The relationship between generative grammar and (relevance-theoretic) pragmatics

ROBYN CARSTON

Abstract

The generative grammar approach to language seeks a fully explicit account of the modular system of knowledge (competence) that underlies the human language capacity. Similarly, the relevance-theoretic approach to pragmatics attempts an explicit characterisation of the subpersonal systems involved in utterance interpretation. As an on-line performance system, however, it is subject to certain additional constraints; this is demonstrated by the way in which matters of computational (processing effort) economy are currently employed in the two types of theory. A sub-module of 'discourse competence' is shown to be compatible with and complementary to the wider system of pragmatic processes.

1 Introduction

This paper grew out of my participation in a panel on formalist and functionalist approaches to language, organised by Frederick Newmeyer for the 16th Congress of Linguists held in Paris in July 1997. My work has been primarily in the field of cognitive pragmatics, specifically within the framework of Relevance theory developed by Dan Sperber and Deirdre Wilson. On this approach, pragmatics is construed as an account of the inferential processes involved in understanding utterances and of the principles that guide and constrain these processes. While most relevance-theorists are positively disposed to the formalist approach to grammar and have taken on board many of its fundamental assumptions, the relationship between the two systems (grammar and pragmatics) and how they interact in language use has never been explicitly spelt out. For the Paris panel, Newmeyer asked me to discuss the relationship between the two endeavours, considering the extent to which certain fundamental conceptions of the generative grammar programme, such as modularity, competence, generativity and notions of least effort and economy of derivation, do or do not carry over into the cognitive pragmatic programme. The position I take here is that while pragmatic theory is like grammatical theory in that it seeks full explicitness and is pitched at the level of

subpersonal mental systems, it is unlike grammatical theory in that it is an account of performance mechanisms rather than of knowledge systems (competence). This difference is reflected in the different ways in which economy considerations operate in the two types of theory. Finally, I consider where the notion of a 'discourse competence' fits into this picture; this concept, which at first seems somewhat anomalous within the system I will have delineated, falls into place once we see that formal linguistic devices (the domain of a grammatical competence system) may encode instructions or procedures concerning appropriate use (part of the domain of pragmatics), rather than conceptual meaning that affects the truth-conditional properties of an utterance. Of course, these are big issues and they invite investigation of much greater depth and breadth than I am able to achieve here.

2 Cognitive pragmatics

The question of which 'generative conceptions' carry over from grammar into pragmatics and which do not depends to at least some extent on the sort of pragmatics one does. I concentrate here on the very cognitively-oriented pragmatics that comes out of the relevance theory framework, with just one or two passing comments about some other approaches. The aim of this sort of pragmatic theory is to account for our capacity to interpret each other's utterances, specifically to account for the inferential processing phase of this capacity, i.e. those processes that take as their input the result of linguistic decoding.

Let us look briefly at some of the phenomena that such a pragmatic theory is expected to account for. Consider the utterance in (1):

(1) Utterance: "I'm tired"
Conversational implicatures (in different contexts):
   a. Let's go home.
   b. You do the washing up.
   c. I need a holiday.

It is fairly easy to imagine three distinct contexts for this utterance in each of which one of the listed propositions, (a) - (c), would be implicated. This is an instance of the basic data of pragmatics: different implicatures that an utterance of a particular linguistic expression may communicate in different contexts. While some utterances convey a single strong implicature of this sort, others may communicate a wider array of weak implicatures. For instance, if the utterance 'I'm tired' was a response to the question

'What are you going to do today?' there would be a range of possible implicatures including: 'I'm going to take it easy', 'I'm not going to do anything very demanding', 'You should treat me gently', etc. In this sort of case, the utterance suggests a conceptual space within which a hearer is to look for implicated assumptions, rather than strongly communicating any single one.

Explaining how conversational implicatures arise is usually seen as the central concern of a pragmatic theory. They are patently not encoded by the linguistic form, they are highly context-dependent and are the result of a non-demonstrative (defeasible) inference process on the part of the hearer. But there is also the crucial matter of working out what proposition the speaker has expressed, what she has explicitly said. While this is partially given by the linguistic form the speaker has chosen, it is also hugely dependent on pragmatic inference. This point is illustrated by the utterance in (2):

(2) Utterance: "It's too cold"
Proposition expressed (in different contexts):
a. The tea is too low in temperature for me to drink it.
b. This novel is written in too unfeeling a way for my taste.

Deciding what the referent of 'it' is, which meaning of 'cold' is operative and working out what it is too cold for, i.e. completing the meaning so as to recover a fully propositional form, are all tasks for the pragmatic system; none of these is given by the linguistic content alone, each is dependent on the particularities of contexts.

Perhaps the most fundamental task of a pragmatic theory is to explain **how the intended context is recognised**; that is, how the hearer is able to work out which of all the assumptions available to his cognitive system at any given time is the set he is intended to use in processing the utterance. Clearly, accomplishing the tasks exemplified in (1) and (2) is dependent on the success of the task of accessing the appropriate context. The relevance-based cognitive pragmatics, developed by Dan Sperber and Deirdre Wilson, aims to account for our ability to perform these tasks (see Sperber & Wilson, 1986/95, 1987). Starting from general cognitive considerations, rather than the specifics of communication, the idea is that any new information is processed by an individual in a context of beliefs and conjectures already available to that individual. If bringing together this new information and this context yields cognitive effects that could **not** have been derived from the new information alone, nor from the context alone, then this information is **relevant in this context**, and **to the individual** who brought this context to bear on this information. Among the types of cognitive effects that make new information relevant are the **addition** of new beliefs implied by the new information in

conjunction with the context, the **strengthening** of existing assumptions, supported by the new information, and the **abandonment** of old beliefs, contradicted by the new information in the context.

