Mixed-Depth Representations for Natural Language Text

Graeme Hirst and Mark Ryan
Department of Computer Science
University of Toronto
Toronto, Canada M5S 1A4 gh@cs.toronto.edu; (416) 978-8747; fax (416) 978-1455

Courier deliveries: Artificial Intelligence Group, Room 283,
D.L. Pratt Bldg, 6 King's College Road, University of Toronto

To appear in *Text-based intelligent systems*, edited by
Paul Jacobs, Hillsdale, NJ: Lawrence Erlbaum Associates, 1992.

February 1992

1 The Limitations of Surface Representations

Text understanding is usually hard for computers and easy for people, so we tend to forget about the times when it's hard for people, too. But even smart, knowledge-based people, and not just dumb computers, can find text understanding extremely difficult. We all know this from our experiences with writing that presents complex ideas—advanced technical papers, for example—and writing that's just plain bad—incomprehensible instructions for assembling a Christmas toy, textbooks that present ideas sloppily, government tax-return guides that try hard to be clear but never quite succeed.[1] So it's no shame if a natural language understanding program, like a human, has to occasionally capitulate and say, in effect, that it cannot fully understand some difficult piece of text.

Now, intelligent text-based systems will vary as to the degree of difficulty of the texts they deal with. Some may have a relatively easy time with texts for which fairly superficial processes will get useful results, such as, say, *The New York Times* or *Julia Child's Favorite Recipes*. But many systems will have to work on more difficult texts. Often, it is the complexity of the text that makes the system desirable in the first place. It is for such systems that we need to think about making the deeper methods that are already studied in AI and computational linguistics more robust and suitable for processing long texts without interactive human help.

A domain that demonstrates some of the most difficult problems is that of searching legal cases for relevant precedents. What makes a legal case a precedent does not necessarily have anything much to do with the domain of the case, but rather the structure of the argument.[2]

Our colleague Judith Dick [1987, 1991] has developed methods for the representation of the texts of judicial decisions in order to permit conceptual searches. Dick's representations are based on Toulmin's [1958] model of argument structures, Sowa's [1984] Conceptual Graph formalism, and Somers's [1987] system of thematic relations, which she has revised and extended. One important

[1] "You write with ease to show your breeding, / But easy writing's curst hard reading."—Sheridan.

[2] "There is a story of a Vermont justice of the peace before whom a suit was brought by one farmer against another for breaking a churn. The justice took time to consider, and then said that he had looked through the statutes and could find nothing about churns, and gave judgment for the defendant. The same state of mind is shown in all our common digests and textbooks. Applications of rudimentary rules of contract or tort are tucked away under the head of Railroads or Telegraphs or go to swell treatises on historical subdivisions, such as Shipping or Equity, or are gathered under an arbitrary title which is thought likely to appeal to the practical mind, such as Mercantile law." (Oliver Wendell Holmes, "The path of the law" (1897). In: MacGuigan, Mark R. *Jurisprudence: Readings and cases.* University of Toronto Press, 1966, 48–62.)

sub-goal of the work is just to find what problems arise in representation that cannot be addressed by conventional techniques. For example, present-day formalisms have great difficulty representing entities whose existence is not definite; but many legal texts are, in fact, discussions of whether some entity (usually an abstract entity such as intent or liability) does or doesn't exist. We have addressed this topic in detail elsewhere [Hirst 1989, Hirst 1991].

2 Mixed and Partial Representations

The dilemma is that on one hand, we have the limitations of raw text databases and superficial processing methods; on the other we have the difficulty of deeper methods and conceptual representations. Our proposal here is to have the best of both, and accordingly we develop the notion of a heterogeneous, or mixed, type of representation.

In our model, a text base permits two parallel representations of meaning: the text itself, for presentation to human users, and a *conceptual encoding*[3] of the text, for use by intelligent components of the system. The two representations are stored in parallel; that is, there are links between each unit of text (a sentence or paragraph in most cases) and the corresponding conceptual encoding. This encoding could be created en masse when the text was entered into the system.[4] But if it is expected that only a small fraction of the text base will ever be looked at by processes that need the conceptual representations, then the encoding could be performed on each part of the text as necessary for inference and understanding to answer some particular request. The results could then be stored so that they don't have to be redone if the same area of the text is searched again. Thus, a text would gradually *grow* its encoding as it continues to be used. (And the work will never be done for texts or parts of texts that are never used.)

So far, this is straightforward. But we can go one step further. The encoding itself may be deep or shallow at different places, depending on what happened to be necessary at the time it was

[3]What we have in mind at present is close to a conventional first-order AI knowledge representation, modified as we describe below; but we use this deliberately 'neutral' term because we don't want to prejudice consideration of other possibilities.

[4]The development of interlingual methods of machine translation also offers interesting possibilities here. Many text bases—laws and regulations in multilingual jurisdictions such as Canada and the European Community, for example—have to be translated anyway. The interlingual representation generated as a by-product of machine translation could be retained and stored in the text base along with the surface text and its translations. Although probably shallower than a regular AI representation, and possibly even erroneous in places, this representation might well serve as a conceptual encoding for many purposes, or as a first step to a deeper encoding.

generated—or on what was possible. Or, to put it a different way, we can view natural-language text and AI-style knowledge representations as two ends of a spectrum.

Our goal is the development of a conceptual representation of text that encompasses this entire spectrum—that is, the representation includes natural language, a first-order (at present, anyway) formalism, and mixtures of both of them at various levels of concentration. Such an encoding could tolerate vagueness, but permit precision. For example, the representation of a sentence might have a piece of natural language embedded in a logical form, or vice versa. Thus, anything that needs (or permits) only partial interpretation or disambiguation may stay in its partially encoded form until dealt with later (if ever).

Some of the aspects of language understanding that would be optional in this representation are: lexical disambiguation; marking case relations; attachment of modifiers of uncertain placement; reference resolution; quantifier scoping; and distinguishing extensional, intensional, generic, and descriptive noun phrases. If all optional parts are omitted, then one just has the surface string. If all are included, one has a full-blown AI-style knowledge representation. Anything in the middle would also be perfectly legal, and be acceptable to the inference and search processes that use the representation. In practice, one would expect to at least include some syntactic markings or bracketings.

For example, if such a system were given a sentence such as *John bought a vxzfl* (where *vxzfl* represents some word or phrase that the system simply can't understand), it should be able to answer the question *Did he pay for it? (Yes)*, and pass gracefully on *Could he eat it? (I don't know)*.

In summary, there are three separate proposals included in this model:

- The usefulness of partial or incomplete encodings.
- Surface natural-language forms as a special case of an incomplete encoding.
- Mixing together encodings of different levels of completeness.

Aspects of the first two of these have already been suggested in various forms in the literature; the third, we believe, is novel. The following sections will discuss each in turn.

3 Incomplete Encodings

The first proposal is that if it is not possible to create the conceptual encoding of a piece of text (a sentence or other fragment), because there is insufficient information, because the text is somehow ill-formed, or simply because it looks like it would require too much work, one may usefully create a *partial* or *incomplete* encoding.

Recognizing word senses and thematic relations. At the lexical level, incomplete encoding means incomplete lexical disambiguation. In earlier work [Hirst 1987, Hirst 1988a, Hirst 1988b], we described the *Polaroid Words* system for lexical and thematic disambiguation. In this system, a variety of information sources, including semantic associations and selectional restrictions, were used to eliminate potential readings of an ambiguous word or case marker until just one was left. If a unique meaning could not be determined, the system was unhappy. Our present suggestion is that it need not be unhappy, but simply report a list of the remaining possible meanings (i.e., a disjunction of possibilities). And if a word that the system encounters is not even in its lexicon, the surface form, marked as such, may be retained.

