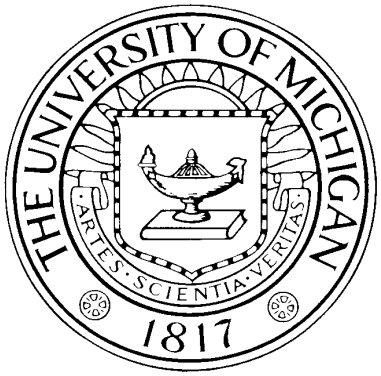
Semantics-Based Reference Resolution in Technical Text Processing: An Exploration of Using the WordNet Database in the Computerized Comprehensibility System

David E. Kieras

University of Michigan



Technical Communication Program
Technical Information Design and Analysis Laboratory
2360 Bonisteel Blvd.
Ann Arbor, MI 48109-2108

Technical Report No. 35 (TR-92/ONR35)

August 30, 1992

This research was supported by the Office of Naval Research, Cognitive Science Program, under Contract Number N00014-88-K-0133, Contract Authority Identification Number NR 442-f002. Reproduction in whole or part is permitted for any purpose of the United States Government.

Semantics-Based Reference Resolution in Technical Text Processing: An Exploration of Using the WordNet Database in the Computerized Comprehensibility System

David E. Kieras
University of Michigan

Abstract

The Computerized Comprehensibility System (CCS) provides an automated copy editing function, generating a "mark-up" of a draft of a technical document by simulating the simpler comprehension processes of a human reader, and then criticizing the text when these simple processes cannot successfully comprehend the material. A key CCS function is criticizing the coherence of the material by tracking which objects are mentioned in the passage. A common comprehensibility problem is that the text mentions a new object using the syntactic structures appropriate for an already-known object. If the reader must make an inference that presence of the new object is implied by earlier-mentioned object, the result is a potential break in the coherence of the text. CCS criticizes all such coherence breaks. However, many such inferences are actually easy for most readers, since only general knowledge is required to make the inference, rather than specialized knowledge about the domain. If so, then the CCS criticism of a coherence break is a false alarm. This report describes exploratory work with an augmented form of CCS, in which the WordNet database is used as a source of general knowledge to allow CCS to make the same kind of general knowledge inferences that human readers do to overcome coherence breaks.

Introduction

This report describes some results obtained by extending the Computerized Comprehensibility System (CCS) described in Kieras (1989, 1990) to make use of the semantic lexicon database developed by Miller and his coworkers (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), called WordNet. CCS is a system which provides an automated copy editing function, by generating a "mark-up" of a draft of a technical document. It has been more completely described elsewhere; here only the basic functions will be summarized. Figure 1 shows the overall structure of the CCS system. CCS attempts to do a full grammatical parse of the sentence structure, followed by an attempt to perform simple reference resolution on each noun phrase. Finally, it integrates the sentence content into a representation of the content of the passage as a whole. A set of criticism rules can comment on poor grammatical structure, inconsistent terminology, and lack of coherence of each sentence with the rest of the passage. As described elsewhere, the advantage of such a system relative to conventional computer-based writing aids is that because it actually attempts to mimic the simpler comprehension processes of a human reader, it can be sensitive to when the writer has made too many comprehension demands upon the reader. For example, if CCS can not resolve a reference, then the writer has apparently expected the reader to perform an inference in order to comprehend the sentence in the context of the rest of the passage.

Simple Reference Resolution in CCS

CCS represents the contents of a passage using a propositional semantic network, based on Anderson's ACT representation (1976). Along the lines of the given-new distinction (Haviland & Clark, 1974, Clark & Haviland, 1977), CCS attempts to identify the *given*, or already known, item in a sentence, and then adds the *new* information

in the sentence to the representation. Thus each noun phrase is matched against the representation of the previous sentences in the passage in order to identify which referent is being referred to. This matching can sometimes be done simply on the basis of the word strings involved, but more generally, it must be done in terms of the propositional representation specified by the noun phrases and passage content. Complex noun phrases such as *the bearings that the oil that the pumps circulates lubricates* are matched recursively; the most interior noun phrase is matched and the results are then used in an attempt to match the next most outermost noun phrase. This process is called *simple* reference resolution because the processing is done strictly in terms of the immediate surface and propositional content of the passage; no semantic knowledge about the word meanings, or general knowledge about the world, is used in this process.