Clearly, the more effects a stimulus has, the more relevant it is. However, achieving cognitive effects involves a cost in the form of processing effort (attention, memory search, inference) and this negatively affects the degree of relevance. So the relevance of an item of information to an individual is a product of these two countervailing factors:

(3) a. *Ceteris paribus*, the greater the cognitive **effects** resulting from processing the information, the greater its relevance.
b. *Ceteris paribus*, the greater the processing **effort** required for processing the information, the lower its relevance.

The Sperber-Wilson view of our general cognitive orientation is given in (4a) below as their First Principle of Relevance. This could do with a measure of unpacking, but I won't attempt that here (see Sperber & Wilson, 1986/95, chapter 2; Sperber & Wilson, 1995), since my focus is on pragmatics and so on one particular sort of information, that is, **communicated information**. This differs from other environmental information in that it standardly raises a definite expectation of relevance. A communicator is implicitly requesting the attention of her audience, since without such attention her communication cannot succeed. So the deal is, in effect, that the addressee gets a certain level of cognitive effects from the utterance such as to make the giving of his attention (the expenditure of his processing effort) worthwhile. This is captured in the Second Principle of Relevance, given in (4b), with the pivotal concept of 'optimal relevance' spelled out in (5):

(4) a. **The First (Cognitive) Principle of Relevance**:
Human cognitive processes are aimed at processing the most relevant information available in the most relevant way.
b. **The Second (Communicative) Principle of Relevance:**
Every act of ostensive communication conveys a presumption of its own optimal relevance.

(5) **Presumption of optimal relevance**
a. The ostensive stimulus is relevant enough for it be worth the addressee's effort to process it.

b. The ostensive stimulus is the most relevant one compatible with the communicator's abilities and preferences.

(Sperber & Wilson 1995, 270)

It is only ostensive communication to which the second principle of relevance applies; this is to be distinguished from the inadvertent transmission of information and from various kinds of covert communication which fall short of the fully overt and mutually apparent nature typical of verbal utterances (for discussion see Wilson & Sperber, 1993). While utterances are the paradigm case of an ostensive stimulus (a stimulus that comes backed by a communicative intention), others are pointing, winking, and other facial and bodily gestures of the appropriately intentional sort.

Let's look at each of the clauses of the presumption of optimal relevance. What the first clause does is set a lower limit on what the addressee can expect from an utterance, that is, sufficient cognitive effects to warrant the request for his attention (effort). The second clause sets an upper limit on effects: the utterance may achieve more than mere adequacy, though the extent of effects is, of course, limited by the speaker's abilities (for instance, her knowledge of the issue being discussed) and her preferences (for instance, how helpful to the hearer she wishes to be). What goes for effects also goes for effort: the first clause guarantees that the addressee won't be required to expend excessive effort, while the second clause goes further than this, promising the least possible demand on hearer effort, subject again to the speaker's competence (for instance, vocabulary limitations) and her preferences (for instance, a dislike of directness). Some of the many factors that determine how much processing effort is required in any instance are the length of the utterance, the frequency of use of the lexical items employed (so, for instance, 'condiments' might take more processing effort than 'salt and pepper', although the latter is longer), and, most important, the accessibility of the assumptions that make up the context needed to derive the intended cognitive effects.

The communicative principle, and, more particularly, the comprehension strategy that follows from it, accounts for the inferences involved in deriving implicatures, such as those in (1), and for the pragmatic aspects of deriving the proposition expressed, as in (2), and for a range of other pragmatic aspects of interpretation (for instance, the speaker's attitude to each of the propositions expressed or implicated). A worked example of implicature derivation is given in section 3.3, where the relevance-based comprehension strategy is spelled out, in the process of comparing and contrasting the way in which economy of effort figures in cognitive pragmatics and in current minimalist generative grammar.

3 Generative grammar and pragmatics

3.1 Subpersonal systems

Relevance-theoretic pragmatics and generative grammar have in common certain fundamental aims and guiding assumptions. They are both located within **cognitive science**, and, as with most cognitive scientific endeavours, they are both looking for explanations at the **subpersonal** level rather than at the person-level, to call on a useful philosophical distinction. As Chomsky (1992, 213) says:

> People (**persons**) ... pronounce words, refer to cats, speak their thoughts, understand what others say, play chess, or whatever; their brains don't and computer programs don't .... just as it is **persons** who take a walk, not their feet.

These person-level activities are not amenable as such to scientific inquiry, though insight can be gained by the scientific study of some of the subpersonal systems that play a crucial enabling role in these higher level abilities. Marr's account of visual algorithms, Chomsky's I-language, and Sperber & Wilson's pragmatics are all cases of theories of subpersonal systems. And in line with this, they both aim at 'generativity', in at least one sense of the term: that is, **full explicitness**, leaving nothing to the intuitions of the reader or user, so that the description or mechanisms specified could be employed by a mindless automaton with the same results as in the human case. Of course, the extent to which Relevance theory has succeeded in providing subpersonal explanations of the aimed for explicitness is another matter (for some discussion, see Carston forthcoming, chapter 1). In the next sections, I move on to some of the differences between the sort of theory this is and the sort of theory that generative grammar is.

