Reasoning, judgment and pragmatics

Guy Politzer

In I. Noveck & D. Sperber (Eds.)
Experimental Pragmatics (pp. 94-115). **London : Palgrave.**

Reasoning, Judgment, and Pragmatics

Guy Politzer

C.N.R.S. , Saint-Denis,

Introduction

In psychological experiments on reasoning, participants are typically presented with premises which refer to general knowledge or which are integrated in an original scenario; then, either they are asked to derive what follows from the premises or they are provided with one or several conclusions and asked to decide whether or not these conclusions follow from the premises. There is always a logical argument underlying the premises and the conclusion, and the aim of such experiments is to study participants' performance with respect to a theoretical model, either normative or, as is more usual nowadays, descriptive. The experiments on judgment do not differ much, except that they look more like a problem to solve, where the final question is a request for a comparison, a qualitative or a quantitative evaluation, etc. The experiment may be administered orally during an interview with the experimenter, but more often it is administered in a written form, using paper and pencil or a computer. Given that there are two interlocutors engaged in a communication, a conversational analysis is appropriate, whether the presence of the experimenter is physically real or mediated by the support of the written messages.

After he has been provided with the instructions and the information that supports the question (the scenario, the argument, the problem statement, etc.) the participant is presented with the target question. Like any utterance, this question must be interpreted. Its meaning generally is not straightforwardly identifiable because the information may be more or less long, complicated (and occasionally conceptually hard). It may also be vague or ambiguous. As for any question, its interpretation is determined by the content of the putative answer: the answer should satisfy the expectation of relevance attributed by the participant to the experimenter. Now, in experimental settings (as well as in instructional settings and more generally in testing situations) the participant is aware that the question put to him is a higher order question, that is, does not implicate 'the experimenter does not know how to find the answer' but rather 'the experimenter knows how to find the answer and she wants to know whether I know how to find it'.

The interpretation of the question is determined in part and revealed by the specific kind of knowledge that the participant chooses to exhibit through his response: this choice is made on the assumption that what is relevant to the experimenter is to know whether the participant has that kind of knowledge. This choice and the underlying assumption reveal in turn the

participant's *representation of the task.* This is why knowledge of the population tested is essential. The range of questions of interest which participants are likely to attribute to the experimenter must be anticipated by the experimenter (another, higher order, attributional process) in the light of the participants' educational and cultural backgrounds. This requires a macroanalysis of the information provided, including the non verbal experimental material (e. g. , does the material used suggest that reaction times will be measured?) Social psychologist had related concerns quite some time ago, albeit more limited and focused on the transparency of the experiment; e. g. , Orne (1962) defined the notion of *demand characteristics* as "the totality of cues which convey an experimental hypothesis to the subject". Only recently did a few investigators of thinking and reasoning (Hilton, 1995 ; Schwarz, 1996, and co-workers) applied the so-called « conversational » approach to the relationship between experimenter and participant, in order to study how participants' expectations and attributions affect their responses.

There is, in addition, another kind of analysis, based on pragmatic theory, which needs to be applied to the sentences used to state the argument or the problem. The output of this analysis is the determination of the interpretation of the premises, conclusion or question which the participant is likely to work out; in a word, it delivers the actual proposition(s) which will be processed during the inferential treatment, taking into account (as will be examplified below) the frame of the task representation. The reason to perform this microanalysis is that it is an essential step to guarantee the validity of the experimental task. Indeed the experimenter is interested in the processing of specific propositions which she expects the participant to recover from the sentences used in the argument or problem statement. Unluckily (at least in the early times of the experimental investigation of thinking) these sentences used to be either ackward and artificial formulations inspired by logic textbooks or sentences expressed in very impoverished contexts; and it was assumed that some kind of literal meaning was communicated and then the associated propositions processed. It is clear that a formal logical argument can be deemed to have been followed or not followed only to the extent that the propositions which constitute it are those which the participant has actually processed. For example, in the study of deduction, the endorsement of a conclusion which does not follow validly from the premises, or the non-endorsement of a conclusion which follows validly can be declared reasoning errors only if it can be ascertained that the participant did construe the propositions (premises and conclusion) in a way that coincided with the formal logical description of the argument.

In brief, knowledge of how people represent reasoning and judgmental tasks and of how they interpret the premises or the questions is an indispensable prerequisite for the investigation of the inferential process proper. The recommendation that experimental tasks should be submitted to a macro- and a microanalysis is made with hindsight. For a long period which ended in the late seventies, psychologists showed little concern about such problems. The reason is that most of them were not yet familiar with the tools offered by pragmatic theory (and at an earlier time pragmatic theory itself was not developped enough to offer such tools). As a result,

many erroneous evaluations of the performance observed in experiments and many unfounded claims about human rationality were made. This will be illustrated by reviewing a number of tasks, some of which have been extremely influential, and by describing some of the experimental work carried out in support of the pragmatic approach just outlined. Studies that concern reasoning (deduction and induction) and judgment (probabilistic and classificatory) will be considered in turn.

Studies of deduction

Quantifiers.

It will be useful to begin with a prototypical case, namely the deductions called *immediate inferences.* They are elementary one-premise arguments in which the premise and the conclusion are quantified sentences which belong to Aristotle's square of opposition. In experiments, participants are presented with one premise such as, e. g. *[on the blackboard] some squares are white,* and asked to evaluate (by 'true', 'false', or 'one cannot know') one or several conclusions provided to them, such as *all squares are white; no square is white,* etc. Whereas performance for contraries *(all...are...* to *no...are...* and vice versa) and for contradictories *(all...are...* to *some...are not...* and vice versa; *no...are...* to *some...are...* and vice versa) is nearly perfect, performance for subalterns *(all...are...* to *some... are...* and vice versa; *no...are...* to *some...are not...* and vice versa) is apparently very poor (around one quarter of the responses coincide with the formal logical response, that is, 'true' from universal to particular sentences, and 'one cannot know' from particular to universal sentences, while a strong majority opt for the response 'false' in both directions. The same obtains for subcontraries *(some... are...* to *some... are not...* and vice versa) to which most people respond by 'true' instead of the formal logical response 'one cannot know' which logic textbooks would prescribe (Begg & Harris, 1982; Newstead & Griggs, 1983; Politzer, 1990).