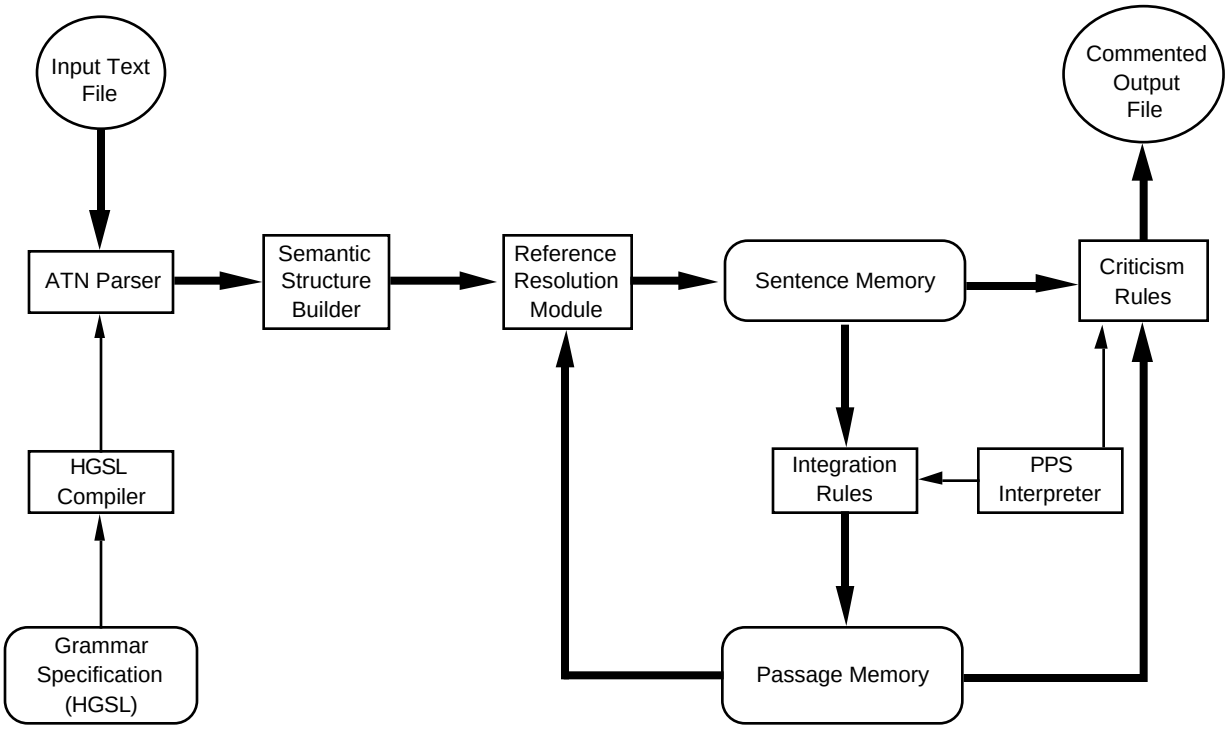


Figure 1. Structure of the Computerized Comprehensibility System (CCS). The work in this report concerns the Reference Resolution Module.

For example, in the simple passage shown in Table 1, the title introduces the *main lube oil system* as the main topic of the passage. The first sentence of the passage refers directly to the system with the identical set of words, *main lube oil system.* However, the second sentence refers to *the lube oil system* which is similar, but not identical, to the phrase *main lube oil system*, but refers to the same object, the *system*. The required matching is more complex than simply matching words; for example, the second sentence also refers to *oil* in the phrase *lubricating oil*. Even though the word *oil* has appeared previously, this referent, the *lubricating oil*, has not previously appeared. Moreover, the third sentence refers to *the oil that the pump circulates*. This *oil* is the same referent as the *lubricating oil* mentioned in the third sentence, and must be recognized as such, even though the form of the noun phrase is completely different — the third sentence describes the *oil* is in terms of being *circulated* by the *pump*, but this description was not a previous noun phrases, but was the main proposition of the second sentence. In processing this passage, CCS isolates each individual noun phrase and attempts to match it against previously mentioned items in the passage. Thus, even simple reference resolution can be complex.

Table 1
An example passage used to demonstrate CCS functions

MAIN LUBE OIL SYSTEM

The main lube oil system consists of a main lube oil pump, an auxiliary lube oil pump, and a duplex lube oil strainer. The function of the lube oil system is to circulate lubricating oil to the turbine and gearbox gears. For example, the bearings that the oil that the pump circulates lubricates support the turbine rotor. The duty officer is responsible for observing the pressure gauges. The lube oil system is critical for operation of the ship.
…

CCS has the ability to match a reference to a referent representation based on whether any proper subset of the previous predicates is mentioned in the to-be-matched noun phrase. Thus *the main system* could be matched against *the main lube oil system.* The restriction is that the head noun, e.g., *system*, must be identical. Although the CCS software had provision for distinguishing between words and the denoted concept, no attempt was made to represent which words designated the same concept; CCS did not recognize synonyms for words. Thus *the plane* would not match a previously mentioned *airplane*. Likewise, CCS had no semantic information, for example that airplanes had wings, that airplanes flew, or that airplanes are a member of the more general class of aircraft.

A basic argument supporting this rather severe limitation on a comprehension system was that an automated copy editing system would not be practical if it had to be stocked with detailed domain knowledge before it could be used. Rather, CCS was defined without such knowledge in an effort to see whether a useful editorial tool could be obtained without any domain knowledge. That is, many of the known problems with the comprehensibility of text are problems at the level of internal sentence structure or internal textual structure. For example, a long-stated problem in technical documentation is inconsistent terminology. Consistent terminology would be characterized by an identical or near identical use of the same string of words to refer to each individual object. Identifying this problem can be done without domain knowledge. In addition, the typical user of a technical document can not be relied upon to have much domain knowledge; clearly if they had considerable domain knowledge it is unlikely that they would be referring to the document at all. All of these considerations led to the initial decision to develop CCS without any domain knowledge.

