Technical Report

UCAM-CL-TR-45
ISSN 1476-2986

Number 45

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Compound noun interpretation problems

Karen Spärck Jones

July 1983

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/TechReports/

ISSN 1476-2986

Compound noun interpretation problems

Karen Sparck Jones

For the designer of an automatic language-processing program, non-lexicalised compound nouns are a problem. Even in specialised domains, where de facto lexicalisation may be fairly rampant, compound nouns are a problem; and they are much more so in ordinary discourse.

The first part of this paper briefly reviews the properties of compound nouns; the second considers what these imply for automatic language processing in general; and the third discusses the particular issues which arise in handling compound nouns in automatic speech processing. The object of the paper is to examine an important problem: it does not pretend to solve it; but while the problems of compound noun interpretation are largely bypassed in today's domain-specific language-processing programs, they will have to be tackled if more powerful and comprehensive programs are to be built.

1. Properties of compound nouns

It is well known that, in English at least, compound nouns can be freely constructed, generating units of, in some cases, surprising length. Certainly pairs of nouns are very common (e.g. "basket lid"), triples occur frequently ("staff tearoom pinboard"), and in 'technical' contexts especially even longer compounds are not unusual ("satellite radio link transmitter", "horse race apprentice training establishment"). When proper names figure, compounds may reach the amazing length of the Gleitmans' "Volume Feeding Management Success Formula Award" (Gleitman and Gleitman 1970; "Volume Feeding Management" is a name).

It is not in fact possible to maintain a principled distinction between lexicalised and non-lexicalised compounds, even within specialised universes of discourse. Some compounds are clearly lexicalised, as may be shown by their becoming single words ("tearoom"), developing meaning extensions having no reference to their underlying structure, etc. However, even those compounds canonised by entries in lexicons differ in the extent to which they are established, and are properly regarded only as representing one end of a spectrum from the firmly established to the totally novel. It may be convenient for the purposes of linguistic discussion to group compounds, and at the same time more satisfactory, to label compounds as established, non-established, or novel (as Warren 1978 does) rather than as simply lexicalised or non-lexicalised. However from both the formal and the programming points of view, a compound is either supplied with an explicit characterisation, as a unit, in a lexicon, or it is not. The problem of compound nouns is that as compounding is a highly productive process, any individual compound may not figure in any particular lexicon, so its meaning has to be constructed by the reader/hearer. From this point of view, a compound recalled as familiar by a human being, though it does not figure in an official lexicon, must be treated as

lexicalised; and equally, a second occurrence of a novel compound in a text may be treated as lexicalised with respect to a lexicon generated by that text. For the purposes of this paper, therefore, I shall simply divide compounds into the lexicalised and the non-lexicalised, in order to focus on the interpretation problems presented by the latter. (Of course a non-lexicalised compound may have a lexicalised constituent: but this will then be assumed to function like a single word.)

Interpreting compound nouns, i.e. providing a meaning representation for them, has three elements. The senses of the constituent words have to be identified; the structure, i.e. syntactic bracketing, of the group has to be determined; and the semantic relations linking the words have to be established. For example, given the noun string "verandah table sprays" in a particular context, say

> We need to do the decorations and flowers and things now: I'll do the verandah table sprays.,

it is necessary to identify the appropriate senses of words, for example 'supportive piece of furniture' rather than 'list of numbers' as the meaning of "table"; to determine that ((verandah table) sprays) rather than (verandah (table sprays)) is the syntactic bracketing of the group, i.e. that "verandah" modifies "table" and "sprays" modifies "verandah table" rather than "table" "sprays" and "verandah" "table sprays"; and to establish the underlying relationships between the words as 'sprays FOR tables IN verandah". However as this example, pushed a little further, shows, the general problem of compound nouns is that, like single words, they are typically ambiguous in isolation, in constituent senses and/or bracketing and/or linking relations, and that quite extensive contextual information may be required to disambiguate them. Thus in the example, the context is in fact quite compatible with sprays on tables, with table sprays for verandahs as opposed to grottoes, and indeed with water fountains rather than floral displays. As this suggests, carrying through the operations required to provide a full interpretation of compounds, given that they cannot be assumed to be simply derivable from the compound's constituents by 'internal' processes, may be a non-trivial matter. (There is no question that providing a meaning representation for a compound implies such full interpretation: thus, for example, we need to know what relations link the words in English compounds to be able to provide the explicit prepositional equivalents required for a French translation.)

