COSC462 Lecture 3: Expressiveness of languages

Willem Labuschagne
University of Otago

Abstract

We continue our exploration of the class of opaque propositional languages by discussing some positive and negative results about expressiveness, and introduce the object-meta distinction.

1 Ineffable information

In the previous lecture we met some properties of opaque propositional languages, for example the property that classical entailment is transitive (if $\alpha \vDash \beta$ and $\beta \vDash \gamma$ then $\alpha \vDash \gamma$). Today we discuss more of these 'metatheoretical' properties, and explain the prefix 'meta'. This lecture focuses on expressiveness, i.e. on the general question of whether an agent's knowledge representation language would allow her to put all the information at her disposal into words.

To avoid complications, we assume during the rest of this lecture that the ontology for every language L_A is (S, V) with $S = W_A$ and V the identity function on W_A, just as in the case of the Light-Fan System. Thus we may speak of states and valuations interchangeably.

There is an important difference between languages that are finitely generated and those with an infinite set A of atoms. To appreciate the difference, we should first review the finite case.

Suppose L_A has a finite set $A = \{p_0, \ldots, p_{n-1}\}$ of n atoms. Then L_A has a finite set W_A of 2^n different valuations (since each valuation is obtained by deciding, for each of the n atoms in turn, whether it is true or false). The information possessed by an agent can be represented semantically by a subset of W_A, corresponding to the states that have been ruled out. There are a finite number of subsets of W_A, in fact exactly 2^{2^n} subsets (because each subset is obtained by deciding, for each valuation in turn, whether it is in the subset or not). The information

represented by a set of excluded states is expressed by a sentence if the sentence has exactly those excluded states as its nonmodels. When A is finite, it is always possible to find such a sentence, and we shall prove this before going on to consider the infinite case.

Definition 1 ***(Normal form)*** *A sentence α is in **disjunctive normal form** if α is of the form $\gamma_1 \vee \ldots \vee \gamma_m$ for some m, where each **disjunct** γ_i is itself a conjunction of the form $\beta_{i1} \wedge \ldots \wedge \beta_{ik}$ for some k, and where every **conjunct** β_{ij} is either an atom or the negation of an atom.*

*If the language has a finite set $A = \{p_0, p_1, \ldots, p_{n-1}\}$ of atoms, then a sentence β is a **state description** iff $\beta = \beta_0 \wedge \beta_1 \wedge \ldots \wedge \beta_{n-1}$ where each β_i is either p_i or $\neg p_i$.*

*If $A = \{p_0, \ldots, p_{n-1}\}$ then a sentence α is in **strong disjunctive normal form (SDNF)** if α is in disjunctive normal form and every disjunct is a state description (i.e. every atom in A appears exactly once in every disjunct γ_i).*

The idea is simpler than the definition may sound. For example, if $A = \{p_0, p_1, p_2\}$ then $p_0 \wedge \neg p_1 \wedge p_2$ is a state description. A sentence in SDNF is just a disjunction of one or more state descriptions (which is why we may read SDNF as 'state description normal form' if we wish).

Theorem 2 *Every sentence α of L_A is equivalent to a sentence in disjunctive normal form.*

If A is finite and α is satisfiable, then α is equivalent to a sentence in SDNF.

Proof. *We build the normal form sentence by looking at $\mathcal{M}(\alpha)$.*

If $\mathcal{M}(\alpha) = \varnothing$ (the empty set), then no state satisfies α, and so we choose as the equivalent sentence a contradiction, say $p_0 \wedge \neg p_0$. Note that the contradiction is in disjunctive normal form, although it happens to consist of a single disjunct.

If $\mathcal{M}(\alpha) \neq \varnothing$, suppose the atoms occurring in α are among $\{p_0, \ldots, p_k\}$. (After all, only a finite number of atoms can go into the construction of α.) If A itself is finite, we may set $\{p_0, \ldots, p_k\} = A$.

We don't care whether states in $\mathcal{M}(\alpha)$ satisfy atoms outside $\{p_0, \ldots, p_k\}$. What we care about is the finite number of different ways (say m) in which the valuations corresponding to states in $\mathcal{M}(\alpha)$ can assign truth values to the atoms in $\{p_0, \ldots, p_k\}$. In other words, we want to go from the states in $\mathcal{M}(\alpha)$ to their associated valuations in W_A and see what these valuations do to the atoms in $\{p_0, \ldots, p_k\}$.