However, CCS pays considerable attention to a serious form of incoherence, in which there is no easy-to-determine relationship between the sentence and the previous content of the passage, because there are apparently no shared referents. When this break in the coherence of the passage occurs, there is often a rather severe demand on the reader's inference-making abilities. For example, the Table 1 passage mentions the *duty officer* and the *pressure gauges* at the end of the first paragraph. Presumably if the reader is a sailor in the U.S. Navy, he or she will probably know what *the duty officer* is. However, unless they have domain expertise, they certainly will not realize what *the pressure gauges* are. Thus an important function of CCS is to point out where there is are such failures of coherence, and which of the referents have not previously appeared in the passage.

The Problem

The problem is that in many cases coherence failures are not useful criticisms of the material because it is reasonable to assume that every reader can easily make the required inference. For example, as shown in the CCS output excerpt in Table 2, CCS comments that the second sentence, referring to *the wings*, has no relationship with the previous material about an *aircraft*, and that *the wings* in the second sentence is a *questionable new referent* which is defined as a definite noun phrase that refers to a textually new reference. That is, it is referred to as if the reader should know about it, being a definite noun phrase, but since this is the first mention, it must be a new referent. Thus, for example, using *the wings* at this point in the passage is incorrect; the readers should not be cued that they

already know about an object that in fact they haven't yet seen. Avoiding this criticism would require rewriting the second sentence to introduce the wings as a new object, such as in an indefinite noun phrase in the sentence predicate, such as *The aircraft has wings that are in a swept-back configuration.*

Table 2
Excerpt from CCS output with input sentences shown in boldface

---

The F-16 aircraft is a high-performance fighter.
~~~~~~
The wings have a swept-back configuration.

The main proposition of this sentence is PROP9:
- REF3 WINGS has relation HAVE
  to REF4 (SWEPT-BACK CONFIGURATION).

NO-KNOWN-REFERENTS
This sentence does not appear to refer to anything previously mentioned,
and so readers may not understand how it relates to the rest of the material.
Be sure that the sentence directly and clearly refers to a previous item.

QUESTIONABLE-NEW-REFERENT
These items were referred to as if the reader already knows about them,
but they could not be matched with something previously introduced:
REF3 WINGS
Check: Can your reader easily figure out what you are referring to?
…

However, *everybody* knows that airplanes have wings. It would certainly be desirable if CCS was "smart" enough to know this, and thus not harass the writer with criticisms of incoherence in such obvious circumstances. In other words, while it would be impractical to give CCS knowledge of every specific domain that technical material might be prepared in, CCS might be much more useful if it made use of general knowledge to understand references; CCS could make the same elementary inferences that all readers would do, and thus not criticize the writer of incoherence in those cases.

Goal of this Work

A common belief is that a useful amount of general knowledge would be gigantic, and thus impractical to incorporate in any real system. However, it can be argued that the general knowledge required to resolve many kinds of reference is in fact very limited, consisting of such simple semantic relationships as part-whole and subset-superset. For example, the coherence inference in the passage shown in Table 2 could be dealt with by simply applying the fact that *airplanes have wings.* Other cases, such as referring to an object by its superset, as in *Lassie is a collie. The dog runs fast*, likewise could be handled by simple set relationships, such as *A collie is a dog*. The WordNet Project (Miller, *et al.*, 1990), sponsored by the Office of Naval Research, has produced a semantic lexicon, in which a very large number of English nouns have been grouped together into synonymous classes, and various simple semantic relationships have been specified between those classes. For example, the relationships that *airplanes have wings* and *a collie is a dog* are represented in this database.

This work was undertaken to determine whether such a database could be used to effectively improve the quality of the coherence criticisms that CCS could produce. This work is just preliminary, and so is not definitive of the potential success or problems of such an approach. However, it does give some initial indications of what problems would have to be solved to effectively make use of such a general semantic lexicon in the context of a text
~~~~~~

comprehension model, or a text critiquing system, such as CCS.

In the remainder of this report, the key technical features of the work will be summarized; this consists first of a description of how a subset of the WordNet database was integrated into CCS, and how the augmented reference resolution process in CCS worked. Then will be presented a summary of some results where criticisms produced by this augmented CCS are compared with those produced by the original version of CCS. Finally, some conclusions and some suggestions for future work will be stated.

Method

Simplification Approach

Ideally, integrating a semantic lexicon into CCS would take the form of rebuilding CCS's lexical representation and processing so that it in fact made full use of its distinction between words and concepts. This would have required rather extensive rebuilding of the system. Instead a simpler approach was used to explore the potential of using the semantic information in CCS. This approach took the form of leaving CCS's processing mechanisms essentially intact, and adding on a component to use the semantic information where needed. If a reference could not be resolved with the original simple processes, the semantic information would then be considered in an effort to relate the unresolvable reference to the previous passage. Thus the use of the WordNet database was strictly as an add-on to the existing CCS mechanisms. A further simplification was that instead of attempting to use the entire, rather large WordNet database, only nouns were considered, which is justified by the fact that most of the coherence issues involved in technical text involve the relationship between noun phrases. The size of this still forbiddingly large database was reduced by using only the subset of the WordNet database that corresponded to the lexicon already used in CCS. This lexicon was originally based upon one developed by the Navy, and consisted of a basic Navy vocabulary of approximately ten thousand words.