3.2 Competence or performance?

Is cognitive pragmatics concerned with a competence system or a performance system, a system of knowledge or a system which, when prompted by an appropriate stimulus, is *activated*, set in motion, as it were, and goes into its standard processing routines? If a competence system, is it part of **linguistic** competence or some other kind of competence? If a performance system, is it a **linguistic** performance system or some other sort of performer?

In one of his few statements bearing on this issue, Chomsky (1980) speaks of 'pragmatic competence' as a component of the mental state of 'knowing a language', that is, as part of linguistic competence. He distinguishes the following: (a) *grammatical competence*: the computational aspects of language, that constitute knowledge of form and meaning, and (b) *pragmatic competence*: knowledge of the conditions for appropriate use, of how to use grammatical and conceptual resources to achieve certain ends or purposes, (Chomsky 1980, p.59, pp.224-225). It seems to follow from the logic of this position that there must be some sort of pragmatic performance mechanisms which put this pragmatic knowledge system to use. One of the few people to pursue this view of pragmatics as a competence system, a body of knowledge about language, is Asa Kasher (1991a, 1991b, 1994). One of his conclusions is that 'pragmatic competence, as such, is independent of communication' (Kasher 1991a, 135). If this is so, then the pragmatics he is pursuing is something quite other than that developed within Relevance theory, whose domain precisely is ostensive-inferential communication. I very much doubt that there is any such pragmatic competence system (see Carston 1997), and Kasher's own work gives backing to these doubts. Despite talking of competence, he ends up distinguishing different types of pragmatics in terms of Fodorian modular input systems and nonmodular central systems (Fodor 1983); for instance, the 'talk-in-interaction' system responsible for turn-taking, sequencing and conversational repair is a module, while 'central pragmatics' responsible for the generation of conversational implicatures, aspects of style and politeness is part of (non-modular) central systems. See Sinclair (1995) for a discussion of the tension in Kasher's account between his strong adherence to a competence approach to pragmatics and his use of the Fodorian concept of modularity.

Others have taken pragmatics to be the performance counterpart of semantic competence. So, in his survey of pragmatics, Horn (1988, 131) sets up the following equation and attributes it to a number of pragmatists (though he does not endorse it himself):

semantics : pragmatics :: competence : performance

While this comes much closer to the sort of pragmatics that emerges from the Relevance theory view, it remains unclear exactly what it amounts to, since much depends on how we understand semantics, whether as linguistically encoded meaning, or truth conditions, or regularities of use, to mention just some possibilities (see Carston 1998). The only one of these ways of construing semantics that might make the analogy work is the first. Just as the parser (a performance mechanism) uses the grammatical knowledge system in its building of syntactic structures, so the pragmatic inferential mechanism uses the encoded

linguistic meaning in arriving at an interpretation of an utterance. But even on this construal the analogy is far from perfect, since the role in pragmatics of this 'semantic representation' (a rudimentary logico-conceptual form), delivered by linguistic decoding, is more analogous to the phonetic input to the parser than it is to the system of grammatical knowledge.

Relevance theory (RT) pragmatics is emphatically not a component of 'linguistic competence'. It is not 'linguistic' because it does not deal in linguistic stimuli alone, but in all ostensive stimuli, that is, stimuli used for intentional communication, and, as far as we know, the non-demonstrative inferential processes it employs are used in information processing quite generally. It is not a 'competence' system, a body of knowledge, but rather it is a doer, a performer, which operates within the constraints of real-time, on-line processing. In this respect, it is on a par with the linguistic parser, a performance system which accesses and uses linguistic competence (or those parts of it that it is able to use). There IS an important relation between the grammar (or I-language) and RT pragmatics but it is far from direct; the point of contact between the language faculty and RT pragmatics is either the output of the parser or of some further performance system interfacing between parser and pragmatics. On the basis of these observations, then, and given just the binary choice (competence or performance?) the conclusion has to be that RT pragmatics is a performance system, though not a **linguistic** performance system in that it is not part of the language faculty.

Perhaps part of the reason that reaching this conclusion is not completely straightforward is that Chomsky and at least some other generativists are sceptical about the feasibility of pragmatics, where pragmatics is conceived of as an account of utterance interpretation. Such a pragmatics is generally taken to involve the **inferential recognition of speaker's intentions** (this is certainly central to relevance-theoretic pragmatics) and for Chomsky, matters involving human intentions may well lie beyond the scope of scientific enquiry: 'General issues of intentionality, including those of language use, cannot reasonably be assumed to fall within naturalistic inquiry' (Chomsky 1995a, 27). It would not be too surprising then if his competence/performance distinction was set up in such a way that it does not offer an obvious place for pragmatics as conceived of here.

3.3 Computational economy considerations

Let's now consider similarities and differences within the two types of theory in the way they employ the notion of 'least effort'. The outline of Relevance theory in section 2 should have made clear the central role that this concept plays in assessments of

relevance, both in regulating the hearer's derivation of the intended interpretation and in regulating the sorts of utterances speakers produce.

