Black, Ezra William

TOWARDS COMPUTATIONAL DISCRIMINATION OF ENGLISH WORD SENSES

City University of New York PH.D. 1987

TOWARDS COMPUTATIONAL DISCRIMINATION OF ENGLISH WORD SENSES

by

EZRA WILLIAM BLACK

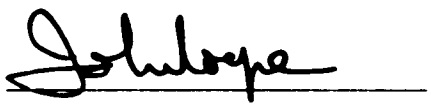
A dissertation submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

1987

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

January 15 1987

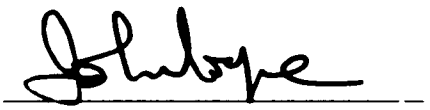
Date



Chair of Examining Committee

January 15 1987

Date



Executive Officer

Prof. Martin Chodorow

Prof. Frederick Damerau

Prof. D. Terence Langendoen

Supervisory Committee

The City University of New York

Abstract

TOWARDS COMPUTATIONAL DISCRIMINATION OF ENGLISH

WORD SENSES

by

EZRA WILLIAM BLACK

Advisor: Professor John A. Moyne

An experiment is conducted which compares three different methods of deciding which of three or four senses characterizes each occurrence of a word for which a Key Word In Context concordance has been constructed. The three methods consist of a dictionary-based approach (DG) where categories intended to classify the words and expressions occurring in each concordance line are simply the subject codes of a major dictionary; an approach (DS1) in which categories are obtained via a frequency analysis of words occurring in the immediate neighborhood of the "node word"--the word in focus--of the concordance, and of "content" words occurring anywhere in a given line; and an approach (DS2) chiefly based on content-analytic categories obtained by closely reading the concordances of a 100-type sampling of words occurring in the 20-25-million-token English text source, consisting of the official proceedings of the Canadian House of Commons. Results are that DG performs extremely poorly--in fact, near-randomly; DS1 and DS2 yield better and substantially similar performances. The conclusion is that domain-general, syntax-based approaches to automatic word sense discrimination and domain-specific, content-analytic approaches need and complement each other.

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to the following individuals:

the chairman and members of his thesis committee, for their intellectual guidance and friendly encouragement: Professor J. A. Moyne, Professor M. Chodorow, Professor F.J. Damerau, Professor D.T. Langendoen;

his managers at the IBM T.J. Watson Research Center, and the Research Staff Members of the Continuous Speech Recognition Group:
Dr. F. Jelinek, for intellectual leadership and for his financial support and encouragement;
Dr. L. Bahl, for assistance and encouragement, and for creating a supportive work environment;
Dr. R.L. Mercer, for his indispensable assistance with the mathematical and computational aspects of this study, and for his encouragement;
Mr. P. Brown, Dr. P. DeSouza, Dr. A. Nadas, for their assistance and encouragement;

Mr. M. Scott, Columbia University Dept. of Computer Science, for his extensive programming and mathematical contributions to this study, and for his enthusiasm;

Dr. J. Cocke, IBM Fellow, for the original idea which began the thinking that led to this thesis, and for financial support and encouragement;

Mr. R. Byrd, Manager, IBM Research, for his generosity in helping the author use his TUPLES package optimally, and for his encouragement;

Dr. D. Johnson, Dr. S. Katz, Dr. J. Klavans, Dr. P. Postal, Research Staff Members, IBM Research, for helpful discussions and friendly encouragement;

Dean R. Smith, Professors M. Gross, J. McH. Sinclair, for helpful discussions and for encouragement;

Ms. R. Albrecht, Professor G. Axel, Ms. A. Black, Mr. A. Black, Ms. C. Cambridge, Ms. E. Czerwczak, Mr. G. Dwornick, Ms. R. Feldman, Mr. R. Feldman, Ms. R. Greenfield, Dr. C. Haspel, Ms. S. Kessler, Professor S. Korner, Ms. J. Savage, Dr. R.C. Plotkin, Mr. R.L. Plotkin, Ms. S. Plotkin, Dr. R. Plotkin, for their friendship, support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF DIAGRAMS

INTRODUCTION

For a considerable number of years, questions as to what constitutes a proper or justified research program within linguistics have been leitmotives of the professional discourse of students of language. Katz (1972), for instance, argues that one has foreknowledge of the rough limits of the subdomains of theoretical linguistic study, within the generative paradigm. On the other hand, de Saussure (1969 pp.149-150) makes the case that it is the theory one develops which itself ultimately determines its own sub-divisions; the point has been reiterated more recently as well (Fiengo (1980)). As another example, Chomsky (e.g. 1980 pp. 24, 218) has argued for what he calls the "Galilean style of inquiry", characterized by the development of abstract mathematical models of particular domains whose constructs and laws are to be considered in some sense more "real" than the sort of quotidian apperceptions on whose basis scientists are wont to found experimental and theoretical conclusions. Among others, Botha (1982) has found this approach unsatisfactory as a complete methodology of language research, insisting that it be at least complemented by efforts which aim at generalizations with more immediate justification in the relevant data.

One element which such programs seem to share is the rather considerable leaps of faith which are required in order merely to get them off the ground. In addition, Gross (1979) has argued for the importance of

a broad coverage of the "brute facts" of syntax as a preliminary to theory-building. While it is certainly not outlandish to reply that the only worthwhile way to counter a theory which accounts for some of the data is to supply one which covers the same data and other facts as well, or explains the same data more parsimoniously, it is nonetheless also true that there are many linguistic applications which demand more comprehensive coverage of "the facts" than such programs, rightly or wrongly, are liable to yield for a long time to come.

Research programs aimed at addressing questions which do not seem susceptible of answer unless attention is devoted not just to crucial cases, but to run-of-the-mill language--such programs have often been considered either impossible, or incoherent, or just uninteresting by theoreticians. But actually, it is possible to adopt as the basis for inquiry questions which, far from being incoherent or impossible, are characterized by procedures of evaluation at least as palpable and demonstrable as those alluded to above. As to their interest, it is well known that a host of practical applications await the solution of the chief difficulties experienced to date in dealing with everyday language use.

Inquiry of the sort just indicated is labelled by many theorists as "research into performance". However, as Moyne (1980) points out,

> What is controversial in linguistics...is that a theory of competence or the grammar of language representing the knowledge of a native speaker, can be formulated for an ideal situation or environment in which there are no external interferences or "noises", and that the investigation will necessarily obtain profound insights into and even a major component of any theory

> of performance. Note, incidentally, that these assumptions are made PRIOR to the investigation into these matters (p.460; capitalization here and throughout replaces italics in original).

The central question underlying the research to be presented here concerns what can be called "predictivity in text". At its most general, the question is this: Given a coherent text, partitioned at any arbitrary point into a beginning and an ending portion, a "known past" and an "unknown future", how may we optimally predict the exact character of the "unknown future"? While any aspects of the past are of course theoretically available as bases of prediction, the only aspect of the future in which we are ultimately interested is its correct representation as a succession of graphemic inscriptions--as printed words. Whichever generalizations defined over spoken or written text eventuate in the most consistent and accurate predictions of the future, are considered ipso facto to be valid.

One generalization which seems intuitively to offer considerable predictive power is the particular operational definition of "a usage of a word" or "a sense of a word" which will be developed below. Assume that a particular word occurs at some position in the past relative to the next word to be predicted. It would be very surprising if the expectations generated by this knowledge were not less accurate than those licensed by the knowledge not just that the word occurs there, but that, say, sense 2 of the four senses recognized for the word, occupies this position of the text.

To take an extreme example, consider the word "will". While in its modal use it is one of the most frequent items in many sorts of text (cf. e.g. Kucera and Francis, 1967), there are certainly a large number of text domains in which its appearance as, say, the "will" of "last will and testament" narrows the range of its likely right neighbors by impressive amounts, as against the uncategorized occurrence of the inscription "will". A more mundane example, and one which does not coincide with part-of-speech differences, is the use of "movement" to mean either "labor movement, student movement, political movement", etc., or "movement of goods", or "movement of (in) events, negotiations" and so forth. It seems clear that at least in the cases cited, predictions of right context will differ appreciably depending upon the sense which occurs. As there are many different ways to apply knowledge gained from sense discrimination procedures to the problem of prediction, and as we will return to this issue in Chapter Three, we will not belabor the issue here. For the same reason, we will not linger over over the points that boundaries among senses are fluid, and that intuitively a number of different sense labels can often be justified for the same occurrence of a word. These issues are treated in Chapters Two and Three.

This thesis will be concerned, then, with describing, analyzing, evaluating and applying a number of approaches to the definition of "usage of a word", as this notion applies to the analysis of text. The description just referred to comprises both the review of relevant literature in Chapter One, and the presentation of new lines of attack on the problem in Chapter Two. The analysis to be carried out rests on certain

general ideas about language and its use, which will be discussed in Chapter Two. Evaluation takes place in Chapter Three, where three distinct methods of word sense discrimination are applied to a particular experimental task, that of disambiguating all the occurrences, within a 20-25-million-word English corpus, of a sample set of words. In the same chapter, conclusions are drawn from the experimental evidence presented as to the directions research must take in order for a definitive solution to the problem of word sense identification to be achieved.