Naturally, the three operations of identifying senses, determining bracketing, and establishing relations, are interrelated: the three processes are mutually selective: thus if furniture figures in house spaces and verandahs are house spaces then "table" means 'board', goes along with "verandah", and implies a relation IN. In practice it seems that in individual cases incomplete interpretations may be acceptable, in that discourse context does not supply enough information to select from alternative equally suitable readings, but that this does not matter as far as an adequate understanding of the text as a whole is concerned. For instance, "meter adjustment screw" may be bracketed either as (meter (adjustment screw)) or as ((meter adjustment) screw), and in one way by the speaker and the other by the hearer, without felt confusion in a conversation about finding or turning the screw. This does imply that each individual has a complete interpretation, but it is also possible, despite the claims that the immediate constituent

structure of compounds is always binary (see Warren, for example), that where there is no contextual requirement for it, some compounds are not fully interpreted by the language user: they may be treated as a simple coordinated string of elements. The same applies to lack of precision in the characterisation of senses and relations.

However these points do not remove the general requirements for sense identification, structure determination and relation establishment in compound processing, and indeed such operations will be involved in the interpretation of compound nouns even where full or refined meaning resolution is not called for.

Compound nouns may of course figure as constituents of complex nominals which include, for example, regular adjectives, as in "fresh verandah table sprays". But other types of constituent cannot in general be expected to contribute systematically to the reduction of interpretive effort for compounds. Indeed a compound like "dirty brown cat food container" shows how adjectives may contribute to ambiguity. In the other direction, the fact that elements of compounds may not be 'pure' nouns but, for example, may be deverbals, like "delivery" in "delivery boy", does not imply that they require different, or are amenable to easier, interpretive processes.

Clearly, it would be nice if the interpretation of compound nouns could be given a principled base or framework in terms of general rules for bracketing and relational construction (which would imply general constraints on sense selection), and attempts have indeed been made to provide such a framework. Some of these have been purely descriptive, for example Warren's investigative study, while others, like Levi's (Levi 1978), have been theoretically motivated in seeking derivational accounts of the formation of compounds. Downing's experiments in novel compound interpretation (Downing 1975, 1977) evaluated both proposed descriptive categorisations and theoretical explanations of compound noun structures.

These analyses have focused on the relations between the members of a compound, and have essentially been concerned with whether these relations are drawn from a finite, and relatively limited, list, and with whether the semantic category membership of nouns imposes constraints on the formation of compounding links between them. The theoretical accounts, like Levi's, seek to show that the presence of these relations, i.e. the fact that they can be 'read into' compounds, is explained by such processes as the deletion of underlying generalised verbs. From the point of view of automatic language processing, such theoretical accounts are of value first, if they can be exploited as the basis of an automatic analyser's meaning representation for compounds, e.g. the compound should be replaced by a construction with explicit pro-verb or pro-preposition, and second, if they can be used to provide guidance for the actual analysis process. However theoretical analyses like Levi's are wanting on both counts (see, for example, Downing's comments (Downing 1977)).

As noted, the main focus of attention in these studies has been the implicit semantic relations linking the compound's constituents. The mechanisms of constituent sense selection have attracted little attention, except where these may depend on word categorisations as a basis for linking, though the syntactic structure of compounds is discussed, for instance, by Warren. Claims have been made for short lists of general relations (or types of relation) underlying if not all

compounds, at least the great majority. Some justification for these claims is provided by the fact that the lists produced have also exhibited substantial overlaps, as is shown, for instance by a comparison between Downing and Warren's lists (their examples):

Whole-part	e.g.duck foot	
Half-half	giraffe-cow	
Part-whole	pendulum clock	
Composition	stone furniture	
Comparison	pumpkin bus	
Time	summer dust	
Place	Eastern Oregon meal	
Source	vulture shit	
Product	honey glands	
User	flea wheelbarrow	
Purpose	hedge hatchet	
Occupation	coffee man	Downing 1977, p.27

Source result	clay bird	
Copula	girl friend	
Resemblance	clubfoot	
Whole-part	spoon handle	
Part-whole	armchair	
Size-whole	22-inch board	
Goal-OBJ	moon rocket	(OBJ is Fillmore's case label)
Place-OBJ	coast road	
Time-OBJ	Sunday paper	
Origin-OBJ	engine noise	
Purpose	coffee cup	
Activity-actor	cowboy	Warren, p.237

