ON MOVING ON ON ONTOLOGIES: MASS, COUNT AND LONG THIN THINGS

Robin P. Fawcett

Computational Linguistics Unit, University of Wales, Cardiff CF1 3EU, UK
e-mail: fawcett@cardiff.ac.uk

Abstract: This paper discusses the principles that should govern the construction of two components of a system for natural language generation (NLG): (1) the ontology - or, rather, as the paper argues, the 'ontological' aspects of a belief system - and (2) the semantic representation of noun senses. It is an interesting fact that many ontologies bear a striking resemblance to a system network, as used in systemic functional grammar (SFG). Furthermore, two major current research efforts in the field of ontology-building are designed to run with a SFG generator: Pangloss, where the generator is Penman, and COMMUNAL, where the generator is GENESYS. It is therefore important to establish a principled approach to the 'division of labour' between the ontology and the equivalent aspects of the model of language - here a system network for the 'meaning potential' of English nouns. (However, the general principles should be relevant to ANY model of language.) The paper summarises (a) the purposes and (b) the structure of (1) a system network for noun senses and (2) the equivalent ontology (based on what we in the COMMUNAL Project judge is required in the next generation of belief systems for NLG). Examples are given of current work on the relevant system network and, more briefly, of the equivalent ontological aspects of the belief system. In particular, reasons are given why it would be inappropriate to give a primary place to the 'mass' vs. 'count' distinction in an 'interlingua' ontology - and even, surprising though it may seem, in a language-specific semantics for English. Finally, it turns out that, in the new perspective presented here, there is no 'component' of the belief system that is 'the ontology', and the reasons for this apparently anomalous position are given.

Keywords: ontology, system network, belief system, knowledge base, semantics, noun senses, natural language generation

1 Some current issues in modelling 'ontologies'

One of the 'givens' of Computational Linguistics (CL) - which is taken here in a sense that includes Machine Translation (MT) - is that any such system needs an ontology.[1] But what, precisely, is an ontology? It seems to be one of those concepts which everyone who works with it instinctively feels they understand, so that the basic assumptions are seldom made explicit. In practical terms, there is a fairly general assumption that an ontology is closely related to, and perhaps isomorphic with, the 'meanings' of the nouns of a language - or of a set of languages, the maximal set being all human languages. But in building a theoretically satisfactory overall model one discovers that there are serious problems with this position, as will be shown in this paper.

There is a long history of work on ontologies for CL, including the important work at Carnegie-Mellon University over many years, that of Dahlgren and her colleagues (e.g. Dahlgren 1988) and the current Pangloss Project, as described in Hovy and Nirenburg 1992, Hovy and Knight 1993 and Knight 1993. We in the COMMUNAL Project have been considering alternative approaches to this aspect of what we term the **belief system**, and I would like to present here, for discussion by the wider NLG community, the principles that we have established, often after years of experimentation, as they relate to these matters. We are currently implementing a system based on these principles.

In many respects, of course, our assumptions are similar to those of others working in this area. But our view is that the next generation of systems in Artificial Intelligence (AI) - and possibly also in MT - will require, as central components, belief systems that represent 'knowledge' (or, more accurately, **beliefs**) of more types and in a more complex manner than in some current systems. As we shall see by the end of this paper, the phenomena that are often handled in terms of an ontology look somewhat different in this new perspective. In relation to some of the issues to be discussed here, then, we are constructing a different overall model from that which appears to underlie much other current work. The purpose of this paper is to set out these ideas, to compare them with those of other researchers on whose work we are seeking to build, and to give some explanation of why we are following the direction that we are. We hope that this will open up further discussion about the next generation of belief systems.

The first step is to be clear about what the issues are. They are (1) issues of levels (which we shall here assume to mean **levels of language**), (2) issues of **components** of the overall system, and (3) issues of the structure and contents of part of the largest component of the overall system, namely the **belief system**. We shall focus particularly on the types of relations that need to be recognised as holding between the 'concepts' in an ontology or rather, between the **generic objects** (in contrast with **specific objects**), such as 'dog', realized as *dogs*, as in I like *dogs*.[2]

2 Why the discussion remains open

Hovy and Nirenburg (1992), in clearing the ground for their discussion of the principles that should guide the construction of an ontology, suggest that

> most ontologies and domain models to date have been assembled based primarily on introspection, and often reflect the idiosyncrasies of the builders more than the requirements of the application (such as MT). Lacking well-founded guided principles, the ontology builder is working in the dark.

This judgement seems a little hyperbolic, in view of the fact that a number of recent ontology-builders have explained their principles as fully as Hovy and Nirenburg do. Thus, while there are aspects of Dahlgren's framework (1988:46f.) that are open to criticism, she in fact gives as detailed an account of her principles as do Hovy and Nirenburg. Bateman (1990, Bateman et al 1990) similarly explains the ideas underlying the 'upper model' in Penman. In fact, as Hovy and Knight (1993) state, the 'ontology base' in Pangloss is in part based on Penman (being a merger of this, the semantic categories from the *Longman Dictionary of Contemporary English* (LDOCE), which is intended as a taxonomy for the nouns of English), and ULTRA (Nirenburg and Defrise 1992) - which itself draws on the LDOCE categories). Nonetheless the main thrust of

Hovy and Nirenberg's comment is surely right, in that the discussion of the principles governing the building ontologies is far from over. Indeed, it may be that the requirements of MT and AI (where NL is used to communicate with a problem solving system) are different (at least at this stage of the development of MT).

Most - and perhaps all - ontology-builders accept that they are in fact working on the basis of natural language - and, typically, that the language is English. In Bateman's words (1990:56) 'the Upper Model represents the speaker's experience in terms of generalized linguistically-motivated "ontological" categories'. He states that '*it is essential that constraints should be found for what an upper model should contain and how it should be organised* [his italics], and he then goes on to suggest that it is the aspects of meaning found in the experiential meta-function in a systemic functional grammar that should guide the construction of the ontology. This application of SFG concepts provides a helpful framework in which to approach the problem. (We might note that distinctions that depend on register, e.g. on tenor (formality), such as that between *cigarette* and *fag* are irrelevant to the ontology.) The reason for depending on NL, then, seems to be that, without it, we would have no guidelines as to how to structure the ontology.

Let us assume for the moment that this position is justified (though I shall suggest some problems with it in sections 5 and 6). This brings us to an important question for those who use the Upper Model or any derived ontology such as Pangloss. This is: if there is to be this 'strong connection to the linguistic system' (Bateman 1990:57), what part of it should that connection be to?

And here we come to an apparent problem for this line of argument. This is that the Nigel model of language around which Penman is constructed lacks a specific network for noun senses. So what are the principles on which these aspects of the Upper Model are constructed? We shall return to this matter after the next section.

3 Two extraneous factors that may lead to differing research assumptions

First, however, let us be explicit about two of the factors that differentiate various research projects in this area, and so, perhaps, the assumptions underlying various conceptual frameworks that have been adopted. The first is the disciplinary coverage of researchers on a project - including the 'home discipline,' as it were, of the leading researcher. Those such as myself whose starting point was linguistics are sometimes shocked at the cavalier way (note the loaded language to express the viewpoint!) in which non-linguists simply adopt the 'senses' of the nouns of English as the starting point for an ontology. AI-minded computer scientists may be equally shocked at the pussyfooting way in which many linguists refuse to recognise the need to move outside the semiotic system of language, even though this is patently necessary in order to build adequate models of how language is used. Often neither really addresses the issue of where the semiotic system of a natural language ends (i.e. the semantics of the language) and where the 'concepts' (or whatever is assumed as the category for 'thinking') begins. In this paper I shall set out a clear and, I believe, defensible position on these issues.