Within the current Minimalist Program, least effort considerations, in the form of certain economy principles, seem to be playing a very central role in the computational system which characterises core grammatical competence. Consider the following quote from Chomsky (1995b, 220):

> The language L thus generates three ... sets of computations: the set D of derivations, a subset D_C of convergent derivations of D, and a subset D_A of admissible derivations of D. FI [the principle of Full Interpretation] determines D_C, and the economy conditions [e.g. 'shortest move'] select D_A.

This has the following implication: two different derivations, made out of the same set of lexical items, generated by operations like 'merge' and 'move', AND fully interpreted (at the interfaces), i.e. convergent, are compared and evaluated with regard to 'cost', that is, the effort involved in their computation, and the more costly (effort-demanding) one is thrown out. For instance, the structures in (6a) and (6b) involve the same lexical items and the same interpretation and, as Marantz (1995) says, **neither of the movements is in itself ungrammatical**:

(6) a. * What did you persuade who to buy t?
b. Who did you persuade t to buy what

But on a comparison between them in terms of economy or least effort, (a) is ruled out as being too costly since it is not the minimal movement possible for a *wh*-element from the starting structure: 'you persuaded who to buy what'. (In fact the derivational comparison required is more global than this would indicate, since it also needs to take account of the cost of the different LF operations involved in the two cases, but this does not affect the point here.)

In order to compare this use of computational economy with that of Relevance theory, I need first to make the background point that there are, quite generally, the following two ways of selecting among possible hypotheses:

(7) a. Set up all the possibilities, compare them and choose the best one(s) (according to some criterion/a).

b. Select an initial hypothesis, test it to see if it meets some criterion/a; if it does, accept it and stop there; if it doesn't, select the next hypothesis and see if it meets the criterion, and so on.

The way in which economy of effort is employed in minimalism seems to be in accordance with (7a); that is, it involves the making of comparisons among hypotheses (derivations). The way in which least effort considerations enter into utterance interpretation, as conceived of by relevance theory, is quite different. There is no process of comparing two or more hypotheses in arriving at the intended interpretation of an utterance.

The comprehension strategy warranted by relevance theory is given in (8):

(8) **Relevance-based comprehension strategy:**
   a. Consider possible cognitive effects in their order of accessibility (i.e. following a path of **least effort**)
   b. Stop when the expected level of relevance is achieved (or appears unachievable)

What the 'expected level of relevance' is varies from situation to situation; it is a function of (the hearer's assessment of) the speaker's abilities and preferences (clause (b) of the presumption of optimal relevance) and of the type of communicative exchange (casual chat, formal lecture, seer's proclamations, etc).

In order to see this comprehension strategy in action, consider the following simple case, focussing on Ann's response to Bob's question:

(9) Bob: Do you want to go to the cinema?
   Ann: I'm tired.
   Possible interpretations:
   a. Ann doesn't want to go the cinema.
   b. Ann does want to go the cinema.

Bob's yes/no question makes it plain that he is looking for an affirmative or a negative answer, that this is **what would be relevant to him** (would have cognitive effects). The proposition expressed by Ann gives him neither, but the utterance comes with a presumption of its own optimal relevance and Bob is thereby licensed to expect that he will get an answer (with accompanying cognitive effects) if he does a bit of interpretive work. Let us suppose that the most accessible assumption to Bob is that when she's tired Ann likes to stay at home. From this implicit premise together with the explicitly

expressed proposition, there follows deductively the relevant conclusion that Ann doesn't want to go to the cinema; this is sufficiently relevant (has enough effects) for Bob, so he stops there. What he doesn't do is try out other less accessible premises, for instance, *if she's tired Ann might like to be taken care of*, on the basis of which, together with further premises, he might eventually arrive at the conclusion that she'd like to be taken to the cinema, a conclusion that might have just as many cognitive effects for him as the conclusion reached on the first interpretation (or even more). The relevance-based comprehension strategy precludes Bob's doing this.

The way in which 'least effort' works here is to prevent any such gathering together of interpretive hypotheses for comparative purposes. Forming several possible interpretations and comparing them to see which achieves the best effects/effort balance, if it were possible at all, would be so effort-consuming that it would be self-defeating. Supposing the level of effects were kept constant, what would be going on would be the expending of a lot of effort to find out which is the least costly interpretation. This is simply ruled out by the way the relevance-based pragmatic criterion works and so it must be if it is to begin to account for the facts of actual on-line interpretation.

The generative endeavour as manifest in the minimalist approach is, in this respect, and myriad others, quite different. As a competence system it is not constrained by temporal or any other on-line processing considerations, so economy considerations can be employed as a means of selecting a subset of derivations from a given larger set. On the other hand, considerations of economy and/or least effort in a parser are more likely to yield a strategy for building syntactic structures which is broadly akin to the relevance-theoretic one for arriving at a full interpretation of an utterance, that is, a strategy of the type in (7b). The parsing strategies proposed by Lyn Frazier and her colleagues, such as 'minimal attachment' and 'late closure', are instances of this type; at those points during on-line structure-building where the grammar allows several possibilities, these strategies ensure that the single most economical course of analysis is pursued (see, for instance, Frazier 1987, Frazier & Clifton 1996).