Let $(x_0, x_1, \ldots, x_k)$ be what one of these valuations does to the atoms. In other words, pick some state that satisfies α, find the valuation assigned to that state by the labelling function V, and list the truth values assigned to the atoms $p_0, \ldots, p_k$ by this valuation.

Now we form a conjunction that imitates the way in which the valuation worked, namely the sentence $\gamma_1 = \beta_0 \wedge \ldots \wedge \beta_k$ where $\beta_i = p_i$ if $x_i = 1$ and $\beta_i = \neg p_i$ if $x_i = 0$.

We repeat this for each way of allocating truth values to $p_0, \ldots, p_k$ that makes α true, thereby producing conjunctions $\gamma_1, \gamma_2, \ldots, \gamma_m$.

Finally we claim that $\alpha \equiv \gamma_1 \vee \ldots \vee \gamma_m$. To see that this must be the case, note that every state that satisfies α must also satisfy one of the γ_i. Conversely, every state that satisfies one of the γ_i must have an associated valuation allocating truth values to $p_0, \ldots, p_k$ in one of the m ways that satisfy α. ■

In the proof we said that we care only about the truth values allocated by a valuation to the atoms actually occurring in α, and would ignore the truth values allocated to all other atoms. In effect we were claiming that the truth value a sentence α gets (relative to a state) depends only on the truth values the atoms in α get, relative to that state. This follows from the lemma below.

Lemma 3 *Let α be a sentence of L_A and suppose $s, s' \in S$ are states that agree on the atoms occurring in α. In other words, if $B \subseteq A$ is the set of atoms occurring in α, then we suppose that s satisfies p iff s' satisfies p, for every $p \in B$.*

Then s satisfies α iff s' satisfies α.

Proof. *Given any sentence α, let B be the set of atoms occurring in α.*

We use induction on the number of steps in which a sentence can be built up from the atoms in B, and show that for every such sentence, it will be satisfied either by both of s and s' or by neither of them.

A sentence built up in 0 steps must itself be an atom in B, say p. By assumption, s and s' both satisfy p or neither satisfy p.

(Induction hypothesis) Assume that if β is a sentence built up from the atoms in B in k or fewer steps, then either s and s' will both satisfy β or else neither will satisfy β.

Now consider any sentence β built up in $k+1$ steps from atoms in B. There are 5 possible cases.

$\beta = \neg\varphi$: State s will satisfy β iff s fails to satisfy φ. Since φ is built up in k steps or fewer, s fails to satisfy φ if and only if s' fails to satisfy φ (by the induction hypothesis), which in turn is the case if and only if s' satisfies β.

$\beta = \varphi \wedge \psi$: State s will satisfy β iff s satisfies φ and satisfies ψ. Since φ and ψ are each built up in k steps or fewer, s satisfies them both if and only if s' satisfies φ and satisfies ψ (by the induction hypothesis), which in turn is the case if and only if s' satisfies β.

$\beta = \varphi \vee \psi$: States s will satisfy β iff s satisfies at least one of φ and ψ. Since φ and ψ are each built up in k steps or fewer, s satisfies at least one of φ and ψ if and only if s' satisfies at least one of φ and ψ (by the induction hypothesis), which in turn is the case if and only if s' satisfies β.

$\beta = \varphi \rightarrow \psi$: State s will satisfy β iff s satisfies ψ or fails to satisfy φ. Since φ and ψ are each built up in k steps or fewer, s satisfies ψ or fails to satisfy φ if and only if s' satisfies ψ or fails to satisfy φ (by the induction hypothesis), which in turn is the case if and only if s' satisfies β.

$\beta = \varphi \leftrightarrow \psi$: State s will satisfy β iff s satisfies both φ and ψ or satisfies neither of them. Since φ and ψ are each built up in k steps or fewer, s satisfies both φ and ψ, or satisfies neither, if and only if s' satisfies both φ and ψ or satisfies neither of them, which in turn is the case if and only if s' satisfies β.