However, as these lists also show, agreement on even very abstract relations expressible by generalised verbs like MAKE or prepositions like FOR is not complete. More importantly, the studies all show, and Downing's experiments particularly vividly illustrate the fact, that virtually all general remarks about compound nouns can only be remarks about tendencies and not about absolutes. The main exceptions are the universal rule that the compound head is the rightmost element and, subject to the caveat raised earlier, the rule that compounds have a binary tree constituent structure. Otherwise, though, for example, left nesting is more common than right in Warren's sample, there are plenty of right nesting examples. Similarly, though very many compounds can be adequately characterised in relational terms using such headings as those illustrated, there are many exceptions; the same applies to noun categories constraining participation in relations. Thus for example if, following Downing, we have the semantic classes Animal and Synthetic Object for head nouns, and find the relation COMPARISON much more common for the former and PURPOSE for the latter, we may still find PURPOSE with Animal heads, as in "beef cattle". Further, as Downing notes, in some cases where general relations can be claimed as applicable, they may be insufficiently indicative of the specific relation between the constituents, which is not, either, simply suggested by, or derived from, the meanings of the linked words.

The main contributions of these linguistically-oriented studies to the design of automatic language analysers are therefore generalisations about sense constraints, syntactic structures, and link relations which can be preferentially matched on text compounds, but which cannot be guaranteed to apply to any particular compound; thus a program's list of general relations, however relevant its members may be for many inputs, does not delimit a set of possible relations, and indeed fails to define the relations holding in a non-trivial number of cases.

2. Automatic compound noun interpretation

The question we now have to consider is just exactly how the interpretation of a compound is established, given that in general, it is not simply derivable from the meanings of its constituent words. There is nothing intrinsic to the meanings of "cardboard" and "system" which tells us that "cardboard system" means 'system for making cardboard', since it could equally well mean 'system made of cardboard'. This is the crucial question for the language-processing program builder which the linguist does not consider.

Interpreting an identified compound is, moreover, not the only problem that compound nouns present in automatic text processing; and it is important to recognise that the processing implications of compounds are not only those of interpreting the compound itself. This is where the problem context supplied by the computational task throws a very different light on compounds from that directed by linguists seeking to describe what they have themselves already understood. Thus one element of the automatic processing problem is that of recognising that some string of words is actually a string of nouns, and that a phrase like "envelope boxes" is not a verb followed by a noun. The (syntactic) categorial ambiguity of many English words makes the sheer recognition of compound nouns an additional burden in analysis.

Processing compound nouns thus implies not only interpreting compounds to provide the explicit meaning representation for them that is the normally required output of an analyser, but also recognising their occurrences, on the very many occasions that they occur in English text. This in turn implies that the treatment of compounds is embedded in, and interacts with, the various processes applying syntactic, semantic and pragmatic information involved in text analysis (whether this is narrowly conceived as parsing, or more broadly as 'understanding'). These processes themselves are mutually interdependent, and can involve not only simple mapping, but more comprehensively inferential, operations.

Given that the 'internal' operations for compound interpretation are themselves interlocked, and must at the same time interact with the 'external' operations on the text surrounding a compound, it is evident that handling compound nouns may not be a simple exercise. Indeed it is in fact the case that interpreting compound nouns can require arbitrarily extensive inference, and further, inference on pragmatic information, i.e. the most exigent form of analyser process. Inference on other types of information, e.g. semantic, could well be required, but the argument, to be illustrated below, is that inference on a body of information which appears to be more extensive and complex than semantic information, namely world knowledge, is sometimes needed.

Moreover, it may be needed for compound interpretation even in supposedly favourable circumstances. Thus even where the senses of the words concerned are quite standard; the bracket structure is quite straightforward, for example because the compound has only two nouns; and most importantly where the linking relation is a 'regular' one, i.e. is in fact a common, standard one like those illustrated earlier, pragmatic inference may be required to supply the compound's interpretation. Moreover this may be the case even where the final meaning representation for the compound need be no more than 'minimal', as might be the case in translation where a 'filled-out' representation, for example explicitly supplying the referents of noun phrases, was not systematically required. Thus pragmatic inference may be called for in apparently less exigent text-based activities, like translation, as well as in more exigent content-based activities, like question answering. Further, pragmatic inference may be needed even where the local text context provides rich information bearing on the interpretation of the compound. Finally, pragmatic inference may be required for any of the three compound noun operations, on senses, groups, and relations (pretending that these are separable), but especially to establish the meaning relations linking a compound's constituents.