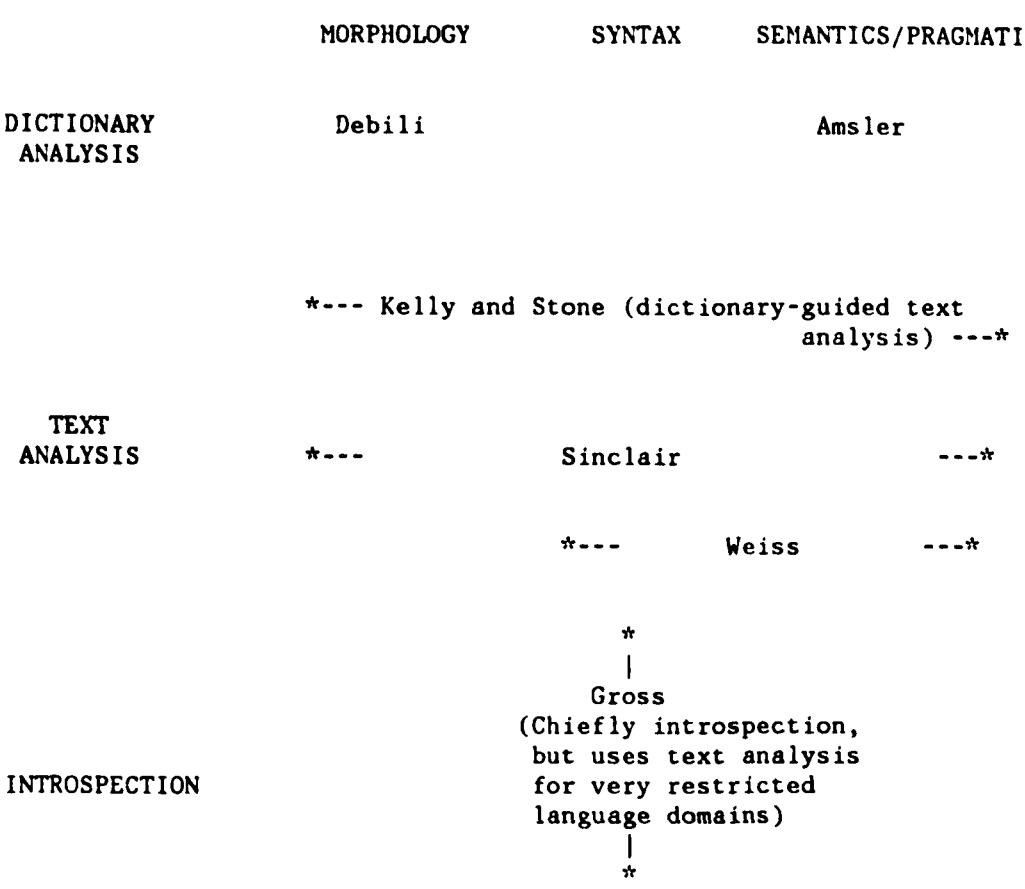
CHAPTER ONE: REVIEW OF THE LITERATURE

The past decade or so has seen about a half dozen serious attempts to attack the problem of word sense disambiguation (1). The purpose of this literature review will therefore be to consider these in some detail, and to relate them to each other.

Before proceeding, however, we wish to acknowledge some research in the field of word sense discrimination which we will not be examining. Work in what might be called the school of conceptually-driven or semantics-driven natural language processing (2) does indeed confront the issue of how word senses may be identified in text. However, the unfortunate fact is that these efforts have not yet progressed to the point where any actual evidence has been produced which might help decide how a real system is to be designed which is able to perform under conditions of relatively unrestricted text input. The work in question includes Granger (1977), Hayes (1975), Small (1980, 1983), and Wilks (1972).

The sense discrimination research we will discuss is that of Amsler, Walker, Debili, Gross, Kelly and Stone, Sinclair, Weiss, and their associates. A rough classification of this work is provided in the chart below.

The labels of the horizontal axis of the chart on the next page are to be understood in an informal way, as mere aids to orienting oneself to the various approaches of the scholars mentioned. To oversimplify somewhat, the chart can be interpreted as saying the following: Debili uses morphological properties of words to determine their senses; he learns these properties by means of dictionary analysis. Amsler uses semantic and pragmatic characteristics of words for disambiguation, and derives these, again, from the dictionary. Kelly and Stone use morphology, syntax, and semantics/pragmatics, and get their information from text analysis guided by dictionary consultation. Sinclair employs all three aspects of words, and relies upon text analysis for his data. Weiss can only be said to be practicing textual analysis, and to be "utilizing" syntax and semantics/pragmatics, in the most latitudinarian sense. However, if he is to be assigned a place in our classification scheme, this would be it. Finally, Gross depends on syntax as the driver of disambiguation, and obtains his facts via introspection, and, in the case of very-restricted-domain studies, via text analysis.



A third dimension to the above informal matrix suggests itself. This is the extent to which each worker sees his approach as appropriate to "English in general", "French in general", or whatever language is under study, "in general". The low end of this scale would be the position that one does in fact have a method applicable to a variety of domains, perhaps all, whatever that would mean, but only to one at a time. For instance, if one utilized content categories, this low-end claim would be that some, most, or all of these classes were more or less particular to the domain with reference to which they were induced. The high end of the scale

would be in effect to deny the relevance of domains to the enterprise of word sense discrimination. In particular, one would claim that one's procedures of sense identification were more or less as good on any one type of text as on any other. The reason this third dimension is not sketched in--besides typographical awkwardness--is that this issue seems either unclear or confused in the work of most of these researchers. The topic will be touched on lightly in the discussions below, and then returned to in later chapters.

Debili tackles word sense discrimination within the context of a larger effort (Debili (1982)). This is a highly interesting attempt to construct, by computer, dependency-grammar representations of input sentences from any domain. Preliminary processors cut the input sentence into nominal and verbal groups or phrases. Then within each such group all potential dependency relations--noun and its adjectival modifier, noun-noun modification, verb-subject, verb-direct object, and so on--are temporarily assigned. In many cases, the heuristics of link identification permit only a single link to be established, or none at all, for each type of dependency relation available. Otherwise, a situation of ambiguity is passed on to the morphological and semantic routines. At this point, the jettisoning of any remaining incorrect dependency links is accomplished. Basically, a dictionary of already-seen and "validated" dependencies is maintained, and the "most valid" of the already-recorded links which are possibilities in the input sentence is accepted as correct. "Validity" involves a simple rating system over "paradigmatic" and "syntagmatic" similarities between a given dictionary dependency and the

proposed link being evaluated. Two words are most related paradigmatically if they are identical, and increasingly less related the more they differ morphologically, while preserving the same root morph. Examples of decreasing "validity" are "affichages" -- "affichage" -- "afficher" and "muraux" --"mural" -- "mur". A proposed syntagmatic dependency is most valid if it is identical to one in the dictionary, i.e. to one already recorded. Otherwise, it either may be considered invalid, or a range of weightings may in effect be established covering each possible pairing of dependency relations. For example, if we have already seen "affichages muraux" (roughly, "things hung up on a wall or walls"), we do wish to be able to induce a close linkage to, say "afficher X sur le mur" ("to hang something up on the wall"), whereas, to take a somewhat unlikely example, we do not want to find any similarity to "des affichages et des murs" ("pertaining to things which are hung on walls and to walls themselves"). In any case, Debili does not pronounce himself as to whether, in practice, he simply sets syntagmatic validity at 0 in the absence of strict identity between dictionary entry and new, putative dependency relation, or whether he sets up some system of weightings. Debili reports success rates in the 90 percentile in choosing the correct dependencies, over a test corpus of the order of magnitude of 10,000 words. This is, then, a novel, and apparently promising approach to a number of thorny problems of computational linguistics. In the mainstream tradition these are known as the issues of "prepositional phrase attachment", "noun-noun modification", and the general problem of treating "discontinuous dependencies". Also note that there is a qualification to be made to the above chart's classification of Debili as a researcher

who relies on the dictionary for his basic information. This is true for his word sense disambiguation efforts, which we are about to describe. But this work is just a small part of the dependency-relation research we have just discussed, and for these purposes, it is clear that he is very much dependent on actual text, and not the dictionary, for his information.

Debili leaves open (p. 244) whether his link-resolution procedures will converge only in a local, or domain-dependent fashion, or whether the convergence will be global, i.e. domain-independent. Of course there is also the possibility of no convergence at all. However, he is able to present results of monotonic increase over successive trials on three texts of several thousand words each, showing that learning was in fact taking place. Actually, with such a small corpus, it is quite interesting that any positive results at all were obtained. Realistic evaluation of Debili's general approach, in fact, awaits its application to corpora several orders of magnitude larger than the ones he used, for the reason just alluded to (3).

It is against the foregoing background that Debili's approach to word sense disambiguation must be seen. Debili obtains the morphological information he uses in disambiguation by processing a large reverse-sorted word list apparently derived from a traditional dictionary and classified in the following manner: those inflectional and derivational forms considered "permissible" with reference to a given major meaning of a word (i.e. a numbered definition rather than one indicated by letter) are

grouped together. By automatically analyzing all listed words into roots, prefixes, and suffixes, Debili arrives, then, at a listing for each major meaning of a word, consisting of all licit sequences of affixes which may combine with the word under a particular major meaning (4).

What Debili does to decide which major meaning is present in a situation where both syntactic and semantic lines of defense have still left an ambiguity--say, in a sentence such as "Le bail expire a la fin du mois", where "expire" can be either "expire(FINIR)", "expire(MOURIR)", or "expire(RESPIRER)"--is the following: Find essentially the maximum of the cartesian products of the "validity" scores of the "word families"--i.e. of the listings of licit morphological variants in the sense of the last paragraph--of the uncertain word, on the one hand, and of the other member of the two-valued dependency relation involved, on the other hand. This latter member is ex hypothesi unambiguous. In the present example, the dependency relation is subject/main verb ("bail"/"expire"). "Bail" is the constant member of the two-valued dependency relation, and "expire" is the polysemous, i.e. variable member of the relation. Now, each member of the morphological word family of "bail" is looked up in the listing of previously-seen dependency linkings to determine whether it has occurred in the relation "subject/main verb" with each of the members of the "word families" of each of the senses of "expire". As a first approximation, we can say that the sense with the most such occurrences is selected as the correct one. However, the actual story is that even occurrences with relations other than the "subject/main verb" relation can count in favor of a sense selection, if it is, roughly, a transform of

"subject/main verb" and if each of the arguments of the relation appeared in the morpholgical form appropriate to that transform of "subject/main verb". So, for instance, we might previously have seen "l'expiration du bail" or "un bail deja expire(PAST PART)", and these instances would have validity values which would help the relevant sense of "expirer".