A second factor which undoubtedly affects the conceptual framework used in any given research project is the time scale set by the sponsors of one's research. Practically all sponsors of CL research expect products that are at least potentially 'applicable' (though often in a sense that is not well defined). For some researchers the time scale may be such that they must work with existing data bases, such as the machine readable version of the *Longman Dictionary of Contemporary English* (LDOCE) (Longman Group 1978) or 'Wordnet' (Miller 1990). This seems to be the case with the Pangloss project (Hovy and Nirenburg 1992, Hovy and Knight 1993, Knight 1993), whose explicit goal is to combine the best features of these sources. This is an ambitious goal and I wish the researchers well. However the well-known fact must be pointed out that, in the last decade and a half, many other researchers have put in a lot of work in trying to make LDOCE usable in a number of ways - and yet so far as I am aware no one yet has found a way to use these data as part of a belief system without an enormous amount of hand-editing. There are important questions that need to be asked about the relationship of these data to ontology building. The answers should relate to an integrated framework that provides appropriately for at least the two linguistic levels of meaning and form, and, outside language, the categories of a belief system ('concepts').

Researchers who are working to a less constrained time-table are perhaps more able to ask such questions. There are arguments for and against each approach, and it is not the purpose of this paper to criticize the work of those who seek directly to exploit existing data bases. Indeed, it may be that such work will in time produce solutions to the problems to be discussed here, by developing EVOLUTION-ARILY into more advanced models. Alternatively, it may be that a significantly different framework is required in order to achieve optimal representations of belief systems; both lines of inquiry should be pursued.

In the COMMUNAL Project our task is to think speculatively about the next generation of belief systems, and about the components, relations and procedures that will be required in it. It is in the nature of research that we shall almost certainly have overlooked some aspects that will strike future researchers as important, but the enterprise is nonetheless worth attempting. Here, however, we shall not try to provide a complete overview, even very briefly, of all of the components that we believe to be necessary in a belief system (for which see Fawcett 1993), but just those aspects that pertain to the concept of the 'ontology'.

4 The intertwined concepts of 'ontology' and 'system network': a brief history

4.1 Halliday's proposal as a starting point

Since ontologies display many of the characteristics of a system network, let us begin with Halliday's seminal paper 'Categories of the theory of grammar' (1961). In it he proposed the concept of lexis 'as most delicate grammar'. He envisaged a model of language in which the earlier choices in a system network would be realised grammatically, i.e. in the **structures** of clause and group syntax and in **grammatical items** - and where these earlier choices would lead on to more 'delicate' choices that would be realised as **lexical items**.

In the following decades Halliday and others did a great deal of work to develop the grammatical aspects of the model - including the important step of integrating intonation with grammar. But what of integrating lexis? While 'grammatical items' such as modal verbs and various types of determiner were modelled in system networks, the concept of system networks for lexical items lexis remained largely unexplored until the mid-70s to the mid-80s. In that period there were several small studies by systemic linguists (though not by Halliday himself) which implemented the concept of 'lexis in networks' (Berry 1977, Fawcett 1980, Hasan 1987), but they were simply illustrative and there was no attempt to explore the implications of a comprehensive treatment of the original concept. (These 'implementations' were linguistic descriptions, not computer implementations.)

Meanwhile, Halliday had added the important notion of 'meaning' to that of 'choice' at the heart of what now came to be called **systemic functional grammar** (SFG), so that the networks of Berry, Fawcett and Hasan were

explicitly proposed as networks of features intended to capture the **meaning potential** of a language. It has always seemed clear to me that, since the networks specify meaning potential, the features in them can be appropriately regarded as **semantic features** (Fawcett 1973/81, 1980, etc). Halliday himself sometimes seems to agree with this position and sometimes to claim that the networks are at the level of form. (See the discussions of his variable position in Butler 1985 and, for example, Fawcett, Tucker and Lin 1993.) For the fullest and best account of the developments in SFG in modelling lexis, we must await Tucker's PhD thesis, currently in preparation.

4.2 Leech's logical semantics

To the best of my belief, the first example of the intertwining of the two concepts of an ontology and a system network occurred in the mid 1960s. It was at about this time that the system networks of SFG were beginning to be semanticized as the meaning potential of a language. Linguists working in the Chomskyan tradition had recently introduced the concept of selectional restrictions - which effectively presuppose semantic features. At this juncture a British linguist, Geoffrey Leech, who was at University College London at the same time as Halliday was developing SFG there in the 1970s, made the interesting experiment of combining the two concepts of system networks and semantic features in his *Towards a Semantic Description of English*, published in 1969 (and later incorporated in his standard text book *Semantics* (Leech 1974/81). This, then, was an early attempt to provide an ontology-based logic for reasoning, and it was done at a level that Leech assumed to be the level of semantics, i.e., presumably, a level within language. (In the view taken in this paper and to be expanded later, reasoning is in fact better modelled as taking place at a level outside and 'above' language - while being heavily influenced by language.) Leech's model included a taxonomy of 'types of object', and it had most of the characteristics of current ontologies (which we shall summarise in Section 6). This work was, in effect, Leech's attempt to combine the relevant parts of a system network with the demands of a reasoning system. It was related to a fairly standard logic, so that the taxonomy could be used for simple reasoning tasks, in some of the ways that current ontologies are expected to.[3] The crucial point, however, is that Leech's network was not in fact a system network, as the term is used in SFG, but an ontology.

One striking visual icon of this difference is the contrast between Halliday's and Leech's notations. They both use the usual systemic notation of a right-opening square bracket to indicate 'or' and a curly right-opening bracket to mean 'and'. But in Halliday's diagrams there is always an arrow pointing right, i.e. from the term (or terms) that is/are the entry condition to the system towards the system itself. And in Leech's diagrams the arrows point from the system to the entry condition. In other words, the two apparently similar structures are to be used in different ways. As we shall see in Section 6, an ontology is typically (but not necessarily exclusively) traversed from right to left, which explains the direction of Leech's arrows. But a system network is intended to be traversed from left to right, i. e. from the less 'delicate' choices to the more 'delicate'; see further below. (Some later systemic linguists, including myself, have followed Winograd (1972) in dispensing with the arrow.[4])

4.3 Dahlgren's ontology

Another ontology with interesting similarities to a system network is found in the important work of Dahlgren (1988). Her description of her 'category cuts' can be expressed directly as a system network, with 'entity' as the entry condition that leads immediately to two simultaneous systems; 'individual' vs. 'collective' and 'abstract' vs. 'real'. Further dependent systems, some of which are entered simultaneously (i.e. as 'cross-classifications'), introduce a total of 37 'category cuts' which, when all possible combinations are counted, generate 4272 potential 'combinations'. (This assumes that 'collective' leads on to 'mass' vs. 'set' vs. 'structure', as implied at one point; at another it does not.)

Interestingly, Dahlgren advises at one point that 'care must be taken when adding cross-classifications at a higher node, ... to avoid proliferating empty terminal nodes', ... [because this] is a sure sign that the proposed cross-classification is spurious.' If this were indeed to be accepted as a major criterion, one would unfortunately have to give her ontology rather low marks. This is because, of her 4272 'terminal nodes', most are unused (all but 57 on one count).[5] In fact, as we shall see in Section 5, Dahlgren's advice is fully appropriate for a systemic linguist who is using the network as a generator - but it is certainly much less relevant to an ontology. Why should this be so?