The need for inference, and specifically for pragmatic inference, can be shown by an example. This deliberately illustrates the favourable case, so the requirement in inference is simply to select the correct relation from alternative 'regular' relations. When to call inference in processing a sentence is clearly an issue for analyser design, and is considered below: for the present we can simply consider the relative contributions to compound noun interpretation which can be made for the example case by syntax, semantics and pragmatics respectively, treating these separately for the sake of clarity, and in the obvious order. The detailed mechanisms involved in each case are of course themselves also issues; however for the purposes of illustration plausible indicative mechanisms are sufficient.

Consider the sentences:

A. It's the fifth hedge bush that needs replacing.

B. This hedge bush is priced at £5.

In the obvious real life contexts "hedge bush" in A means 'bush in hedge' and "hedge bush" in B means 'bush for hedge'. That is, we have underlying general relations of the 'regular' type listed earlier, which we may label LOCATION and PURPOSE respectively, and which we may say are representable by the pro-prepositions IN and FOR. How, then, do we establish that LOCATION is the relation for A, and PURPOSE for B?

In this example syntax does not have to do much, as the compound is only a two-word one, beyond marking the rightmost element as the compound head. As noted, syntax will not generally select the correct bracketing for compounds with more than two words, as in

The hedge bush customers like variegated privet.,

where we can have either ((hedge bush) customers) or (hedge (bush customers)). The problem of demarcating compounds is brought out by a comparison between sentences like this and slightly more complex ones. In a simple declarative sentence with one noun phrase to the left of the

main verb, the string of nouns involved, like "hedge bush customers", forms a single unit. However in non-simple sentences with more than one noun phrase to the left of the first verb, we cannot guarantee even to find the heads of all the noun phrases if there are more than two nouns. For example, given

The hedge bush customers like costs £10.,

we can only be sure that "customers" is the noun head for the subsidiary verb "like", but we do not know, because of bracketing ambiguity, whether "hedge" or "bush" is the head for "costs". (But notice that the categorial ambiguity of "like" allows an alternative syntactic parsing with a four-noun compound analogous to "glue factory employees trip".)

Of course syntax can get rid of some word category options, for instance "hedge" as a verb in these cases; but as the example shows, syntax will typically not do anything for sense selection or relation derivation, and indeed cannot be expected to do so directly, even if these are well-related to any bracketing, because they are not syntactic matters.

Turning now to the contribution of semantics, and specifically of semantic pattern mapping using, for example, semantic primitives, this may, but will not necessarily, give us sense selection and bracketing. For illustrative purposes we assume a Wilks-style (Wilks 1975) preferential approach (heavily simplified). Thus suppose we have dictionary entries as follows, with semantic primitive formulae defining each word sense. For the senses of "hedge" and "bush" relevant to both sentence A and sentence B we have

bush1 (PLANT) 'plant'

hedge1 (GOAL((WHERE PART)BOUND))
((PLANT HAVE)(SUBJ(LINE STRUCT))) 'row of plants',

and for other senses of the words we have, say

bush2 ((THING POSS)((IN PART)STUFF)) 'lining'

bush3 (WORLD PART) 'maquis'

hedge2 (SUBJ(LINE STRUCT)) 'abstract boundary'.

Pattern matching exploiting the commonality of PLANT could pick up and prefer (which we will treat as unequivocally select) 'bush1' and 'hedge1' for both sentences. It is, however, easy to see cases where we would not get such clear preferences, for example where there are more than two nouns, as in "hedge bush world".

But semantics as illustrated can do nothing to identify LOCATION and PURPOSE as the relations for sentences A and B respectively. The source for these derivations in any pattern-mapping rules which would explicitly construct the relations as their outputs would have to be contained in the formulae for the words concerned. But the candidate source of this information here, HAVE in the formula for 'hedge1', is ambiguous with respect to LOCATION and PURPOSE, so there is no selection of relations by pattern matching. How, therefore, is this to be achieved?

For the purposes of the example, we assume that this is the only question about relations, and that no help in the selection is

forthcoming from elsewhere in the sentence through pattern mapping. Thus if we assume that we began semantic analysis with a list of 'regular' semantic relations, including, for instance,

PART-WHOLE
PRODUCT
LOCATION
ACTIVITY
PURPOSE,

we assume here that semantic pattern mapping can eliminate all of these except LOCATION and PURPOSE as possible relations for "hedge bush", either because they have no lexical sources in the sentences, or because their sources are less preferred (i.e., for the purposes of the present argument, rejected). We also assume that pattern matching over the rest of the sentence in each case does nothing to distinguish LOCATION and PURPOSE: this cannot be demonstrated here, but is sufficiently plausible, as the further discussion of the example below indirectly suggests.