To be precise, we may formalize Debili's calculation as:

$$S = \operatorname{argmax}_{S} \sum_{w \text{ in } f_a} \sum_{w' \text{ in } f_b} D(w,w')$$

where a = the unambiguous word, i.e. "bail" in the example above; where b = the ambiguous word, that is, "expire" here, f_a is the word family of the unambiguous word, i.e. of "bail" here; f_b^S is the word family of a sense of the ambiguous word, i.e. "expire(FINIR)", "expire(RESPIRER)", etc.; w is a member of f_a; w' is a member of f_b^S; and D(w,w') is the dependency function--in this case "subject/verb", i.e. w is "subject of" w'. We take as the sense of b the maximum value for S (i.e. argmax) in the above formula.

Now, as Debili himself points out, there are problems here. Basically, either the "word families" of the various senses must be disjoint, or else somehow it must already be known which sense was involved in previously-seen relevant text sequences. That is, the fact of the matter

is, for example, that the Larousse Dictionary lists as the senses of "expiration" both the RESPIRER and the FINIR meanings. Therefore, there is no way a computer would know that it had seen the FINIR sense when it recorded "l'expiration du bail", since it could also be the RESPIRER sense. This is a very serious problem, and, as just pointed out, Debili is fully aware of it: "Il est clair cependant que ce mecanisme ne sera possible que si les familles considerees ne sont pas entierement polysemiques, puisqu'il faudrait alors amorcer manuellement le processus en enregistrant prealablement dans la connaissance la relation qui convient" (p. 253). From these comments, one surmises that Debili feels that full disjointness is not necessary for his purposes, since he will have enough data over those morphological forms which do differ from sense to sense to tip the balance. This is a highly speculative notion, since his experimentation with this procedure involved literally only a few cases. The same conclusion therefore applies to Debili's sense resolution mechanism as was reached with regard to his dependency-selecting algorithm (of which, in fact, the former is a part): No acceptable evaluation can be made until it is run on a corpus many orders of magnitude larger than that it has so far encountered.

If for Debili, "sense disambiguation is morphology", for Gross and his associates, "sense disambiguation is syntax". Gross (1985) presents an interesting and provocative orientation toward sense discrimination in English (and, by implication, in French and other Romance idioms as well). To illustrate his approach to the problem, which relies upon syn-

tactic subcategorization in differentiating senses, Gross steps through the envisioned automatic syntactic analysis of the utterance:

(D) Two men cleaned the offices, then, they waited for the janitor

In particular, he considers the issue of resolving the reference of the word "they" as it occurs in D. Prima facie, of course, "they" could be co-referential with the noun phrases "Two men" or "the offices", or with some third noun phrase referred to earlier in the discourse; or else it could refer arbitrarily, as it were, as in (to cite Gross's example):

They (= the Government) are wasting our money. (Gross 1985 p. 2)

The last two possibilities are set aside, so that attention may be focused on the two above-mentioned noun phrases. By disambiguating "offices" and "waited", then, and by applying a variety of data stored in the lexical listings of the two words, Gross is able categorically to rule out "offices" as a possible antecedent of "they". Later, D is expanded and is subjected to various lexical substitutions in order to display further properties of the tack being taken with regard to disambiguation.

What, then, is Gross's approach? Each word of his lexicon corresponds to a line of a lexicon-grammar, which is a two-dimensional matrix, apparently on the order of 32,000 x 400 (Gross (1982, 1985)) for French at present, and of uncertain size for English (6). Columns of the lexicon-grammar are labelled with possible syntactic properties of an

entry word. Most of these consist of intra-clausal syntactic environments in which the candidate word can or cannot appear in a given slot; these environments can be sequences of syntactic categories to the right and/or left of the word, or they can be sequences of lexical items to the left and/or right of the word, or, finally, they can be combinations of the latter two possibilities. In addition, certain properties of one or more nouns involved in the environment-- properties which some linguists have called syntactic, others semantic--such as "concrete", "animate", and "human", are labels of still other columns. Each cell is marked with a "+" or a "-", depending on whether the team of linguists conducting the research finds the clause represented by the cell to be acceptable or unacceptable. (For a description of Gross's methods of linguistic team investigation, see Gross (1984).)

The details of Gross's use of facts about "offices" and "waited" to resolve the reference assignment problem mentioned above are as follows: If "they" = "offices", then "offices" is the subject of the verb "waited". So the lexicon-grammar entry for "wait" is consulted (or would be analyzed by the computer program being simulated). Only verbal entries for "wait" need to be scanned, of course, because of the "-ed" ending it bears in D. Moreover, the three entries requiring an "-ing" ending for "wait" are also not examined. (These are "keep (someone) waiting", "No Waiting", and "waiting game".) The columns of the lexicon-grammar then allow us to eliminate the following entries for "wait" as not accepting an "-ed" terminally:

Max can't wait (*can't have waited, *couldn't have waited) for the results.

Wait and see (*I waited and saw, *I have waited and (have) seen)

Just (you) wait (*Just (you) waited)

While you wait (*While you waited).

Two additional entries depending upon the presence of certain adverbs immediately after "wait" ("waited") can be eliminated as possible matches for the instance under analysis, since the requisite adverbs are not found. They are as follows, where abbreviations are: N0, first noun phrase; N1, second noun phrase; V1, verb phrase; W, variable word string:

N0 wait(ed) around/about/behind/in/up for N1

and

N0 wait(ed) out Adverb for N1.

Similarly eliminable are:

N0 wait(ed) on N1

and

N0 wait(ed) upon N1.

The unique entry left in the lexicon-grammar for the spelling "waited" is:

N0 waited for N1 to V1 W,

and in fact D does feature "for" immediately after "waited". The only remaining indeterminacy left with regard to "waited" is that the

lexicon-grammar tells us that "wait(ed) for" takes both human and non-human subjects (Gross 1985 pp.1-7).

Now consider the cases where "they" is non-human, namely where its antecedent is "the offices". All compound nouns involving "offices" are barred from consideration because the necessary companion words do not occur in the appropriate places in the environment of "offices" (e.g. post offices, doctor's offices, box offices, divine offices). "Offices" is the object of "cleaned" and, if "offices" is the antecedent of "they", then "offices" must be able to serve as subject of "waited". Since the lexicon-grammar entry for "cleaned" stipulates that it must have a concrete object, only one of the six entries for "offices" is permissible in this role. The six (excluding compound nouns) are: have an office (a physical office); work in an office (an abstract idea); the office celebrated Bob's promotion; N0 hold, take an office; N0 propose his offices to N1; N0 performed an office. It is the first of these entries which remains alive. It is claimed at this point, without elaboration, that 'If we look up the entry for "wait", we see that no entry of "office" can be the subject of "to wait" ' and that 'hence, the ambiguity is resolved: "they" cannot refer to "offices", and (D) is not ambiguous. "They" = "two men" ' (Gross 1985 p. 7). (7)

This is an impressive example of the use of one approach to sense disambiguation in natural language processing, and it is not hard to imagine similar methods being used for predictive purposes: for instance, instead of asking for the antecedent of "they", we might have assumed we

had seen, of D, only "Two men cleaned the offices, then, they" and have interrogated the lexicon-grammar as to the identities of "possible next verbs". The present study focuses on word use rather than syntactic environment as a vehicle for sense disambiguation. However, it is of some interest that words were sometimes found that were susceptible of nearly complete disambiguation on the basis of syntactic environment. An example is "due". If we assume the following six senses for "due", then most of them may be identified, in the main, via a particular pattern of subcategorization (8):

1. due respect, due regard, and similar expressions
   (= due + abstract noun)

2. due (totally) to the fact that..., due (more or less) to circumstances...
   (= due (+ degree adverb) + to + non-verb)

3. payment (immediately) due; payment is due
   (= noun (+ adverb) + due; or = copula + due)

4. give someone their (rightful, just, etc.) due
   (= possessive pronoun (+ adjective) + due)

5. due to start, due for completion, due date
   (= due + to + verb; or = due + for + nominalized verb; or = due + time noun)

6. a debt of gratitude is due someone; the respect due someone
   (= (gratitude, respect, where credit is)...+ due)

Returning to Gross (1985), the sense definitions which were set up there and whose usefulness was demonstrated, follow from the bold position which Gross adopts on the question of what criteria define a word sense. Lyons (1977 pp. 553-555), in discussing the notions polysemy and homonymy, notes that there are, to begin with, two radical positions which one can

take on the question of the relation of these two ideas. On the one hand, one can claim that there is no such thing as a homonym, i.e. that all meanings associated with a given inscription are its alternate senses--the word is polysemous. On the other hand, it might be asserted that each new sense of an inscription betokens a new word--a homonym--which for one reason or another is the same physically as a set of words whose meanings resemble its own to varying degrees. Gross's point of view is a version of the latter position.

Gross proposes "that dictionary entries be characterized by their contexts...from a strictly syntactic point of view" (Gross, 1985, p.3). He argues that one must set up a separate lexicon-dictionary entry, i.e. a new row, whenever it is the case that two environments of a given inscription are not related transformationally. One example, from Gross (1984, p.378), illustrates three of the senses which would be set up for the French verb "voler" (to fly):

<Sense 1, EB>

Cet avion vole a une altitude de 1.000 m.
(This plane flies at an altitude of 1.000 (sic) m.)
=
Cet avion est en vol a une altitude de 1.000 m.
(This plane is in flight at an altitude of 1.000 .)...

<Implied(EB): Cet oiseau vole a une altitude de 1.000 m.
(This bird flies at an altitude of 1,000 m.)...>

*Cet oiseau est en vol (E + lent)
(This bird is in slow flight.)...

<Here a transformational relationship exists when the subject is a plane, or a concrete inanimate object perhaps--Gross does not make any specific claims here--which fails to hold when the subject is a bird, or an animate object. EB>

Notre avion a vole directement de Lyon a Nice.
(Our plane flew non-stop from Lyon to Nice.)