Syntactic analysis. At the syntactic level, incomplete encodings can be used when it is not possible to determine a unique parse for part of a sentence. Such a situation occurs, for example, when a modifier such as a prepositional phrase or relative clause has two (or more) seemingly permissible points of attachment, and when lexical ambiguity permits more than one distinct structure for a clause (as in *Time flies like an arrow*).

A number of writers have suggested methods for the representation of the set of choices in such multiple parses. For example, Church's YAP parser [1980] was able to 'pseudo-attach' nodes; that is, if a unique parent couldn't be determined, the node kept a list of the possibilities. Seo and Simmons [1989] propose the more general method of 'packed parse forests', graphs that compactly represent a number of trees. An algorithm for finding each tree in the forest is given.

Metzler *et al.* [1989] suggest a slightly different kind of partial representation for a parse tree. Rather than even trying to decide where a prepositional phrase or the like might be attached, their 'Constituent Object Parser' simply assumes attachment to the right-most available node. For Metzler *et al.*'s application, which is information retrieval by matching the dependency trees created by the parser, the consequent imprecision is easily tolerated (the same simplifying assumption being made for both the target text and the query).

Except for Metzler *et al.*'s, the expectation is usually that such representations are just a stage on the way to choosing one of the options. But if the choice must be (indefinitely) postponed, a system could continue to use the incomplete representation as is—provided subsequent processes so permit. We will discuss this point in section 5.

Semantic interpretation. Most of the time, any incompleteness at the lexical or syntactic levels will give rise to a corresponding incompleteness in semantic interpretation.[5] And of course even a representation that is complete at the lower levels can give rise to an incomplete semantic interpretation.

Perhaps the most important ambiguity at this level is in the scope of quantification. Alshawi and van Eijck [1989] have described a representation in which quantifier scoping and certain modifiers are unresolved. (As with some of the incomplete syntactic representations, this is intended just as an intermediary step, not a possibly-final form; that is, it is a *logical form* in the sense made precise by Allen [1991].) And Hobbs [1983] has proposed a 'scope-neutral' representation of quantification that would be amenable to subsequent inference processes.

Ambiguities of intension and description also occur at this level. As far as we are aware, the only notation that even describes all the possible readings is that of Fawcett and Hirst [1986], but there is no attempt to permit incompleteness.

Discourse structure and pragmatics. Full text comprehension includes resolving anaphors, understanding the role of each sentence in the discourse, and deriving pragmatic inferences such as presuppositions and conventional implicatures. In practice, an intelligent text retrieval system will probably not need to figure out every last ounce of the writer's intent. But certainly anaphor resolution will be necessary, and perhaps also the recognition of some simple indirect speech acts such as asserting by asking a question.

An unresolvable anaphor or definite reference can be treated much as an unresolvable lexical ambiguity—that is, regarded as a disjunction of possible referents. Unresolved discourse connec-

[5] The exceptions are cases where two distinct possibilities end up 'saying the same thing'. For example, if the PP in *Nadia kissed the boy in the park* is attached to the VP, then the kissing took place in the park, and we may infer that the participants were in the park, too. If it is attached to the object NP, then the boy was in the park, and we may infer that Nadia and the kissing were, too. So either way, everything is in the park, and it's not actually necessary to worry about which PP attachment is correct; they both are. Such cases are not as rare as one might expect, but it's certainly not worth a system's time to check for them especially.

tions are somewhat harder, as the possibilities are open-ended, and it seems likely that if a unique connection is not found, then almost anything is possible. The connection then must just be recorded as 'I don't know', rather than as a small disjunction.

Interaction between levels. Information available at one level of an NLU program will often permit the resolution of uncertainties at another. Hirst [1987, 1988a] describes an architecture for a system in which the flow of information supports this to the greatest extent possible. We assume such a mechanism here, so that incompletenesses may indeed be only temporary. The difference here is that nothing *requires* an incompleteness to be patched up.

4 Natural Language as its Own Representation

If for some reason, the system could create only very incomplete representations at all levels of analysis for a particular sentence, then the representation of the sentence would be effectively little more than its unaltered surface form. In fact, a system might sometimes decide that a hard-to-process fragment is better left wholly in its surface form. Despite our earlier remarks concerning the limitations of surface representations (section 1), such forms would not necessarily be useless. After all, information retrieval systems have for many years operated on surface forms, with some fair success (but see Dick [1991] for discussion of the limitations).[6] Moreover, some recent research in AI has used natural language as a form for inference [Kayser *et al.* 1987, Jayez 1988] and as an interlingua in machine translation [Schubert 1988, Guzmán 1988]. We don't have space here to discuss the advantages and disadvantages of these approaches (but see Hirst [1993]), but we just need to point out that, contrary to the impression one gets from some AI research, natural language can be used, at least in some ways, as a knowledge representation. Indeed, it is the only knowledge representation we have so far that meets the fundamental requirement of having the expressive power of natural language.

[6] "We actually made a map of the country, on the scale of *a mile to the mile!*" "Have you used it much?" I enquired. "It has never been spread out, yet," said Mein Herr: "the farmers objected; they said it would cover the whole country, and shut out the sunlight! So now we use the country itself, as its own map, and I assure you it does nearly as well."—Lewis Carroll, *Sylvie and Bruno Concluded* (1893), chapter 11.

5 Mixed-depth Encodings

If the representation of a sentence or text is incomplete in different ways in different places, the result is a *mixed-depth encoding*. For example, one might obtain a vaguely scoped logical form with a piece of surface text and a disjunction of attachments embedded in it. In fact, a mixed-depth encoding could be quite a mess, as we will see in the examples below.

If such representations are to be useful, it will be necessary to devise inference and search methods that can operate upon them. For example, given our earlier example, *John bought a vxzfl*, it should be possible to infer that if the *vxzfl* was bought, it was paid for, even if the description of it is missing or incomplete. (One approach to this is exemplified by Granger's FOUL-UP program [Granger 1977], which used expectations in context, including a library of scripts for stereotypical situations, to perform such inferences.) As well as the usual kinds of matching and inference, these processes should be able to continue refining the encodings—that is, removing some of the uncertainties—if they find themselves able to do so.

One possible weakness of incomplete encodings is a vulnerability to a sort of 'snowballing' of incompleteness. Since understanding a sentence generally requires understanding the preceding text, one might find each sentence understood a little less than the one before it, until the system eventually becomes completely confused. This is not uncommon in undergraduates, and there's nothing to prevent it happening to any NLU system that is given a text that far exceeds its abilities. So the proposal here crucially depends on incompleteness not being resorted to too often. Mixed-depth representations are intended to add flexibility, not to act as a substitute for intelligence.

This proposal is to be distinguished from others that involve multiple kinds of representation. For example, Sparck Jones [1983] has proposed the simultaneous use of several different representations of text, each optimized for different aspects of the tasks of the system that uses them. While such representations may be 'partial', in the sense that they won't all contain full information about the text (*cf.* Sloman [1985]), they are not incomplete in our sense; that is, each is as fully refined as intended. In addition, the representations are not fragments mixed together; each individually covers the full text.

6 An Example

In order to demonstrate what mixed-depth representations might look like, we now present an extended example. A relatively complex paragraph is presented, along with a parse of each sentence and a semantic interpretation. We will assume that neither the parser nor the interpreter can cope with the full complexity of the text, and they therefore resort to incomplete encodings where necessary.