3.4 Modularity

Certain other properties of generative grammar have been claimed to carry over to pragmatics. Consider the following quote from Larry Horn:

> Pragmatics itself may be viewed as internally modular and interactionist, in the sense that the conceptually distinct subcomponents (suborientations) of pragmatic analysis may be simultaneously called upon within a single explanatory account of a given phenomenon, just as autonomous but interacting grammatical systems may interact to yield the simplest, most

> general, and most comprehensive treatment of some linguistic phenomenon (cf. the deconstruction of passive in Chomsky 1982).
>
> (Horn 1988, 115)

This does not apply in any obvious way to RT pragmatics, which runs the inferential phase (post-linguistic decoding) of utterance interpretation on a single principle. There is no parallel with the way in which the passive construction arises, epiphenomenonally, from the conspiring together, as it were, of several distinct grammatical principles, each of which enters also into the generation of a range of other grammatical structures.

Horn's own pragmatic system runs on two principles, the Q (quantity) principle and the R (relevance or informativeness) principle, given in (10), which do not seem to interact in anything like the way that grammatical principles do. Each accounts for a discrete set of cases of pragmatic inference, which are labelled as such, in (11) and (12):

(10) a. *Q-principle:* Make your contribution sufficient; say as much as you can (given R).
b. *R-principle*: Make your contribution necessary; say no more than you must (given Q).

(11) **Q-implicatures:**
a. Some of the kids were sick >> Not all of the kids were sick
b. I slept in a car last night >> The car was not mine
c. He ate six cakes >> He ate at most six cakes

(12) **R-implicatures:**
a. John and Sam went to Paris >> John and Sam went to Paris together.
b. Mum shouted and the kids wailed >> The kids wailing was a consequence of Mum's shouting
c. I broke a finger yesterday >> The finger was mine

(Horn 1984)

The Q-implicatures, which involve the negation of some stronger proposition than the one expressed, follow from the Q-principle, and the R-implicatures, which involve a strengthening, to a subcase of the proposition expressed, follow from the R-principle. In fact, the two principles appear to make dead opposite predictions, and there is, as yet, no adequate account of when one rather than the other comes into operation, so again we do not seem to have crosscutting, interacting modular principles comparable

to those postulated in generative grammar. (See also Levinson 1987 and forthcoming, for a discussion of three allegedly interacting pragmatic principles.)

However, some very recent developments in the way that the central cognitive system is construed indicate that although the pragmatic system itself seems not to consist in interacting orthogonal principles, there may be a way in which Horn's talk of a **modular** interaction can be given some substance. Sperber (1994) has advocated a thorough-going modularity of mind, in which individual concepts (with their own inferential procedures and proprietary data bases of encyclopedic information) constitute distinct modules. Furthermore, within this picture, there may well be a distinct module specialised for utterance comprehension, a pragmatics module. Cognitive pragmatics does seem to have at least some of the essential properties of Fodorian modules: domain specificity, its own specific principle (the communicative principle of relevance), and a preprogrammed course of development. But whether it could be said to be informationally encapsulated is very questionable; I don't know how to reconcile the marshalling of contextual assumptions, which may be accessed from a range of sources, including current perception and long term memory store, with the view that an essential property of a module is its encapsulation from at least some significant body of the information available in the overall system. But that is a matter that may be resolved with some recasting of the concept of modularity so that it can encompass both domain-specific systems of perception and of central cognition. It is simply too soon to tell. (For further discussion, see Carston 1997.)

4 'Discourse competence' and pragmatics

In this last section, I would like to look at the view that there is a 'discourse competence' system, which is a component of the grammar, a view propounded for some time by Susumo Kuno (1980, 1987) and Ellen Prince (1985, 1988, 1997). The idea is that there is a 'linguistic pragmatic competence', so that a 'generativist' approach to discourse analysis is called for. This is discourse analysis construed as 'the principles underlying a speaker's choice of a particular syntactic or referential option in a context and the principles underlying a hearer's understanding of it.' (Prince 1988, 166-67). Prince recognises that there is a whole sheaf of aspects of utterance understanding that are not a part of linguistic competence; she explicitly mentions conversational implicatures as falling outside the domain of linguistic competence, hence outside the domain she is looking at.

Let us look at the phenomena she has in mind as belonging in the domain of discourse competence. An instance of the **syntactic** cases she considers is the following:

(13) a. They found Eichmann.
b. It was they who found Eichmann.

Here we have two sentences of English with rather different syntactic structures but which seem to have the same propositional (truth-conditional) content. What governs a speaker's choice of one over the other and what difference does one rather than the other make to the hearer's understanding? It is not a difference in the truth-conditional content communicated. The clefting in (b) imposes a certain structure on the information it encodes, a kind of presupposition-focus structure, making it appropriate in certain contexts and not in others. For instance, (13b) is appropriate in a context in which it has been established that Eichmann has been found and there is a question about who found him; (13a) would not be appropriate in such a context. That the cleft construction partitions the information in this way is something speakers and hearers **know**; it is plausibly part of their linguistic competence. And the case is clinched by a cross-linguistic comparison of structures that function in the same way as the English *it*-cleft, because it appears to be the case that the functional counterparts in other languages are very different from the English structure, that is, they are arbitrary and language-specific, which is generally seen as the mark of knowledge that belongs in linguistic competence. Any feelings of iconicity we may have about the *it*-cleft are an illusion. The English cases are compared with the corresponding Yiddish structures in (14):