Database Format

The WordNet database is represented as ASCII text files intended to be read by string-processing programs written in the *c* language, and is rather cryptically coded. Since CCS, like most AI programs, is written in LISP, a basic technical issue was converting the WordNet database into a form that was idiomatic to LISP programming, and at the same time was more "human-readable" to facilitate the work. LISP programs were written to parse the WordNet database and convert it into a form corresponding to a semantic net. The basic data format consists of a record for each concept, which corresponds to a WordNet *synset* or set of synonymous terms. Each concept has pointers to the words that can be used to refer to that concept, and pointers to related concepts, such as supersets, subsets, parts, or wholes. All of the simple semantic relationships in WordNet were included in this representation, but only the subset/superset and part-whole relations were used in this work.

Table 3 gives an example of a few entries in this representation. The file containing these entries can be simply read by a LISP program which automatically represents each word or concept as a symbol, by virtue of LISP's built-in mechanisms, and the relations between these concepts and words can be represented as properties and attributes using LISP's property-list feature. Together with LISP's built-in symbol referencing system, this approach provides direct access from one point in the semantic network to another simply by using the GET function in LISP. Thus the expression (GET '^N-AIRPLANE '<<) returns the symbol ^N-AIRCRAFT.

Table 3
Sample of reformatted WordNet database

```
-------------------------------------------------------------------------------------
(^N-AIRCRAFT (AIRCRAFT) <M ^N-FLEET >P ^N-SKELETON3 >> ^N-HELICOPTER >> ^N-GLIDER
>P ^N-FUEL_GAUGE >> ^N-DRONE3 >P ^N-CABIN2 >P ^N-COCKPIT2 >> ^N-LIGHTER-THAN-AIR_CRAFT
>> ^N-AIRPLANE >P ^N-AIRCRAFT_ENGINE << ^N-VEHICLE)

(^N-AIRCRAFT_CARRIER (AIRCRAFT_CARRIER CARRIER FLATTOP ATTACK_AIRCRAFT_CARRIER)
>P ^N-FLIGHT_DECK >P ^N-ARRESTER << ^N-WARSHIP)

(^N-AIRCRAFT_ENGINE (AIRCRAFT_ENGINE) <P ^N-AIRCRAFT << ^N-ENGINE2)

(^N-AIRDOCK (AIRDOCK HANGAR REPAIR_SHED SHED) <P ^N-AIRPORT << ^N-BUILDING3)

(^N-AIR_FILTER (AIR_FILTER) >> ^N-FILTER_TIP <P ^N-VENTILATOR << ^N-FILTER2)

(^N-AIRFOIL (AIRFOIL AEROFOIL) >> ^N-WING6 >> ^N-VERTICAL_TAIL >> ^N-STABILIZER
>> ^N-RUDDER >> ^N-ROTOR_BLADE >> ^N-FLAP5 >> ^N-ELEVATOR >> ^N-HORIZONTAL_STABILIZER
>> ^N-AILERON << ^N-DEVICE2)

(^N-AIR_HAMMER (AIR_HAMMER JACKHAMMER PNEUMATIC_HAMMER) << ^N-HAMMER5)

(^N-AIR_HOLE (AIR_HOLE) << ^N-HOLE8)

(^N-AIR-INTAKE (AIR-INTAKE) <P ^N-CARBURETOR << ^N-DUCT2)

(^N-AIRLINE2 (AIRLINE) << ^N-TRANSPORTATION_SYSTEM)

(^N-AIRLINE (AIRLINE) << ^N-HOSE3)

(^N-AIRLINER (AIRLINER) >P ^N-SEAT5 >P ^N-GALLEY << ^N-AIRPLANE)

(^N-AIRLOCK (AIRLOCK AIR_LOCK) << ^N-CHAMBER2)

(^N-AIR_PASSAGE (AIR_PASSAGE AIR_DUCT AIRWAY) >P ^N-VENT2 >> ^N-UPCAST >> ^N-SNORKEL2
>> ^N-DOWNCAST << ^N-DUCT2)

(^N-AIRPLANE (AIRPLANE AEROPLANE PLANE) >P ^N-WING6 >P ^N-WINDSHIELD >> ^N-TURBOJET
>> ^N-SEAPLANE >P ^N-RADOME >> ^N-PROPELLER_PLANE >P ^N-POD2 >> ^N-MONOPLANE >P ^N-LANDING_GEAR >> ^N-JET3 >P
^N-FUSELAGE >> ^N-FIGHTER4 >P ^N-ESCAPE_HATCH >P ^N-COWL
>> ^N-BOMBER >> ^N-BIPLANE >> ^N-AMPHIBIAN >> ^N-AIRLINER << ^N-AIRCRAFT)

(^N-AIRPLANE_PROPELLER (AIRPLANE_PROPELLER AIRSCREW PROP) <P ^N-PROPELLER_PLANE
<< ^N-PROPELLER)

(^N-AIRFIELD (AIRFIELD LANDING_FIELD) >P ^N-TAXIWAY >P ^N-RUNWAY >> ^N-AUXILIARY_AIRFIELD >P ^N-APRON2 >> ^N-
AIRSTRIP >> ^N-AIRPORT <P ^N-TRANSPORTATION_SYSTEM << ^N-FACILITY5)
-------------------------------------------------------------------------------------
```