The answer is that Dahlgren's taxonomy is not intended as a system network, and so it should not be thought of as operating by being traversed from left to right (e.g. as if to generate noun senses). An ontology is in fact typically used deductively, i.e. working from right to left TO REASON ABOUT OBJECTS (as Leech's arrows remind us). So the supposed 'overgeneration' of Dahlgren's ontology has no practical consequence - unless one wishes to use the ontology inductively. (In other words, many systems regularly use reasoning of the nature of 'If X is a rose then X is a flower', while there seems to be less demand for inductive reasoning. We shall return to the topic of reasoning in Section 6.

I cite this case in order to show the need for clear principles in the construction of both (1) system networks at the level of semantics, i.e. within language, and (2) ontological relations in a belief system.

We turn now to natural language generation systems that are based on SFG. There have been many of these, and many others that have incorporated significant aspects of the SFG approach to language (see Matthiessen and Bateman 1991 and Fawcett, Tucker and Lin 1993 for overviews of the use of SFGs in NLG). Here we shall consider just the two major SFG natural language generators: Penman, which is to be used as the generator for the current Pangloss project, and GENESYS, which is the generator in the COMMUNAL Project. Given the significant place of SFG generators in NLG, it is important to be clear about the theoretical framework for what such enterprises are attempting. In particular, it is important to understand the relation between the system network and the ontology.

4.4 The Penman Upper Model

The Penman Project was the first large SFG-based generator, and the main work that established the structure and nature of what Halliday has called the **lexicogrammar** was done in the very early eighties (based fairly closely on Halliday's work of the seventies). In that period there was much more emphasis in linguistics as a whole on syntax and much less on lexis than there is now, and Halliday's programmatic use of the term 'lexicogrammar' was a far-sighted pointer to where work would be needed in the future. But Halliday himself has always approached language from the grammatical rather than the lexical end and, under his guidance, Mann and Matthiessen naturally worked first on the grammatical structures and items. Unfortunately, when the time came to extend the model to lexis, the sponsors of the project (working within the traditional framework of the 'grammar-vocabulary' distinction) required the Penman team NOT to implement Halliday's concept of an integrated 'lexicogrammar', but to build instead a traditional lexicon which could be shared with the parser that was being contributed by another research team (working at Bolt,

Beranek and Newman). There were some very interesting consequences of this decision.

First, recall that the choices in the system networks were at this time being increasingly seen as the 'meaning potential', and thus as semantic choices. In much of the detailed description in Halliday 1985, for example, the term 'semantic' abounds. And the concept of 'realization' itself - the process through which the choices in the system network are realized as structures - is one that explicitly invokes the two levels of 'meaning' and 'form'. Thus if one does not represent the lexicon in system network form - as of course Penman does not - there is a major gap in the overall system network of the language at just the place where one would wish to locate a system network of **noun senses** (to complement the list of **noun forms**). For the same reason the Penman network lacks the meanings of all the other major word classes. This has made it harder than it is in the COMMUNAL framework (see below) to follow the increasingly strong trend in linguistics to give a more central place to lexis than it had in the 60s and 70s. (Fawcett Tucker and Lin 1993 show how lexis is handled in COMMUNAL, and important current modifications are exploiting yet further the value of having integrated lexicogrammatically realized networks of meanings - e.g. in verb complementation, and the many other cases where syntactic realizations depend upon lexical choice.)

The Penman response to the position in which they found themselves was a sensible compromise. The development work that might have gone into the meaning potential of lexical forms went partly into the separate lexicon (building into it the relevant features from the grammar), and partly into the Upper Model (UM). The main function of the UM is to serve as an 'abstraction hierarchy' (Bateman 1990:57), and so to provide the usual functions of an ontology for property inheritance, etc. But at the same time each concept in it is 'known to Penman', in the sense that 'it is possible to state for each concept in the UM the fragments of grammatical or lexical realization that will be used to realize it' (Bateman et al. 1990:5).

However, the section of the UM corresponding to noun senses seems surprisingly small for this task - at least, as described in Bateman et al 1990. It contains just over 30 categories (half being specialised within 'spatial-temporal'). The idea is that these are 'common core' concepts and that users of the UM will add to these as required.[6] (It is interesting to compare these with the early semantic COMMUNAL categories, summarized in Section 5.)

In Penman, then, the ontological categories map directly to (1) grammatically realized options in the system network, and (2) the standard lexicon. The job does get done. But the great advantage of a SFG - namely that it can equally easily capture generalisations across large and small classes (including one-member classes) - is not exploited. Thus Penman fails to utilize the many advantages of a fully integrated semantic network that provides for direct and integrated realizations in grammar and lexis (and indeed also in intonation or punctuation).

At the time when it was developed, in the very early 80s, Penman was a pioneering breakthrough. Its influence has changed the face of NLG. But the fact remains that it does not contain a lexicogrammar, in the SFG sense - just a grammar. The historic significance of Penman must not distract us from recognising that it has not yet explored Halliday's original proposal to integrate grammar and lexis in one great net-work. Yet this, as we in COMMUNAL have found, is a concept which - with the important modification that we do not restrict lexical meanings to the 'most delicate' parts of the network - brings enormous advantages of power and flexibility to the modelling of language.

4.5 Alternative research strategies

If one's goals are (1) to build a SFG generator and (2) to relate it out to a representation of those ontological relations required for reasoning, the following question arises: 'Is it (a) necessary and (b) desirable to have two separate layers of network, representing (1) the semantic stratum within language and (2) the ontological relations in the belief system?'

There are two research routes that may lead to a sound answer. The first is to try having just one layer of network, and then to move on to two if it turns out to be desirable or necessary. The second route to an answer is to start with two networks, one for each level, and then, if there turns out to be inadequate motivation for maintaining two, to abandon one or conflate the two of them. In effect, Penman had the first strategy forced upon them, while we in COMMUNAL have followed the second. As we considered the purposes - and so the desirable structural characteristics - of the two potential components, it became increasingly clear that there are advantages in including them both.

The next two sections therefore set out the purposes and consequent structural organisation of (1) the system network for noun senses currently being implemented on a very large scale in COMMUNAL, and (2) the ontological aspects of the matching part of the belief system. While the discussion is naturally exemplified from the COMMUNAL Project, the principles are of general relevance to any researcher working in this area. Here we shall restrict ourselves to the 'core' area of 'objects', including abstract and event-like objects, i.e. that part of the belief system that corresponds to the senses of nouns.

5 The purposes and structure of a system network for noun senses

5.1 The purposes of a system network for noun senses

My work in linguistics and NLP over the last couple of decades has taught me that one of the most important lessons to learn, when trying to model language, is:

DO NOT TRY TO DO TOO MUCH WORK AT ANY ONE LEVEL.

Thus the key to modelling language successfully is to have a sufficiently holistic theory and, to be able to recognize the appropriate level - or component - at which each particular type of work should be done.

Once one commits oneself to having a separate component to handle reasoning (including property inheritance, etc) that is OUTSIDE LANGUAGE (even though, as I have always insisted, its internal structures are strongly INFLUENCED by language), then it becomes immediately clear that the system networks inside the language system are NOT in fact well-suited for use in reasoning. The reason for this is very simple: the design features that are required in the structure of the relevant parts of the network are different from the design features required for ontologies.

So what are the purposes of this part of the system network? Its primary purpose is very simple. Just as the well known systems for transitivity, mood, and theme, etc generate clauses, so it is the purpose of this network to generate the nouns which will expound the heads of nominal groups.[7] There is an important subsidiary purpose for those noun senses to which participant roles ('argument structure' to some) are attached, i.e. as with verb senses; compare the roles attached to *die* and *death*, to *ascend* and *ascent*, etc, but we cannot discuss these here. The other subsidiary purposes will be introduced in considering why the structure should have the form proposed below.