It should be understood that our example representation is constructed by hand, and is not the output of any actual system. Our point is the general nature of the representation, rather than the strengths and weaknesses of any particular system of parsing or semantic interpretation. Consequently, we have constructed our example by making reasonable assumptions about the abilities and limitations typical of state-of-the-art parsers and interpreters operating under time pressure.

We have chosen Dick's [1991] interpretation of Sowa's conceptual graphs (CGs) [Sowa 1984] as a typical first-order knowledge representation scheme upon which to base our extensions for representations that are incomplete or varying in depth. Readers unfamiliar with the notation should still find its meaning fairly obvious; the abbreviations for case roles that we use (following Sowa and Dick) are shown in Table 4.1. We chose CGs merely for generality and readability, and have no particular commitment to this notation.

For reasons of space, we do not show any attempt at processing the discourse relations in the text; rather, we simply show the CG output for each sentence or clause. In practice, of course, an interpreter would have to tie these graphs together with discourse-level relations such as `ELABORATION` or `EXAMPLE`; this process is vulnerable to the same kinds of uncertainties, and leading to the same kinds of representational gaps, as sentence interpretation.

6.1 Incompleteness

There are two ways that both the parse and the semantic interpretation can be incomplete:

1. A structure in the original sentence may be left *uncoded* if there is no appropriate representation for it.

2. A structure in the original sentence may have more than one encoding because the system cannot choose a single encoding. We call such a structure *multicoded.*

Table 1: Abbreviations for case roles used in our conceptual graphs.

`AGNT`	Agent: Relates an action to the entity performing it.
`ATTR`	Attribute: Relates an entity to one of its properties.
`BENF`	Beneficiary: Relates an action to the entity for whom it was performed.
`CAUS`	Cause: Relates a state to its cause.
`CHRC`	Characteristic: Relates an entity to an inalienable or characterizing property.
`DUR`	Duration: Relates an action to the time period over which it occurs.
`LOC`	Location: Relates an action to the place at which it occurs.
`MANR`	Manner: Relates an action to the manner in which it is carried out.
`POSS`	Possession: Relates an entity to another entity it possesses.
`PTNT`	Patient or theme: Relates an action to the entity upon which it is performed.
`TEMPL`	Temporal location: Relates an action to the point in time over which it occurs.

Even if the parse produces a complete encoding, it is still possible that the semantic interpretation be incomplete.

In order to control the number of uncoded elements, we assume that default assumptions can be made in certain situations. For example, if one sense of an ambiguous word is much more common than the others, that sense is chosen in the absence of any positive evidence for the alternatives (*cf* Hirst [1987]).

When the parser or semantic interpreter is faced with a structure that it cannot uniquely encode, it goes through a list of possible encodings for the structure and discards those that are not possible given what it has already been able to interpret. If there are no possibilities left, the structure is left uncoded. If there is more than one possibility left, and no unique default applies, the structure is multicoded with the remaining allowable defaults. This method of eliminating the impossible and taking what is left as the only possible encoding is used in such systems as Polaroid Words [Hirst 1987, Hirst 1988a] and constraint grammars [Karlsson 1990, Karlsson *et al.* 1991].

6.2 The Example in Detail

In this section, we present the parse trees and multiple-depth semantic interpretations for each sentence of the example paragraph.

The parse trees are drawn in a simple, linear fashion. Syntactic elements are underlined and labelled. If an element cannot be identified, it is labelled with a question mark. If the parser has to make a guess about the role of an element, that element is underlined with a dotted line, and its name is followed by a question mark. Once the parser makes a guess about the role of an element, all of the higher-level constituents that the element is taken to be part of are also treated as guesses. Key relationships between parts of a sentence are identified with arrows. If the relationship is ambiguous or the result of a guess, the arrow is dotted. The top-level structure of each sentence is `S`; in order to keep the parse trees as simple as possible, we have omitted this level from the graphs of longer sentences (sentences (4), (5), and (7) below).

In the conceptual graphs, uncoded elements are denoted by a question mark followed by the problematic surface string in quotation marks. Unspecified conceptual relations are represented as question marks. We add to Sowa's notation the disjunction operator "`or`", which is used to indicate alternatives that the interpreter cannot decide between. For example, "`(AGNT or PTNT)`" is an *uncertain representation* of a relation that might be `(AGNT)` or might be `(PTNT)`; the interpreter can't decide which. This is distinguished from the standard CG operators "`(OR)`" and "`|`" [Sowa 1984, p. 118–119], which are *certain representations*, and can be used to represent sentences that explicitly talk about disjunctive possibilities.

We have followed the style of Sowa's linear notation rather than network diagrams for CGs. This means that when an instance of a concept or structure participates in more than one relation, it might have to be written more than once. We do this by assigning it a name to the instance (using the notation *concept*`=#`*name*), and then, when necessary, using the name. For example, `[SAILOR: {*}=#ss]` denotes a set, named `#ss`, of sailors. A subsequent occurrence of `[SAILOR: {*} #ss]` would refer to the same set. It should be understood, however, that in an implementation there would be only one copy of the structure, with pointers deployed as necessary. Similarly, when literal strings of the original text are incorporated into the conceptual structures, these could, in an implementation, be pointers to the text rather than copies of it.

Our names for relations are mostly taken from Sowa's catalogue of conceptual relations [Sowa 1984, p. 415ff]. Our names for concepts are, for ease of reading, mostly suggestive English words.

Our example text is taken from a review[7] of *Dreadnought*, a book by Robert K. Massie that describes British naval history in the period leading up to the First World War. The following

[7] Elson, John. "When Britannia Ruled", *Time*, **138**(19), 11 November 1991.

paragraph describes the state of the Royal Navy prior to its Edwardian revival:

> Without warships, Britain was perilously vulnerable to blockade or invasion. But Britannia's capacity to rule the waves, as Massie also points out, was somewhat illusory; the Royal Navy during much of Victoria's reign was largely unfit for combat. Weighed down by moribund traditions that Winston Churchill acidly defined as "rum, sodomy, and the lash," British tars were ill fed and worse led. While their social-climbing officers fopped and preened, sailors spent long days at sea scrubbing decks and polishing brightwork, or wielding cutlasses in boarding drills as if they were still in the age of sail. Meanwhile, gunnery practice was cursory even though naval bombardments were ludicrously inaccurate. In 1881, for example, eight British battleships fired 3,000 rounds at forts guarding the Egyptian city of Alexandria and scored precisely 10 hits.

We will show the parse tree and semantic interpretation for each sentence in turn, illustrating various aspects of incomplete and mixed-depth representation.

Sentence (1)

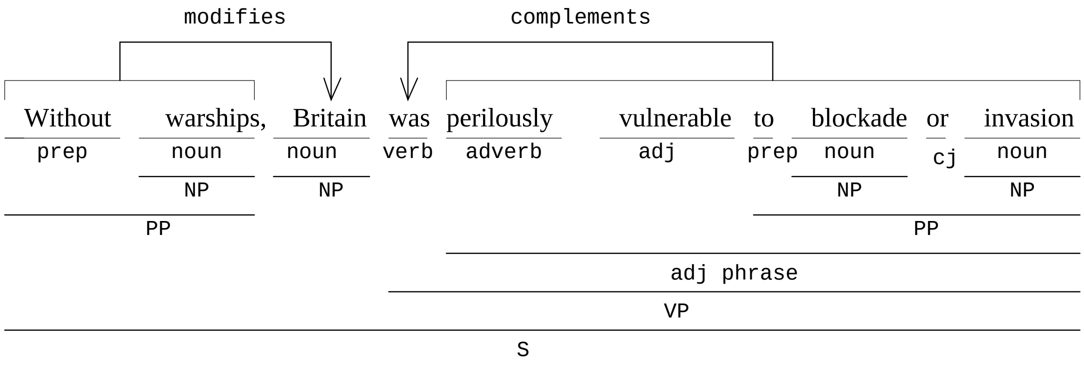
(1) Without warships, Britain was perilously vulnerable to blockade or invasion.