Key: Each entry is of the form:
(<concept> <list of synonyms for the concept> <semantic relation> < related concept> ...)
Concept labels are prefixed by "^N-". The relations are:
<</ >> = subset/superset, <P/>P = part-of/has-part, <M/>M = member-of/has-member.

The complete noun database was reduced to correspond to the CCS lexicon with a program that noted which concepts in the semantic database corresponded to words in the lexicon, and then incorporating all of the semantic relationships and concepts needed to connect the lexicon words together. For example, device and wing would be related by a set of intervening set and and part relations. The intervening concepts for the subsets and parts were included in the reduced semantic database. The result was a semantic representation that included all of the semantic information available about the words in the lexicon, and also had considerably more concepts and words, namely those that related the lexicon words together.

Semantics-Based Reference Resolution Mechanisms

As mentioned before, the simple reference resolution process in CCS attempts to relate a noun phrase back to the previously introduced items in the passage. Note that indefinite noun phrases, such as *a magnetron*, actually introduce a *new* referent, and the reader is not normally expected to attempt to identify this item with a previously mentioned item. Thus CCS only attempt to resolve definite noun phrases (those starting with the article *the*) because these are normally a textual instruction to the reader to attempt to make such a connection. In the augmented CCS, the standard simple reference resolution process was first attempted, and then any definite noun phrases that remained unresolved (the *questionable new references*) were subject to a semantic-based search.

The basic strategy of the semantics-based reference resolution was to find a connection through the semantic relationships between the unresolved definite noun phrase and some other item already mentioned in the passage. This process essentially simulated a spreading activation search through the semantic network. First, the semantic relations attached to the head noun of the unresolved noun phrase were examined, and the associated concepts retrieved and put into working memory. It was found necessary to set an arbitrary limit of 100 such retrievals in order to stop the system from getting lost in futile searches. Then a test was performed to determine whether any of those concepts were appropriately related to the head nouns of previously mentioned items in the passage. If not, the semantic relationships between the last set of concepts retrieved for the unresolved noun would then be followed and and a new set of concepts placed in working memory, and the test repeated. If the concepts were appropriately related, then the new reference was designated as a resolved reference, and a proposition added to the passage representation to show the relationship between this *implied* referent and the previously existing referents.

A rather drastic simplification was made; only the head noun information in both the unresolved noun phrase and previously mentioned noun phrases was used in the reference resolution. Because the modifiers in the noun phrases were ignored, this simplification turned out to produce a great many false results, as will be described below. However for purposes of testing the approach, this provides a very liberal test in that it allows the system to make use of *any* relationship found through the semantic network, regardless of whether the relationship is actually the correct one.

There were three semantic relationships tested for in the reference resolution process. If these relationships were found, then the new referent could be taken as implied by a previously mentioned referent. In the *same-concept* relationship, the head noun of a new referent refers to a concept that a previous head noun also refers to, and thus the previous referent implies the new referent. This computation allows reference to an item by a synonym, but since only the head noun was used, many false resolutions resulted.

The second type of relationship, *implied sub/superset*, involved chaining through superset or subset relations, so that a previous item could be referred to in the new referent noun phrase by either a superset concept or a subset concept. Note that normally, referring by a superset is well-defined, as in *Lassie is a collie. The dog is brave.* in which *dog* designates a superset of *collie*. But reference by the subset is logically questionable; for example, *Rover is a dog. *The dachshund is fat.* is unacceptable, because the reference *the dachshund* is not an accepted way to refer to the class of *dog*; in fact, this usage is a way to convey new information in certain settings (see Haviland & Clark, 1974; Clark & Haviland, 1977, for more discussion). But in most situation, it should probably be expressed as *Rover is a dog. The dog, which is a dachshund, is fat.* However, in military text, there appear to be many cases where reference by subset appears, as in *The T-38 is a supersonic aircraft. The fighter ...*, in which *aircraft* is technically a superset (once or twice removed) of *fighter*. While this might again be simply a device to convey new information, it must followed set relationships. Since in this exploration it was desirable to give CCS every opportunity to resolve references, both reference by subset and superset was allowed.

The third relationship, *implied part*, was part-whole relations, with possible intervening subset-superset relations. If the new referent was a subpart of a previously mentioned item, or a part of a subset or superset of a previous item, then it was accepted as an implied referent. For example if an *aircraft* had been mentioned, and the unresolved reference was *the propeller*, the resulting relationship would be that a *propeller* is part of an *airplane* and *airplanes* are a subset of *aircraft*. Thus mixtures of set relationships and part-whole relationships were accepted as implying a the existence of the new referent.