5.2 The structure of a system network for noun senses

What, then, should the internal structure of a system network for generating nouns be like? The answers given here are derived from the experience of developing very full

sub-networks for large areas of the network for the 'cultural classification' of 'things' in English - i.e. for the classification of objects provided by the noun senses of English. As an indication of the coverage, consider the fairly well-developed sub-network for 'artefacts' that are 'for human consumption' (i.e. food and drink): it has 137 systems and it generates 330 noun senses. Another quite well-developed area is that of plants, where there are 130 systems that generate 246 noun senses. (This reflects a layman's taxonomy of different trees and flowers; a specialist could well have three to six times as many.) An area currently under development, 'use_of_land', includes 'built_up_area', 'countryside', 'for_travelling', 'for_recreation', 'wasteland' and 'for_dividing_land', and it has so far 135 sub-systems, generating 370 noun senses. Many other areas are of similar size.

We can give some idea of the overall taxonomic structure by saying that 'use_of_land' is one of fourteen types of 'artefact'. The 'substantial' (see below) features in the first few systems are as follows (where a system is represented as 'x -> y/z'): thing -> physical_thing / abstract_thing / event_thing; physical_thing -> living_thing / non_living_thing; living_thing -> plant / creature; creature -> human_cr / non_human_cr; human_cr -> individual_hum / group_hum; non_living_thing -> thing_as_object / general_physical_phenomenon / thing_as_substance; thing_as_object -> artefact / natural_object; general_physical_phenomenon -> energy / weather / colour; event_thing -> event_as_process / complex_event. The network is already very large, and it is growing all the time.

What, then, are the principles on which it is constructed? The simplest 'network' for the noun senses of English would consist of just one large system, containing one long list of the noun senses, each realized by a noun form - i.e. a single, massive system. Why not do this? The main linguistic evidence is the existence of hyponymic and contrastive relations between words - or more strictly, between word senses. Thus the word-form *cat* is a hyponym of *mammal*, because the meaning 'cat' is systemically dependent on the meaning 'mammal'. And a system network is equally appropriate for giving formal expression to another major type of relationship between words, e.g. as specified in standard works on semantics such as Lyons 1977: that of 'contrast'; the sense of *dog* is in contrast with that of *cat* The evidence that such hyponymic and contrastive relations are anchored firmly within language - and not simply in some 'higher' component of the belief system - is the existence at the level of form of nouns such as *thing, object, animal, mammal, person*, etc; see the note on the 'as such' type of feature in 5.5.

The overall structure is therefore taxonomic. Within this framework, we turn now to a 'guideline' criterion that derives directly from the practical experience of building ontologies.

Our experience in COMMUNAL suggests that it is advisable to avoid simultaneous entry conditions, i.e. right-opening curly 'and' brackets. Here Dahlgren's point about overgeneration (4.3 above) DOES apply. (We shall meet the formalism when we consider ontological relations in Section 6, because in that type of network they do have a role to play.) The problem in generation is that such parallel systems lead to parallel pathways through the network, and so, almost inevitably, to permitting more co-selections than are in fact possible. For example, the distinction between 'male' and 'female' is realized lexically in the case of only some animals. In COMMUNAL we simply allow the repetition of such systems, or gather the relevant features together into a disjunctive entry condition (as in Figure 2), whichever is easier in practical terms for the maintenance of the network as it is expanded. The key point is that, if a network designed to generate nouns allows one to go down pathways - and so to choose a set of features - FOR WHICH THERE IS NO REALIZATION, it is a bad system network.[9] (But note that, while simultaneity tends to lead to problems in networks for lexis, it has a valuable role in modelling grammatically-realized meanings - and, as we shall see in Section 6, ontological relations.)

The large network for noun senses in COMMUNAL therefore has essentially the structure illustrated in Figure 1. Neutrally, this structure can be described as: 'If 'a' then 'b' or 'c'.' More typically, in a SFG framework, we express it as: 'If 'a' is chosen (by a speaker or a computer generator) then either 'b' or 'c' must be chosen.' Each feature may become the entry condition to a subsequent system, thus building up a system network, so it might continue: 'If 'b' is chosen, then 'd' or 'e' must be chosen, and if 'c' is chosen then 'f' or 'g' or 'h' must be chosen ....' and so on.

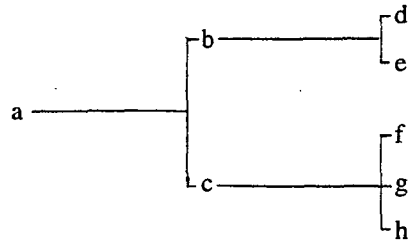

Figure 1: A simple system network

However, there is one type of complex entry condition which we have found to be occasionally useful in the network for noun senses: the disjunctive entry condition. This has the form shown in Figure 2:

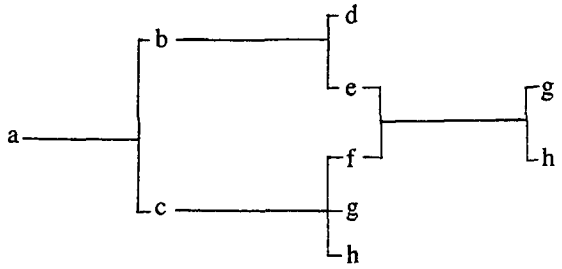

Figure 2: A disjunctive entry condition

Here is a (simplified) example of this phenomenon from the current network (where "/" signals 'or' and "->" signals 'entry to a system'):

```
tomato_plant_c / pepper_plant_c / cucumber_plant_c /
    strawberry_plant_c ->
        95%     plantness_explicit (66.325)
        5%      plantness_implicit.
```

This representation of a system network (in the COMMUNAL format) is to be read as follows:

'If you select either 'tomato_plant', etc (which come from the 'vegetable' section of the 'ground_crops' network) or 'strawberry_plant' (which comes from the 'fruit' sub-network), then there is a choice between having the concept of 'plantness' made specific and leaving it implicit.' So, if the choice is to make it explicit, forms such as *tomato-plant* will be generated, and, if implicit, forms such as *tomato* - but in the gardener's sense of 'tomato plant', as in *I've watered the tomatoes this morning.*

Another valuable characteristic of the COMMUNAL system networks is the use of probabilities. As you can see, the weighting is 95% to 5% towards 'explicit' (unless this is overruled by the discourse planner). The reason for this strong weighting is that there is another type of object

that is commonly referred to by the same word-form, namely the fruit of the tomato plant. (The choice of 'plantness_explicit' triggers Realisation Rule 66.325, which adds to the head of the nominal group the item *plant*.)

A third criterion used in constructing the network relates to its value in specifying what in Chomskyan linguistics are termed 'selectional restrictions'. Some might wish to argue that these really belong in the belief system rather than in language itself. However, it turns out that it is quite simple, if one constructs the network on appropriate criteria, to enable very many of the types of restriction that it was hoped to capture in transformational models of the 1960s and 1970s to be captured in a system functional grammar. Moreover, since we use probabilities in the grammar in any case (for purposes that we cannot go into here), we can capture selection restrictions as relatively strong or weak 'preferences'. Thus there seems to be no reason why we should not capture such phenomena in both the **system network** within language and in the **belief system**. If we can do this, it brings the advantage, when TESTING the system, that the language components can be run independently of the belief system. They can be set to generate sentences randomly as we test the lexicogrammar, and still generate sentences that sound fairly plausible.[10]

The way this works is, in outline form, as follows. Suppose we have just generated a clause, with an Agent as Subject and *ask* as the main verb. We want to restrict the choices when the network is re-entered, so that, if a nominal group with a noun at its head is to be generated to fill the Agent, it will have one that is plausible as an 'asker'. We do this by specifying that, on re-entry to the network, the following features are chosen: [thing, concrete, living, creature, 99.9% human / 0.1% non_human, whole_hum, 99% individual_hum / 1% group_hum]. At two points, you will notice, there is not an absolute 'preselection' of just one of the features in the system, but a **preference** for one over another. Thus the feature [human] is shown to be a thousand times more probable than [non_human] - thus allowing for talking computers and talking animals, such as the white rabbit in *Alice in Wonderland*. Rather similarly, individual humans are shown as being a thousand times more likely to be the Agents in a process of 'asking' than are groups of humans - though groups, such as a committee, may well ask questions on occasions.