To decide which of LOCATION and PURPOSE applies to each of sentences A and B, i.e. is at least the most preferred reading, inference is required, and specifically pragmatic inference. The distinction between semantic and pragmatic inference is not a clearcut one, indeed is problematic. Hence it could be argued that the detail which follows illustrates only semantic inference, and not pragmatic inference. But I believe that the knowledge in question is primarily, or mainly, pragmatic knowledge about the real world which is not in any obvious sense contained in the meanings of the words involved; it is factual knowledge about the referents of the words, not linguistic knowledge about the words. However even if a case could be made for the information involved being semantic rather than pragmatic, it must be emphasised that the procedures involved in manipulating it are indubitably inferential. Thus any direct pattern mapping operations on pragmatic information would not be adequate either.

The pragmatic inference processing necessary for the full interpretation of "hedge bush" for the two example sentences can only be shown very schematically, with many simplifications for convenience and with many important matters left unexamined. Thus we assume 'global' world knowledge facts, 'local' facts tied to specific concepts for which individual words are triggers, and production-style inference rules supporting conclusions of varying degrees of certainty, indicated here in abbreviated and standardised form by the occurence of P (for "possibly/probably") on the rule. Thus for example the system's global facts include

G1 X in outdoor location deteriorates

and the local facts, associated with the lexical entries for word senses, include

L1 'hedge' delimits open space

L2 'hedge' is structure of plants

L3 'bush' is plant.

(The single quotes are used here simply to flag the conceptual correlates of the key source text words.) The rules include:

```
R1   X needs renewal P==> X has deteriorated

R2   Y delimits space P==> Y in outdoor location

R3   Y is structure of X P==> X is part of Y

R4   X is part of Y  )
   & X in Z location ) P==> X is in Y,
   & Y in Z location )      in -->> LOCATION

R5   X has price P==> X is for sale

R6   X is for sale P==> W wants to buy X

R7   W wants to buy X P==> W wants to use X

R8   X is part of Y P==> W can use X to form Y

R9   W can use X to form Y  )
   & W wants to use X      ) P==> W wants to use X for Y,
                                  for -->> PURPOSE.
```

Then to establish that LOCATION is the relation relevant to sentence A, and necessarily concentrating on "hedge" and "bush" and hence allowably in medias res as far as the rest of the sentence is concerned, we assume that an initial or derived representation of "replacing" has already been obtained as 'X needs renewal'. (Note that, whatever the source of this representation, i.e. word sense formula or local fact, it has to be given a form appropriate to the pragmatic inference operations.) Then, making substitutions for variables via specific word-derived triggers like 'hedge', 'bush', etc. in appropriate facts and rules (indicated at the left), we proceed as follows, to emerge with 'bush LOCATION hedge':

```
R1  X needs renewal      P==> X has deteriorated
    X = 'bush'           P==> 'bush' has deteriorated

G1  X in outdoor location deteriorates
    X = 'bush'           P==> 'bush' in outdoor location
                                has deteriorated

L1  'hedge' delimits space

R2  Y delimits space     P==> Y in outdoor location
    Y = 'hedge'          P==> 'hedge' in outdoor location

L2  'hedge' is structure of plants

L3  'bush' is plant

R3  Y is structure of X P==> X is part of Y
    Y = 'hedge', X = 'bush'
                         P==> 'bush' is part of 'hedge'

R4  X is part of Y  )
  & X in Z location )    P==> X is in Y,
  & Y in Z location )         in -->> LOCATION
    X = 'bush', Y = 'hedge' Z = outdoor
                         P==> 'bush' is in 'hedge',
                              -->> 'bush' LOCATION 'hedge'.
```

In this case, we are driving the inferential chain to construct the relation by the outdoor idea. It must of course be emphasised that

the chain given is the ultimately triumphing one: many other possible paths will have been started and abandoned in searching the space of available facts and rules.