It would be a mistake to attribute poor logicality to participants in such experiments. Assuming that participants process the sentences as if they were uttered in a daily conversation (rather than using the conventions of logicians which require a literal interpretation), the microanalysis applied to quantifiers suggests that people add the scalar implicature *not all* to *some.* If this is so, all the data are coherent. A universal sentence (e.g. , *all... are...)* and its particular counterpart *(some... are...)* being contradictory under the interpretation of the latter as *some... but not all are...,* the inferences that involve these two sentences will lead the reasoner to the conclusion 'false'. And similarly, both particular sentences being equivalent to *some... are... but some... are not...,* the reasoner concludes 'true' when one is a premise and the other the conclusion.

As this example shows, pragmatic theory provides the conceptual tools to identify the propositions actually processed by participants in psychological experiments. It could be argued that, in return, the tasks used by psychologists can provide useful tools to test some claims made

by pragmatic theory. As far as quantifiers are concerned, one of these claims is that the hearer's awareness of the speaker's epistemic state can affect his interpretation of *some.* If the speaker is known to be fully informed, the choice of the weaker item on the scale does convey an implicature based on the fact that the stronger item which is more informative or more relevant was not chosen; but if he is known to be not fully informed, then the choice of the weaker item may as well be attributed to lack of knowledge, and the implicature is less likely to be generated. Consider now the following situation. A radar operator is describing the screen. Some participants are told that the operator is working without time pressure and with certainty, i. e. , she is omniscient, and some others that she is working with time pressure and uncertainty (non-omniscient). Consider the statement, *some spots are large.* When she is omniscient the use of *some* may license the implicature *not all* for the reasons seen above. But when she is not, it cannot be ruled out that all the spots are large. In an experiment (Politzer, unpublished) that used this scenario, the frequency of restrictive interpretations of *some* could be inferred on the basis of the conclusions that participants endorsed (such as *all spots are large).* When the speaker was assumed to be omniscient, the rate of restrictive interpretations was around 75 percent; but when she was assumed to be non-omniscient it dropped on average to 50 percent. This difference was reliable and it was observed in a within- as well as in a between-subjects design, which bears out the general pragmatic prediction. One might wonder why the restrictive interpretations did not collapse altogether. This seems to illustrate one limitation of the paper-and-pencil methodology, namely the difficulty for participants to exploit mental states attributed to fictitious characters. Given the artificiality of the manipulation, one might even regard its effect as impressive.

Conditional reasoning.

For many years, studies of propositional reasoning have focused on "conditional reasoning", that is, two deductively valid arguments:

- Modus Ponendo Ponens (MP): *if A then C; A; therefore C,* and
- Modus Tollendo Tollens (MT): *if A then C; not-C; therefore not-A,*

and two invalid arguments, which are the fallacies of:

- Affirming the Consequent: *if A then C; C; therefore A,* and
- Denying the Antecedent: *if A then C; not-A; therefore not-C.*

Nearly everyone endorses the conclusion of MP. For example (instantiating A with *it rains,* and C with *Mary stays at home),* given *if it rains Mary stays at home* and *it rains,* most people instructed to consider the premises as true endorse the conclusion *Mary stays at home.* However, not everyone endorses the conclusion of MT: knowing for sure that *if it rains Mary stays at home,* and that *Mary does not stay at home,* only about two thirds conclude *it does not rain.* Performance on the two invalid arguments seems even less satisfactory: given that *if it rains Mary stays at home,* and that *it does not rain,* around one half of the people endorse the conclusion *Mary does not stay at home,* although this does not follow deductively. And

similarly, from the premises *if it rains Mary stays at home,* and *it does not rain,* around one half of the people incorrectly endorse the conclusion *Mary does not stay at home.* These are robust observations (Evans, Newstead & Byrne, 1993).

Invalid arguments. Do all people who endorse the conclusion of the invalid arguments commit a fallacy? Let us first consider the microanalysis of the task.

Ducrot (1971) proposed a principle (similar to Grice's first maxim of quantity), which he called the *law of exhaustivity,* "give your intelocutor the strongest information that is at your disposal and that is supposed to be of interest to him", from which it follows that there is a tendency to comprehend a limited assertion as the assertion of a limitation; in particular, *if it rains Mary stays at home* suggests that it is only in case it rains that Mary stays home, which explains the interpretation of *if* as a sufficient-and-necessary condition (or *biconditional* for short).

Geis & Zwicky's (1971) used the now often quoted example *if you mow the lawn, I'll give you five dollars* to show that in some contexts a conditional sentence suggests an *invited inference,* in the present case the obverse of the original sentence, *if you don't mow the lawn, I will not give you five dollars.* This inference was hypothesised to follow from a *principle of conditional perfection,* but Lilje (1972) questioned that there is such a principle. He objected that the inference crucially depends on the circumstances, as shown by the example in which the target sentence would be a reply to *"How can I earn five dollars?"* In such a context, there are alternative antecedents (clean up the garage or whatever) that prevent mowing the lawn from being a necessary condition. Nevertheless, Geis & Zwicky's paper was very influential, so that the conditional reasoning task was the first reasoning task to be examined from a pragmatic point of view (Taplin & Staudenmayer, 1973; Staudenmayer, 1975; Rips & Marcus, 1977). There are more recent theoretical treatments of conditional perfection (Horn, 2000; van der Auwera, 1997); without entering the technical debate, it will be assumed that the interpretation of *if* as a biconditional stems from an implicature which the hearer may generate on the basis of his knowledge base, given the aim of the conversational exchange.

This leads us to the macroanalysis. Braine (1978) was among the first psychologists to stress the differences between 'practical reasoning' which uses premises as they are comprehended in daily verbal exchange, and formal reasoning which requires a special attitude in order to set aside implicatures. That there are individual differences in interpretation of the conditional which can be related to educational background (among other factors) was demonstrated by the results of a truth-table task (Politzer, 1981). In such a task, given a conditional sentence *if A then C,* participants are asked to choose which of the four possible contingencies (A and C; A and not-C; not-A and C; not-A and not-C) they judge to be compatible with the sentence. The choices made by Arts students were characteristic of a biconditional interpretation (A and C; not-A and not-C) more often than the choices made by Science students; these in turn had more often the formal interpretation (all cases except A and

not-C). Clearly the Science students (even though they were untutored in formal logic) were more apt to represent the task as a formal game using literal meaning.