Now since α is built up from the atoms in B by some finite number of steps, the result follows. ■

In passing we may remark that disjunctive normal form is not the only normal form — we introduce *conjunctive normal form* in the exercises.

Now let us pause to reflect on the significance of the theorem. It actually tells us more than merely that we can rewrite any sentence in disjunctive normal form. Instead of starting with the set $\mathcal{M}(\alpha)$, we could start with any set $X \subseteq W_A$ of valuations — say, the complement of some set of excluded valuations. As long as we are able to restrict consideration to some finite set of atoms, we will be able to duplicate the construction in the proof to arrive at a sentence in disjunctive normal form which is satisfied by exactly the (states associated with the) valuations in X. Since we no longer have a starting sentence α to serve as focus, the only way to justify restricting attention to a finite set of atoms is to have a set A that is itself finite. So we get:

Corollary 4 *If A is finite and if $\overline{X}$ is any set of excluded valuations with complementary set X of included valuations, then there is a sentence $\gamma_1 \vee \ldots \vee \gamma_m$ in SDNF such that $s \in \mathcal{M}(\gamma_1 \vee \ldots \vee \gamma_m)$ iff $V(s) \in X$.*

Thus, if A is finite then any information (reflected by a division of S into two subsets) can be expressed by a sentence of the language. Thus if A is finite then the agent can always say what she thinks.

The infinite case is different — if A is infinite, there exist subsets C for which no sentence α can be found such that $\mathcal{M}(\alpha) = C$, and an agent may therefore have information that cannot be expressed in the knowledge representation language.

You may wonder whether the problem is just that we are restricting ourselves to a single sentence α, and would disappear if we allowed ourselves a set of sentences, say Γ. After all, we have defined what we mean by the set of models of a set Γ of sentences: $\mathcal{M}(\Gamma) = \{v \mid v \in \mathcal{M}(\gamma)$ *for all* $\gamma \in \Gamma\}$. Well, it turns out not to matter whether we stick to a single sentence or allow ourselves the freedom of taking a set of sentences — there is still some information in the infinite case that cannot be expressed declaratively (i.e. that cannot be expressed in the knowledge representation language). In the field of Artificial Intelligence, there is a well-known distinction between *procedural* knowledge and *declarative* knowledge, which is closely related to the distinction between iconic representations and symbolic representations. One may know how to ride a bicycle or hit a tennis shot, but be unable to explain to anyone how to do it. "Knowing how" means that an agent has an iconic representation of something (like a feeling in the body that when the bicycle's handlebars turn the body must lean inwards). Being unable to explain it means being unable to transform the iconic representation into a symbolic representation. In logic, we try to keep things simpler than in psychology, so that the iconic representations are states (or sets of states) and the symbolic representations are sentences (or sets of sentences).

This brings us to the first limitative theorem of logic, which I like to call the Ineffability Theorem. It is possible to prove this by a general cardinality argument, but we shall give a more direct and satisfying proof that actually exhibits a set C which is 'ineffable'.

Definition 5 ***(Ineffability)*** *A subset $C \subseteq S$ is called ineffable if there is no set Γ of sentences of L_A such that $\mathcal{M}(\Gamma) = C$.*

Theorem 6 ***(Ineffability Theorem)*** *Let $A = \{p_0, p_1, \ldots\}$. Then there exists an ineffable set C of states/valuations.*

Proof. *Pick any valuation $w \in W_A$. For example, you may take w to be the valuation such that $w(p_i) = 1$ for all $p_i \in A$.*

Let C be all the remaining valuations, i.e. $C = \overline{\{w\}}$.

We claim that there is no set Γ of sentences of L_A such that $\mathcal{M}(\Gamma) = C$, where $\mathcal{M}(\Gamma)$ is the set of all valuations each of which satisfies all the sentences in Γ. For if there were such a Γ, it would have to contain at least one sentence which is satisfied by every valuation in C but not satisfied by w.

Take any sentence $\gamma \in L_A$. We show that if γ is satisfied by all the valuations in C then γ is satisfied by w as well.

Suppose γ is satisfied by all the valuations in C.

Let $\{p_0, \ldots, p_k\}$ be such that every atom in γ is among $p_0, \ldots, p_k$.