=

Notre avion a fait un vol direct de Lyon a Nice.
(Our plane made a non-stop flight from Lyon to Nice.)

Les oies volent sur des distances sans s'arreter
(Geese fly over distances without stopping)

=

*Les oies font des vols sur des distances sans s'arreter.
(Geese make continuous flights over distances)

<This is a second transformation permissible under the first set of circumstances noted above, and illicit in the second such case. EB>

Other uses of the verb "voler"--"to fly" will have to be distinguished, as for example

Max vole sur Boeing 321
(Max flies a Boeing 321)

<Implied(EB): *Max est en vol sur Boeing 321
(*Max is in flight on a Boeing 321)

*Max fait des vols sur Boeing 321
(*Max makes flights on Boeing 321s)>

<Here the subject is human, and neither of the above two transformational possibilities is allowed. Hence a third sense of "voler" is deemed to exist. EB>

From our point of view (see Introduction) the important questions concerning this stance of Gross's are whether it will lead to effective methods of sense identification, and if so, whether such senses, once located, are of the correct degree of generality to allow correct prediction of linguistic "events" in the surrounding text. One can only speculate as to the answers to these questions in the absence of actual testing on large amounts of data, and a fortiori of a running system permitting such experimentation. Neither of these conditions is currently met, to our knowledge. Hence, while intuitively it seems that this approach has a good deal to contribute to the sense resolution problem, a more solid verdict awaits the method's implementation.

Gross and his associates are involved in a long-range ongoing research program aimed at the development of lexicon-grammars, and view this effort as a prerequisite to progress in the statistical analysis of text (Gross 1985, pp.10-11). It seems fair to surmise that, while he considers such a tack to be necessary, he would not count it sufficient. He states: "The proposal that dictionary entries be characterized by their contexts is a classical idea that we have approached from a strictly syntactic point of view," and follows this with this footnote: "I. Melchuk et alii 1984 have approached a similar question from a semantic point of view" (Gross 1985 p. 3). The work to be presented here approaches the issue from a third and complementary vantage point--that of word usage.

The example of "due" given above, and others like it, indicate that syntactic environment can be a powerful tool in word sense discrimination.

Gross (1985), along with other, related work (Freckleton (1983), Gross and Tremblay (1985)) is indicative of the serious progress which Gross and his associates have already made in this direction , and of continuing effort underway to deepen and extend this success (9).

Of the research efforts reviewed in this chapter, Kelly and Stone (1975) is probably the one most closely related to the work to be reported here. Nonetheless, its title, Computer Recognition of English Word Senses, is rather confusing, in that it suggests that it is a method for distinguishing word senses or usages which is presented. In fact, their work is principally concerned with part-of-speech differentiation in text, a different problem and one which, unlike that of sense discrimination, is at present well in hand (10).

The problem confronted by Kelly and Stone arose within the context of the operation of a suite of programs previously developed by Stone and his associates whose purpose was to pass through a text, assigning to each appropriate word a label representing a particular content-analytic category (11). They found early on in the application of the system that its utility was sharply reduced by its lack of ability to assign parts of speech and word senses (Kelly and Stone (1975, p.1). What they therefore did was to construct a preprocessing program to perform such assignments.

Drawing from a corpus of about six million words within the domain of interest to the content analysts they wished to serve-- namely behavioral science--they selected a sample of 510,976 words (Kelly and Stone

(1975, p. 5). They then restricted their efforts to the part-of-speech labelling and/or sense disambiguation of 1815 words. While they state (p. 10) that these 1815 items are those with a frequency of 20/510,976 or better, their actual dictionary entries are stated in terms of lemmas (12). Therefore it is not clear what the inclusion criterion really amounted to.

Approximately eight individuals--undergraduates, graduate students, Kelly, and Stone--together worked from a KWIC concordance of the items selected, over the 510,976-word corpus (13). The first step in the data analysis carried out consisted of establishing senses for each word. General orientation to this task was gleaned by consulting an unabridged dictionary's definition; at this point attention was turned exclusively to the concordance itself as the source of information as to the use of the words under scrutiny. In the case of senses, as opposed to parts of speech, it was found that the dictionary served in only the most general way to guide the formulation of definitional criteria. Even where the dictionary senses could be adopted as such, "the set of senses is itself relative to our pragmatic aims--i.e. we ask which of 'the senses' of this entry seem useful, or worth discriminating. To this end we were aided by our accumulated knowledge of what kinds of distinctions are important to content analysis work" (p.10). Presumably, by "content analysis work" is meant content analysis work in the behavioral sciences. As we will see in Chapter Three when we examine some of Kelly and Stone's actual dictionary entries and compare them to senses we found useful in analyzing a content area quite different from the behavioral sciences, the influence

of the domain of application was quite radical in terms of the partitioning of senses, as compared either to a standard dictionary definition or to partitions from other domains.

As noted above, this effort was mostly devoted to part-of-speech labelling: "something like 60-70% of the distinctions we make tend to follow part-of-speech lines" (p. 11). However, within-part-of-speech distinctions were handled to some extent, and this is what we will be interested in.

The goal of the analysis of a given word was to come up with an ordered set of disambiguation rules which bear on the word itself--i.e. look for its morphological characteristics--and on any number of words within a window of word-plus-and-minus-four-words. Any word in this range can be tested by a rule for part of speech, for membership in one or more of a set of semantic categories to be discussed below, and for Boolean combinations of these conditions. Passage or failure of a given test was allowed to determine either assignment of a specific word sense number, or a jump to a subsequent rule in the set, or even to a rule in the set for some other word.

As suggested above, senses were simply stipulated, as a function of the meaning differences which, it was believed, a behavioral scientist would find revealing or important, and after having read one dictionary's listing of meanings for general orientation. It is noteworthy that Stone and Kelly remark that "Partly because we distinguished so few senses,

agreement on sense identifications ran very high for us, around 95%" (p. 14). For the purpose of word sense discrimination, Kelly and Stone created sixteen semantic categories, which we list: Animate, Human (Male, Female, Kinship), Collective, Abstract Noun (Abstract, Time, Distance), Social Place, Body Part, Political, Economic, Color, Communication, Emotions, Frequency, Evaluative Adjective, Dimensionality Adjective, Position Adjective, and Degree Adverb. In Chapter Two we will return to these categories in order to discuss them in relation to the ones we adopt for our research. The pros and cons of this category set, from the authors' point of view, are stated as follows: "This part of the system is much less satisfactory...Many plausible candidate areas are not represented at all in the system, and some of the marker categories we do include are broad and ill-defined (particularly ABS). Nevertheless, these categories have proved useful, when liberally supplemented with tests for individual words. If we were to redesign the system, however, this is the part that would merit the most effort" (p. 22).

Actually, there is only one statistic which we are given to enable us to form some idea of the performance of the word sense discrimination mechanism as such, as opposed to the part-of-speech-labelling aspect of the program. An experiment is reported in which the test corpus consisted, rather bizarrely perhaps, of 125,000 words sampled from the same six-million-word corpus as the training data came from--(Recall that the latter consisted of about 511,000 words.)--plus 60,000 words sampled from the Brown Corpus (Kucera and Francis (1967)). Thus the total sample size is about 185,000 words of running text. About a third of the test cor-

pus-- 64,253 tokens (14)--was evaluated manually as to whether the parts of speech or senses assigned by the computer were correct. Then contingency tables were constructed for each of the 671 words (14) thus evaluated. These were simple matrices where the columns corresponded to the actual senses as manually scored, and the rows to the computer-assigned senses. A perfect performance by the computer would therefore mean that all tallies occurred on the diagonal. All tallies occurring in other cells indicate errors. For instance, the table for "last" looked like this:

		ACTUAL SENSES					
		1	2	3	4	5	Totals
COMPUTER ASSIGNED SENSES	1	80	2				82
	2	10	24				34
	3			11			11
	4	13	2		7		22
	5					2	2
Totals		103	28	11	7	2	151

(Kelly and Stone 1975 p. 40)

The five senses of "last" which are in question, are presumably the following (Kelly and Stone (1975, p. 189): 1. (adjective) previous, past, prior, most recent; 2. (adjective, adverb) final, finally; 3. (idiom) at long last; 4. (verb) endure, remain; 5. (adjective) "lasting"--enduring,

remaining. This table is of no particular interest to us in and of itself, since it ranges over parts of speech as well as word senses. However, the next set of numbers will be interesting, and in fact constitutes the lone evaluative statistic referred to above.

Summary measures were defined over the entire group of 671 such tables. One was kappa (Cohen (1960)), which is a simple variant of raw percent agreement, in which chance agreement is corrected for by subtracting the sum of the expected frequencies for the diagonal cells from both the numerator and the denominator of the raw success ratio (Kelly and Stone (1975, p. 41). Among other statistics, kappa was used to determine rate of accuracy in sense prediction, depending upon "Entry Type". By this term is meant that words can be sorted into three classes. First, there are those whose "senses" each have a different part of speech--as a fanciful example, take "hog" as either the animal, the verb meaning to maintain exclusive control of some resource, and the idiomatic expression "hog tied". Second, there are words, such as "forget", for instance, whose alternate senses are confined to a single part of speech: e.g. "Forget it"--Don't bother about it--as against "He forgot to do his chores"--fail to remember. Finally, there are words which are intermediate to these two extremes. That is, there is some duplication of parts of speech, but there is also within-part-of-speech sense choice. "Last", as defined above, is such a word. Kelly and Stone speak of Syntactic, Semantic and Mixed Entry Types. Now, the statistics which interest us use "Entry Type" as independent variable, and as dependent variable employ what may be

thought of as "degree of accuracy" of the set of words of a given Entry Type.