Now an important point is that under a biconditional interpretation of the conditional premise the two fallacious arguments become valid: from *if A (and only if A) then C; not-A*, the conclusion *not-C* follows; and similarly from *if A (and only if A) then C; C,* the conclusion A follows. Consequently, if a participant endorses the conclusions of the two invalid arguments while construing the conditional sentence as a biconditional, one cannot talk any more of committing a fallacy because under such an interpretation the arguments become valid. It follows that the only way to know whether people commit a fallacy, and if so, how often, is to present a conditional premise of which the implicature is cancelled. In order to do so, Rumain, Connell, & Braine (1983) presented a control group of participants with the invalid arguments made of a major premise such as *if there is a dog in the box, then there is an orange in the box* and the appropriate minor premise, *there is no dog in the box* (for the argument of Negation of the Antecedent) or *there is an orange in the box* (for the argument of Affirmation of the consequent); the fallacies (namely, concluding *there is not an orange in the box* and *there is a dog in the box*, respectively) were commited 70% of the time. The experimental group was presented with the same two premises together with an additional conditional premise such as *if there is a tiger in the box, then there is an orange in the box* indicating that there may be an orange without a dog. This aimed at cancelling the implicature *if there is not a dog, then there is not an orange* that is held responsible for the biconditional interpretation and therefore for the fallacies. Indeed, participants in this group committed the fallacies only 30% of the time, presumably because the cancellation of the implicature gave way to the conditional interpretation. (The question of the residual 30% of fallacies is beyond the scope of this chapter). This kind of manipulation has been widely replicated and generalised to various contexts (Byrne, 1989; Manktelow & Fairley, 2000; Markovits, 1985).

Valid arguments and credibility of the premises. While it is established that performance on the invalid conditional arguments crucially depends on the interpretation of the major conditional premise, in the past twelve years a number of experimental manipulations have revealed interesting effects on the endorsement of the conclusion of the two valid arguments.

Cummins (1995; Cummins, Lubart, Alksnis, and Rist, 1991) studied these arguments with causal conditionals. She demonstrated that the acceptance rate of the conclusion depends on the domain referred to in the major premise. For example, of the two following arguments:

If the match was struck, then it lit; the match was struck; therefore it lit, and

If Joe cut his finger, then it bled; Joe cut his finger; therefore it bled,

people are less prone to accept the conclusion of the first. The variable which was manipulated is the number of "disabling conditions" that are available. Disabling conditions are such that their satisfaction is sufficient to prevent an effect from occurring (and their non-satisfaction is therefore necessary for the effect to occur, e. g. , dampness of the match, and superficiality of the cut, respectively): the acceptance rate was a decreasing function of their number.

Thompson (1994, 1995) obtained differences in the endorsement rate of the conclusion with causals as well as non-causal rules such as obligations, permissions and definitions by using conditionals that varied in 'perceived sufficiency' (estimated by judges). A sufficient relationship was defined as one in which the consequent always happens when the antecedent does; for example, the following sentences were attributed high and low sufficiency, respectively: *If the licensing board grants them a license then a restaurant is allowed to sell liquor. If an athlete passes the drug test at the Olympics then the IOC can give them a medal.* She observed that the endorsement rate of the conclusion was an increasing function of the level of perceived sufficiency.

Newstead, Ellis, Evans, and Dennis (1997) and Evans and Twyman-Musgrove (1998) used as a variable the type of speech act conveyed by the major conditional premise; they observed differences in the rate of endorsement of the conclusion: promises and threats on the one hand, and tips and warnings on the other hand constituted two contrasted groups, the former giving rise to more frequent endorsements of the conclusion than the latter. (These classes of conditionals were investigated in the seventies by Fillenbaum, 1975, 1978). They noted that the key factor seems to be the extent to which the speaker has control over the occurrence of the consequent, which is higher for promises and threats than for tips and warnings.

George (1995) manipulated the credibility of the conditional premise of MP arguments. Two groups of participants received contrasted instructions. One group was asked to assume the truth of debatable conditionals such as *If a painter is talented, then his/her works are expensive* while another group was reminded of the uncertain status of such statements. As a result, 60 percent in the first group endorsed the conclusion of at least three of the four MP arguments, but only 25 percent did in the second group.

While each of these authors has an explanation for his or her own results separately, it will be proposed that there is a single explanation along the following lines (Politzer, 2003).

(i) conditionals are uttered in a background knowledge, of which they explicitly link two units (the antecedent and the consequent), keeping implicit the rest of it, which will be called a *conditional field;*

(ii) the conditional field has the structure of a disjunctive form, as proposed by Mackie (1974) for causals. The mental representation of a conditional *if A then C* (excluding analytically true conditionals) in its conditional field can be formulated as follows :

$$[\,(A \,\&\, A_1 \,\&\, A_2 \,\&\ldots)\ \vee\ (B \,\&\, B_1 \,\&\, B_2 \,\&\ldots)\ \vee \ldots\,] \rightarrow C\,.$$

A is the antecedent of the conditional under consideration; B is an alternative condition that could justify the assertion of *if B then C* in an appropriate context. (The fact that alternative antecedents like B and its conjuncts may not exist, or may be assumed to not exist, is at the origin of the *if not-A, then not-C* implicature considered above, but this is not our current concern). We focus on the abridged form,

$$(A \,\&\, A_1 \,\&\, A_2 \,\&\ldots) \rightarrow C\,.$$

While (A & A_1 & A_2 &. . .) is a sufficient condition as a whole, each conjunct A_1 , A_2, . . . separately is necessary with respect to A. These conjuncts will be called *complementary necessary conditions* (henceforth CNC). Each of the CNC's has its own availability, and this availability is part of what specifies the conditional field.

(iii) it is hypothesised that in asserting the conditional *if A then C,* the speaker assumes that the necessity status of the conditions A_1 , A_2, . . . is part of the cognitive environment, and most importantly that the speaker has no reason to believe that these conditions are not satisfied. The formula can be rewritten as:

$$\{A_1 \ \& \ A2 \ \&. \ . \ .\} \ \& \ A \ \rightarrow \ C,$$

where the braces indicate that the CNC's are tacitly assumed to hold. This is justified on the basis of relevance: in uttering the conditional sentence, the speaker guarantees that the utterance is worth paying attention to. But this in turn requires that the speaker has no evidence that the CNC's are unsatisfied, failing which the sentence would be of little use for inferential purposes. (In making this assumption, one must accept that the implicature concerns not a single constant, such as A_1, but a variable A_i).

In brief, conditionals are typically uttered with an implicit *ceteris paribus* assumption to the effect that the normal conditions of the world (the satisfaction of the CNC's that belong to the cognitive environment) hold to the best of the speaker's knowledge. Suppose now that for some reason the satisfaction of the CNC can be questioned. This typically occurs when it has high availability. The conditional sentence no longer conveys a sufficient condition and consequently the conclusion of the argument does not follow any more. This explains the results of the foregoing manipulations. For the sake of simplicity the formula can be rewritten as:

$$\{A_1\} \ \& \ A \ \rightarrow \ C.$$

Formally, from

$$\text{if } (\{A_1\} \ \& \ A) \text{ then } C; \quad A,$$

C follows, whereas from

$$\text{if } (A_1 \ \& \ A) \text{ then } C; \quad A,$$

C does not follow.