To get PURPOSE for sentence B, we assume we already have representations for "is priced at" as 'X has price', and achieve the relational interpretation as follows, relying on the idea of means:

```
R5  X has price              P==> X is for sale
    X = 'bush'               P==> 'bush' is for sale

R6  X is for sale            P==> W wants to buy X
    X = 'bush'               P==> W wants to buy 'bush'

R7  W wants to buy X         P==> W wants to use X
    X = 'bush'               P==> W wants to use 'bush'

R8  X is part of Y           P==> W can use X to form Y
    X = 'bush', Y = 'hedge'
                             P==> W can use 'bush' to form 'hedge'

R9  W can use X to form Y )
  & W wants to use X      ) P==> W wants to use X for Y,
                                 for -->> PURPOSE
    X = 'bush', Y = 'hedge'
    X = 'bush', Y = 'hedge' P==> W wants to use 'bush' for 'hedge'
                                 -->> 'bush' PURPOSE 'hedge'.
```

No claim is made for these specific rules, and alternatives, e.g. a simpler R4, could easily be proposed. However these are the kind of rules that actual artificial intelligence systems depend on, and carping at individual rules does not subvert the principle illustrated. The example is also, as noted, heavily simplified, in that much detail is necessarily involved in particular in the manipulation of the lexical trigger characterisations, representative of local facts, to link them to rule characterisations - the example simply and completely unrealistically assumes direct verbal identity, and in the application of the whole body of rules, of course including many more than those listed, not only to a single input sentence characterisation, but to alternative characterisations. There are major issues of knowledge representation and search control which are deliberately disregarded here. The important point about the illustration is rather that to establish compound noun meaning representations, we need inference. Even if less rules, perhaps only one, are applicable in each case, they are still inference rules. Moreover it is very likely that to achieve sufficient flexibility in inference we will have to work with generally-applicable rather than specially-applicable rules, i.e. with less clearly focused rules, and hence with longer rather than shorter chains to make connections.

It is also necessary to point out that it is no use assuming that one could avoid the need for explicit inference by relying on world knowledge structures like frames, referring to particular universes of discourse and effectively 'canning' many individual pragmatic connections, so that fitting 'hedge' and 'bush' into different frames for the two sentences A and B would automatically select the relevant linking relations. Even if there were, say, GARDENING and NURSERY frames, it is not obvious that these would exclusively select for sentences A and B

respectively: that is, it is not the case that LOCATION and PURPOSE are specific to those two frames. Both could hold for a single frame: thus NURSERY would be the frame for the sentence

> We're looking for hedge bushes because two of our hedge bushes have died.,

so even if the frame simplified other interpretive operations, inference would still be required to select the different relations linking "hedge" and "bush" in the two occurrences of "hedge bushes" in this sentence. Pragmatic inference, in other words, is still needed.

As noted earlier, the requirement for pragmatic inference in compound interpretation clearly raises the issue of analyser design: what sort of structure should a language analyser have? If inference, and especially pragmatic inference, may be required to interpret compounds and thus, in the limit, even to demarcate them, what follows for the way in which syntactic, semantic and pragmatic processing is organised within the overall framework of the analyser? Simple analyser designs tend to be sequential, with syntactic processing followed by semantic processing, and this in turn by pragmatic processing, with inference in practice primarily, though not necessarily, associated with the latter. This is the classic 'three-box' model, tacitly assumed in the example just considered. (It is ordinarily not justified by any claims about psychological modelling.) A model like this assumes that not calling inference (at least extensively) until the pragmatic stage will not inhibit syntactic and semantic processes altogether, even though they cannot be guaranteed, or even expected, to deliver complete interpretations.

Thus the syntactic processor could deliver compound nouns as a string, effectively equivalent to all possible alternative bracketings, and the semantic processor would seek whatever sense selection could be achieved directly by semantic pattern mapping on possible bracket structures, and by any inference processes exploiting semantic information. These semantic operations would, it should be noted, have to provide not merely for sense selection and the indication of any consequent preferred bracket structures, but for the explicit construction of any linking relations justified by the available semantic information. But semantic processing could not be guaranteed to deliver a unique, or single preferred, interpretation, and further pragmatic pattern mapping could still not resolve the compound even for regular relations, and in general not for idiosyncratic ones. Depending on pragmatic inference for the final resolution would therefore imply that a serial analyser would be carrying forward alternatives until the last stage of processing, with all the overheads that this involves.

The trouble with sequential processing is that it does not achieve enough in intermediate steps, so that many options are necessarily held open along the line. On the other hand, attempting to do everything in parallel, so that, for example, syntactic parsing and pragmatic inference are working together, imposes a heavy overhead in control complexity, and can in practice lead to 'lockstates' where a pair of operations are looking for one another's outputs as inputs. For the builders of analysis programs, the question of when to call inference, and particularly pragmatic inference presupposing a large knowledge base, is an extremely serious one. This is emphasised by the fact that we

cannot necessarily assume, as we did in the example, that establishing the relation between the members of a compound, even when it involves inference, can be tidily organised to try a list of regular relations first, and then only if applying these fails, to consider the possibility of some more idiosyncratic relation.