We turn now to an important issue that arises as a result of taking this position, and we shall illustrate it first from English.

5.3 Problem case 1: the 'count' versus 'mass' contrast

One important effect of deciding to organize the network around meanings (rather than around grammatical correlates of word forms) is to put in perspective the contrast between 'count' and 'mass' noun senses that is so dominant in English (and many other European languages). If, for example, we are stating preferences for the Affected entity in a process of 'eating', it is not important whether what is eaten is mass or count. The COMMUNAL network for food classifies types of food on semantic criteria - so that, for example, vegetables that typically occur with a meat course are placed together, with the mass noun sense 'cabbage' next to the count noun senses of 'potato' or 'pea'. Moreover, the network also includes a way of showing that, while 'potato' occurs regularly as either singular or plural, it is rather unusual to talk of a single pea. The result is that the system 'knows' that it is a thousand times more likely that "peas" will be generated than "pea". (Note that the system networks in COMMUNAL not only have probabilities, but the grammar can change these when required. For a more detailed account of this, together with a full worked example, see Fawcett, Tucker and Lin 1993.)

However, there is not in fact a neat 'count' versus 'mass' distinction in English at all. It is one of those useful generalisations which hold for 95% of the time, but which, if one commits oneself to it, leads to considerable trouble when the idea is extended to the whole of language. It is certainly not a category that can be extended, on the basis of a system network for English, even to a close European language such as French. The illogicality - in physical 'number' terms - of items such as *furniture* and *cutlery* is just the well-known visible tip of quite a large iceberg. First, note the 'plural-only' nouns, such as *police, staff* and *contents*. Then there are the 'plural-preferring' items, with varying strengths of preference, as for example between *pea* and *sprout*, and *pebble* and *leaf*. There are also the 'pair-only' nouns such as *trousers, scissors,* and *binoculars* - and, even though the word denotes two garments, of which only the bottom half conforms to the pattern, there is *pyjamas*. As an example of the dissonance within one relatively small semantic field between the grammatical criterion of 'number' and the semantic classification of noun senses, consider the field of clothing. Suppose the problem is that of stating the preferences for the entity that is to complete a clause such as *He went home and put on .....* It could be a nominal group with, at its head, (1) a mass noun such as *some warm clothing*, or (2) a plural-only noun as in *some warm clothes*, or (3) a singular noun such as *a warm jersey,* or (4) a pair-only noun such as *some warm trousers*.

The COMMUNAL solution to this problem is as follows. Every feature in the system network for which there is a realization is given a suffix '_c' (for 'count' things) or '_m' (for 'mass' things) or '_pl' (for plural only' things) or '_pair' (for 'pair only' things). Those labelled '_c' lead into the system for NUMBER, where there is a choice between [singular] and [plural], while for the others there is no choice: they are either [mass] or [plural]. Those for which a RELATIVE preference for [plural] has been stated will enter a version of the NUMBER system for which the probabilities have been re-set according to the strength of the preference associated with than noun sense, i.e. the lexicogrammar 'knows' that for 'pea' the probabilities of choosing 'plural' are very much greater than for 'cabbage'.

Consider too the question of where to place the two senses - realized as count and mass nouns - of *cloud*. In this case the differences between the two appear to derive mainly from the fact that one is an individual entity and the other is a non-discrete mass of 'stuff'; they are not differentiated by function. So the system network shows them to be distinguished as follows:

cloud -> cloud_as_individual / cloud_as_mass.

Notice that this is a rather different matter from 'oak' as 'individual' and as 'mass'; in the first case the referent is a tree (with *oak-tree* as a possible variant, if 'treeness_explicit' is selected), and it is located with the rest of the trees as a type of plant, while in its other sense it is a material whose function is to be used for making things, and it is located with other types of wood, fairly close to 'iron', etc.

In the COMMUNAL system it is even possible to build in the preference for meanings realized in expressions such as *a pair*, followed by *of*, as in *a pair of trousers*. This is because *trousers* rather than *pair* is treated as the head of the nominal group. The words *a pair* are generated as a special nominal group expressing 'quantity' that fills the quantifying determiner. It is because the noun sense of the nominal group is generated on the first pass through the network for the object in question - i.e. because the lexically realized meaning is integrated with the grammatically realized meaning in COMMUNAL - that part of the realization of the choice of 'trousers' can be to express a preference for the way in which the quantifying determiner is filled - here by the embedded nominal group of *a pair*.

Thus the COMMUNAL way of handling these various aspects of 'number' in English can reflect accurately the

true, complex nature of 'number' and 'quantity' in English, while at the same time maintaining the semantic relationships of hyponymy and contrast in the network, and so making possible the expression of preferences.

5.4 Problem case 2: long thin things and other such grammatically realized categories

The issue is in fact much wider than that of whether 'count' vs. 'mass' should be a primary distinction in English and related languages. How strong a candidate it is for a generalized 'interlingua' ontology that is to accommodate all the languages of the world? Consider Chinese, with its well-known classifier system, in which the mass-count distinction plays no part. Then think about Swahili, with its *ki- vi-* class of non-living things, its *m- wa-* class for humans, its *u-* class for abstract things, etc. Japanese, as it happens, has a special set of cardinal determiners, whose form depends on the semantic class of the noun: i.e. whether the object is human or a small thing or even, it would seem, a long thin thing. Thus, if the thing concerned is a flower (*hana*), a tree (*ki*), a pen (*pen*), a pencil (*enpitsu*) or a river (*kawa*) - all long thin things - the determiner meaning 'one' will be *ippon*. But if the thing is a human it is *hitori*, and if a non-human creature it is *ipipi* - and so on, for many more classes of thing and for many more cardinal determiners. The semantic generalisation that unites those things that require *ippon* appears to be simply that they are all 'long thin things'. If this seems strange to the investigator from a European background, consider how odd it must seem to the investigator of English from outside Europe who finds that, in English and related languages, there exists a basic distinction which affects many aspects of the semantics and syntax of the nominal group, between things that are and things that are not 'countable'.

However, it must be emphasised again that, important as the distinction is at many points, it is not the basic organising principle of the semantics of English noun senses. While there is, of course, a distinction between 'substances' and 'objects' in the taxonomy implied in the COMMUNAL network, and while all substances are 'mass', it is not the case that all 'objects' are 'count' - as we saw in section 5.3.

In COMMUNAL, then, our semantic system network gives no weight to grammatically realized contrasts such as 'count' versus 'mass', and we use instead the semantic criteria such as those that help us to state the preferences associated with a given participant role such as agent.

Is there a price to be paid for all of the advantages outlined in the preceding sections? The answer seems to be that there isn't. If the mechanism for handling NUMBER in a way that is dependent on noun senses (in 5.3) required one to work from a visual representation of the wiring, it would be tedious in the extreme But in the computer it is a simple matter - and this has in turn suggested a simple representation for the written version.