Compare two arguments defined by different conditionals such that one has less available CNC's (or disabling conditions in terms of causality) than the other, like *If Joe cut his finger, then it bled* against *If the match was struck, then it lit:* in the first case, the low availability of the CNC's makes it more likely that their satisfaction goes unchallenged than in the second case. This analysis generalises to the non-causal sentences like the 'licensing board' or the 'athlete' scenarios above. In fact, it makes a step towards the formalisation of the concept of credibility of a conditional sentence: once the antecedent and the consequent have been identified as related to each other, the conditional is all the more credible as there are fewer CNC's whose satisfaction is questionable. There are close links between this claim and the classic view that belief in a conditional is measured by the conditional belief of the consequent on the antecedent, and it can

be formally demonstrated that the former is a specification of the latter (Politzer & Bourmaud, 2002).

In the experiments mentioned above, there is an interesting case where the epistemic implicature is reinforced. This is the case of the Evans et al. manipulation mentioned earlier: the speaker of a promise or a threat warrants the satisfaction of CNC's, which he is not in a position to do when uttering a tip or a warning. The difference is one between a warrant "to the best of one's knowledge" and a warrant of full knowledge that renders the conditional more credible.

Finally George's manipulation (mentioned earlier) of the level of credibility of the conditional is another way of questioning the satisfaction of CNC's: by asking to asssume the truth of such conditionals, participants were invited to dismiss CNC's acting as possible objections like *the painter must be famous,* whereas stressing the uncertainty of the statement is a way to invite them to take such objections into account.

Valid arguments and nonmonotonic effects. There are other means of cancelling the implicature and this is what gives rise to nonmonotonic effects to which we now turn. Nonmonotonic deduction is defined by the following property: consider a proposition Q that is deducible from P; Q is not necessarily deducible from the conjunction of P with another proposition R, contrary to the case of classic deduction.

Byrne (1989) asked one control group of participants to solve standard arguments such as, for MP:

If Mary meets her friend, then she will go to a play;
Mary meets her friend;
therefore: (a) *Mary will go to a play;* (b) *Mary will not go to a play;* (c) *Mary may or may not go to a play.*

As is commonly observed, nearly every participant chose option (a). An experimental group was asked to solve the same arguments modified by the addition of a third premise, *if Mary has enough money, then she will go to a play.* The result is that fewer than 40 percent in this group chose option (a) and the others chose option (c). A similar effect was observed with MT. Notice the special structure of the argument: the third (additional) premise was a conditional that had a necessary condition in its antecedent; since it had the same consequent as the major premise, it contained a necessary condition for the consequent of the major premise (in fact, a CNC) and served as a means of introducing it in the context. The result has been replicated many times with rates of non-endorsement varying from one third to two thirds, depending on sentence type and population.

Within the proposed framework, the additional premise raises doubt on the assumption of satisfaction of a CNC in the main conditional. This is made possible by using the CNC in the antecedent of another conditional: in uttering "if Mary has enough money . . " the speaker implicates that she does not know whether or not Mary has enough money, so cancelling the implicature that accompanies the main conditional. This now has decreased credibility and the

conclusion follows with a level of credibility inherited from the premises. This is why in an all-or-none format of response, a majority of people choose option (c).

This explanation has testable consequences. One, by replacing the additional conditional sentence with a categorical sentence that expresses doubt, such as *it is not sure that she has enough money*, it should be possible (i) to simulate the effect (a decrease in the rate of endorsement of the conclusion); (ii) and to bring this rate of endorsement in fact to zero since the doubt stems from an explicit statement and no more from an implicature that may not always be generated. This is precisely what was observed (Politzer, in press). Two, when participants are given a chance to evaluate the conclusion, the proportion who find it doubtful should be about the same as the proportion who chose option (c) above; again this is what was observed.

Another consequence is that it should be possible to manipulate the credibility of the major conditional premise by introducing various degrees of satisfaction of the CNC's and observe correlated degrees of belief in the conclusion. This was tested by Politzer & Bourmaud (2002) who used different MT arguments such as:

If somebody touches an object on display then the alarm is set off;

the alarm was not set off;

therefore: *nobody touched an object on display* (to be evaluated on a five-point scale ranging from certainly true to certainly false).

This was a control; in the three experimental conditions, degrees of credibillity in the conditional were defined by way of an additional premise that provided information on a CNC:

High credibility: *there was no problem with the equipment;*

Low: *there were some problems with the equipment;*

Very low: *the equipment was totally out of order.*

The coefficients of corrrelation between level of credibility and belief in the truth of the conclusion ranged between .48 and .71 and were highly significant. This result supports the proposed theoretical approach all the more as the kind of rule used was not limited to causals but included also means-end, remedial, and decision rules.

Nonmonotonocity is highly difficult to manage by Artificial Intelligence systems because of the necessity of looking for possible exceptions through the entire data base. What I have suggested is some kind of reversal of the "burden of the proof" for human cognition: at least for conditionals (but this could generalise) looking for exceptions is itself an exception because the conditional information comes with an implicit guarantee of normality.

Hypothesis testing

Some people are professionally trained to test their hypotheses; they may be scientists or practitioners such as detectives, medical doctors or technicians specialised in trouble-shooting. But how do lay people behave when they have to put a hypothesis to the test? One of the classic laboratory tasks used to answer this question was designed by Wason (1960). The situation resembles a game played between the experimenter and the participant. The experimenter