It should be noted that there is no attempt to guide the search by the general context of the discussion, meaning that any connection between the new item and previous ones is accepted. Thus, for example even in the context of an *airplane*, *fighter* in the sense of *boxer* may be considered, and might even be the identified connection if the passage contained any related referents, such as *combatant*.

CCS was augmented with the semantic network and additional production rules to perform the semantics-based

reference resolution. Some example output from the augmented CCS is shown excerpted in Table 4. The relevant semantic net is shown graphically in Figure 2. Notice that the references to parts, supersets, and subset of the initially mentioned aircraft are successfully resolved.

Table 4
Excerpt from CCS output illustrating implied reference processing; input sentences in boldface

```
-------------------------------------------------------------------------------------------------------------------------------------------------------
The F-1 is an aircraft.
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
The wing is long.

IMPLIED-PART
Assuming that these newly introduced items are part of previously mentioned items:
New REF3 WING (concept: ^N-WING6) is part of REF2 AIRCRAFT (concept: ^N-AIRCRAFT)
Check: Is this correct?
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
The flaps are big.
IMPLIED-PART
Check: Is this correct:
New REF5 FLAPS (concept: ^N-FLAP5) is part of REF3 WING (concept: ^N-WING6)
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
The airplane is expensive.
IMPLIED-SUBSET
Assuming that these newly introduced items refer to previously mentioned items:
New REF6 AIRPLANE (concept: ^N-AIRPLANE) is included by, and refers to, REF2 AIRCRAFT
(concept: ^N-AIRCRAFT)
Check: Is this correct?
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
The fighter is essential.
IMPLIED-SUBSET
Check: Is this correct:
New REF7 FIGHTER (concept: ^N-FIGHTER4) is included by, and refers to, REF6 AIRPLANE
(concept: ^N-AIRPLANE)
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
The vehicle has wheels.
IMPLIED-SUPERSET
Assuming that these newly introduced items refer to previously mentioned items:
New REF8 VEHICLE (concept: ^N-VEHICLE) includes and refers to REF2 AIRCRAFT
(concept: ^N-AIRCRAFT)
Check: Is this correct?
-------------------------------------------------------------------------------------------------------------------------------------------------------
```

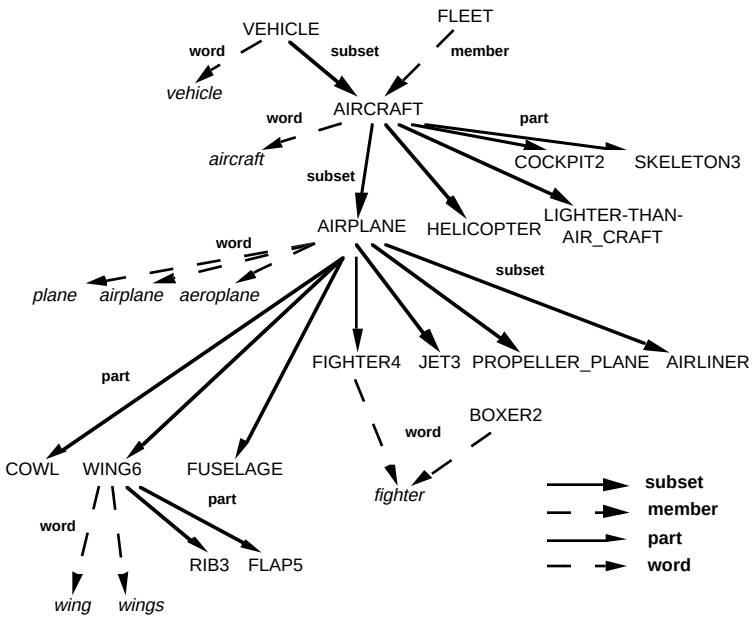


Figure 2. Graphic representation of a portion of the semantic net relevant to the Table 4 example built from the

WordNet data base.

Results

The example in Table 4 shows that the implied reference resolution mechanism operates correctly, within its limitations and simplifications. The results reported here were an effort to determine how often such a mechanism might be useful in processing technical prose of the type CCS is intended for. The measure of usefulness would be how often CCS would be able to identify a reference as being an implied reference, instead of criticizing it as a failure in coherence, and how often this identification would be correct. Due to the simplifications in the process, it was expected that there would be many incorrect identifications, so the question is whether there would be many correct ones.

The augmented version of CCS processed a set of passages, and then the number of questionable new referent was counted. As mentioned above, these are definite noun phrases that could not be resolved using the simple reference mechanisms, and so look as if the reader is expected to be able to identify them with some item earlier in the passage that did not in fact appear. For some of these cases, the implied reference mechanisms succeeded in identifying a relation between the questionable reference and a given referent in the passage. These cases were tabulated in terms of what type of relation was involved, and whether the relationship was correct, that is, whether the identified relationship expressed a reasonable semantic relation between the new and given referent. The processing of references in passage sentences that could not be parsed are not included.