The data are summarized in the following table:

Kappa vs. Entry Type

		Entry Type			
		Syntactic	Mixed	Semantic	Percent/Totals
kappa	0-.7	23.3%	37.9%	53.3%	34.6%
		58	125	49	232
	.7-.9	34.5%	37.3%	27.2%	34.9%
		86	123	25	234
	.9-1.0	42.2%	24.8%	19.6%	30.6%
		105	82	18	205
Totals		249	330	92	671
Percent		37.1%	49.2%	13.7%	100%

(gamma = -.331)

This table may be interpreted as follows: There were 671 words for which accuracy data were gathered, i.e. for which the quotient actual correct predictions/possible correct predictions was determined (simplifying, by not taking account of the kappa correction). Of these 671, 92 were of the Semantic Entry Type. And 49, or 53.3%, of these 92 had an

accuracy score somewhere between 0% and 70%. Contrast this with 23.3% for the Syntactic Entry Type. As might be expected, the Mixed Entry Type was intermediate between these two points, with a score in the same circumstance of 37.9%.

Unfortunately, we obtain very little information from the above table about the actual performance of Kelly and Stone's system. This is because, for one thing, we are not given even an average number of choices per entry. If there are always just two choices, for instance, and if the distribution of samples of the two choices is always exactly even, then of course chance would produce a correct selection 50% of the time. In this case, depending upon the distribution of words in the frequency band 0% - 70%, and upon the number of concordance lines featuring each word, performance could actually be near random in the case of the Semantic Entry Type. On the other hand, if there are typically a great many choices, and if the 0% - 70% distribution is concentrated in the higher values, we might consider the reported performance rather impressive. Unfortunately, we are substantially at sea when it comes to judging data of this sort, even if they were formally susceptible of interpretation. This is because we have no benchmarks--the field of word sense identification is, at least from the point of view of quantification, essentially virgin.

Under these circumstances, the remarks of those who generated the data, and who therefore are most familiar with them, take on added importance. In commenting on this performance, Kelly and Stone state that

the above result, as well as others like it, "verifies our impression of the degree to which this dimension <viz. Entry Type, EB> governs the difficulty of entries in the current system" (p.44). This means that within-part-of-speech sense differentiation is much more difficult for their system than across-part-of-speech "sense" discrimination. However, recall that for our purposes the only problem which counts is the former one, since as remarked above, the task of part of speech labelling is well explored at the present time, and in any case not our focus of interest in this study. So, from the viewpoint which interests us here, Kelly and Stone feel their system faces a thorny problem in word sense identification. This may be an accurate impression. However, it is possible that some or even much of the effect might be due to the introduction into the test corpus of the 60,000 words of Brown Corpus material. This represents not simply a single new discourse domain, but on the order of a dozen such. It seems reasonable to assume, and in fact is one of the points of the present research, that the pool of senses does not remain constant when we pass from domain to domain. It would have been quite interesting therefore to have seen a repeat of this test without the Brown Corpus sub-sample.

Turning to the work of Sinclair and his associates, we may note that in Sinclair (1985a), the following set of statements is made, and labelled "Hypothesis": "There is a close correlation between the different senses of a word and the structures in which it occurs. 'Structure' includes lexical structure in terms of collocations and similar patterns. 'The sense of a word' includes the contribution that a word may make to a

multi-word lexical item" (p.2). Now this assertion may be taken in at least three ways. First, it may put forward the very general claim that there is SOME formal correlate of "word sense" in text, perhaps of a probabilistic nature. Interpreted in this manner, the "hypothesis" does nothing more than state the problem of word-sense discrimination in written corpora. Or alternately expressed, it begs the question as to which particular aspects of "structure" indeed correlate with word sense, and to what extent. In particular, how would the notions "lexical structure" and patterns "similar" to collocation, be made explicit? Second, the quoted sentences may be construed as making some definite claim to be worked out in the text to follow. In fact, Sinclair does argue and illustrate a single point in the remainder of his article--namely that "Conclusion...It seems that there is a strong tendency for sense and syntax to be associated" (p.12). The syntactic concepts cited in the body of the text comprise garden variety, traditional grammatical notions, such as transitive and intransitive verb, nominalization, and object. So we are entitled to interpret the "syntax" of "sense and syntax", cited above, as what we learn about in traditional grammar books. Thus, the second possible construction would be that sense is strongly correlated with syntactic structure. Third, there is the possibility of a compromise between the two interpretations already offered. This would be that it is the more general proposition that is affirmed, and the exclusive attention devoted to syntax in the text is due to its choice as an example. That is, the power of syntax as a correlate of word sense is intended as an illustration of the rightness of the general claim.

To summarize our three possible construals of Sinclair's claim, they are: 1. There are formal correlates of sense in text. 2. Syntax and sense are associated. 3. The syntax-sense correlation is one example of many correlations of formal factors with sense. Now let us note the interrelations of these three construals, with particular attention to the consequences of verification or falsification of each. The first really needs to be made more precise in order to be susceptible either of demonstration or refutation. The second can be argued for, say, by showing syntactic structure--taken, for instance, as the rules of some particular traditional grammar book--leads to some particular frequency of "right" predictions about word sense for some sample of words relative to a specified corpus and a specified purpose. Although it probably could never be disproved, it could be rendered very uninteresting by the failure of the data to cooperate with such analyses of a wide range of cases. As to the third interpretation, notice that it reduces to the second one. For if the case were made for the predictivity of syntax, that would in fact whet one's appetite for investigations into the degree of control of other "structural" variables over word sense. In a similar way, construal one also reduces to construal two, which serves to make the earlier claim precise. (In the absence of such a move, the first claim seems to lack empirical content.) But then construal one becomes empirically indistinguishable from construal two. So, in sum, the only work to be done by way of argumentation is the demonstration or refutation of construal two.

Now in fact, it does appear that Sinclair gives us reason to consider construal two likely. This fact in turn lends support to construals one and three. Restricting himself to "the most cautious figures", Sinclair is able to claim that the "alignment of sense and structure" can be quantified at the level of "practically 70%" for the one word he presents. Instances of this word-- actually the lemma "yield"--are drawn from the 7.3-million-word Birmingham data base of English, and number 125. On the basis of the classification tables presented in the text, we arrive at the score of 87/125 for this correlation: 81 "major uses" of "major senses" and 6 well-behaved uses of the two-word verb "yield up" (p. 13). This also converts to 70%. (Sinclair's fraction was 75/108.) The figure of 70% seems to represent the contribution of any one predictor of sense identity, taken by itself. We will comment further on this point in light of the experimental results to be presented in Chapter Three.

In Chapter Three we will also look at some of Sinclair's categorizations of instances of "yield". Here it might simply be pointed out that it is with reference to a definite, stated purpose that Sinclair stipulates definitions of "yield". The purposes in question are lexicography and the teaching of English as a second language. Additionally, he is well aware of the fluidity of boundaries among senses of a word, a point to be addressed in Chapter Two. For instance, in presenting the third "major meaning" of "yield", he states that it "is less easy to pin down, since it tends to merge with <the second meaning, EB>...I have not located a suitable dictionary definition" (Sinclair (1985 p.4)).

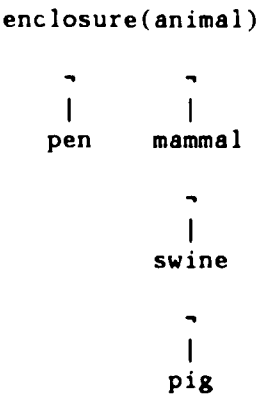
While (Sinclair 1985a) does not demonstrate or illustrate construal one, much other work of Sinclair and his associates over the past two decades can be viewed as preparatory to research in this direction. Sinclair (1966) set forward a precise interpretation of the Firthian (15) (and apparently originally Palmerian; see Seal (1981)) notion of collocation, or lexical co-occurrence within a stated window; this was developed in a computational study of some 135,000 words (Sinclair et. al. (1970), Jones and Sinclair (1974)). In 1973 work was begun on a collocationally-oriented text-processing system called CLOC (Reed (1977), Reed and Schonfelder (1979), Reed (1984), Sinclair (1985b)), which has received extensive elaboration in the period since then, and now appears adequate to support a major effort to supply hard justification to construal one above. The resources of the Birmingham researchers as far as computer-readable English text is concerned, are formidable, ranging into the tens of millions of words of running text, spread over a variety of discourse domains. Whether this last characteristic of their data base is an advantage or not is a question to be reflected on in the light of the next two chapters. But the point is that Sinclair and his associates are now in a position to make good the extremely interesting and provocative hypothesis advanced in Sinclair (1985a); there is reason to believe such an effort will in fact come to pass (J. Sinclair, personal communication), and it is to be hoped that it will.

Next we consider the work of Amsler and his associates. In Amsler (1980), structurally-determined keywords of definitions of the Merriam-Webster Pocket Dictionary (G. and C. Merriam Co. (eds.) (1971)) are

disambiguated essentially by hand. That is, concordances are labelled, as was done by Kelly and Stone. The resulting information is used to automatically construct lexical hierarchies based on the text of the Merriam-Webster definitions. Further work is suggested in which each definition would be disambiguated in its entirety, in the sense that each of the words it contained, and not simply its "keywords" or "head words", would be assigned a sense number. This work would proceed automatically, in a bootstapping process by which each correctly disambiguated definition would constitute additional data of the store from which word-sense discrimination decisions were made (16). In addition, the conjecture is made that such a fully disambiguated dictionary could serve to distinguish word senses in free text (Amsler 1980 p. 123).