(14) a. ... zey hobn gefunen aykhmanen
... they have found Eichmann
b. ... *dos* hobn zey gefunen aykhmanen
... this have they found Eichmann
= it was they who found Eichmann

(Prince 1988, p.169)

The *dos*-construction, which is the functional equivalent of the English *it*-cleft, is quite different from it in structure: it is a **single clause**, with the lexical item *dos* meaning 'this' in leftmost position, and the functionally focussed *zey* ('they') is not syntactically isolated or highlighted in any way. Furthermore, the *dos*-construction is absolutely parallel in form to another single clause structure in Yiddish, the *es*-construction, which differs from it formally only in having *es* (meaning 'it') in place of *dos*; this structure has a totally different discourse function: it is used when the sentence is 'athematic' (or non-presuppositional), carrying the fewest possible assumptions about shared knowledge. The formal similarity and functional dissimilarity of the two structures reinforces the arbitrary, hence linguistic, nature of the knowledge involved in their use.

Prince has looked at a wide range of other syntactic structures too, including *wh*-clefting, gapping, topicalization, and left-dislocation, reaching the same general conclusion: 'such form-function correlations must lie squarely within the domain of linguistic competence, attributable neither to common sense reasoning nor to "iconicity"' (Prince 1997, 119). And she has made comparable points regarding equivalent referential options :

(15) a. Last week I read a book and I met an author.
b. Last week I read a book and I met the author.

The choice of the definite or the indefinite clearly makes a difference to understanding but that difference is not, according to her, truth-conditional (Prince 1988, p.172). It is a pragmatic difference, the definite NP signalling that the hearer is already familiar with, or can easily infer, the entity in question, while the indefinite NP signals that the entity in question is not already available to the hearer. Again, she looks at a range of data from languages other than English and finds that they all mark this distinction but that they do so in various different ways. So this looks like another area of linguistic knowledge which falls within the domain of 'discourse competence'. (See Prince (1992) for more detailed discussion of the discourse functions of different types of definite NPs.)

Not everyone agrees with Prince that the interpretive effects of these elements of linguistic form are all to be explained in terms of arbitrary linguistic encodings. For example, Sperber & Wilson (1986/95, 202-217) argue instead for 'a natural linkage between linguistic form and pragmatic interpretation' without any intermediate level of semantic or pragmatic description. Their idea is that the way in which the truth-conditional content of an utterance is organised syntactically (and also phonologically) directly affects its on-line processing. By employing such devices as contrastive stress, *it*-clefting, left-dislocation and other preposing structures, a speaker can ensure that the hearer allocates different degrees of processing effort to different parts of the overall information encoded in the utterance, and is thereby directed towards the primary loci of cognitive effects. This is an exciting idea but one which calls for considerable further research, since Prince's careful analytical work on cross-linguistic data does seem to provide strong support for the view that particular discourse functions are encoded by particular syntactic structures and that the nature of these encodings varies considerably across languages.

A further wrinkle in the picture comes from the observation made by Seuren (forthcoming) that clefting (and contrastive accent) makes a clear truth-conditional difference under certain intensional operators. So while (16a) and (16b) seem to be truth-conditionally equivalent (but each appropriate in a different context), (16c) and (16d) are not truth-conditionally equivalent as is shown by the consistency of (16e).

(16) a. It's Bill who has won the BMW.
b. It's the BMW that Bill has won.
c. Pip is annoyed/surprised that it's BILL who has won the BMW.
d. Pip is annoyed/surprised that it's the BMW that Bill's won.
e. Pip is annoyed/surprised that it's BILL who has won the BMW, not that it's the BMW that Bill has won.

And one might have similar doubts about the truth-conditional equivalence of the definite/indefinite referential options in (15).

However, although there are many issues to be resolved and perhaps some sorting out of different cases to be done (for instance, it may be that clefting affects truth conditions while left-dislocation does not, etc.), Prince has made a strong case which has not yet, to my knowledge, been challenged head-on. Assuming then, for the time being, that there are indeed truth-conditionally equivalent syntactic structures (and referential options) which arbitrarily encode differences in discourse function, what interests me is a striking parallel between Prince's accounts of these and a line of work currently going on under the label of 'procedural semantics' within the framework of relevance theory. This work has focussed mainly on a particular group of lexical items, sometimes called *discourse connectives*, whose meaning also does not appear to affect truth-conditional content:

(17) a. Bill's not coming. But Jane is.
b. Bill's not coming. So Jane is.
c. Bill's not coming. After all, Jane is.

Diane Blakemore has analysed the connectives here as encoding **not** concepts, **not** truth-conditional meaning, but directives, bits of guidance to the hearer about the sort of inferential relations he should be looking for between the propositional content of the utterance they introduce ('Jane is coming') and the context (here the immediately preceding utterance). Intuitively, the three connectives in the examples in (17) indicate three distinct inferential relations, which Blakemore (1987) examines in detail. The idea is that the semantics of these connectives acts as a constraint on the pragmatic processing required to reach the intended interpretation. For instance, 'after all' signals that the proposition it introduces is evidence for (can be used as a premise in an argument supporting) the preceding assertion ('Bill's not coming'). (See also Blakemore 1990 and forthcoming, and Wilson & Sperber 1993.)