5.5 'Special features' in the system network

In a full system network for noun senses, the relations between features are not all of the 'subcategorization type that the systemic notation typically signifies. The extensive work that has been done in COMMUNAL over the last few years has produced a small set of supplementary (or 'special') features whose function is to express relations between the 'substantial' features that represent noun senses. In principle, all of those shown in Figure 3 can occur between any two substantive features. Here, then, 'x' stands for a feature such as 'human', and 'a', 'b' and 'c' are the next 'substantive' features. Possible realizations of the selection of the feature (in some cases in features dependent on substantive features) are given in Figure 3.

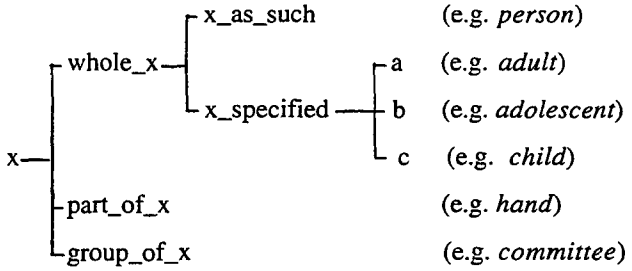

Figure 3: Some types of features that are not related by 'be-type' relations

'Substantive' features are those that translate types of object in the belief system. In most cases none, or only one, of the special features are needed between substantive features. Note in particular the 'x_as-such' feature. This provides for the generation of those nouns that are used for lexical substitution such as *thing*, *stuff* and, more delicately, *plant* and *animal*. It is the existence of such forms that provides the evidence that such less delicate meanings exist in natural languages, and that the intra-linguistic level of semantics does indeed require a network for it to be adequately represented.

There are possible variations of the feature [x_specified] which it is important to note. The **type** of specification may be spelled out more fully, with a feature corresponding to each type that leads to its own dependent network. Thus in some cases (such as [human]) we may find [x_as_such] vs. [x_specified_by_form] vs. [x_specified_by_role].

The somewhat dense description given above is intended to give the flavour of the criteria that are guiding the construction of the very large system network for noun senses in COMMUNAL.

There is one last benefit that this approach brings. It is a practical rather than a theoretical one. This is that the process of constructing the network, and so of deciding which types of 'special feature' are needed, goes some way in preparing the ground for constructing the equivalent ontological aspects of the belief system. And it is this to which we now turn.

6 The purposes and structure of the ontological aspects of a belief system

6.1 The purposes of an ontology

What are ontologies for? They are as they are, of course, because of the functions that they are required to perform. In the case of an ontology, the purpose is NOT to represent the meanings of the nouns of a language, but to facilitate reasoning. Thus lexically prominent register distinctions such as that between *fag* and *cigarette* would be modelled as denoting the same **generic object,** and so share the same 'concept'. This is because the same **events** (or 'propositions') are attached to it, whatever degree of formality is selected in the TENOR system. It is the latter that requires recognition in the belief system, and not a difference of concept.

Ontologies are used in two principal ways, the second being dependent on the first. The first is for the type of reasoning known as **entailment** (see, for example, Leech 1969 and 1974/81). Through entailment a reasoner can infer from the belief (or proposition) that Object 79 is a member of the set of objects denoted by the form *dog* (and by the sense 'dog') that it is also a member of the set of objects to which the sense 'mammal' pertains. In layman's language, *A dog is a mammal.* The crucial point is that, typically, the **directionality of the reasoning** is from the more delicate category to the less delicate. For example, in terms of Figure 4 below, if an object is 'd' it is also 'b'

and 'a'. This directionality is of course different from the typical use of a system network in generation.

The second main use of ontologies is an extension of this. It is for the **inheritance** of what are commonly called 'properties'. (However, as we shall see, the term 'property' is potentially misleading). Thus, if such 'properties' are attached to a category at one node of the ontology, they can then be assumed to hold for any given object to which a logically dependent category applies. So, if we build into a belief system the proposition that 'land mammals typically have four limbs', then we can infer that, because a dog is a land mammal, it too typically has four limbs - and so on. (More strictly, as we shall shortly see, the relationship is between the generic thing that corresponds to that node; in the COMMUNAL model of logic *dogs* is a referring expression as much as is *this dog*.)

These types of entailment-based reasoning are important. But it is equally important to recognize that they are only one of a variety of types of reasoning that are regularly used in real life situations. Moreover, the concept that all such beliefs about categories of object are handled by inheritance is not a matter that is entirely beyond dispute. It is arguable, for example, that the category of 'human' is so prominent in our perception of the world that we build sets of beliefs around it - and that we do not use an ontological structure in order to inherit, every time that we refer to a human, all of the many propositions that relate to the many other, ever more general, categories that are superordinate to 'human' - such as 'mammal', 'creature', 'living' and 'concrete'. And, once one allows for this possibility, there is no clear way of knowing where to stop. There may be many such nodes that have attached to them large sets of beliefs (or 'propositions') that are held by the system, and which may indeed involve the redundant repetition of beliefs that are attached to less delicate concepts. Who can say whether it is more economical (or more elegant, more efficient, or whatever) to store a large (but not infinite) set of propositions many times over at the various nodes where they are most often needed - the 'basic types', in the terms of Rosch (1978) - or to store each of them just once but to have to perform a multiple act of entailment reasoning, involving multiple searches back down the tree, every time one uses the belief that a dog needs air to breath, or is a 'creature', or is a 'concrete' object? A particular dog-lover, for example, may have his/her set of primary beliefs about spaniels attached to 'spaniel', rather than to 'dog', and so on (cp. Reiter and Dale 1992). In other words, while we are undoubtedly capable of performing the quite complex type of reasoning involved in inheritance, it may be that it is not the backbone of all reasoning that it has sometimes been assumed to be.

The argument is not that inheritance has no place in our reasoning, but (1) that it may have a much less central place than has generally been supposed, and (2) that other types of reasoning are probably equally or more important. A belief system of the type assumed here is 'object-oriented' in the sense that it consists of a vast number of **specific objects** and **generic objects** (with each of the former linked by a 'be-an-instance-of' relationship to one or more of the latter) such that one of the many things that the system believes about any generic object is what other generic object - or objects - that generic object is itself a type of.

In the above discussion, we have been assuming that we are considering generic objects, e.g. 'cats in general'. I assume that there would be general agreement that the relationship between 'mammal' and 'cat' is what we might term a 'be-a-type-of' relationship, while that between a specific instance of a cat, such as our family cat Timmy, is a 'be-an-instance-of' relationship. Specific objects are related to concepts via the belief that 'Timmy is an instance of a cat'. In other words, we need to distinguish between (1) 'Timmy is a cat' and (2) 'Cats are mammals'.

I suggested earlier that the use of the term 'properties' to refer to the 'propositions' attached to categories can be misleading. Consider the simplest model of what 'properties' should be attached to categories (the 'frame' approach). All there is space to say here is that there are clearly limitations as to what can be expressed in such limited structures, and, like others, we in COMMUNAL are exploring richer alternatives. In our case we are experimenting with a specially developed logical form for the representation of both (1) complex beliefs about, say, the mating habits of dogs, and (2) the belief that dogs are mammals - and indeed the belief that dogs are typically pets ('multiple inheritance').

The purpose of ontological relations, then, is to facilitate reasoning. But our prediction is that in future there will be less emphasis on inheritance and more on other logical relationships. Given these purposes, we turn now to the question of the structure of ontologies.

6.2 The structure of ontologies

There seems to be a general agreement that an ontology has the general form of a 'taxonomy' or 'hierarchy' (in one sense of that overused term) or 'tree' (which is equally overused). The type of 'tree' required here is thus a paradigmatic tree (rather as is a system network) where, in the simplest model, a pathway that consists of a list of features chosen as one traverses the network corresponds to any one (aspect of) an object. In the simplest type of taxonomy, then, 'a' is subcategorized as 'b' or 'c', 'b' as 'd' or 'e', 'c' as 'f', 'g' or 'h', and so on; see Figure 4.