The results and examples will be described separately for each passage. The general result is that many of the questionable new referents were identified as related to previously mentioned referents in the passage, but most of the identified relationships were incorrect, in that they were not semantically reasonable.

A Navy Rate Training Manual Excerpt

Ejectors is the text from about 2 pages of the Navy Machinist Mate 2 &3 manual. It is very typical technical descriptive text that describes the air ejectors used on steam propulsion plant condensers. It contains only a little principles or physical theory; it is mostly a description of the typical system structure associated with air ejectors. Table 5 summarizes the results.

The few cases of correct semantic relations appear to be just fortuitously correct. The incorrect relations found are often due to only matching on the head noun. But note that while this simplification produced many false alarms, it also failed to make many hits. Using context would have blocked some of the false results, but would not have produced any more hits. The coverage of the database is apparently rather spotty, containing word usages that are unusual in a technical context.

A Historical Text

Air war is a 48-sentence passage prepared in collaboration with Bruce Britton for an evaluation study (in progress) of CCS, and is based on one prepared by Britton for recall studies. It had been modified from Britton's original so that all of the sentences could be parsed correctly enough for CCS to produce reasonable criticisms. *Air War* is a discussion of the Johnson Administration's Vietnam War policy of bombing North Vietnam. It is written in a somewhat formal style, with the subject matter being historical, and concerned with policy and administrative decision-making rather than purely technical content. It was chosen for this study after the purely technical passages yielded very few correct implied reference solutions; perhaps the WordNet database does not have enough technical content, and so the less specialized subject matter of *Air War* might engage more of the WordNet database. Table 6 summarizes the results.

The literary style of this passage apparently produced a lot of variation in reference forms, which the same-concept mechanism was able to compensate for through its simple-minded matching only on the head noun. While it was

often correct, this was in fact due to a relatively small number of distinct words being used with different modifiers. Certain very vague and abstract words, such as *significance* and *sense*, produced many false connections, and some of the identified relations were especially out of context. Despite the relatively nontechnical content of this passage, many appropriate connections were not available. For example, the topic of *war* certainly implies *the enemy* in the sentence ...*could not defeat the enemy in the field*, which could not be resolved. The problem is that the relation between *war* and *enemy* is not categorizable in terms of the simple semantic relations; some more complex relationship, such as *action-participant* would be required. Of course, not all references could be resolved on semantic information, such as the clearly "episodic" knowledge required to resolve *the Tonkin Gulf incident* in the context of *The Vietnam War*.

A Pilot's Flight Manual

T-38 Flight Control is the text from about 3 pages of the T-38 Flight Manual (essentially the "owner's manual" for the T-38 supersonic trainer aircraft). It is typical technical descriptive text, and describes the flight control surfaces of the airplane and their associated cockpit controls. It was modified very slightly to increase the number of sentences that would parse successfully by correcting some very idiosyncratic sentence structures, and it was given an overall title of *T-38 Airplane* because the original excerpt made no mention of an airplane being involved until very late in the passage, which severely limited the implied reference searches. Since the same-concept relation did not work very well, it was disabled in the test described here. Table 7 summarizes the results.

This passage contained a few cases which were textbook examples of implied references, due to the database having exactly the required relations. However, the database also failed to include some lower-level information, such as the above-mentioned fact that throttles have quadrants, but even less specialized concepts that switches and controls have *positions*.

Overall Results

Table 8 totals the statistics across the three passages. Using the semantic relations allowed about half of the questionable references to be resolved, but roughly only a fifth of the relations were reasonably correct, and most of these were due to one passage, *Air War* in which varied forms with the same head noun were used in a way that would not be applicable to most technical text. The many incorrect relations could be suppressed by a more sophisticated approach to searching and matching implied references, but it is disappointing that there were not more correct relations found. This can be attributed to the fact that the semantic net, containing the subset of the WordNet database that was related to the CCS lexicon, was not rich enough (at least in the abridged version used here). For example, many of the required part-whole relationships were not present in the technical passages, and the less specialized knowledge involved in *Air War* was also not present.

Table 5
Results for Ejectors *passage*

Total number of sentences and headings:	48
Sentences and headings parsed:	33
Referents constructed:	166
Questionable references:	35
Questionable references not resolved:	12
References resolved via semantic relations:	23
Correct relations identified:	4
Incorrect relations identified:	19

Same-concept relation

Correct: 3 cases

Example: *the most commonly used air ejector* is the same as *the air ejector* introduced in the first sentence of the passage, *The air ejector removes air and noncondensable gases from the condenser*.

Incorrect: 11 cases

Example: *the gland exhaust condenser* is not the same as *the condenser* introduced in the first sentence of the passage.

Part-whole and sub/superset relations

Correct: 1 case

Example: *the steam* in the sentence *Figure 6-13 shows the flow of the steam, air, and noncondensable gases in one type of air ejector unit.* is correctly associated as a subset of *substance* in the previous sentence *The flow of a substance from a higher pressure area ...* .