chooses a rule to generate sequences of three numbers. The aim of the game for the participant is to discover this rule. In order to do so, the participant can use two kinds of information. The main source of information is the result of tests which he carries out as follows: he submits triples to the experimenter who replies every time by 'yes' (the three numbers obey the rule) or 'no' (they do not). (ii) The second source of information is an initial example of a sequence conforming to the rule provided by the experimenter at the beginning of the game: this sequence is *2, 4, 6.* When the participant thinks he has discovered the rule, he states it; in case he is wrong, the game may continue for another cycle until the rule stated is correct or the participant gives up. The rule which the experimenter follows is *three increasing numbers* (integers). It is usually observed that the majority of participants state at least one incorrect rule and that failure is not uncommon. More strikingly, the incorrect rules proposed by participants often express one of the salient features of the initial exemplar (2, 4, 6), such as *even numbers, increasing by the same interval, or increasing by two* and it seems difficult for them to eliminate such hypotheses. This is especially interesting from a pragmatic point of view because the triple *2, 4, 6* has very salient features; given that it has been specially selected and presented as an instance by the experimenter, participants are thereby invited to assume that its features are relevant; but unluckily for the participant, these features overdetermine the rule (the numbers need not be even, they need not increase by two, etc. in order to follow the rule actually used), so that one can consider that the whole situation is deceptive. As every teacher knows, it is misleading to offer an example of a concept that is too specific. This analysis made on theoretical grounds (Politzer, 1986) has received support from the results of a recent experiment performed by Van der Henst, Rossi, & Schroyens (2002). In their experimental procedure, the *2, 4, 6* instance was not presented to participants as resulting from a deliberate choice made by the experimenter, but rather as the output of a computer program which randomly generated instances of the rule: the authors observed that the erroneous first solutions diminished by one half, and that the mean number of rules proposed as solutions diminished by one third, presumably because the salient features are not presumed to be relevant if they are the result of a random, non-intentional process.

The 2, 4, 6 task is not the only inductive task that deserves pragmatic scrutinising. One of the most extensively investigated tasks in the psychology of reasoning, also due to Wason and also designed to study hypothesis testing behaviour, is the four-card problem (or selection task) in which participants are required to select the information that they think is necessary in order to test whether a conditional rule is true or false. Studies by Sperber, Cara & Girotto (1995) and Girotto, Kemmelmeier, Sperber & van der Henst (2001) show that the task as understood by the experimenter is rather opaque to participants. Ironically, the comprehension mechanisms pre-empt any domain-specific reasoning mechanism, so that the task cannot be considered as one of reasoning in the strict sense.

Studies of probabilistic judgment

There is a huge psychological literature on probabilistic judgment that dates back to the sixties. The conclusion which has been retained, especially among philosophers and economists, is that performance is poor and often reveals irrational judgments. This widely shared opinion is essentially due to the work of Kahneman and Tversky (1982 for an overview). Whether they are right or wrong is not an issue to debate here; instead, it can be argued that their demonstration is often unconvincing because in too many cases they grossly neglected the pragmatic analysis of their experimental paradigms. Two of these tasks, posssibly the most famous ones, the Linda problem and the Lawyer-Engineer problem, will be discussed.

The conjunction fallacy (the Linda problerm). In a typical version of the experimental paradigm, participants are presented with the following description:

> Linda is 31 years old, single, outspoken and very bright; she majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations (Tversky & Kahneman, 1982).

They are then asked to decide which of the following statements is the most probable:

- Linda is a bank teller (B);
- Linda is active in the feminist movement (F);
- Linda is a bank teller and active in the feminist movement (B+F).

Whatever the response format (multiple choice, rank ordering, etc) over 80 percent judge B+F to be more probable than B, in apparent violation of a fundamental axiom of probability theory which requires that the probability of a conjunction be no more probable than that of any one of its conjuncts. The authors take this result as evidence for the use of the representativeness heuristic, that is, an assessment of the degree of correspondence between a model and an outcome: being a 'feminist bank teller' (B+F) is more representative of the description because it has one common feature with the description, which 'bank teller' (B) is lacking. This explanation is appealing if only because of its simplicity but it cannot be accepted before a pragmatic analysis of the task has been made. Now, from this point of view, there are two main problems with the task.

The first problem is that the crucial options have an obvious anomaly: in comparing two items B vs (B and F), there are two permissible construals for B in the first option, viz. an inclusive construal *(B whether or not F)*, and an exclusive one that carries an implicature *(B but not F)*.

The claim that the implicature is licensed by the juxtaposition of the two options was supported by the results of the following manipulation (Politzer & Noveck, 1991; see Dulany and Hilton, 1991 for a similar approach). Keeping constant a scenario that depicted a very brilliant and determined student, two formulations of the options were presented to two experimental groups as follows:

The first group had clearly nested options (and for this reason it was hypothesised that conjunction errors would be less frequent than in a Linda-type control):

1) Daniel entered Medical School.

2) Daniel dropped out of Medical School for lack of interest.

3) Daniel graduated from Medical School.

The second group had the same options, but with the explicit mention of the inclusion structure of the questions introduced by *and,* which was predicted to trigger an implicature attached to option one:

1) Daniel entered Medical School.

2) Daniel entered Medical School and dropped outfor lack of interest.

3) Daniel entered Medical School and graduated.

Indeed, while 77 percent committed the error on the Linda-type control the rate of errors collapsed to 31 percent for the first control, but as predicted it increased significantly to 53 percent for the second control.

The second problem with the task is even more basic; it revolves around the task representation. From a computational point of view, Linda's profile is useless: all the necessary and sufficient logical information is given in the options. But participants normally assume the description to be relevant and one obvious way to satisfy this is to consider the task as a test of one's sociological or psychological skills and the description as a source of information that provides a theme together with the necessary evidence for or against the answer to a question (the possibility that Linda is a feminist): the *and-not* interpretation of option (B) is then constrained.

This point is important in relation with the between-subject task. In this variant of the task, only one statement is presented: B to one group, and B+F to the other, and participants are asked to estimate the probability of the statement. As B+F is rated as more probable than B, many investigators have been convinced in favour of the representativeness theory. But what this demonstrates is only that participants are enclined to try to render the description relevant to the question asked: they identify the kind of activity which provides greater relevance to the description of the character and like when one has to imagine what could be the best end of a story, it does not have to be the most probable event - rather, it generally is not.