Figure 4.1(a) contains the parse tree for sentence (1). There are no ambiguities or guesses in this parse. The conceptual graph for sentence (1), shown in Figure 4.1(b), is not complete, because the semantic interpreter cannot find a way to connect the phrase *to blockade or invasion* to the graph for the rest of the sentence. Thus, the interpretation of this phrase forms its own separate graph. Notice also that the `(OR)` relation in this graph expresses a disjunction in the semantic content of the original sentence, not in its interpretation.

Sentence (2)

(2) But Britannia's capacity to rule the waves, as Massie also points out, was somewhat illusory;

Figure 4.2(a) contains the parse tree for sentence (2). This shows that the parser cannot determine whether the dependent clause *as Massie also points out* modifies *Britannia's capacity* or the sentence as a whole.



(a)

```
[STATE:[COUNTRY:Britain] -> (ATTR) -> [VULNERABLE] <- (ATTR) <- [PERILOUS]] -
     (CAUS) -> [[COUNTRY:Britain] -> (~POSS) -> [WARSHIP:{*}]].

[[BLOCKADE] <- (OR) -> [INVASION]].
```

(b)

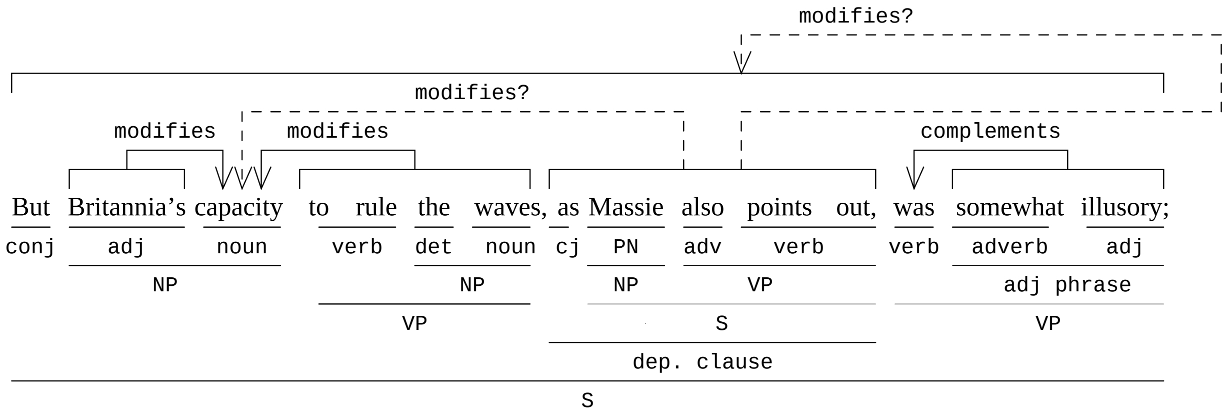
Figure 1: Parse and interpretation of sentence (1)

The conceptual graph for sentence (2), shown in Figure 4.2(b), is almost complete. But the ambiguous attachment of the dependent clause *as Massie also points out* is reflected in the CG by the disjunction in the `PTNT` ("patient") relationship for `POINT_OUT`. The patient of `POINT_OUT` is either `[COUNTRY:Britain]` or `[CAPACITY: #cap]` depending on whether the dependent clause modifies the whole sentence or the NP *Britannia's capacity*. And while the parser identifies *also* as a modifier of *points out*, the semantic interpreter cannot identify the exact relationship, so the relationship is left unspecified in the graph. The same happens for *somewhat* as a modifier of *illusory*. And the idiom *to rule the waves* has not been recognized, the interpreter representing it quite literally.

Sentence (3)

(3) the Royal Navy during much of Victoria's reign was largely unfit for combat.

Figure 4.3(a) contains the parse tree for sentence (3). There are no ambiguities or guesses in this tree. But we shall assume that the interpreter cannot determine if the *Royal Navy* is the agent or the patient of the action implicit in the phrase *unfit for combat* (*cf. The food was*

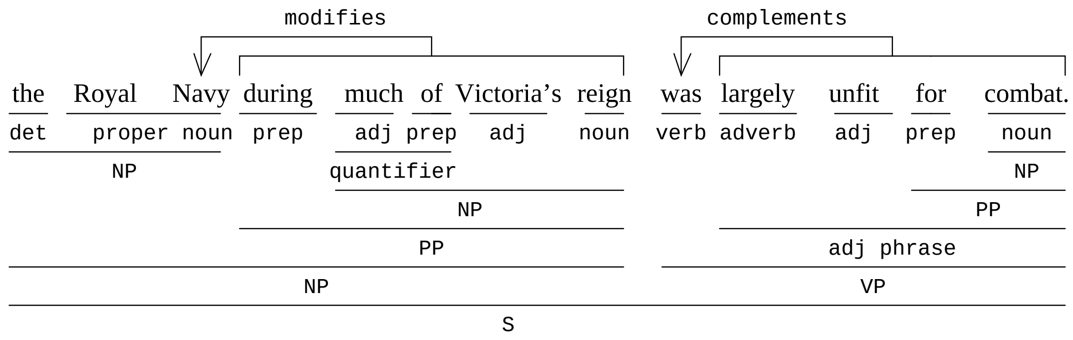


(a)

```
[POINT_OUT] -
     (AGNT) -> [PERSON:Massie]
     (PTNT) -> [[[COUNTRY:Britain] -
                    (POSS) -> [CAPACITY:=#cap] -
                                   (CHRC) -> [RULE]
                                   (PTNT) -> [WAVE:{*}#]
                                   (ATTR) -> [ILLUSORY] -
                                        (?) -> [MODERATE]]
               or
               [CAPACITY: #cap]]
     (?) -> [ALSO].
```

(b)

Figure 2: Parse and interpretation of sentence (2)



(a)

```
[[COMBAT] -
     (AGNT or PTNT) -> [NAVY:Royal Navy] -
          (ATTR) -> [UNFIT]
          (DUR) -> [[[PERSON:Victoria] -> (POSS) -> REIGN]]:@much]].
```

(b)

Figure 3: Parse and interpretation of sentence (3)

unfit for consumption). The CG therefore shows the disjunction of the alternatives. In addition, the interpreter has not taken the time to consider the subtleties of the possessive in the phrase *Victoria's reign*, and has simply encoded it as the reign that Victoria possesses, just as if it were a physical possession like *Victoria's carriage* or *Victoria's nose*. The resulting conceptual graph is shown in Figure 3(b).

Sentence (4)

(4) Weighed down by moribund traditions that Winston Churchill acidly defined as "rum, sodomy, and the lash," British tars were ill fed and worse led.