Amsler provides an example of how this process would proceed (pp. 125-129). It is the meanings of "pig" and "pen" in Bar-Hillel's classic poser for machine processing of language, "The pig is in the pen" (Bar-Hillel (1964)), which Amsler undertakes to disambiguate. We seek to recover the interpretation under which pig means "a swine" and "pen" means "a small enclosure for animals". First, the "is a" relationship between "pig" and "animal" is discovered by moving through the hierarchy of definitional head words already created by Amsler ("pig" ISA "swine" ISA "mammal" ISA "animal"). "Pen" offers two such paths: "pen" ISA "enclosure" ISA "thing", and "pen" ISA "writing instrument" ISA "instrument" ISA "thing". Assuming the full disambiguation of words employed in definitions--Amsler's proposal to this effect is noted above--and now making the further move of supposing the full text of a definition to be

available for triggering moves through the conceptual hierarchy, we can take advantage of the fact that "for animals" forms part of the target definition of "pen". The desired senses of "pig" and "pen" are now related, since the element "animal" of the definition of "pig" is an argument of the "enclosure(animal)" portion of the target definition of "pen":



(diagram from (Amsler 1980 p. 126)). Amsler goes on to show that, with reference to the Merriam-Webster definitional structure, the situation is more complicated than this "naive picture" would suggest, and to define relations over elements of his projected conceptual network which enable the desired disambiguation to be carried out in this actual and more complex environment (pp. 127-129). In the course of this explication, he points out that sense pairings other than the one focused on so far are possible within the definitional structure of his data base: "The existence of these (i.e. of additional, EB) senses means that far more than the simple ambiguity perceived by naive speakers would be considered, which is not necessarily a handicap in light of other possible contexts" (p. 127). In particular, he draws our attention to the pairing of the

senses "pig...-- a casting of metal (as iron or lead) run directly from a smelting furnace into a mold" and "pen...--a small place of confinement or storage". He continues: "What is desired is a set of ranked alternatives from which a selection might be made on the basis of higher-order context variables. If one were discussing the operation of a steel mill, then 'iron' would already be highly likely to provide a context for (the appropriate sense of "pig", EB) being selected"; "The most common interpretation of 'pen' (in the sentence 'The pig is in the pen', EB) would be as... a small enclosure for animals', with "pig" being either ('SWINE' or 'a young swine', EB)" (p. 127).

Several remarks are in order here. One is that Amsler's brief remarks on the possibility of developing a dictionary-based general-text disambiguator are merely schematic; that is clear, for instance, from the fact that no characterization is given of the "higher-order variables" which would effect the "selection" from among "a set of ranked alternatives" as to sense. Further, it is far from obvious how the notion "discussing the operation of a steel mill" or its congeners are to be made precise. And finally, one can agree with Amsler in the abstract that "the most common interpretation(s)" of "pig" and "pen" in "The pig is in the pen" are as indicated just above. That is, it is intuitively likely that if, say, a hundred randomly selected native speakers of English were asked to rank the Merriam-Webster definitions of these two words, used in the specified context, along the scale "commonness of use" or something similar, with no further explanation given, most would respond along the lines indicated by Amsler. But--and this is a key point of interest in

the chapters to follow--it is just not clear what this insight buys within the context of actual natural language processing applications. The question is how one determines the likely place of a word--"pig", "pen", or what have you--within a discourse or discourse type. This problem is touched on in Amsler's reference to unnamed "higher order variables" which will "make a selection" from "a set of ranked (sense) alternatives".

The above remarks should not be conceived of as a criticism of Amsler's suggestion. The aim of his research was to so process the Merriam-Webster definitions as to bring out the hierarchical relations between individual word senses, using head words of definitional phrases. This he did in exemplary fashion. Suggestions for extension of his results and for their application to new and frankly uncharted areas necessarily have a speculative character. Certainly it would be of extreme interest and value to see the results of a working-out of a research program such as Amsler suggests.

Amsler, along with D. Walker, presented a related idea for a research program in Amsler and Walker (1985). Word senses can be distinguished, they reported, in the following manner: Let each word within a paragraph be assigned all its possible "subject categories" as listed in the computer-readable version of the Longman Dictionary of Contemporary English (17). The category most frequently represented among the words of the paragraph is the theme or subject area it covers. Accordingly, wherever one of the possible senses of a word within the paragraph is assigned the "theme category", it may be selected as correct. Actually,

this account of their approach is probably oversimplified, since they report using some 1600 categories compounded in some way from the roughly 120 subject categories of Longman's Dictionary. Unfortunately, the author has no clue as to how such compounds might be formed. We count 124 "major" subject headings, and 251 additional "minor" such headings, for a total of 375 possible labellings. It is possible that Amsler and Walker have in mind combinations of these with the 32 "semantic categories" such as "Human", "Liquid", etc. But this is mere speculation. Amsler and Walker added that they would like to extend this approach to further, and to more detailed domains, but were not sure how to proceed.

A sampling of these subject categories, where in the interest of clarity we let the major categories begin with capital letters and the minor ones with small letters, is: Baseball, Building, car building, bricklaying, carpentry, plastering, plumbing, Beauty Culture, cosmetics, hairdressing, perfumery, Basketball, Bible,...Numismatics, currencies, Occult, alchemy, palmistry, astrology, spiritualism, Occupation, medical profession, royal rank,..., Transport, Tobacco, Nonautomotive Vehicles, Water Sport, swim clothing, swimming, Winter Sport, curling, ice skating. As should be clear, the minor categories cited each belong with the closest preceding major category.

It seems clear that, while this method of confronting the word sense discrimination issue has something to offer, it is not by any means comprehensive in scope. First, much of lexical ambiguity has nothing to do with subject area distinctions. For example, it is hard to imagine how

subject-category-counting would lead to a triage of occurrences of the word "due" into the senses listed above, e.g. "due respect" vs. "due to the fact that..." vs. "due and payable" vs. "give someone their due" vs. "due to start". Second, in the case where senses are all comprehended by a given subject area, no distinctions are available under this method. So, if baseball is the paragraph topic, no choice will be possible between "ball" as the object which players at bat attempt to hit, and "ball" as in "base on balls" or "one ball and no strikes". Third, paragraph topic or theme is often overridden as a controller of sense selection by "figurative use of language" or by simple sense frequency. Even in legal discourse, for example, the most frequently-occurring sense of "will" is likely to be the modal one, as opposed to the meaning assignable to the word in "last will and testament". And it would certainly not be outlandish, say in an informal discourse on labor relations, for the word "break" to be used not as "a respite, as 'a short break', 'a coffee break'", but rather in the locution, "give (someone) a break", meaning "cede or show tolerance toward someone" (18). Fourth, in cases where the part a word plays within a multi-word expression is crucial to sense assignment, as in, say, "post office", "office building", "good offices"--in such cases, the suggested procedure will be effective in proportion to the representation in the dictionary of the relevant locutions. And Amsler and Walker themselves indicate (Amsler and Walker (1985)) that a comparison of multi-word phrases found in The New York Times, with those found in Webster's Seventh Edition (19), a much larger dictionary than Longman's, revealed only about a 25% overlap. Of course it is unknown how many of such phrases would constitute contexts of sense

resolution, but such lack of similarity between the two sources is at least not encouraging for the approach under discussion. Fifth--and this point will be returned to in later chapters--the sort of "discourse-general" meaning distinctions embodied in the Longman categories, or in those of any other similar work, may just not be relevant to the particular type of text submitted for sense discrimination. For example, consider the corpus the author has studied in connection with the research presented in this thesis, the text of the proceedings of the Canadian House of Commons. About a hundred randomly selected occurrences of the word "deserves", taken from a 20-25-million-word source, can be quite neatly and "naturally" partitioned into the uses, "X deserves something good", "X deserves something bad", and "Y deserves to become the object of discussion or analysis", for X a human, Y an issue. One looks in vain among the Longman categories for the ones which would effect such a dividing-up.

While this analysis of the Amsler/Walker disambiguation scheme may seem discouraging, it must be kept in mind that we do not know for sure what the effectiveness of any method will be until it is implemented and tested. Moreover, there may be applications for which precisely such a "domain-general" reservoir of senses is necessary to success. And finally, it seems unlikely that any well-developed, data-oriented approach to word sense resolution would fail to be useful when used in combination with other, complementary methods. For example, the syntax-based procedures of Gross, discussed above, might serve to fill in gaps left by Amsler/Walker's processor, and vice-versa.

We will now briefly present the work of Weiss (1973), to which we return in Chapter Two for more detailed consideration. Weiss confronts sense differentiation from the point of view of the discipline of information retrieval. He designs and tests a program embodying the following procedure: He wishes to learn the contextual concomitants of sense-label assignments associated with a set of input sentences all featuring some word w, and originating in any sort of corpus. To this end, he will induce an ordered set of sense-determination rules on the basis of an initial, "training" corpus C1, and apply these to a "test" corpus C2 drawn (randomly?) from the same initial set of citations, in such a way that C1 and C2 are disjoint.

Starting with the empty set of rules, his procedure examines each sentence/label pair of C1, and performs one or more of the following operations: it adds a rule or rules; it deletes a rule or rules; and/or it adds a deleted rule to a running list of "prohibited new rules". What is a rule? There are two sorts--template and contextual rules. Template rules are of the form, "word x occurs in the current sentence within two words to the left or right of word w whose senses are being disambiguated". Contextual rules say: "word y occurs in the current sentence within 5 words to the left or right of word w". There are two additional differences between template and contextual rules: 1. Template rules are ordered, en bloc, before contextual rules. 2. Contextual rules do not count "function words"--in some sense of this term which need not occupy us--but template rules count all words.

When meeting a new sentence/label pair, the procedure first attempts to apply, in order, every rule it knows, until a rule is satisfied. It then matches the label predicted by rule with the actual label of the sentence. If there is a match, it simply proceeds to the next sample. Otherwise, it traces back to the offending rule, and both deletes it and adds it to the list of rules which cannot be coined in the future. There is a tendency for the useful rules within each set--the templates and the contextuals--to "rise to the top of the stack" as incorrect rules are deleted.

Once the final sentence of C1 is examined, attention shifts to C2. Now, in test mode, the only function remaining of the three utilized on C1 is rule application. That is, each sentence of C2 is labelled according to the rule set derived from C1. An additional function is added, particular to the present phase: tallying of right and wrong answers. Accuracy scores are determined on the basis of the following figures: the total T of samples in the data set (= N); the number C of correctly resolved ambiguities; the sum I of incorrectly resolved ambiguities; and the number U of unresolved ambiguities. The two evaluative statistics employed are "resolution recall" RR = C/T, and "resolution precision" RP = C/(C + I).