3. Compound noun interpretation in speech processing

If we now consider automatic speech processing, what difference does this make to the analysis of compound noun processing requirements just presented for written text? Specifically, does speech processing present any new problems, and does it remove any old ones?

There is no doubt that prosody influences compound interpretation, just as it contributes to the interpretation of any other discourse elements. It has been claimed, for instance, that stress is always on the first constituent, but the reality seems to be more complex than this, though still rule-based (see Cooper and Paccia-Cooper 1980). The difficulty is that there are a number of specific prosodic mechanisms, and equally a number of linguistic functions for which prosodic mechanisms may be exploited, with, so far, no clearly established correlations between specific mechanisms (either alone or in combination) and functions. What seems fairly clear is that for compound nouns, the contribution of prosody is to assist with bracketing, including the bounding of the entire unit, and that prosodic mechanisms contribute only indirectly, and in a very limited manner, to the selection of senses and derivation or construction of relations. Further it seems at least probable that prosody, while eliminating many possible bracketings, cannot be guaranteed to deliver a single bracketing.

Thus in comparison with written text, speech provides additional clues for the interpretation of compounds. However to set against this, speech calls for extra processing effort in the identification of segment limits, especially at the phrase and word levels, i.e. in the segmentation of the continuous acoustic signal and evaluation of the resulting elements.

In automatic speech recognition little progress has been made in developing and applying models of prosodic mechanisms for purposes other than phrase and word recognition. The treatment of compound nouns in automatic speech programs can therefore currently rely on no more information than is available for written text.

The other side of the coin is indeed of more significance. This is the effect of compound nouns on the identification of the speech units. The general claim made by participants in the ARPA Speech Understanding Research Project (see, for example, Woods et al 1976) was that unit identification is materially assisted by the application of top-down knowledge, for example syntactic and semantic knowledge embodying constraints on expectations about the occurrences of words. The problem thus presented by compound nouns is: what kind of contribution can they make to the supply of information for top-down application in interpreting the input speech signal? We saw earlier that the general property of non-lexicalised compounds is their relative unpredictability, and especially their 'pattern unpredictability'. This suggests, therefore, that information given by constituents of compounds is likely to be a relatively weak source of predictive constraints on the

verbal identification of other members of the compound and hence, indirectly, of constraints on the identification of other words in the sentence.

Or rather, it appears that more effective predictions can only be provided as part of an extensive inferential process, and more importantly, cannot be made until an entire compound, including its head, is available. Even then, the complexity of carrying out inferences using only tentatively-identified words, in an exceptionally exigent analysis by synthesis process, is manifest.

The completeness requirement on the compound does not arise if the speech interpretation is done autonomously, and not in concurrent real time. In this case, illustrated by ARPA projects (e.g. Woods et al), top-down information can be applied 'simultaneously' over the whole input string. (The complexity involved in doing this with compounds of course remains.) However if we take the long-term need for concurrent speech-processing programs seriously, compound nouns are manifestly a challenge.

This is clearly shown by reference to such investigations of human speech processing as Marslen-Wilson and Tyler's (Marslen-Wilson and Tyler 1981). Summarising a series of experiments, Marslen-Wilson and Tyler argue that human beings show abilities to recognise words correctly, even before thay have been completely uttered, which necessarily imply not only that extensive top-down work is being done on the speech signal as it arrives, but that this work involves inference, including pragmatic inference.

This argument would appear to imply that compound noun interpretation can be carried out in an essentially predictive, i.e. strictly expectation-driven, manner, relying only on long-term memory and local context. This is very strong, perhaps too strong to accept. Thus if we consider a compound noun variation on the experiment summarised on p.325 of Marslen-Wilson and Tyler, namely the text beginning

> John is writing this amazing program. The program /stɑɹ .../...,

it is much less clear that there are only two 'obvious' candidate word completions for /stɑɹ ..., in the way that there are two obvious candidates, on context meaning grounds, for /stɑɹ ... given Marslen-Wilson and Tyler's text

> John was trying to get some bottles down from the top shelf. To reach them he had to /stɑɹ .../...

and the candidate list

stag	stagger	stack
stalactite	stagnate	stand
stamina	stammer	
stance	stamp	
standard	stampede	
standoffish	stab	
static		
statistic		
statue		
stature		
statute		

stanza,

namely "stack" and "stand". There is indeed only one candidate completion for /stœn.../, i.e. when the additional nasal has been uttered, namely "stand". In the compound example case, and considering only completing nouns, all of

stack
stacks
stacking
stamping
standard
standards
standing

are plausible candidates for /stœ..., and would be accepted as such, as it were post hoc, by the participants in a discussion of programs, and all of

standard
standards
standing

are completion candidates for /stœn..., as the example list of texts below shows:

The program stack is held in ...
stack checker works by ...
stacks are stored by ...
stacking routine ...
stamping ground for novices is ...
standard is so high ...
standards panel is going to have ...
standing committee will never ... .