The base rate fallacy (the Lawyer-Engineer problem). In this paradigm, participants are told that a panel of psychologists has written personality descriptions of 30 engineers and 70 lawyers (the associated proportions provide what is called the *base rates).* A description that is assumed to have been chosen at random and that coincides with the stereotype of an engineer is presented; one group of participants is asked to estimate the probability that the person described is an engineer; another group is asked to do the same based on the reversed base rates: 70 engineers and 30 lawyers. Provided some technical assomptions are satisfied, standard probability theory requires that the estimate given by the first group should be lower than that of the second group. The first study reported by Tversky and Kahneman (1973) showed no difference, hence the widely held belief that 'people are insensitive to the base rates'; however, more recent studies have shown that people do take base rates into account, although "not

optimally or even consistently" (Koehler, 1996). Tversky & Kahneman's explanation for their results is again based on the representative heuristic: people would base their judgment exclusively on the extent to which the description fits the stereotype. This explanation is again problematic because it does not take into account the participants' representation of the task. In a recent series of experiments (Politzer & Macchi, in press) it was hypothesised that people view the task as a request to exploit a psychological description that is assumed to be relevant. If that is the case, the neglect of base rates should be relative and could be suppressed in an experimental condition where no psychological description is provided but instead the psychological characterisation is provided in a single statement to the effect that the person's description is typical of an engineer: in this way, the outcome is available (in order to let the representativeness heuristic operate, if at all) but the details are missing in order to suppress the interpretation of the task as one of extraction of a psychological profile from such data. In being told that the description is typical, these participants receive a near answer to the question, which makes it lack relevance; consequently, they reinterpret the question as a request for an unconditioned probability, which enables them to render both the statement of typicality and the base rate information relevant and to fulfill the task, so that most of them should give the base rate as their response. This is what was observed (85 percent used the base rate exclusively while the rate of its use in a control group was 17 percent). It seems therefore that the paradigm could be better described as showing that people have difficulty in combining information from two sources, the base rates and the individuating information, and that they focus on the one that maximises relevance. Previous research has shown that when the psychological description is uninformative (that is, completely non diagnostic between the engineer and the lawyer stereotypes), they rely entirely on the base rates.

Class inclusion and categorisation

Class inclusion in children. One of the most thoroughly investigated paradigms in developmental psychology during the period that goes from the sixties to the eighties, and which nowadays is still subject to debate is class-inclusion, initially created by Piaget (Piaget & Szeminska, 1941; Piaget & Inhelder, 1959). In a typical experiment, the child is presented with the picture of five daisies and three tulips, and then asked, "Are there more daisies or more flowers ?" The rate of what is considered the correct response, "more flowers", reaches the 50% value only around 8 or 9 years of age. This highly robust result is puzzling given the well-documented precocity in the acquisition of lexical hierarchies. We will consider in turn the interpretation of the interrogative sentence and the representation of the task.

First, the microanalysis indicates that the relation of hyperonymy-hyponymy between *flower* and *tulip/daisy* licenses the use of *flower* to refer to either all the flowers or a subclass of them. Indeed, it can be demonstrated that in the experimental setting, *flower* is indeterminate between an inclusive sense (all the flowers) and an exclusive sense (tulip). This was done as

follows. Two groups of 6- and 7-year-old children were presented with the picture. The control group was just asked (i) to first point to the flowers, (ii) and then to the daisies; in contrast, the experimental group was asked the same questions in the reversed order. Whereas in the control group 90% of the children asked to point to the flowers pointed to all the flowers, in the experimental group half of the children pointed to all the flowers and the other half pointed to the tulips. This demonstrates that *flower* apparently had become completely indeterminate in the context of *daisy*. Half of the children decided that *flower* must refer to the flowers that are not daisies presumably because the word *daisy* had just been used; the other half were not able or not willing to make this decision.

Consequently, the standard class-inclusion question is ambiguous because the lexeme *flower* can receive either its inclusive/hyperonym or its exclusive/hyponym interpretation. It follows that many children may compare the daisies with the tulips (which is well documented), a comparison that is not intended by the experimenter though semantically permitted, and pragmatically justifiable under one representation of the task as we will see shortly.

If this explanation is correct, it should be possible to enhance performance by disambiguating the question. This was done in another experiment that used a double disambiguation procedure. Firstly, 5- to 8-year-old children were requested to "point to the flowers" and then to "point to the daisies" (as in the previous experiment). Secondly, they were asked a modified class-inclusion question in which all three terms appeared: "Are there more tulips, or more daisies, or more flowers?" The 5-year-olds reached the 50% rate of success (control: 6%) and the 7- and 8-year-olds were very close to the 100% rate (control: 30%). Two other experiments showed that each disambiguating procedure is effective separately but less than in combination. In brief, the disambiguation of the question has revealed that children acquire inclusion three to four years earlier than previously claimed.

But still a major question remains to be answered: Why do children change their response to the standard question when they are about 8 or 9 years old? The answer is that the younger choose the exclusive interpretation of *flowers* (tulips) and the older the inclusive interpretation (all the flowers). But again, why ? This question leads us to the macroanalysis and the representation of the task. So long as the child attributes to the experimenter an interest in knowing whether he can count (one of the great achievements during that period) the relevant comparison is between the tulips and the daisies (this response is likely to produce the more cognitive effects: you will know that I know how to count). But when the child has progressed enough in the development of metacognitive skills such as logical necessity (Cormier & Dagenais, 1983; Miller, Custer & Nassau, 2000) and awareness of semantic ambiguities (Gombert, 1990) he can attribute to the experimenter an interest in these abilities, and the relevant comparison shifts to comparing all the flowers and the daisies, which yields the "correct" response. In brief, this overview of an old paradigm in the study of logical development shows once again that the verbal material and the speaker/experimenter - hearer/participant relationship must be pragmatically scrutinized.

Categorisation: mathematical hierarchies. Although the approach taken here is focused on laboratory tasks, the analysis that has been proposed can help identify some sources of difficulty in learning mathematical concepts; more specifically the application of the foregoing analysis of the inclusion question to lexical hierarchies reveals a tension between the use that is made of them by the lay person/student on one hand and the scientist/teacher on the other hand.

We noticed earlier that the standard class-inclusion question is ambiguous because the lexical unit *flower* can receive either its inclusive/hyperonym interpretation or its exclusive/hyponym interpretation. This case is reminiscent of markedness: opting for the inclusive rather than the exclusive meaning amounts to opting for an unmarked rather than a marked interpretation. This is at the basis of riddles such as *"What animal barks but is not a dog?"* the solution of which is blocked if *dog* is interpreted as unmarked but transparent if *dog* is interpreted as contrasting with *bitch.* (This ambiguity is sometimes referred to as *privative).* Now, as far as mathematical hierarchies are concerned, the speaker's freedom to use an ambiguous lexical unit is constrained by the register of the communication. In daily life, it seems that the items on such hierarchies are essentially used exclusively; for instance, *square* contrasts with *rectangle,* (which in turn contrasts with *parallellogram,* etc.), which means that for a "naive" person no square is a rectangle. On the contrary, in the mathematical vocabulary, items on the same hierarchy are used inclusively: a square is a special rectangle (which in turn is a special parallellogram, etc.); hence technically all squares are rectangles. Similarly, for the layman integers are not decimal numbers although mathematically they are. In brief, whenever two items are compared, the subset-to-set relations generated by the folk hierarchy and the mathematical hierarchy are logical contraries (note 1). It follows that a crucial difficulty in the learning of these classifications lies in the student's capability to shift appropriately from his familiar classification to the technical one (Politzer, 1991). The cognitive difficulty is illustrated in Figures 1a and 1b which show both hierarchies for elementary geometry.