Figure 4.4(a) contains the parse tree for sentence (4). We assume that the lexicon does not recognize *sodomy* or *lash*, but the parser can guess that both words are nouns, and that the entire phrase *rum, sodomy, and the lash* is a conjunctive NP. The structures in the sentence that include this NP are also recorded as guesses in the parse tree. In addition, we assume that the lexicon does not recognize *tars*, knowing *tar* only as a mass noun meaning black, sticky stuff; it guesses that *British tars* is an NP. Finally, it guesses that the introductory phrase *Weighed . . . lash* modifies *tars*.

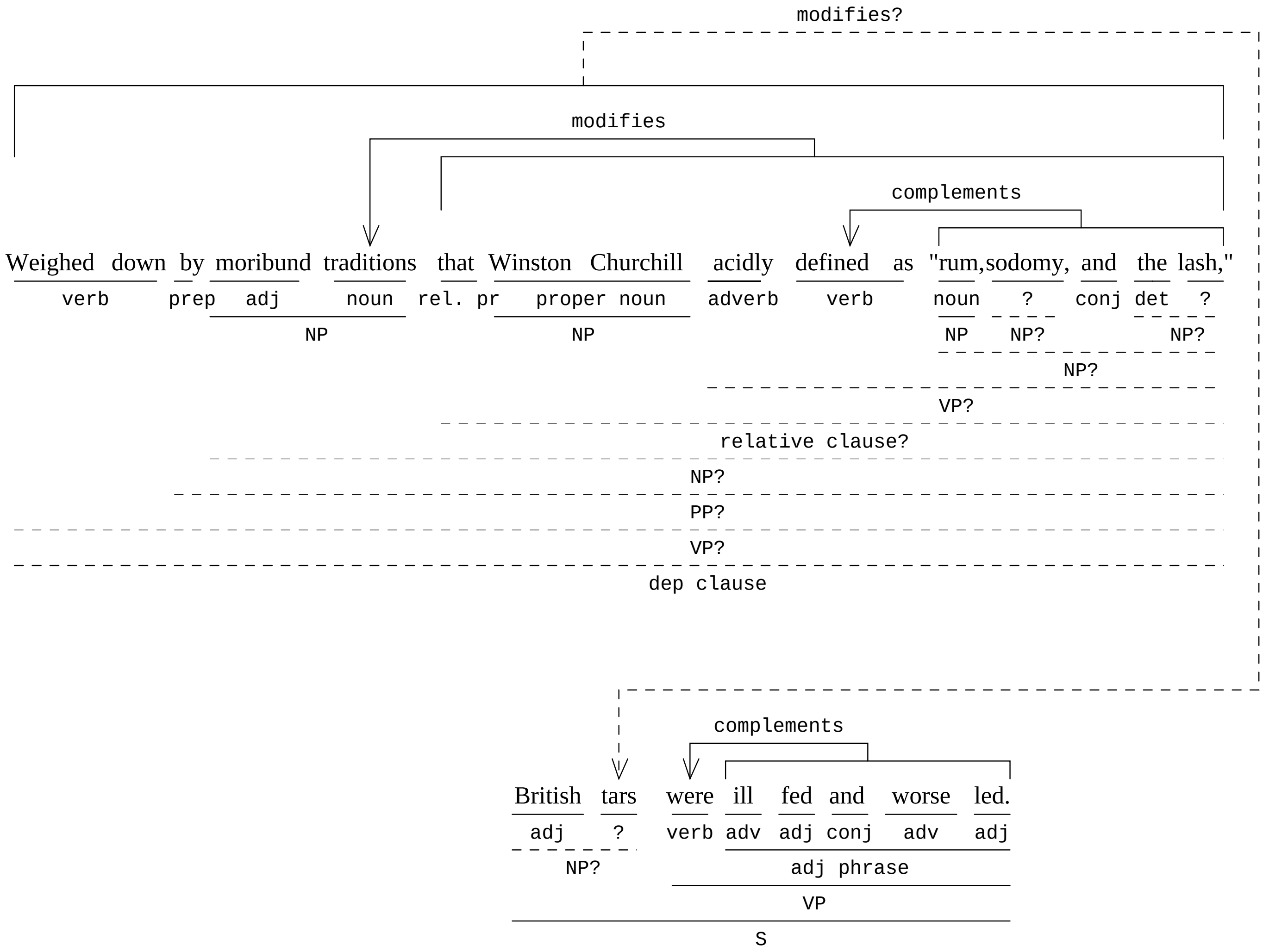
Because of these uncertainties, the conceptual graph for sentence (4), in Figure 4.4(b), has some significant holes in it. To begin with, the focus of the sentence, *tars*, has to be left uncoded. The interpreter makes the default choice that this element represents the patient of the verb *weighed down* and is what is modified by the phrase *ill fed and worse led*. In the phrase *rum, sodomy, and the lash*, only the noun *rum* is encoded. The knowledge representation scheme makes the default choice that this phrase has a conceptual relationship with the verb *defined*, but it cannot specify which relationship.

Sentence (5)

(5) While their social-climbing officers fopped and preened, sailors spent long days at sea scrubbing decks and polishing brightwork, or wielding cutlasses in boarding drills as if they were still in the age of sail.

Figure 4.5(a) contains the parse tree for sentence (5). The lexicon does not recognize the neologism *fopped*, nor does it have an entry for *preened* as an intransitive verb (a usage that is not included in the *Oxford Advanced Learner's Dictionary*). The parser guesses that both words are being used as intransitive verbs. The lexicon also does not have an entry for *brightwork*, but the parser guesses that it is a noun. The parser cannot determine the attachment of the dependent clause *as if they were still in the age of sail*. It guesses that this clause modifies the verb phrase *wielding cutlasses in boarding drills*.

The conceptual graph for sentence (5), shown in Figure 4.5(b), also has many holes and uncertainties in it. To begin with, the interpreter guesses that the agent of the unknown words *fopped* and *preened* is *their social-climbing officers*. We shall assume also that it falls short in its representation of the clause *as if they were still in the age of sail*. The phrase *the age of sail* is not recognized as an idiom, nor can the system determine the relationship between `[AGE]` and `[SAIL]`. Thus, it leaves the relationship unspecified. This uncertainty, plus the great polysemy of the word *in*, means that, in turn, the relationship between this phrase and the pronoun *they* has to be represented as the disjunction of a `TEMPL` relationship and a `LOC` relationship; the interpreter can't determine whether the preposition *in* in the phrase *in the age of sail* flags a physical location or a period in time, so it records both. And lastly, it cannot determine whether *they* refers to the officers or the sailors. (The representation of *as if* as a manner relationship between an action and a state labeled as counterfactual is clearly inadequate, but that is a separate issue in semantics.)