Let $v \in C$ be the valuation such that $v(p_i) = w(p_i)$ for all $i \leq k$ and $v(p_i) \neq w(p_i)$ for all $i > k$.

Now v satisfies γ iff w satisfies γ (because the truth value a sentence gets is determined by the truth values assigned to the atoms actually occurring in that sentence, not the truth values assigned to other atoms — recall the lemma above).

And so we are faced by two choices: either v does not satisfy γ, contradicting our choice of γ, or else w also satisfies γ.

Thus there is no sentence satisfied only by the valuations in C. ■

The Ineffability Theorem was first proved in Brink C and Heidema J (1989): A verisimilar ordering of propositional theories: The infinite case, Technical Report TR-ARP-1/89 (Technical Report Series of the Automated Reasoning Project, Research School of Social Sciences) Canberra: Australian National University. A more general account of related matters may be found in Peppas P, Foo N and Williams M-A (1992): On the expressibility of propositions, *Logique et Analyse* 139-140:251-272.

You may care to reflect on this theorem and its broader significance. Some say a picture is worth a thousand words. The theorem says that some pictures can't be put into words at all, at least not completely, because there may be an infinite number of facts about the picture that one could be interested in. As the cognitive scientist Stevan Harnad puts it (in a chapter called "Category induction and representation" in Harnad S (editor): *Categorical Perception: The Groundwork of Cognition*, pp535-565, Cambridge University Press 1987):

> Words obviously fall short when they leave out some critical feature that would be necessary to sort some future or potential anomalous instance; but even if one supposes that every critical feature anyone would ever care to mention has been mentioned, a description will always remain essentially incomplete in the following ways:
>
> (a) A description cannot convey the qualitative nature of the object being described (i.e. it cannot yield knowledge by acquaintance), although it can converge on it as closely as the describer's descriptive resources and resourcefulness allow. (Critical here will be the prior repertoire of direct experiences and atomic labels on which the descriptions can draw.)

(b) There will always remain inherent features of the object that will require further discourse to point out; an example would be a scene that one had neglected to mention was composed of a prime number of distinct colors.

(c) In the same vein, there would be all the known and yet-to-be-discovered properties of the prime numbers that one could speak of — all of them entailed by the properties of the picture, all of them candidates (albeit far-fetched ones) for further discourse "about" the picture.

(d) Finally, and most revealingly, there are the inexhaustible shortcomings of words exemplified by all the iterative afterthoughts made possible by, say, negation: for example, "the number of limbs is *not* two [...]" The truth of all these potential descriptions is inherent in the picture, yet it is obvious that no exhaustive description would be possible. Hence all descriptions will only approximate a true, complete "description".

The point this quotation attempts to make is that in connection with any iconic representation — any picture, say — there are potentially an infinite number of facts that might be of interest to an agent. This provides for so many ideas the agent may potentially wish to express that even a large knowledge representation language with infinitely many atoms, one for each fact of interest, can't express them all.

On the other hand, let us carefully remind ourselves of the key assumption made by the Ineffability Theorem — that A is infinite. As long as we are interested only in a specific finite number of basic facts about the system (which would lead us to build a knowledge representation language with a finite number of atoms each expressing one of those basic facts), we will indeed be able to express our ideas in L_A. It is only when we are unwilling to place such a finite limit on the facts we regard as being relevant that our information may outstrip the available knowledge representation language, even if that language has been equipped with infinitely many atoms and the agent uses infinite sets of sentences.

Exercise 7 *1. Consider the Light-Fan System with $S = \{11, 10, 01, 00\}$ and $A = \{p, q\}$. For each of the following sets of states, give a sentence in SDNF whose set of models coincides with it:*

- $\{11\}$
- $\{11, 10\}$
- $\{11, 10, 01\}$

- $\{11, 10, 01, 00\}$

2. *Let us say that a sentence α is in **conjunctive** normal form if $\alpha = \gamma_1 \wedge \ldots \wedge \gamma_m$ for some m, where each conjunct γ_i is of the form $\beta_0 \vee \ldots \vee \beta_k$ for some k, and where every β_i is either an atom or the negation of an atom.*

   Suppose $A = \{p_0, p_1, p_2\}$. For each of the following, give two equivalent sentences, one in disjunctive normal form and the other in conjunctive normal form:

   - p_0
   - $p_0 \vee p_1$
   - $p_1 \wedge p_2$
   - $p_0 \leftrightarrow p_2$
   - $((\neg p_0) \vee p_1) \rightarrow p_2$
   - $(p_0 \rightarrow p_1) \rightarrow p_2$

3. *We know that the Ineffability Theorem does not hold for finitely generated languages. Take $A = \{p_0, \ldots, p_n\}$. Trace through the proof of the Ineffability Theorem and find out where it breaks down.*

4. *Suppose the agent is a mathematician contemplating the set of natural numbers $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$. And suppose she is interested in whether a number is prime or not. As knowledge representation language take L_A where $A = \{p_0, p_1, p_2, \ldots\}$ is the infinite set of atoms in which p_i says "i is prime". Of course, in the 'actual state' of the system, some of these atoms will be true and other false. But for the moment focus on something else. Take $S = W_A$, the set of all valuations, and let V be the identity function on S. Give an example of an idea the agent would be unable to express in L_A. Prove that the idea cannot be expressed in L_A.*

   (Hint: Try "Not all the natural numbers are prime".)

2 Metalanguage

We have described, in English augmented with some mathematical symbols, the construction of propositional languages L_A and we have investigated some properties of such languages. The formal language L_A is called the *object-language* while the language in which we talk about L_A, namely English + math, is called the *metalanguage*. The meta-object

distinction is an important one, and we pause to reflect on it. Why should there be a distinction at all?

Think informally about the problem of deciding whether an English sentence is true or false. Some English sentences don't qualify to get truth values (for example: questions or commands). The sentences that do get a truth value are declarative, *i.e.* are sentences that describe the way things are, like 'The fan is on'. But it turns out that some declarative sentences resist our attempts to decide whether they are true or false. The classic example is: 'This sentence is false'.

What is this sentence talking about? Itself, apparently. Is it giving an accurate description of the way things are? Well, that's not so easy to decide. If the description is accurate, then the sentence itself must be false, which is another way of saying that it is not accurate. On the other hand, if the description is not accurate, then the sentence can't be true, so it must be false, which means it is indeed accurately describing itself. Either way we have a contradiction. And the *Liar Paradox* is not the only example of this phenomenon.

Is there some way to pick out the declarative sentences to which truth values can be assigned without the above difficulties? Our first thought is to single out those sentences that do not refer to themselves. However, Alfred Tarski showed that this doesn't do the trick. He devised a version of the Liar Paradox in which sentences refer to one another but no sentence refers to itself. Imagine a book. It needn't be a very thick book, as long as it has at least two pages. Imagine that page 1 has written on it only the sentence: 'The sentence on page 2 is true.' And imagine that page 2 contains only the sentence 'The sentence on page 1 is false.'

Is it clear that we have a version of the Liar Paradox? If the sentence on page 1 is true, then what it says about the sentence on page 2 must be accurate, and so the sentence on page 2 must indeed be true, and so the claim made by the sentence on page 2 about the sentence on page 1 must be accurate, and thus the sentence on page 1 must be false! Similarly, if the sentence on page 1 is false, then it follows that it must be true. This is precisely the sort of double contradiction characteristic of the Liar Paradox.

Let us summarise the position. Not all English sentences can be provided with truth values. There is no obvious way to pick out the sentences of English which can. And any collection of English sentences in which it is possible to talk about the truth and falsity of all the sentences in the collection will inevitably permit versions of the Liar Paradox. So how can we get a grip on the question of which English sentences it makes sense to regard as either true or false?

Tarski's idea is that we should not be too ambitious. Forget about characterising in one fell swoop the collection of all those declarative sentences that can safely bear truth values. Start small instead, and try to characterise some sets of declarative sentences that are safe (*i.e.* won't generate paradoxes). One way to get a safe fragment of English is to build it up recursively. This is done more easily and cleanly if we do it mathematically, i.e. formally. One defines a formal language by saying what its alphabet is and then giving a grammar according to which well-formed strings may be built. A propositional language like L_A is an example. We shall, in later lectures, look at more ambitious fragments of English (or at least the formal languages that correspond to them). Once we have a formal language like L_A it can be used to pick out safe fragments of English by associating with the atoms some simple English sentences that we regard as unproblematic. For example, we could think of p_0 as standing for 'The light is on' and p_1 for 'The fan is on' and this picks out the fragment of English that consists of sentences built from these two atoms by means of the usual connectives. The formal language is simple enough for us to see that paradoxes won't arise in the corresponding fragment of English (as long as we're not silly enough to take p_0 to be 'This sentence is false').