There are many similarities here to the cases Prince is interested in: the form-function relation of these connectives is arbitrary (non-iconic), the forms across languages that encode these 'discourse functions' are quite distinct and specific to the particular language, and the contribution of these elements to utterance meaning is

non-truth-conditional. In short, there is quite a range of linguistic forms, both lexical items and syntactic structures, which have these properties and native speakers' knowledge of how they work is certainly a part of their linguistic competence. Nothing substantive hangs on whether we call this subcomponent of linguistic competence 'discourse competence', as Prince does, seeing it as that part of pragmatics which is properly linguistic, or 'procedural semantics', as relevance theorists do, seeing it as a component of 'linguistic semantics' (competence). On the latter approach, the term 'pragmatics' is reserved for that part of utterance meaning which is recovered by inferential processes wholly dependent on the guidance of a general principle of communication. These processes take as their input the whole of decoded meaning, both that which contributes to the truth-conditional content of the utterance and that which encodes information about its linkup with context.

It is not too surprising that a linguistic system that gets used for ostensive-inferential communication should develop devices with these discourse functions. As Blakemore has pointed out, these are just the sort of effort-saving devices you would expect to be provided by a code which is subservient to a relevance-driven inferential processing mechanism, a mechanism which is geared to derive cognitive effects at least cost to the processing resources of the system. An interesting question, though somewhat tangential to this paper, arises here: does the language of thought have anything comparable to these truth-conditionally irrelevant structures and lexical items? It seems unlikely since their function is to constrain and guide the processes involved in communication/discourse. The question is the more interesting if you believe that we think in natural language, at least some of the time. This is a position that Chomsky sometimes seems to take (Chomsky (1975, 57), (1986, 14, endnote 10), (1993, 48)) and that some other linguists and philosophers certainly take (for instance, Smith (1982), and Carruthers (1996, 1998)). Carruthers (1996) argues not merely that we sometimes think in natural language, but that sometimes we *must* think in natural language. This strong claim is that, of natural necessity, conscious human propositional thought involves images of *natural-language sentences*, a view at odds with that of Jerry Fodor (1975), that the language of thought is an innate universal 'mentalese', distinct from specific natural languages. Among the many issues raised by Carruthers' position is whether or not such natural language structures as *it*-clefts and left-dislocations, arguably dedicated to particular discourse functions, could also be structures for thinking with, and whether or not discourse connectives and markers, analysed as processing directives to addressees, have any any role to play in thinking.

5 Summary

I've addressed some of the questions that arise out of a consideration of the relationship between generative grammar and cognitive pragmatics, specifically

relevance-theoretic pragmatics. The deepest foundations of the two theories are similar: they aim to give fully explicit accounts of subpersonal cognitive systems that play a role in the person-level activities of expressing thoughts and understanding utterances. They are, though, quite different types of theorising, the one a study of a system of knowledge, the other of a performance mechanism. Many differences follow from this distinction. I demonstrated one of these in some detail, the employment of economy of effort considerations in the two theories; in the competence theory, they provide a means for selecting one derivation from an antecedently given range; in the performance theory, they determine an order in the accessing of interpretive hypotheses, so that less accessible (more effort requiring) ones usually do not enter the picture at all.

Pursuing another issue raised by the competence/performance dichotomy, I have looked at the idea that there is a discourse or pragmatic component of linguistic competence. There is strong evidence that languages contain a wide variety of forms which encode procedures or directives which guide the hearer in his task of processing an utterance and reaching the intended understanding. Describing what these devices encode is part of the bigger task of formulating an account of linguistic knowledge (grammar). It is intimately tied to the complementary account of pragmatic performance, since these encodings play a shaping role in the accessing of contextual assumptions and in establishing particular inferential relations between the conceptual content of the utterance and the context.

In his recent book on debates between formalists and functionalists in language studies, Newmeyer (1998) presents the work of Kuno and Prince, on the one hand, and of the relevance-theoretic pragmatists, on the other hand, as apparently at odds with each other (although he suspects that they may ultimately be 'integrable at some level of abstraction'). According to the former camp, there is a discourse component to the grammar; according to the latter, pragmatic phenomena are a function of general central systems' inference, so that 'discourse competence cannot be on a theoretical par with grammatical competence' (Newmeyer, 1998, p.66). I maintain that there is absolutely no clash or tension here, but rather two distinct and complementary domains of enquiry, the one a part of the study of linguistic competence (and so on a theoretical par with grammatical competence), the other an account of pragmatic performance (and so not on a theoretical par with grammatical competence).

References

Blakemore, D. (1987). *Semantic Constraints on Relevance*. Oxford: Blackwell.

Blakemore, D. (1990). Constraints on interpretations. *Proceedings of the 16th Annual Meeting of the Berkeley Linguistic Society; Parasession on the Legacy of Grice*, 363-370.

Blakemore, D. (forthcoming). Discourse and relevance theory. In D. Schiffrin, D. Tannen & H. Hamilton (eds.) *Handbook of Discourse Analysis*. Oxford: Blackwell.

Carruthers, P. (1996). *Language, Thought and Consciousness: An Essay in Philosophical Psychology*. Cambridge: Cambridge University Press.

Carruthers, P. (1998). Thinking in language? Evolution and a modularist possibility. In P. Carruthers & J. Boucher (eds.) *Language and Thought: Interdisciplinary Themes*, 94-119. Cambridge: Cambridge University Press.