Weiss obtains the following results, where "DEGREE", "TYPE", and "VOLUME" are the actual words used to test his procedures:

WORD	T	C	I	U	RR	RP
DEGREE	180	160	12	8	0.89	0.93
TYPE	180	164	4	12	0.91	0.98
VOLUME	180	152	15	13	0.84	0.91
TOTAL	540	476	31	33	0.88	0.94

(Weiss (1973 p. 40)).

RR is number right over number seen; RP is number right over number both seen and pronounced upon. As Weiss observes, both statistics are revealing. Now these numbers are, of course, impressively high. On the other hand, there is the following problem with Weiss's approach: If C1, his training corpus, is too small he loses the capability of predicting over C2, i.e. over text he has not seen before. But the larger C1 gets, the greater the probability that, even "by chance", a single instance of some word considered criterial for sense S1 will occur in a sentence labelled S2, or S3, etc. And in this case the correct rule not only would be erased, but would be prohibited from ever reappearing. This defect--potentially a very grave one--is a function of the logistic, all-or-nothing aspect of Weiss's rules. A method similar to his, i.e. "in the same spirit", but probabilistic in nature, rather than logistic, would not suffer from this flaw.

Another important disadvantage of Weiss's approach is the absolute lack of carry-over from one word to the next. The notion of generalization simply does not occur within his framework. It is easy to imagine that if one wanted to disambiguate not three words, as he does in his test, but 3000 words, and if one, prudently, relied on considerably larger training data sets than Weiss does, both the storage and the CPU time required to effect his procedures would be prohibitive.

Summing up this literature review, we have examined a number of scholars' orientations toward the problem of computational discrimination of word senses. Gross's approach is strongly syntax-based. To judge from actual examples cited, so is Sinclair's, although in principle he favors a much broader definition of "structure" than simply "syntax". Debili relies on the facts of morphology and of dependency relations to provide the cutting edge which will separate one sense from another. Amsler, in Amsler (1980), suggests using "conceptual" categories derived from dictionary definitions, as input to a procedure which in effect dynamically discriminates senses. In Amsler and Walker (1985), he and Walker opt for dictionary-derived "subject" classes as tools for establishing sense differences. Weiss's is an attempt at machine learning, and subsequent recognition, of "words as patterns". Finally, Kelly and Stone appeal to their knowledge of the discourse domain from which text will be drawn for analysis, both in restricting the range of interesting senses and in shaping the definitions themselves of the sense to be recovered automatically.

In Chapter Two we will lay the theoretical groundwork for our own research into computational word sense resolution. In addition, we will return to some of the work discussed above, in order to apply the perspective developed to issues which arise when one or another of these research orientations is adopted.

FOOTNOTES TO CHAPTER ONE

1. See Wilks (1972) for a review of some early approaches to word sense disambiguation which date back to the first wave of interest in machine translation.

2. Overall reviews of approaches to natural language within artificial intelligence may be found in Moyne (1982) and Moyne (1985).

3. Debili's is in some sense a collocationally-based approach; independently, Sinclair and Damerau (personal communications) have informally estimated the minimum corpus size for meaningful collocational research to be of the order of magnitude of 10 million running words.

4. For a different and deeper approach to morphological analysis via computer, see Byrd et. al. (1986).

5. Debili makes use of the notion of "lexical function" developed by I.A. Mel'chuk and his associates. For a comprehensive introduction to these functions, as well as a useful overview of lexical classification in general, see Evens et. al. (1983).

6. Freckleton (1985) presents 8000 lines of lexicon-grammar in a study on English idioms. In Gross (1985), examples of English lexicon-grammar entries are cited from Salkoff (1983), which includes, as an appendix, roughly 350 lines of lexicon-grammar. Gross (1984), in citing work in progress on English lexicon-grammars, also cites Machonis (1982); it is unknown to the author whether Machonis has a lexicon-grammar fragment of English, and if so, of what size.

7. The only two "+human" entries for "offices" would be the second and third ones above. "The offices waited" would be awkward, and probably unacceptable, on both readings, but it is not evident how this is formalized within the lexicon-grammar. However, this is a minor point.

8. As will be evident shortly, Gross would probably disagree with the particular sense definitions used here. However, the point here is to demonstrate the effect of syntactic environment in sense disambiguation, not to argue for the specific partitioning established, or even for the method used in establishing them. Also note that many would claim we stray into the realm of semantics in referring to "time nouns" in sense 5. Finally, the criterion for sense 6 is clearly collocational. But the claim was that most of the senses could be syntactically identified most of the time.

9. There is one aspect of Gross's work which invites comment, but which does not have much to do with the topic of this thesis.

Gross may be interpreted as claiming (Gross (1985, p. 3)) that language users employ facts of subcategorization as a basis for distinguishing word senses. If this were right it would provide a startling extension, and hence justification, of some interesting speculations made, on the basis of restricted but interesting experimental evidence, by Landau and Gleitman (1985). These are that the motor driving the human infant's acquisition of vocabulary consists of the combination of its developing concepts of objects, events and qualities, and their registering strict subcategorization restrictions on the words used by those around them--that is, a combination of situation, interpreted according to certain predetermined canons, and syntax. The missing piece of this argument is the rather plausible claim that different phrase types normally or typically are the vehicles for the expression of different concept types. Now, if Gross is right about the syntactic basis of word sense differentiation, then at least the suggestion of an answer would be at hand to the conundrum concerning how a sparse exposure to word occurrences might eventuate in a human being's ability to manipulate and understand these same words in a broad variety of contexts, both syntactical and conceptual. The notion of transformation would, on this account, also have a clear role to play in accounting for the extension of syntactic facility, since in establishing his separate dictionary entries, Gross defines as variants of a single item those constructions which can be considered transformationally related (in Harrisian terms) on grounds independent of the facts concerning the single word (spelling) under analysis (Gross (1985, pp. 3-6)).

10. Part-of-speech labelling by computer has been performed, for instance, by the Continuous Speech Recognition Group of IBM's T.J. Watson Research Center, using the Viterbi algorithm (on which see e.g. Forney (1973)), and achieving a degree of accuracy of prediction in the high 90 percentile for a category set on the order of 30. A different approach to part-of-speech labelling, also quite successful and implemented over a much larger set of categories, on the order of 200, is being taken by G. Leech and his associates at the University of Lancaster, England (Atwell (1973), Beale (1985a, 1985b), Garside and Leech (1985)). There are quite a number of other programs in existence which perform this task with a fairly high level of accuracy.

11. For presentations of content analysis, see Budd et. al. (1967) and Rosengren (1981).

12. In lexicography, a lemma is a citation-form of a word. For example, in English, "yield" might serve as a lemma for "yield", "yields", "yielding", "yielded". Depending on how the term "word" is defined in a particular instance, both the nominal and verbal senses of "yields", or simply the verbal one, would form part of this listing.

13. A KWIC concordance (Key Word In Context) for some word w is a file each record of which features w at roughly the same field, often set off on both sides by one or two padded blank characters. To the left and right of w in a given record is the sequence of words or other character strings which, respectively, precede and follow w in a particular line of the text over which the concordance has been constructed.

14. The figure 64,253 tokens arises in the following manner: In the interest of keeping their summary statistics meaningful, Kelly and Stone decided to analyze only those types with a frequency of 10 or more in the sample corpus. This yielded a set of 671 types, to which there corresponded exactly 64,253 tokens in the sample text.

15. On Firth's approach to linguistics, see Firth (1957), Palmer (ed.) (1968), Firth (1966). A treatment of semantics partially in the Firthian tradition is Lyons (1963). An interesting critique of Firthian linguistics in general, from the transformational generativist perspective, is Langendoen (1968). A critique of the notion of collocation specifically is Mitchell (1971).

16. For work which continues in the direction of dictionary-processing research taken by Amsler, see Chodorow et. al. (1985). Here semantic hierarchies are extracted from Webster's Seventh New Collegiate Dictionary by automatic and semi-automatic methods.

17. The print version of this, which unfortunately does not list these subject categories, is Longman Group Limited (eds.) (1978).

18. These definitions are the author's, and are informal and intended only for the purpose of illustrating the point at hand.

19. Merriam and Merriam (eds.) (1963)

CHAPTER TWO: THEORETICAL FOUNDATION

Our discussion of the theoretical foundations of the research to be presented in Chapter Three begins with background sketches of five ideas of Wittgenstein which inform our thinking about this research. Following this, the notions are applied as we consider both the rationale for our own research decisions, and the strengths and weaknesses of related work or programs in the light of what Wittgenstein's ideas lead us to expect concerning the characteristics of text.

The five ideas mentioned above are numbered I1-I5 for ease of reference in the following discussion; one further idea, which is our own extension of Wittgenstien's framework based partially on a notion of Goodman's, is labelled "W".

The thrust of the philosophy of "the later Wittgenstein", i.e. of Wittgenstein post-1929, is that by focusing our attention on language, we can furnish to philosophers a method ultimately permitting them to rid themselves of groundless "philosophical puzzlement". It is such puzzlement--the situation of being trapped in philosophical dilemmas of one's own contriving--which Wittgenstein attributes to the bewitching influence of ways of understanding inherent in our language. Only through study of and reflection upon the way language works, and in particular, the way it works its sorcery on thinkers, can philosophers find release

from the prison of their chimerical conundrums. "Philosophy, as we use the word, is a fight against the fascination which forms of expression exert upon us" (Wittgenstein 1965a p. 27).

II. Meaning As Use

Wittgenstein is concerned to combat those views of language which treat it as a calculus. These theories are represented for his purposes by that of St. Augustine, but the important contemporaneous theories of this sort which are also under attack, directly or indirectly, notably include Frege's, Russell's, and his own previous position as expressed in the Tractatus Logico-Philosophicus (1),(2).