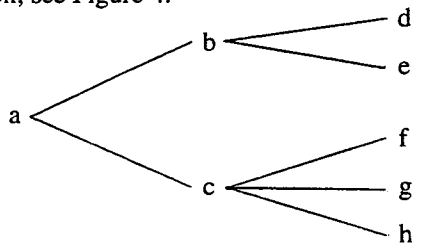

Figure 4 : Relations in the simplest type of ontology

Such a network shows that, if an object satisfies the conditions for being an 'h', it follows that it is also a 'c', and so also an 'a'. So far, this structure is similar to the simplest type of system network, as shown in Figure 1.

This 'simplest structure' for an ontology is in practice augmented in various ways in all of the ontologies that I know of. The essential addition is as in Figure 5:

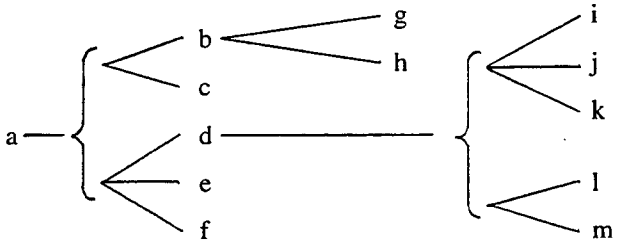

Figure 5: Simultaneity in ontological relationships

This is to be read as, 'a' is subcategorized as 'b', 'c' or 'd', and also as 'e' or 'f'; 'b' is subcategorized as 'g' or 'h', and 'd' as 'i', 'j' or 'k', and also as 'l' or 'm'.

Here the right-opening curly brackets do NOT signify that you must follow all the designated paths, as they would in a system network - because this type of network is designed to be used from right to left, for the types of entailment and inheritance outlined above.

Dahlgren (1988:46f.) discusses the mathematical

properties of ontologies, and she quite rightly points out that, while binary ontologies 'have simplifying mathematical properties, they are not likely to be psychologically real' (e.g. suggesting that we operate with 'fish' vs. 'bird' vs. 'mammal', etc). She also points out the need for cross-classification, in that we classify animals, let us say, both in terms of their 'types' (a concept that itself needs to be unpacked, at least for some purposes), but also in terms of their 'role' in humankind's scheme of things ('tame' vs. 'wild' animals, probably with other such subcategories as well). Thus Dahlgren and others operate with the formalism shown in Figure 5.

In the cases of other ontologies there is even more freedom, in that any feature can be linked by a line that represents a 'be-type' relationship to any feature to its left. And in some ontologies, of course, other relationships such as 'be-a-part-of' are used.

However, there is an important sense in which the above diagrams are misleading - at least in relation to the current COMMUNAL framework. This is because we consider that the relations do not hold, in fact, between the 'concepts' in the ontologies, but between generic objects. The relationship that we have shown in Figures 4 and 5 by a line can also be expressed by the following:

e (e111, [ca:o222, pr:be_type, at:o333]).
o(o222, [cc:dog, qt:all]).
o(o333, [cc: mammal]).

This states that event no. 111 has a carrier (a 'participant role'), which is object no. 222, a predicate, which is 'be_type', and an attribute (another participant role), which is object no. 333. Object 222 is then defined as the class of 'all dogs', while object no. 333 is defined as 'mammals'. There are several aspects of the meaning of this representation of a belief which are assigned by default - most importantly, that the time position of the event is 'past, present and future', that the 'confidence level' is so high as to be interpretable as '100% confident'. Taken together, they state, in a natural language translation: *All dogs are mammals.*

This event-based representation is used in order to enable the system to carry out reasoning on inputs to and outputs from the system that have the sorts of annoyingly messy complexity associated with natural language - representations of time position, usuality and quantification, for example. In the belief systems of the future it will no longer be possible, in our view, to depend on simple data structures such as frames. The need to incorporate more sophisticated representations of time, of modality and of other such phenomena related to events demands that information be stored in the form of some type of 'predicate logic', e.g. as in the (simplified) example above.

The key point is that the representation, like the minimal operational syntactic unit of the clause, is based on the concept of an EVENT (our equivalent term to 'proposition' and 'eventuality' in other frameworks). It uses categories that reflect idealised aspects of systemic functional grammar, so it is a 'systemic functional logical form' (SFLF).

Note, finally, that the relationship that it expresses is NOT one that holds between two concepts, but between two referring expressions, each with its SFLF structure, and each of which refers to a generic object.

To adopt this position leads to a reappraisal of the status of diagrams representing ontological relationships such as those in Figures 4 and 5. Those relations are, in the approach advocated here, simply one event type among many within the mass of beliefs that the system holds about dogs. In other words, the 'fact' (i.e. the confidently held belief) that dogs are mammals is just one of many things that the system assumes that it 'knows' about the generic object of 'dogs'.

7 Summary and conclusions

The COMMUNAL Project (Fawcett, Tucker and Lin 1993) has demonstrated the immense advantages that follow from having a unified system network for all meanings, whether realized grammatically or lexically. Thus Halliday's original 1961 insight was well-founded, the only major modification needed being that lexically realized meanings are not necessarily 'most delicate', in the sense of 'at or near the terminal leaves of the system network'. (Meanings realized in intonation should similarly be integrated into the overall network, and the way in which this is done in COMMUNAL is described in Fawcett 1980.) In other approaches some equivalent intralinguistic level of 'linguistic meaning' would surely be of similar value.

It would be interesting to compare the COMMUNAL 'division of labour' in modelling lexical meaning with other current proposals, such as those of Reiter and Pustejovsky, but that is beyond the scope of this paper. What is clear is that information about the purposes of entities, how they come into being, etc is all handled through events attached to the equivalent generic object in the belief system.

The two types of network - the system network of noun senses and the ontological relations in the belief system - serve different purposes. One is to generate nouns (and to control related structures such as those associated with participant roles) and the other is to facilitate reasoning. In the first it is preferable to avoid right-opening brackets - but in the second such 'cross-classifications' (or their equivalents in an 'event-based' logical form) are vital to certain types of reasoning. Since the two structures serve different functions, both have their place in a holistic model of language. (I say 'structures', because it is probably unhelpful, in the new framework, to think of the equivalents of ontological relations in the belief system as a 'network'.)

In the previous section we have seen how the COMMUNAL Project has begun to develop a new way of representing - and so thinking about - the relations found in standard ontologies. Since these are now seen as a part of the belief system, the suggestion is that we no longer need to think in terms of 'an ontology', in the sense of a taxonomy of 'concepts' - but rather in terms of relations between generic objects. We think that these fit more naturally into the more complex, event-based reasoning that we judge will be needed for the systems of the future. So perhaps there is no such thing as an ontology, but instead a set of mutually referring beliefs, each represented as an event?

If this approach is on the right lines, or even if it is only partly on the right lines, it is time to be moving on to the exploration of the next generation of belief systems. These will make different assumptions about the nature of the relationship of word forms to word senses, and word senses to the related aspects of the belief system - and perhaps too about the nature of 'ontological' relations.

Notes

1 The research reported here was supported by grants from DRA Malvern (under contract no. ER1/9/4/2181/23), from the University Research Council of International Computers Ltd, and from Longman, and by the University of Wales, Cardiff. I am grateful to my colleagues Gordon Tucker, Yuen Lin and Ulrich Gysel for the many useful discussions which have contributed to the emergence of the view presented here. The main debt is to Gordon Tucker; together we have hammered out the sense in which we now interpret Halliday's original challenging concept of 'lexis as most delicate grammar'. I am also grateful to the three anonymous reviewers of this paper; the remaining infelicities being, as always, the author's.