Carston, R. (1997). Relevance-theoretic pragmatics and modularity. *UCL Working Papers in Linguistics* 9, 29-53.

Carston, R. (1998). The semantics/pragmatics distinction: a view from Relevance Theory. *UCL Working Papers in Linguistics* 10, 303-329. Longer version in K. Turner (ed.) 1999. *The Semantics/Pragmatics Interface from Different Points of View* (CRiSPI 1), 85-125. Elsevier Science.

Carston, R. (forthcoming). *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell.

Chomsky, N. (1975). *Reflections on Language*. Fontana.

Chomsky, N. (1980). *Rules and Representations*. Oxford: Blackwell.

Chomsky, N. (1986). *Knowledge of Language*. New York: Praeger.

Chomsky, N. (1992). Explaining language use. *Philosophical Topics* 20, 205-231.

Chomsky, N. (1993). *Language and Thought*. Moyer Bell.

Chomsky, N. (1995a). Language and nature. *Mind* 104, 1-61.

Chomsky, N. (1995b). *The Minimalist Program*. Cambridge, MA.: MIT Press.

Fodor, J. (1975). *The Language of Thought*. Sussex: The Harvester Press.

Fodor, J. (1983). *The Modularity of Mind*. Cambridge, MA.: MIT Press.

Frazier, L. (1987). Sentence processing: a tutorial overview. In M. Coltheart (ed.) *Attention and Performance XII*. Hillsdale, N.J.: Lawrence Erlbaum.

Frazier, L. & Clifton, C. (1996). *Construal*. Cambridge, MA.: MIT Press.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (ed.). *Meaning, Form and Use in Context (GURT '84)*, 11-42. Washington: Georgetown University Press.

Horn, L. (1988). Pragmatic theory. In F. Newmeyer (ed.). *Linguistics: the Cambridge Survey, vol. I*, 113-145. Cambridge: Cambridge University Press.

Kasher, A. (1991a). Pragmatics and Chomsky's research program. In A. Kasher (ed.). *The Chomskyan Turn*, 122-149. Oxford: Blackwell.

Kasher, A. (1991b). On the pragmatic modules: A lecture. *Journal of Pragmatics* 16, 381-397.

Kasher, A. (1994). Modular speech act theory: Programme and results. In S. Tsohatzidis (ed.). *Foundations of Speech Act Theory*, 312-322. London: Routledge.

Kuno, S. (1980). Functional syntax. In E. Moravcsik & J.Wirth (eds.). *Current Approaches to Syntax. Syntax and Semantics 13*, 117-136. New York: Academic Press.

Kuno, S. (1987). *Functional Syntax: Anaphora, Discourse, and Empathy*. Chicago: Chicago University Press.

Levinson, S. (1987). Minimization and conversational inference. In: J. Verschueren & M. Bertuccelli-Papi (eds.) *The Pragmatic Perspective*, 61-129. Amsterdam: John Benjamins.

Levinson, S. (forthcoming). *Presumptive Meanings: The Theory of Generalized conversational Implicature*. Cambridge, MA.: MIT Press.

Marantz, A. (1995). The minimalist program. Chapter 7 in G. Webelhuth (ed.) *Government and Binding Theory and the Minimalist Program*. Oxford: Blackwell.

Newmeyer, F. (1998). *Language Form and Language Function*. Cambridge, MA.: MIT Press.

Prince, E. (1985). Fancy syntax and "shared knowledge". *Journal of Pragmatics* 9, 65-82.

Prince, E. (1988). Discourse analysis: a part of the study of linguistic competence. In F. Newmeyer (ed.) *Linguistics: the Cambridge Survey, vol. II*, 164-182. Cambridge: Cambridge University Press.

Prince, E. (1992). The ZPG letter: subjects, definiteness, and information status. In W. Mann & S. Thompson (eds.) *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, 295-325. Amsterdam: John Benjamins.

Prince, E. (1997). On the functions of left-dislocation in English discourse. In A. Kamio (ed.) *Directions in Functional Linguistics*, 117-143. Amsterdam: John Benjamins.

Seuren, P. (forthcoming). Presupposition, negation and trivalence. *Journal of Linguistics*.

Sinclair, M. (1995). Fitting pragmatics into the mind: some issues in mentalist pragmatics. *Journal of Pragmatics* 23, 509-539.

Smith, N. (1982). Speculative linguistics. Inaugural lecture at University College London.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. Hirschfeld & S. Gelman (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture*, 39-67. New York: Cambridge University Press.

Sperber, D. & Wilson, D. (1986/95). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Sperber, D. & Wilson, D. (1987). Precis of *Relevance: Communication and Cognition*. *The Behavioral and Brain Sciences* 10.4, 736-54.

Sperber, D. & Wilson, D. (1995). Postface. In D. Sperber & D. Wilson *Relevance: Communication and Cognition*, second edition, 255-279. Oxford: Blackwell.

Wilson, D. & Sperber, D. (1993). Linguistic form and relevance. *Lingua* 90, 1-25.

Generative grammar is a linguistic theory that regards linguistics as the study of a hypothesised innate grammatical structure. A sociobiological modification of structuralist theories, especially glossematics, generative grammar considers grammar as a system of rules that generates exactly those combinations of words that form grammatical sentences in a given language. The difference from structural and functional models is that the object is placed into the verb phrase in generative grammar. This