The formal language (and also, if we wish, the fragment of English associated with it) is called an *object*-language, because it is the thing to which our attention and study is directed. English, enriched by whatever symbols we need, is the *metalanguage* in which we talk about the object language, where the Greek 'meta' means 'with'. If we were talking about French, but were speaking in English, then English (augmented by including the names of French sentences) would be the metalanguage and French the object language. For example, you might tell me 'In French one uses double negatives, as when one says "*Je ne parle pas français*".' The sentence as a whole belongs to the metalanguage, and this metalanguage sentence mentions a sentence of the object-language (quoted and written in Italics).

The distinction between object-language and metalanguage is a special case of a more general distinction between the levels at which one may operate. Think about rule-based behaviour, such as playing a game. Consider the rule 'The only rule is that there are no rules'. Can you see that this is another form of the Liar Paradox? Are there any rules or not? To avoid the paradox, we need only distinguish between the object-level to which the rules apply and the metalevel at which the rules are formulated (and changed). It immediately follows that rules don't apply to themselves. At the metalevel we may have zero, one, or more rules governing behaviour at the object-level. The paradox 'The only rule

is that there are no rules' is resolved by interpreting it as being not a rule but a declaration that there are no rules at the metalevel governing object-level behaviour.

To illustrate the distinction between levels more concretely, consider the painful subject of cricket. The object-level is the level at which behaviour is expected to conform to the rules and where Brett Lee is permitted to viciously hurl leather spheres at the heads of innocent batsmen. The metalevel is the level to which we ascend after the game when we draft a petition to the International Cricket Council to change things so that everyone has to bowl underarm.

Here are more everyday examples of the distinction. Consider the duties of a chairperson. The members of a committee are allowed to become totally involved in debating an issue, whereas the chairperson functions at the metalevel to ensure fair play and that the discussion remains relevant. If a chairperson gets involved in the discussion without relinquishing the chair to someone else, then something has gone wrong — the object- and metalevels are being confused. A chairperson who fails to maintain the distinction between levels invites conflict. Conversely, establishing a distinction between levels can resolve conflict. Consider for a moment the conflicts that arise between partners. There is often the temptation to be trapped in an exchange of recriminations of the form 'You did so and so!' 'But you did such and such!' A guidance counselor would break the circle of blame and lift the discussion to a metalevel from which the problem can be seen in perspective. This is why people seek counselling and why counselling often works.

Finally, consider the paradox of democracy. Suppose a political party is elected democratically, in free competition with other parties. And suppose the party then proceeds to change the rules by outlawing other parties and effectively ruling out the key democratic concept, namely that it should be possible to replace the government without the use of violence. Does the meta-object distinction suggest a way around this paradox? One approach is to have a constitution which restrains government. Of course, one has to make it impossible for the government to change the constitution on their own. In the USA, Congress may suggest changes to the constitution but these changes have to be ratified within a specified period by sufficiently many states, so that a broad consensus outside Congress is necessary before changes are made. In effect, the constitution has been placed at the metalevel, where it constrains object-level behaviour of the government.

Returning to logic, we (the logicians) stand at the metalevel using the metalanguage. In this metalanguage we can describe agents and their iconic representations of systems (i.e. semantic structures like val-

