006)

    She had an uncle who predicted that she will be a musician and a genius.
    And she always said, well, at least I became a musician.
    (Originally Hebrew, Reshet Bet radio, Mar. 17, 2006)

The implicit message in (8a) is that Israeli Knesset Member (and former prime minister) Netanyahu is not quite a patriotic Israeli, possibly, that he cannot be trusted as a leader. In order to arrive at this conclusion, we need to rely on a host of background assumptions: as is well known (to Israelis), years before he became a politician, Netanyahu had left Israel and lived in the US for quite a number of years, where he changed his name from the Israeli-sounding *Binyamin Netanyahu* to the American-sounding *Benjamin Nitai*. On the assumption that the name one adopts testifies to the national identity one aspires to, there is possibly an even stronger message here, that Netanyahu is not actually loyal to Israel but, rather, to the US. No wonder Knesset Member Ronnie Bar-On (the speaker in (8a)) prefers an implicit over an explicit mode. As we point out below, one can always deny
what one did not say explicitly. Next, consider (8b). Being a musician certainly does not preclude being a genius. Yet, it’s clear that Drora Chavkin (the she cited in the example) implied that she was not a genius. But she didn’t quite say it. There is a difference between explicitly stating some information and conveying it by triggering an inference (a self-humoristic effect in this case). So, the answer to our question of why rely on inferences rather than on decoding is that (a) inferences save on time and effort (sometimes), and (b) they are indirect.

1.2 Generating implicatures

We need a “science of the unsaid,” as Levinson (2000b) calls inferential pragmatics theories. Grice’s (1975, 1989) idea was that we use the same four maxims that inform our cooperative behavior in general as guiding principles in inference drawing as well. Interlocutors’ working assumption is that cooperative speakers do abide by the maxims. If so, should they seem to violate one of the maxims, and blatantly so, rather than take it as a breakdown in the communication (as is done in the atypical (3)–(6)), addressees assume that there was a special speaker communicative intention behind the maxim flouting. That communicative intention is a pragmatic inference, based on what was said explicitly and contextual assumptions the speaker intends the addressee to consider in the computation of the inference. Those inferences which fall under the communicative intention of the speaker are what Grice termed conversational implicatures ((8a) triggers a particularized conversational implicature, (8b) a generalized conversational implicature – see below). Let’s consider a few examples to see how the mechanism of speaker-generated implicatures works.

Consider (8) again. We can now explain why it is that speakers generated the specific implicatures we noted above. In both cases we have violations of the maxim of Quantity. The speaker in (8a) provides too much information. While vaguely relevant to the topic, Netanyahu’s American name constitutes superfluous information at the current stage of the discourse. Once the addressee figures that the speaker is being cooperative, and that he has a specific intention in being too informative, the way is paved for deriving the implicature that the added information is relevant in that if Netanyahu has an American name, he may have an American identity and loyalty. Similarly, in (8b), where Chavkin provides too little information (confirming the prediction about her becoming a musician, but remaining silent on the question of her being a genius). Assuming that this is no accident (or performance error), the speaker is seen as avoiding the second question, thereby implicating nonconfirmation.

3 Note that unlike the inferred conclusion specified above, that Netanyahu cannot be trusted, the background assumptions leading to this inference are not spelled out not because the speaker wants to be indirect about them, but because he doesn’t want to articulate that which can be more easily inferred (the efficiency motivation alluded to before).
Ironical interpretations are more often than not implicatures generated because Quality (truthfulness) has been flouted:

(9) S: I’m going upstairs now to the regional committee’s office . . . to see what’s going on. **Only nine years** they’ve had it, no less ((it = S’s application for a building permit)). (Lotan 1990: 4)

Obviously, for a builder, as S is, to wait nine years to get a building permit is not a short time. He doesn’t really mean ‘**only** nine years,’ which conventionally implicates that one would have expected it to take longer. Since Quality was breached, but the speaker is taken as cooperative, he is taken to intend to implicate something, in this case, that he means quite the opposite of a short, negligible waiting time: nine years is much too long for the committee to delay its decision about a building permit.

Next, the following is a case where relevance is flouted, the speaker, a job candidate anxious to get the job, generating an implicature, which renders her reply relevant after all (and see again (1a)):

(10) **BOSS:** You have small children. **How will you manage long hours?**
    **HD:** I have a mother. (Originally Hebrew, June 14, 1996)

Surely, the fact that HD has a mother does not seem a relevant answer to the boss’s question about how she will be able to stay at work after regular hours. But since we assume that HD is cooperative, we see the flouting as an indication that she is generating an implicature, in this case, that ‘her mother will take care of her children when she needs to stay late at work.’

Manner violations too can serve as a basis for implicature generation, as can be seen in (11a):

(11) a. Let us look at the **racial, or rather, racist** themes in the argument for population control.
    (Pohlman, *Population: A clash of prophets*; Du Bois, 1974; ex. 8)

b. ~Let us look at the **racist** themes in the argument for population control.

(11a) is a case of a repair, where the speaker corrects himself. Supposedly, he said **racial** when he really meant **racist**. When we compare the original (a) version with the contrived (b) version, it’s quite clear that (a) is not as brief as (b). Now, in spontaneous conversation, we might see this as a performance error, and not attribute any significance (implicatures) to the repair. But since the example is taken from a carefully edited written source, argues Du Bois (1974), we cannot ignore the fact that the writer here chose the longer over the shorter version. We then attribute to the writer additional implicated interpretations, perhaps that he is not comfortable in asserting an unmitigated version of his strong term (**racist**).

---

4 We are here glossing over the need to access just the right contextual assumptions for generating the intended implicature. For example, in the case of (10) one could theoretically render HD’s utterance relevant by assuming that her mother will come to replace her at work when she has to work late. Of course, this is not what she implicates.