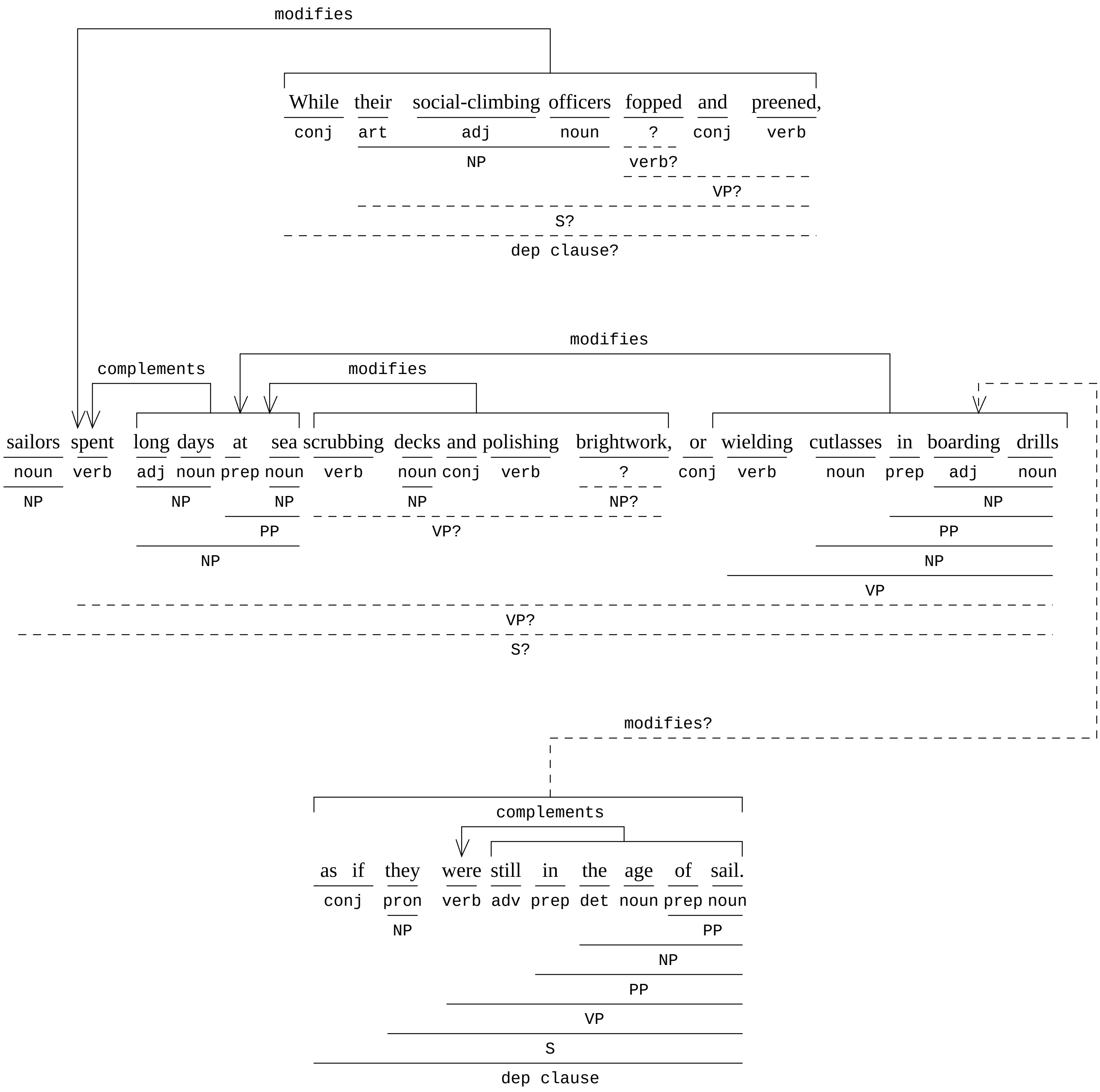


(a)

```
[WEIGH_DOWN] -
     (PTNT) -> [ ? ‘tars’ ] -
          (CHRC) -> [BRITISH]
          (ATTR) -
               [[[FED] -> (ATTR) -> [BADLY]] and [[LED] -> (ATTR) -> [BADLY]]]
     (AGNT) -> [TRADITION:{*}=#trad] -> (ATTR) -> [MORIBUND]]
[DEFINE] -
     (PTNT) -> [TRADITION:{*} #trad]
     (AGNT) -> [PERSON:Winston_Churchill]
     (?) -> [[RUM] and [? ‘sodomy’] and [? ‘the lash’]].
```

(b)

Figure 4: Parse and semantic interpretation of sentence (4)

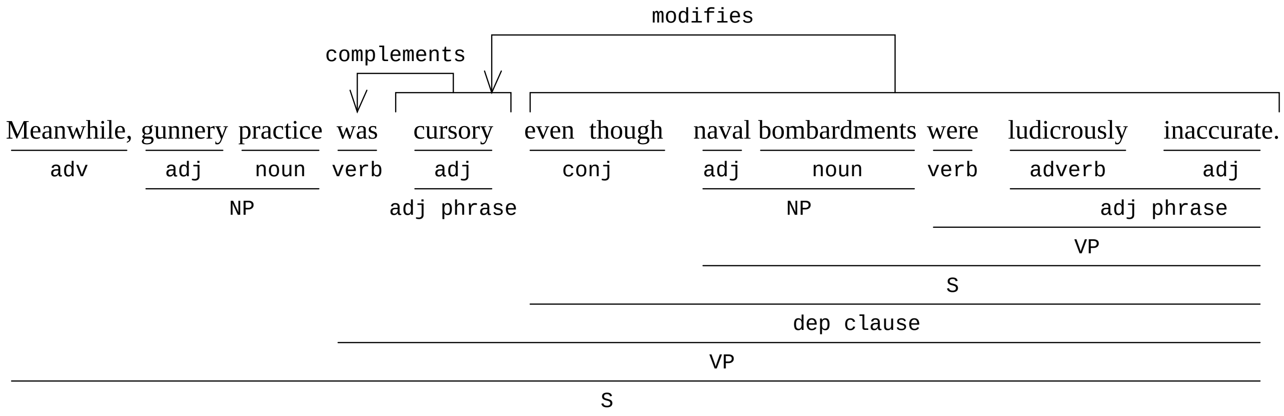

modifies
While their social-climbing officers fopped and preened,
conj art adj noun ? conj verb
NP
verb?
VP?
S?
dep clause?
modifies
complements
modifies
sailors spent long days at sea scrubbing decks and polishing brightwork, or wielding cutlasses in boarding drills
noun verb adj noun prep noun verb noun conj verb ? conj verb noun prep adj noun
NP NP NP NP NP? NP
PP VP? PP
NP NP
VP
VP?
S?
modifies?
complements
as if they were still in the age of sail.
conj pron verb adv prep det noun prep noun
NP PP
NP
PP
VP
S
dep clause

(a)

```
[ ? 'fopped']
[ ? 'preened'] -
     (AGNT) -> [OFFICER:{*}=#offs] -
          (ATTR) -> [SOCIAL_CLIMBING]
          (POSS) -> [SAILOR:{*}=#ss]
[SPEND_TIME] -
     (AGNT) -> [SAILOR:{*} #ss]
     (LOC) -> [SEA]
     (PTNT) -> [DAY:{*}] -> (ATTR) -> [LONG]
[SCRUB] -
     (AGNT) -> [SAILOR:{*} #ss]
     (PTNT) -> [DECK:{*}]
[POLISH] -
     (AGNT) -> [SAILOR:{*} #ss]
     (PTNT) -> [ ? 'brightwork']
[WIELD] -
     (AGNT) -> [SAILOR:{*} #ss]
     (PTNT) -> [CUTLASS:{*}]
     (LOC) -> [DRILL:{*}] -> (CHRC) -> [BOARDING]
     (MANR) -> [COUNTERFACTUAL:
                    [[SAILOR:{*} #ss] or [OFFICER:{*} #offs]] -
                         (TEMPL or LOC) -> [AGE] -> (?) -> [SAIL]].
```

(b)

Figure 5: Parse and interpretation of sentence (5)



(a)

```
[BOMBARDMENT] -
     (CHRC) -> [NAVAL]
     (ATTR) -> [INACCURATE] <- (ATTR) <- [LUDICROUS]
[PRACTICE] -
     (CHRC) -> [GUNNERY]
     (ATTR) -> [CURSORY].
```

(b)

Figure 6: Parse and interpretation of sentence (6)

Sentence (6)

(6) Meanwhile, gunnery practice was cursory even though naval bombardments were ludicrously inaccurate.