uations). For use by the agent, an object-language may be constructed. This object-language typically corresponds to some well-behaved fragment of the metalanguage and is a set of strings generated recursively from a set of atoms. An agent may or may not be able to say everything it wants in the object-language, i.e. may or may not be able to express all its iconic information about the system symbolically. Tarski showed that some metalogical things can never be said in the object-language — in order to avoid versions of the Liar Paradox, the object-language must be constructed in such a way that it is impossible to speak, in the object-language, about the truth or falsity of all object-language sentences. Some agents (humans!) can speak both the object-language and the metalanguage. Other agents (e.g. a software agent doing e-commerce on the web) may be limited to manipulating the strings of the object-language. An agent locked into the object-level would be able to say a sentence like '$p_0 \rightarrow \neg p_1$' but would not be able to say metalanguage sentences such as '$p_0 \rightarrow \neg p_1$ is satisfied by valuation 100' or '$p_0 \models \neg p_1$'. The connectives $\rightarrow$ and $\leftrightarrow$ are available to agents at the object-level, but only agents capable of functioning at least partly at the metalevel would be able to use the symbols $\models$ and $\equiv$. (Recall the exercises met previously in which we contrasted $\equiv$ with $\leftrightarrow$ and $\models$ with $\rightarrow$.)

The moral of the story is that the metalanguage is always stronger and more expressive than the object-language.

Summary 8 *We proved a simple limitative theorem about the lack of expressiveness of an object-language L_A that has an infinite set of atoms. We then discussed the object-meta distinction due to Alfred Tarski, and touched informally on Tarski's theorem about the indefinability of truth (no language can escape the Liar Paradox if it is so expressive that we can talk in it about the truth or falsity of all its sentences). The desire to avoid the Liar Paradox is one reason why agents use formal, recursively constructed, object-languages for knowledge representation rather than using a natural language such as English. These limitative results are similar in spirit to the better-known (but frequently misunderstood) incompleteness theorems of Gödel.*

3 Glossary

- **conjunct** — in a conjunction $\alpha \wedge \beta$, each of the sentences α and β is called a conjunct.
- **disjunct** — in a disjunction $\alpha \vee \beta$, each of the sentences α and β is called a disjunct.

- **disjunctive normal form** — a sentence which is the disjunction of conjunctions, for example $p \vee (q \wedge r)$, where the disjunct p is in effect a conjunction of a thing with itself, and the other disjunct is the conjunction $q \wedge r$.

- **effable** — what, if we were feeling mischievous, we might call information that can be expressed by sentences of the knowledge representation language, in contrast to 'ineffable' information.

- **ineffable** — information, reflected by a set $\overline{X}$ of excluded states and its complementary subset X of included states, that cannot be expressed in the knowledge representation language L_A because there is no sentence α such that $\mathcal{M}(\alpha) = X$ (and in fact no set Γ of sentences such that $\mathcal{M}(\Gamma) = X$).

- **logician** — a peculiar creature often observed to be quietly muttering to himself and chortling. Those with acute hearing will discover that the creature is muttering something ridiculous like "If I asked you a yes/no question, would your answer be no?" In groups, logicians ask one another "Would you like tea or coffee?", whereupon they reply in chorus "Yes!" and roll around cackling with crazed glee.

- **metalanguage** — the language, typically English + mathematical symbols, in which the logician can speak of both the system and the agent; often a very rich language, with not only the virtues but also the defects of expressiveness (such as the capacity to express forms of the Liar Paradox) and thus a language to be used carefully — this is why logicians may often develop pompous or pedantic habits, and are to be found on rainy days hunched over cups of coffee arguing about the correct placement of parentheses or commas.

- **object-language** — the recursively defined fragment of the metalanguage in which an agent expresses symbolically some or all of the information the agent has gained and initially represented iconically; up to this point we have restricted consideration to opaque propositional object-languages, but in due course we will look also at transparent propositional languages and then at first-order languages.

- **state description** — a very handy kind of sentence that in effect lists all the facts characterising a state, and which exists only in languages having finite A.

- **SDNF (state description normal form)** — a disjunction of state descriptions, in effect listing all the models of the sentence.

- **Tarski, Alfred** — a very clever logician born in Poland in 1901, who invented the idea of model and gave the first definition of satisfaction. Tarski was originally named Teitelbaum, but changed his name and emigrated to the USA, where he taught at the University of California at Berkeley. His doctoral students had a hard time — he liked to work throughout the night, and he smoked like a chimney. But they learnt a lot and became influential logicians. In the history of logic, there are perhaps six names that stand out: Aristotle, who started it all; George Boole; Gottlob Frege; Bertrand Russell; Kurt Gödel; and Alfred Tarski. Very possibly, Tarski was the greatest and most influential of them all.