The point is that it is only the complete word itself, or at least something much more like it than in Marslen-Wilson and Tyler's example, that supplies the information the interpretation of the text containing it requires. This interpretation may well involve inference for the given compound elements, as indicated earlier, and in accordance with Marslen-Wilson and Tyler's general invocation of inference. However the point to be emphasised here is that, even with inference, speech processing may not be real-time deterministic, or even nearly so, but that there may be significant local non-determinism. Marslen-Wilson and Tyler's argument is indeed not for determinism, but for top-down processing seeking to constrain selection from the acoustic alternatives as much as possible. But with compounds the ability to apply top-down constraints may be very limited or even completely inhibited. Thus the need to wait until the complete compound has been received before being able to carry through the semantic and pragmatic processing is well illustrated by a comparison between

The program stack checker ...
The program stack checking mechanism ...
The program stack ckecking outcome ...,

even where prosody would indicate which of the alternative bracketings for each was being initiated, for example for the second

(program((stack checking)mechanism))

or

((program stack)(checking mechanism))

or

.....

The problem which then arises is that of achieving useful word identification within the string. More generally, the complexity of combining, and, as necessary, switching between, top-down and bottom-up processing of the speech signal is shown by reference to the fact that "The program /stæ ..." could be completed by verbs as well as nouns, as in

The program stacks variable values in ...
staggers about the database like ...
stands or falls by ... .

Conclusion

It is clear that characteristing compound noun interpretation processes, and further, modelling them sufficiently precisely to drive language processing, and even more, speech processing, programs, is a major enterprise. Moreover the fact that specifically pragmatic inference may be arbitrarily required for compound interpretation, irrespective of whether it is also required for other interpretation functions, clearly raises the issue of language-processing system design, even in relation to efficiency, and without any reference to cognitive modelling. Specifically, in the 'three box' model, the fact that compounds cannot be fully resolved until the final step implies the possibility of at least local uncertainty, and hence of carrying implicit or explicit alternative analyses through successive processing steps. Moreover, it is far from obvious how the necessary pragmatic inference processes are to be supplied, even if they can be acceptably isolated in a pragmatic module.

The business of testing proposed pragmatic knowledge bases and inference operations in the context of analysis programs allowing more dynamic interaction between pragmatic and other operations, with the aim of reducing input uncertainty by bringing as much, different, information to bear as soon as possible, is therefore not just the computational linguist's challenge, but his nightmare.

Acknowledgement

This work was supported by the U.K. Science and Engineering Research Council. I am grateful to Bran Boguraev, Ted Briscoe and John Tait for discussions.

References

Cooper, W.E. and Paccia-Cooper, J.P. 1980 Syntax and Speech. Cambridge, Mass.: Harvard University Press.

Downing, P.A. 1975 Pragmatic Constraints on Nominal Compounding in English. Master's Thesis, University of California at Berkeley.

Downing, P.A. 1977 "On the creation and use of English compound nouns."

Language 53, 810-842.

Gleitman, L.R. and Gleitman, H. 1970 Phrase and Paraphrase: Some Innovative Uses of Language. New York: Norton.

Levi, J.N. 1978 The Syntax and Semantics of Complex Nominals. New York: Academic Press.

Marslen Wilson, W.D. and Tyler, L.K. 1981 "Central processes in speech understanding." Philosophical Transactions of the Royal Society of London B 295, 317-332.

Warren, B. 1978 Semantic Patterns of Noun-Noun Compounds. Gothenburg Studies in English 41, Göteborg: Acta Universitatis Gothenburgensis.

Wilks, Y.A. 1975 "A preferential, pattern-seeking, semantics for natural language inference." Artificial Intelligence 6, 53-74.

Woods, W.A. et al 1976 Speech Understanding Systems. Report 3438, 5 vols., Cambridge, Mass.: Bolt, Beranek and Newman Inc.