Figure 4.6(a) contains the parse tree for sentence (6). There are no ambiguities or guesses in this parse tree. The conceptual graph for the sentence, in Figure 4.6(b) is relatively complete, but the causal relationship between the two clauses has been lost. The sentence is represented as two unrelated conceptual graphs, one with the head `BOMBARDMENT` and the other with the head `PRACTICE`.

Sentence (7)

(7) In 1881, for example, eight British battleships fired 3,000 rounds at forts guarding the Egyptian city of Alexandria and scored precisely 10 hits.

Figure 4.7(a) contains the parse tree for sentence (7). There are no ambiguities or guesses in this parse tree. But the sentence is quantificationally ambiguous: Did the eight battleships fire 3,000 rounds each or 3,000 in total; and did they score 10 hits each or 10 in total? In standard conceptual graphs [Sowa 1984, p. 118], the notation `[BATTLESHIP: dist{*}@8]` means each element of the set of battleships acts independently, giving the first reading of each pair, while `[BATTLESHIP: {*}@8]` means all the battleships act together as a single agent, giving the second reading of each pair. We represent the uncertainty by the disjunction `[BATTLESHIP: (dist or set){*}@8]` (adding the word "`set`", implicit in Sowa's notation, to explicitly disjoin with "`dist`"). The CG is shown in Figure 4.7(b).

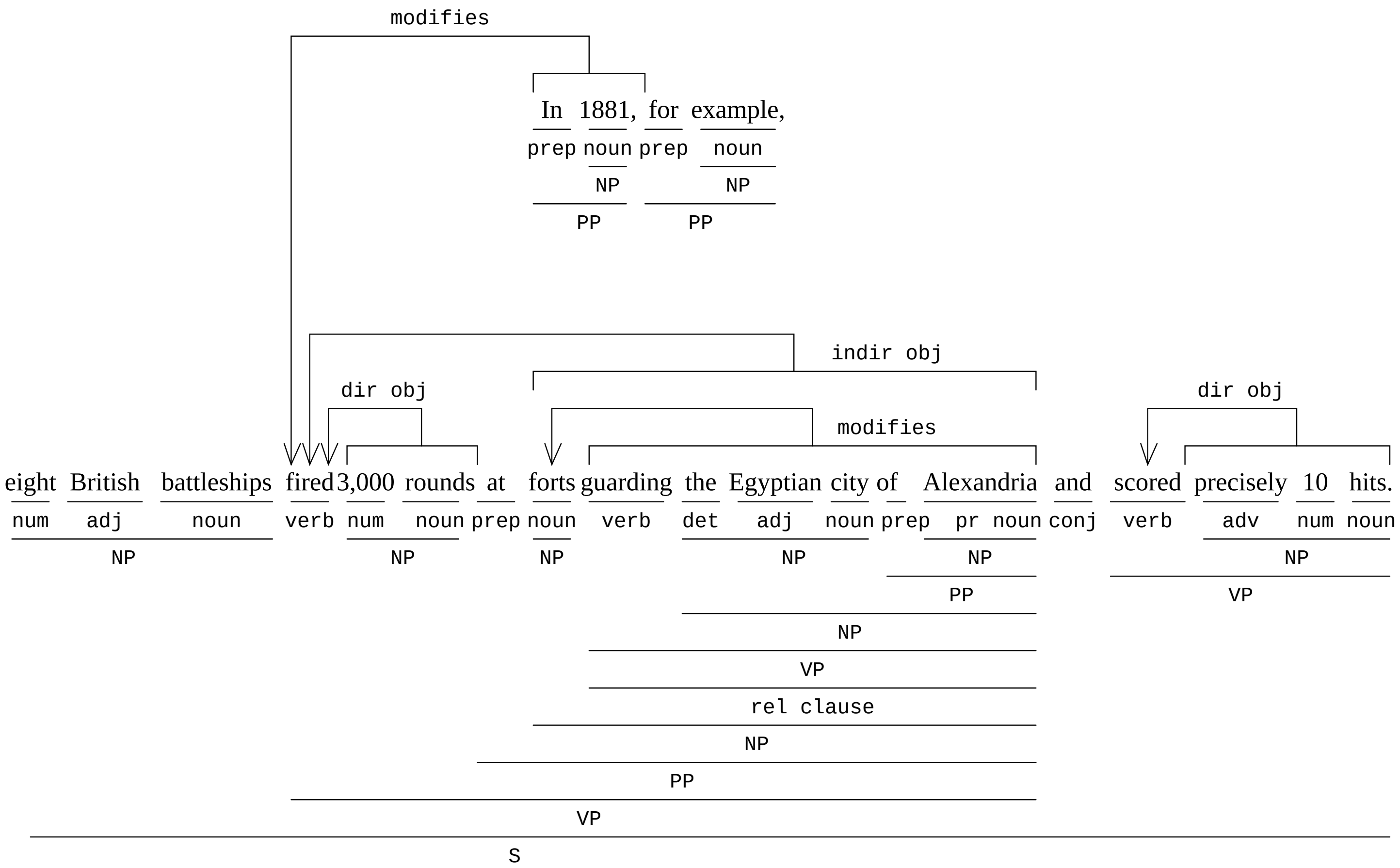
6.3 Using This Representation

The purpose of a representation such as we have shown is, of course, to permit conceptual retrieval by matching queries that are similarly represented (*cf.* Dick [1991]). The matching process might take into account inheritance of properties and even inference of arbitrary complexity. While the development of such processes is a topic of continuing research, it should be clear that the general principles of matching processes that operate on complete representations will carry over to our mixed-depth partial representations.

6.4 What We Have Shown

Our example demonstrates a number of points about incomplete and multiple-depth representations.

Many inaccuracies are benign: for example, the representation of the NP *Victoria's reign* as a possession of the person *Victoria* in sentence (3), rather than as a time period. Thus, if this representation were used to retrieve entities belonging to Victoria, her reign would be one of the entities retrieved, which is incorrect but probably harmless. On the other hand, because of inaccuracy, the conceptual graph for sentence (3) would probably *not* match queries about the nineteenth century, a more serious error. Such errors are the price paid for being unwilling or unable to analyze the text as deeply as an ideal language understander could. Nevertheless, a partial representation is better than none at all. (And often, redundancy in the text will come to the rescue. In this example, other sentences in the text, including sentence (7) of the same paragraph, would presumably match such a query.)



(a)

```
[FIRED] -
     (AGNT) -> [BATTLESHIP:(dist or set){*}@8 =#ships] <- (CHRC) <- [BRITISH]
     (PTNT) -> [ROUND:{*}@3000]
     (BENF) -> [FORT:{*} =#ff]
     (TEMPL) -> [DATE:1881]
[GUARD] -
     (AGNT) -> [FORT:{*} #ff]
     (PTNT) -> [CITY:Alexandria]
[SCORE] -
     (AGNT) -> [BATTLESHIP:(dist or set){*}@8 #ships]
     (PTNT) -> [HIT:{*}@10].
```

(b)

Figure 7: Parse and interpretation of sentence (7)

Metaphors can be taken literally. The CG for sentence (2) has *rule the waves* represented as a literal ruling over waves rather than the intended meaning: military (and perhaps economic) domination of the seas. The representation is not correct, but it does provide some relevant information, the notion of British rule.