2 The term 'typic' has been used in preference to 'generic' in some of our earlier writings because we could give to this new term a has a clear definition - leaving 'generic' for 'generic (=

genre) structure'. Here 'generic' will serve equally well.

3 The fact that Leech could not propose a workable way to map the structures of language onto those that he inherited from standard logic is significant, but he can hardly be censured for failing to solve a problem that has also defeated all others who have attempted it.

4 There are occasions in natural language processing when it may be desirable and/or necessary to use system networks by working from right to left, e.g. as proposed (1) by Patten (1988) for some aspects of generation and (2) by O'Donoghue (1994) for the higher stages of parsing. But these are adaptations of the central concept for specific purposes, and they do not affect the main point that system networks are designed to be used in generation, working from left to right.

5 While Dahlgren has only 57 'category cuts', some of her terminal nodes such as 'animal' bring together a number of these and then lead off into further subsystems - and so effectively to further 'terminal nodes', such as 'vertebrate' vs, non-vertebrate', and then 'mammal' vs. 'bird' vs.'fish'. Thus there are in practice a few more than 57 'terminal nodes'. But this does not affect the more general point being made here.

6 Our experience in COMMUNAL - working at both the semantic and the belief system levels - suggests that the development of even quite delicate areas of the network raises problems requiring a high level of expertise and experience for a satisfactory solution, and we believe that this experience is mirrored elsewhere. One can foresee that clients might be wise to subcontract the work of ontology-building back to the Penman team.

7 Contrary to what is still a quite widespread assumption, there is no need for a separate network from that for the nominal group to deal with the rank of the word; the nouns at the head of a nominal group are simply the realization of one part of its meaning (see Fawcett, Tucker and Lin 1993).

8 These have ontological equivalents in the COMMUNAL belief system, so that interested persons may wish to compare this slicing of the universe of objects with those assumed for other ontologies. It would be an interesting exercise to get the creators of alternative taxonomies to explain their reasons for foregrounding their primary distinctions.

9 A pseudo-remedy that some systemicists use is to add complex wiring that prevents the realization of all but those combinations of features for which there is a realization (e.g. 'dog' plus 'female' but not 'hare' plus 'female'). But this to try to correct poor systemic modelling at too late a stage; it is the task of a system network to constrain possible co-selections.

10 The random generation of sentences is frowned upon by some NLG researchers, but it serves a valuable role in developing large generators, because it tests the availability of the lexicogrammatical resources, and so has a role in solving the 'expressibility' problem (Meteer 1992).

References

Bateman, J.A., 1990. 'Upper modelling: organizing language for natural language processing'. In *Procs of 5th International Workshop on Natural Language Generation*, Pittsburgh, pp. 54-61.

Bateman, J.A., Kaspar, R.T., Moore, J.D., and Whitney, R.A.. 1990. *A General Organization of Knowledge for Natural Language Processing:the Penman Upper Model.* Technical Report, USC/ISI Information Sciences Institute, Marina del Rey, California.

Berry, M., 1977. *Introduction to Systemic Linguistics, Vol 2: Levels and Links.* London: Batsford.

Butler, C.S., 1985. *Systemic Linguistics: Theory and Application.* London: Batsford.

Dahlgren, K, 1988. *Naive Semantics for Natural Language.* Boston: Kluwer.

Fawcett, R.P., 1973/81. 'Generating a sentence in systemic functional Grammar'. University College London (mimeo). Reprinted in Halliday, M.A.K., and Martin, J.R., (eds.) 1981 , *Readings in Systemic Linguistics.* Batsford, pp. 146-83.

Fawcett, R.P., 1980. *Cognitive Linguistics and Social Interaction: Towards an Integrated Model of a Systemic Functional Grammar and the Other Components of an Interacting Mind.* Heidelberg: Julius Groos.

Fawcett, R.P., 1990. 'The computer generation of speech with semantically and discoursally motivated intonation'. In *Procs of 5th International Workshop on Natural Language Generation*, Pittsburgh, pp. 164-73a.

Fawcett, R.P., 1993. 'Language as program: a reassessment of the nature of descriptive linguistics'. In *Language Sciences* 14.4: pp. 623-57.

Fawcett, R.P., 1994. 'A generationist approach to grammar reversibility in natural language processing'. In Strzalkowski, T., (ed.), *Reversible Grammar in Natural Language Generation.*, Dordrecht: Kluwer, pp. 365-413.

Fawcett, R.P., Tucker, G.H., and Lin, Y.Q., 1993. 'How a systemic functional grammar works: the role of realization in realization'. In Horacek, H., and Zock, M., (eds.), 1993, *New Concepts in Natural Language Generation,* London: Pinter, pp. 114-86.

Halliday, M.A.K., 1961. 'Categories of the theory of grammar'. In *Word 17*, pp. 241-92.

Hasan, R., 1987. 'The grammarian's dream: lexis as most delicate grammar'. In Halliday, M.A.K., and Fawcett, R.P. (eds.) 1987, *New Developments in Systemic Linguistics, Vol 1: Theory and Description*, London: Frances Pinter, pp. 184-211.

Hovy, E.H., and Knight, K., 1993. 'Motivating shared knowledge resources: an example from the PANGLOSS collaboration'. In *Proceedings of the IJCAI Workshop on Shared Knowledge*, Chambery, France.

Hovy, E.H., and Nirenburg, S., 1992. 'Approximating an interlingua in a principled way'. In *Proceedings of the DARPA Speech and Natural Language Workshop,* Hawthorne, NY.

Knight, K., 1993. 'Building a large ontology for machine translation'. In *Proceedings of DARPA Human Language Conference*, March 1993.

Leech, G.N., 1969. *Towards a Semantic Description of English.* London: Longman.

Leech, G.N., 1974/81. *Semantics.* Harmondsworth: Penguin.

Longman Group, 1978. *Longman Dictionary of Contemporary English.* London: Longman.

Lyons, J., 1977. *Semantics* (Vols 1 & 2). Cambridge: Cambridge University Press.

Matthiessen, C.M.I.M., and Bateman, J.A., 1991. *Text Generation and Systemic Functional Linguistics.* London: Pinter.

Meteer, M., 1992. *Expressibility and the Problem of Efficient Text Planning.* London: Pinter.

Miller, G., 1990. 'Wordnet: an on-line lexical database'. In *International Journal of Lexicography* 3(4) (Special Issue).

Nirenburg, S., and Defrise, C., 1992. 'Application-oriented computational semantics'. In Johnson, R., and Johnson, M., (eds.) *Computational Linguistics and Formal Semantics.* Cambridge: Cambridge University Press.

O'Donoghue, T.F., 1994. 'Semantic interpretation in a systemic grammar'. In Strzalkowski, T., (ed.), *Reversible Grammar in Natural Language Generation.*, Dordrecht: Kluwer, pp. 415-447.

Patten, T., 1988. *Systemic Text Generation as Problem Solving.* Cambridge: Cambridge University Press.

Reiter, E., and Dale, R., 1992. 'A fast algorithm for the generation of referring expressions'. In *Proceedings of COLING-92: 14th International Conference on Computational Linguistics* (Nantes). 1992. Morristown NJ: Bell Communications Research.

Rosch, E., 1978. 'Principles of classification'. In Rosch, E., and Lloyd, B.B., (eds.). *Cognition and Categorization.* New Jersey: Erlbaum.

Winograd, T., 1972. *Understanding Natural Language.* Edinburgh: Edinburgh University Press.