[Insert Figures 1a and 1b about here]

Conclusion

From a methodological point of view, the experimental study of thinking is among the most difficult in cognitive psychology to carry out. This is the area where the representation of the task interferes the most with the thought process under study, to such an extent that the task may be devoid of validity if no precaution is taken. It has been argued that precautionary measures should include two kinds of analysis. One, that has been called *macroanalysis,* aims to determine the task representation, that is the participant/student's attributions to the experimenter/teacher about the latter's expectations regarding the former's knowledge or performance. This is based on the content of each task, taking into account the specificity of the relationship between experimenter/teacher and participant/student which creates a special element of pretense in their communication. The other, that has been called *microanalysis,* takes into

account the result of the first and aims to determine the disambiguations, referential assignments and implicatures which the participant/student works out on the way to his final interpretation of the premises, questions, problem statement and the like. When such analyses yield interpretations that are at variance with the experimenter's intended meaning it is possible to write up an alternative formulation or to design an alternative task whose validity is no more questionable and to compare performance on this new task with the initial one. In the past, many unwarranted conclusions have been drawn from participants' seemingly poor performance in terms of human irrationality. The experimental method that compares initial and modified materials on the basis of pragmatic theory plays a crucial role to redress the balance.

References

Begg, I. & Harris, G. (1982). On the interpretation of syllogisms. *Journal of Verbal Learning and Verbal Behavior, 21,* 595-620.

Byrne, R. M. J. (1989). Supressing valid inferences with conditionals. *Cognition, 31,* 61-83.

Cormier, P. , & Dagenais, Y. (1983). Class-inclusion developmental levels and logical necessity. *International Journal of Behavioral Development, 6,* 1-14.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory and Cognition, 23,* 646-658.

Cummins, D. D. , Lubart, T. , Alksnis, O. , & Rist, R. (1991). Conditional reasoning and causation. *Memory and Cognition, 19,* 274-282.

Ducrot, O. (1971). L'expression en français de la notion de condition suffisante. *Langue Française, 12,* 60-67.

Dulany, D. E. , & and Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition, 9,* 85-110.

Evans, J. St. B. T. , Newstead, S. E. , & Byrne, R. M. J. (1993). *Human reasoning. The psychology of deduction.* Hove: Lawrence Erlbaum.

Evans, J. St. B. T. , & Twyman-Musgrove, J. (1998). Conditional reasoning with inducements and advice. *Cognition, 69,* B11-B16.

Fillenbaum, S. (1975). IF: Some uses. *Psychological Research, 37,* 245-260.

Fillenbaum, S. (1978). How to do some things with *if.* In J. W. Cotton & R. L. Klatzky (Eds.), *Semantic factors in cognition* (pp. 169-214). Hillsdale, N. J. : Lawrence Erlbaum.

Geis, M. L. , & Zwicky, A. M. (1971). On invited inferences. *Linguistic Inquiry, 2,* 561-566.

George, C. (1995). The endorsement of the premises: Assumption-based or belief-based reasoning. *British Journal of Psychology, 86,* 93-111.

Girotto, V. , Kemmelmeier, M. , Sperber, D. , & van der Henst, J.-B. (2001). Inept reasoners or pragmatic virtuosos? Relevance and the deontic selection task. *Cognition, 81,* B69-B76.

Gombert, E. (1990). *Le développement métalinguistique.* Paris: Presses Universitaires de France. [English translation: *Metalinguistic development.* University of Chicago Press, 1992].

Hilton, D. J. (1995). The social context of reasoning : Conversational inference and rational judgment. *Psychological Bulletin, 118,* 248-271.

Horn, L. R. (1972). *On the semantic properties of logical operators in English.* Bloomington: Indiana University Linguistics Club.

Horn, L. R. (2000). From *if* to *iff* : Conditional perfection as pragmatic strengthening. *Journal of Pragmatics, 32,* 289-386.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237-251.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*, 1-17

Mackie, J. L. (1974). *The cement of the universe.* Oxford University Press.

Manktelow, K. I. , & Fairley, N. (2000). Superordinate principles in reasoning with causal and deontic conditionals. *Thinking and Reasoning, 6,* 41-65.

Markovits, H. (1985). Incorrect conditional reasoning among adults: Competence or performance? *British Journal of Psychology, 76,* 241-247. Ò

Miller, S. A. , Custer, W. L. , & Nassau, G. (2000). Children's understanding of the necessity of logically necessary truths. *Cognitive Development, 15,* 383-403.

Newstead, S. E. , Ellis, M. C. , Evans, J. St.B. T. , & Dennis, I. (1997). Conditional reasoning with realistic material. *Thinking and Reasoning, 3,* 49-76.

Newstead, S. E. , & Griggs, R. A. (1983). Drawing inferences from quantified statements: A study of the square of opposition. *Journal of Verbal Learning and Verbal Behavior, 22,* 535-546.

Orne, M. T. (1962). On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *American Psychologist, 17,* 776-783.

Piaget, J. , & Inhelder, B. (1959). *La genèse des structures logiques élémentaires.* Neuchâtel: Delachaux et Niestlé. [English translation: *The early growth of logic in the child: classification and seriation.* London: Routledge & Kegan Paul, 1964].

Piaget, J. , & Szeminska, A. (1941). *La genèse du nombre chez l'enfant.* Neuchâtel: Delachaux et Niestlé. [English translation: *The child's conception of number*. London: Routledge & Kegan Paul, 1952].

Politzer, G. (1981). Differences in interpretation of implication. *American Journal of Psychology, 94,* 461-477.

Politzer, G. (1986). Laws of language use and formal logic. *Journal of Psycholinguistic Research, 15,* 47-92.

Politzer, G. (1990). Immediate deduction between quantified sentences. In K. J. Gilhooly, M.T.G. Keane, R.H. Logie, & G. Erdos (Eds.), *Lines of thinking: Reflections on the psychology of thought* (pp. 85-97). Vol 1. London: John Wiley.

Politzer, G. (1991). L'informativité des énoncés: contraintes sur le jugement et le raisonnement. *Intellectica, 11,* 111-147.