Even completely uncoded text can still be useful in certain situations. In the fourth sentence of the example, the lexical units *sodomy* and *the lash* are unknown, but the entire unit *rum, sodomy, and the lash* can still be identified as text that the person *Winston Churchill* used to define the *traditions* that are mentioned earlier in the sentence. An information retrieval system can still present such a text as a whole in answer to a query about naval traditions (or about quotations by Winston Churchill).

The incomplete representations preserve ambiguities that are present in the text itself. In sentence (7), for example, the phrase *eight British battleships fired 3,000 rounds ... and scored precisely 10 hits* is quantificationally ambiguous. The representation retains this ambiguity.

6.5 What We Haven't Shown

By choosing judiciously our "reasonable assumptions about the abilities and limitations typical of state-of-the-art parsers and interpreters operating under time pressure", we have tried to show, in one short example, many of the things that might occur in a mixed-depth representation. But we cannot show all that might arise. While many uncertainties in parsing are only a question of where to attach a modifying constituent, as in sentences (2) and (5), it can sometimes be the case that the parser cannot decide between two completely different structures for much or all of a sentence or clause: the declarative and imperative parses of *Time flies*, for example. The resulting interpretation might be the disjunction of two completely separate CGs; we have not shown such a case here. Nor have we shown disjunctive intepretations of ambiguities of intension and description, where CGs cannot always even represent all the possibilities (*cf.* Fawcett and Hirst [1986]). And we have not even attempted to show any kind of discourse processing.

7 Conclusion

The ultimate criterion for adequacy of a mixed representation would be its usefulness, and, in particular, the degree to which inference and understanding processes can be developed or modified to accept such forms. At present, research in the area is very preliminary; a number of representations that are incomplete in one way or another have been proposed, but there has been no previous attempt at integration. However, we believe the basic idea to be a promising one worthy of further development, and have showed, by means of an example, how this might proceed.

Acknowledgments

We are grateful to Judy Dick for many discussions on the role of representations of text in information retrieval, and to Stephen Regoczei, Karen Sparck Jones, Paul Jacobs, and Yorick Wilks for comments on earlier versions of this paper. Our work is supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Information Technology Research Centre.

References

[Allen 1991] James Allen. Natural language, knowledge representation, and logical form. Technical report 367, Department of Computer Science, University of Rochester, 1991.

[Alshawi and van Eijck 1989] Hiyan Alshawi and Jan van Eijck. Logical forms in the Core Language Engine. In *Proceedings of the 27th Annual Meeting, of the Association for Computational Linguistics*, pages 25–32, Vancouver, June 1989.

[Church 1980] Kenneth Ward Church. *On memory limitations in natural language processing.* MSc thesis / TR 245, Laboratory for Computer Science, MIT, June 1980. Distributed by the Indiana University Linguistics Club, Bloomington, IN.

[Dick 1987] Judith Dick. Conceptual retrieval and case law. In *Proceedings of the First International Conference on Artificial Intelligence and Law*, pages 106–115, Boston, May 1987.

[Dick 1991] Judith Dick. *A conceptual, case-relation representation of text for information retrieval.* Doctoral dissertation, Faculty of Library and Information Science, University of Toronto. Published as technical report CSRI-265, Computer Systems Research Institute, University of Toronto.

[Fawcett and Hirst 1986] Brenda Fawcett and Graeme Hirst. The detection and representation of ambiguities of intension and description. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 192–199, New York, June 1986.

[Granger 1977] Richard H. Granger, Jr. FOUL-UP: A program that figures out meanings of words from context. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, pages 172–178, Cambridge, Mass., August 1977.

[Guzmán 1988] Iván Guzmán de Rojas. "ATAMIRI—Interlingual MT using the Aymara language." In Maxwell *et al.* [1988], pages 123–129.

[Hirst 1987] Graeme Hirst. *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press (Studies in natural language processing), 1987.

[Hirst 1988a] Graeme Hirst. Semantic interpretation and ambiguity. *Artificial Intelligence*, *34*(2):131–177, March 1988.

[Hirst 1988b] Graeme Hirst. Resolving lexical ambiguity computationally with spreading activation and Polaroid Words. In Steven Small, Garrison Cottrell and Michael Tanenhaus (eds). *Lexical ambiguity resolution*, 73–107, Morgan Kaufmann, Los Altos, CA, 1988.

[Hirst 1989] Graeme Hirst. Ontological assumptions in knowledge representation. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 157–169, Toronto, May 1989. Morgan Kaufmann Publishers, San Mateo, CA.

[Hirst 1991] Graeme Hirst. Existence assumptions in knowledge representation. *Artificial Intelligence*, *49*:199–242, May 1991.

[Hirst 1993] Graeme Hirst. Natural language as its own representation. In preparation.

[Hobbs 1983] Jerry R. Hobbs. An improper treatment of quantification in ordinary English. In *Proceedings of the 21st annual meeting of the Association for Computational Linguistics*, pages 57–63, Cambridge, MA, June 1983.

[Jayez 1988] Jacques Jayez. *L'inference en Langue Naturel: Le Problème des Connecteurs; Représentation et Calcul.* Hermès (Langue, raisonnement, calcul), Paris, 1988.

[Karlsson 1990] Fred Karlsson. Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, vol. III, pages 168–173, Helsinki, August 1990.

[Karlsson *et al.* 1991] Fred Karlsson, Atro Voutilainen, Arto Anttila, and Juha Heikkilä. Constraint grammar: A language-independent system for parsing unrestricted text, with an application to English. In *Workshop Notes from the 9th National Conference on Artificial Intelligence (AAAI-90): Natural Language Text Retrieval*, Anaheim, July 1991.

[Kayser *et al.* 1987] D. Kayser, P. Fosse, M. Karoubi, B. Levrat, and L. Nicaud. A strategy for reasoning in natural language. *Applied Artificial Intelligence*, *1*(3):205–231, 1987.

[Maxwell *et al* 1988] Dan Maxwell, Klaus Schubert, and Toon Witkam (eds). *New Directions in Machine Translation.* Foris (Distributed Language Translation 4), Dordrecht, 1988.

[Metzler *et al* 1989] Douglas P. Metzler, Stephanie W. Haas, Cynthia L. Cosic, and Leslie H. Wheeler. Constituent object parsing for information retrieval and similar text processing problems. *Journal of the American Society for Information Science*, *40*(6):398–423, 1989.

[Schubert 1988] Klaus Schubert. The architecture of DLT—Interlingual or double direct? In Maxwell *et al.* [1988], pages 131–144.

[Seo and Simmons 1989] Jungyun Seo and Robert F. Simmons. Syntactic graphs: A representation for the union of all ambiguous parse trees. *Computational Linguistics*, *15*(1):19–32, March 1989.

[Sloman 1985] Aaron Sloman. Why we need many knowledge representation formalisms. In Max A. Bramer, (ed). *Research and Development in Expert Systems*, pages 163–183, Cambridge University Press (British Computer Society workshop series), 1985.

[Somers 1987] Harold Somers. *Valency and Case in Computational Linguistics.* Edinburgh University Press (Edinburgh Information Technology Series), 1987.

[Sowa 1984] John Sowa. *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley, Reading, MA, 1984.

[Sparck Jones 1983] Karen Sparck Jones. Shifting meaning representations. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pages 621–623, Karlsruhe, August 1983,

[Toulmin 1958] Stephen Toulmin. *The Uses of Argument.* Cambridge University Press, 1958.