Incorrect: 8 cases

Example: the *valve* in the sentence *When you open the make-up feed valve* was incorrectly identified as an electrical component (via the British *valve* = vacuum tube) that is related to the previous *condenser*, which is an obsolete synonym for capacitor, an electrical component.

Table 6
Results for Air War *passage*

Total number of sentences and headings:	48
Sentences and headings parsed:	48
Referents constructed:	313
Questionable references:	71
Questionable references not resolved:	44
References resolved via semantic relations:	27
Correct relations identified:	10
Incorrect relations identified:	17

Same-concept relation

Correct: 10 cases

Example: *The primary objective* was correctly identified with *the objective* in the previous ...*serious differences arose over both the objective and the methods to be used.*

Incorrect: 4 cases

Example: *The beginning* in *From the beginning, Rolling Thunder was hedged with restrictions...* was identified with *the source* in *...Hanoi as the source of the continuing problem in the south.*

Part-whole and sub/superset relations

Correct: 0 cases

Incorrect: 13 cases

Example: *The face* in *...would not risk its fragile and limited industrial base in the face of overwhelming American power* was interpreted as a part of the human body, and was circuitously associated with *the extension* in *...over the extension of the war...* as being a body part.

Table 7
Results for T-38 Flight Control *passage*

Total number of sentences and headings:	91
Sentences and headings parsed:	67
Referents constructed:	287
Questionable references:	41
Questionable references not resolved:	17
References resolved via semantic relations:	24
Correct relations identified:	3
Incorrect relations identified:	21

Part-whole and sub/superset relations
Correct: 3 cases
Example:The *flaps* in the sentence *The wing flaps are electrically controlled by a flap lever.* are recognized as a part implied by the object mentioned in the title: *T-38 Airplane*

Incorrect: 21 cases
Example: The *quadrant* in *The wing flap lever is located on the throttle quadrant of each cockpit.* is not recognized as part of the equipment in an airplane cockpit, but is incorrectly associated via the concept *measure* with *the amount* in the previous sentence ... *by increasing the amount of horizontal tail deflection ...* .

Table 8
Overall results totaled across the three test passages

Total number of sentences and headings:	187
Sentences and headings parsed:	148
Referents constructed:	767
Questionable references:	147
Questionable references not resolved:	73
References resolved via semantic relations:	74
Correct relations identified:	17
Incorrect relations identified:	57

Same-concept relation (two passages only)
Correct: 13 cases
Incorrect: 15 cases

Part-whole and sub/superset relations
Correct: 4 cases
Incorrect: 42 cases

Conclusions

This work shows that a natural-language processing system such as CCS can make use of a semantic database such as WordNet. However, in the test cases here, a broad, shallow, general database such as WordNet does not seem to have enough of even the unspecialized knowledge to substantially reduce the number of unresolvable questionable references. That is, while knowledge that throttles have quadrants is quite specialized, the notion that wars have enemies is not. Making WordNet more complete in the relevant ways would require a more complex set of semantic relations, and would also require it to take on a specialized flavor; for example, in the technical domain of airplanes, the database would have to include the complete list of airplane parts that it is reasonable to assume the typical reader knows. This would be a long list, but the remarkable thing is that as large as the WordNet noun database is, it apparently has only a few facts about each technical domain; the result is a large, general, semantic lexicon that does not apply very much in any single domain. Thus, a WordNet-style database would have to be much larger, and constructed with much more of an eye towards technical coverage, in order to form a basis for the types of semantic reference resolution explored in this work.

Two possible routes for further work are possible. First, many of the cases where the semantic relations were correct would be amenable to a simpler treatment; for example, identifying *primary objective* with a previously mentioned *objective of the war* could be done with an extension to the current simple reference resolution process, and CCS could make an appropriate comment asking the writer to check whether this is the intended meaning. Thus these results suggest some extensions to the current no-semantics approach in CCS. In fact, these results imply that the original judgment to try a no-semantics approach was a good one, given that most of the passage references could be handled with the original simple resolution process, and of those references that couldn't, a large semantic database was not very useful.

A second approach would be more scientifically interesting. The lexicon in technical domains could be characterized more systematically, and the semantic databases for technical domains could be developed. This could then done by hand, but an interesting possibility would be to develop natural-language processing software capable of processing the definitions in a technical glossary to construct a WordNet-style list of simple semantic relations automatically. The mechanisms in CCS that can parse many technical sentences and resolve references might make good foundations for such software.

References

Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed.), *Discourse processes: Advances in research and theory*, Vol. 1. Norwood, NJ: Ablex.

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, **13**, 512-521.

Kieras, D.E. (1989). An advanced computerized aid for the writing of comprehensible technical documents. In B. Britton & S. Glynn (Eds.), *Computer Writing Environments: Theory, Research, and Design.* Hillsdale, NJ: Erlbaum.

Kieras, D.E. (1990). The computerized comprehensibility system maintainer's guide (Tech. Rep. No. 33, TR-90/ONR33). Ann Arbor: University of Michigan, Technical Communication Program.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Five papers on WordNet. (CSL Report 43), Princeton, Princeton University, Cognitive Science Laboratory,