Politzer, G. (1993). *La psychologie du raisonnement: Lois de la pragmatique et logique formelle.* [The psychology of reasoning: laws of pragmatics and formal logic. Ph. D. thesis. University of Paris VIII].

Politzer, G. (2003). Premise interpretation in conditional reasoning. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment, and decision making* (pp. 79-93). London: Wiley.

Politzer, G. (in press). Uncertainty and the suppression of inferences. *Thinking and Reasoning.*

Politzer, G. , & Bourmaud, G. (2002). Deductive reasoning from uncertain premises. *British Journal of Psychology, 93,* 345-381.

Politzer, G. & Macchi, L. (in press). The representation of the task : The case of the lawyer-engineer problem in probability judgment. In V. Girotto & P. N. Johnson-Laird (Eds.), *The shape of reason : Essays in honor of P. Legrenzi.* Hove : Psychology Press.

Politzer, G. & Noveck, I. (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research, 20,* 83-103.

Rips, R. J. , & Marcus, S. L. (1977). Suppositions and the analysis of conditional sentences. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 185-220). Hillsdale, N.J. : Lawrence Erlbaum.

Rumain, B. , Connell, J. , & Braine, M. D. S. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: If is not the biconditional. *Developmental Psychology, 19,* 471-481.

Schwarz, N. (1996). *Cognition and communication.* Mahwah, N.J. : Lawrence Erlbaum.

Sperber, D. , Cara, F. , & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 52,* 3-39.

Staudenmayer, H. (1975). Understanding conditional reasoning with meaningful propositions. In R.J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults.* (pp. 55-79). Hillsdale, N. J. : Lawrence Erlbaum.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition, 22,* 742-758.

Thompson, V. A. (1995). Conditional reasoning: The necessary and sufficient conditions. *Canadian Journal of Experimental Psychology, 49,* 1-60.

Tversky, A. , & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84-100). Cambridge: Cambridge University Press.

Van der Auwera, J. (1997). Conditional perfection. In A. Athanasiadou & R. Dirven (Eds.), *On conditionals again* (pp. 169-190). Amsterdam: John Benjamins.

Van der Henst, J.-B. , Rossi, S. , & Schroyens, W. (2002). When participants are not misled they are not so bad after all : A pragmatic analysis of a rule discovery task (manuscript in preparation).

Footnote

1. One might argue that this phenomenon is but a particular case of scalar phenomenon, by which the use of *rectangle* on the scale implicates *not square*, a higher item on the scale (Horn, 1972). However, while it is easy to imagine or observe in daily life utterances that exhibit literal meaning on various scales (quantifiers, modals, frequency terms, etc.) it seems debatable that this happens with mathematical classifications. Whether *rectangle* can refer to a square in a non mathematical context is an open question that could be answered empirically. In the absence of evidence to the contrary, it is assumed that there is no lexical unit in ordinary English to refer to the set of figures that conjoins the squares and the rectangles. That it is so is understandable: in daily life, it is the exclusive contrast that is useful; the inclusive contrast has only a metacognitiven theoretical interest, which justifies its scholar use.

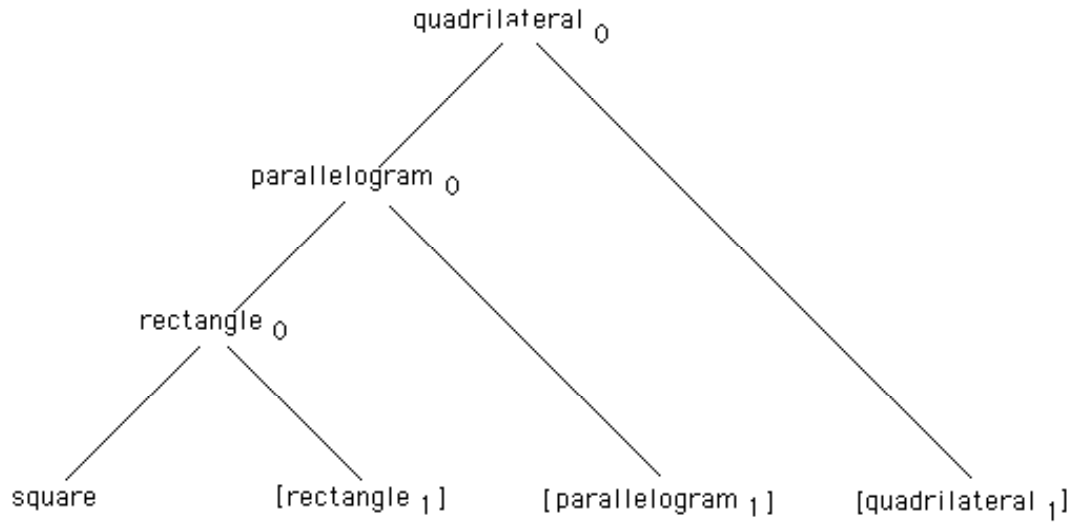

Figure 1a: The scientific hierarchy

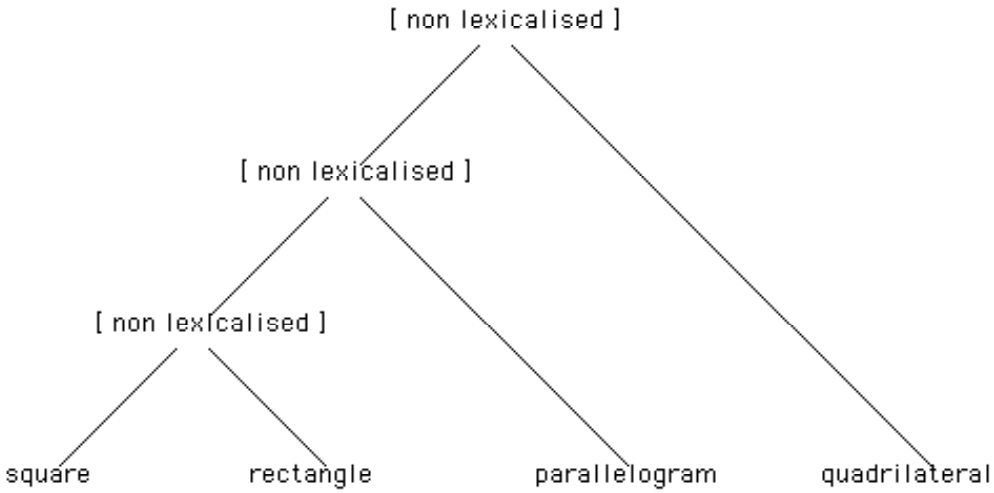

Figure 1b: The naïve hierarchy