First, the language-as-calculus school fails to understand the term "meaning". Because in paradigmatic cases of the use of substantives, an object is indicated, we all tend to make the mistake of looking for the use or meaning of a sign, as if it also were an object (Wittgenstein (1965a, p.5)). Presumably if meaning is an object, it must either be a mental object or one apperceived mentally. However, if we actually examine instances of use of some particular word, and especially of philosophically crucial words, such as "wishing", "thinking", "understanding", and "meaning", we find that we justify or explain our employment of these terms in a large variety of ways, only some of which involve mental experiences occurring contemporaneously with the utterance of the word. Moreover, it is not usually the case that we can identify any single definitional criterion, or a group of such, which we can with certainty call central or core criteria. But this does not mean that we do not use

in a clear manner a word for which we are unable to state necessary and sufficient conditions of use. While it is always possible to introduce further clarification of a usage, for example by defining each of the terms used in one's explanation of the usage, such a process has no principled ending. In any case, clarification continued to this extent is otiose, since the purpose of an explanation of usage is to allow the hearer to correctly employ the term himself, and no matter how precise a rule we provide, he/she must apply the rule to the situation in which he/she desires to use it. But rules do not come with procedures of application. Hence, there will always be the possibilty of errors of application. "Doctors will use names of diseases without ever deciding which phenomena are to be taken as criteria and which are symptoms; and this need not be a deplorable lack of clarity. For remember that in general we don't use language according to strict rules--it hasn't been taught us by means of strict rules, either. WE, in our discussions on the other hand, constantly compare language with a calculus proceeding according to exact rules" (Wittgenstein (1965a, p.25)).

I2. Forms of Life and Language Games

In order to further clarify what Wittgenstein means when he says: "For a LARGE class of cases--though not for all--in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language" (Wittgenstein (1968, Section 43)--we need to examine briefly the notions "form of life" and "language game".

The term "form of life" occurs just five times in the Philosophical Investigations (Wittgenstein (1968)). It is considered an extremely difficult item of the later Wittgenstein's vocabulary to pin down, and one can find greatly different interpretations of it in the literature (Phillips (1983)). Three of the occurrences are: "...to imagine a language means to imagine a form of life" (Section 19); "What has to be accepted, the given, is--so one could say--FORMS OF LIFE" (Wittgenstein (1968, p. 226)); "One can imagine an animal angry, frightened, unhappy, happy, startled. But hopeful? And why not?...Can only those hope who can talk? Only those who have mastered the use of a language. That is to say, the phenomena of hope are modes of this complicated form of life. (If a concept refers to a character of human handwriting, it has no application to beings that do not write.)" (Wittgenstein (1968, p. 174)). Baker and Hacker (1985a, p. 48) define "form of life" as follows: "A form of life is a given unjustified and unjustifiable pattern of human activity (part of human natural history...). It rests upon, but is not identical with, very general pervasive facts of nature. It consists of shared natural and linguistic responses, of broad agreement in definitions and in judgements, and of corresponding behaviour."

The notion "language game" is closely linked to that of "form of life" (ibid.). Language games are either "imagined" or "natural". Imagined language games are short descriptions by Wittgenstein of "languages" with extremely sparse resources compared to natural language, but which are useful in allowing one to focus on the significance in natural language of these particular features (or of the lack of them). In de-

scribing a language game, Wittgenstein typically provides data concerning a context, (e.g., for the famous "builders" example of Section 2, building activities), a speech community (e.g. builder and assistant), a vocabulary (e.g. "block", "pillar", "slab", "beam"), criteria of understanding (e.g. bringing the requested stones in the appropriate order), and the use of the vocabulary (e.g. the words are called out when the corresponding stones are wanted) (Baker and Hacker (1985a, p.26). A natural language game is an isolation of some aspect of natural language which can be focused on for analysis, and is named by analogy to the imagined language game. Its use is successful insofar as it allows demystification for the philosopher of the aspect of language in focus (Baker and Hacker (1985b, pp. 55-56)).

Now, to relate "form of life" and "language game" with "meaning as use", we need only say that a word or sentence of a natural language (or of an imagined language game) is no more or less than a "move" within a "game" set up by the conjuncture of a particular form of life and a given language game. That is, if we scrutinize actual practices of explanation of meaning, such as in the (informal, naturalistic) teaching of language, and if we examine actual conditions under which one is adjudged to have grasped the meaning of an expression, or to know its meaning, we find several things. First, there are numerous sorts of explanation, including paraphrase, the citation of examples of use, presentation of hypothetical situations of use, and genus-and-differentia definition. They all apparently can eventuate in the hearer's knowing the expression being explained, they all are language-internal and in most cases invoke context

or situation. Second, the actual criteria we use in determining whether one knows or has grasped the meaning of an expression is the extent to which he/she can use it in appropriate situations, at appropriate times, and in response to normal requests, directives, and so on. That is, an examination of our actual practices in dealing with meaning shows that the meaning of an expression is just how one is supposed to use it.

I3. Vagueness

One way to approach the notion of vagueness would be to say that the meanings of words, expressions, and sentences are "fuzzy", have ill-defined borders. However, this manner of speaking conceals an incoherence. An examination of natural conditions of linguistic explanation and judgments as to whether one knows or understands particular words or sentences reveals that explication of meaning via necessary and sufficient criteria is both unnecessary--not the practice, not what is meant by "explaining the meaning of a word or expression"--and, in any case, impossible: there is never a principled end to explanations, in the sense that one can always ask for further clarification. But it is only as against such a criterion of definiteness that the notion "vagueness" makes any sense. Hence, what we wish to say is not, for instance, that "the conditions of use, or the meaning, of such-and-such word are vague and hard to pin down", but rather that it is incoherent to ask for precise definitions. "Vagueness", like "inexactness", are simply terms of criticism (Baker and Hacker (1985b, Chapter XI)). "If I tell someone 'stand roughly here'--may not this explanation work perfectly? And cannot every other one fail too?...'Inexact' is really a reproach, and 'exact' is

praise. And that is to say that what is inexact attains its goal less perfectly than what is more exact. Thus the point here is what we call 'the goal'. Am I inexact when I do not give our distance from the sun to the nearest foot, or tell a joiner the width of a table to the nearest thousandth of an inch?" (Wittgenstein (1968, Section 88)).

I4. Stipulative Definitions

What about situations where definitions are considered to be precise, or where precise definitions are needed for some purpose? "Occasionally, especially, e.g. in legal philosophy, philosophers seek a regimentation (e.g. of the concept of legal right) for more or less practical purposes. But this, again, is not an explanation of our term in current legal discourse, but a stipulation to be justified by certain purposes" (Baker and Hacker (1985a, p. 33)). "What still counts as a game and what no longer does? Can you give the boundary? No. You can DRAW one; for none has so far been drawn. (But that never troubled you before when you used the word 'game'.)" (Wittgenstein (1968, Section 68)). "And do we know any more about it ourselves? Is it only other people whom we cannot tell exactly what a game is?--But this is not ignorance. We do not know the boundaries because none have been drawn. To repeat, we can draw a boundary--for a special purpose. Does it take that to make the concept usable? Not at all! (Except for that special purpose.)" (Wittgenstein (1968, Section 69)).

Special-purpose definitions (the only precise ones) should be evaluated, then, according to their usefulness for the task they were created

for. One sets up such definitions according to one's current requirements, and other things being equal, according to one's preferences.

I5. Family Resemblance

Wittgenstein (1965a, p.28) says: "There are words with several clearly defined meanings. It is easy to tabulate these meanings. And there are words of which one might say: They are used in a thousand different ways which gradually merge into one another. No wonder we can't tabulate strict rules for their use." Later, in Wittgenstein (1968, Sections 66, 67), he introduces the famous example of the term "game", and the concept of "family resemblance" among expressions. This is the notion that the different uses of a word--say "game"--relate to one another in the manner of the members of a large extended family. Some share particular facial features, others a way of walking perhaps, others still a characteristic coloring, and so forth. But as one glances from one to the other of such an assemblage, no single characteristic--and, importantly, no "minimal set" of characteristics, either--strike one as criterial for membership in the clan. It is noteworthy that in the first citation just above Wittgenstein speaks of cases where family resemblance is a useful model or analogy, and of cases where this is not so. This point of view was borne out in our close analysis of 100 words preparatory to the experiment to be presented in Chapter Three.

W. Worlds

In treating Wittgenstein's later philosophy as a background for linguistics, one is conscious of at least one needed supplementation of his ideas. Consider the relation between language games and forms of life. The former only have meaning when related to, when seen and experienced within the context of the latter. Change the form of life, and you change the language game: "If a lion could talk, we could not understand him" (Wittgenstein (1968, p.223)). However, once we start looking at language use either for its own sake, or with reference to special purposes such as disambiguating word senses, it becomes important to confront contexts less all-encompassing than that of forms of life, yet which have the effect of changing language games when we pass from one to the next. In shifting from a philosophical treatise to a corpus of advertisements for health food products, say, the language game of attempting to persuade a reader of our point of view is, in a certain sense, played in both contexts: that is, we have not changed forms of life. What it is to assent to a claim, for instance, is a given in both cases; in neither case is there any question of expecting that each claim made is always meant as its opposite, and so on. However, there is also a reasonable sense in which language games do not remain the same across this shift in context. In the case of the philosophical work, for instance, all human beings might be grouped together as "individuals", "knowers", "subjects"; but in the context of the health food advertisements, the class of people might resolve into "manufacturers", "agents of manufacturers", "experts", and "customers". Then, in the one case, "persuasion" will mean trying to bring the reader to assent to claims at an abstract level concerning himself/herself, as well as everyone else, qua

"knowers", "subjects", etc. But in the other case, "persuasion" is close to wheedling and cajoling, or to urging and coaxing, or even to menacing and intimidating; it is carried on by the "manufacturer" or his/her agent, often with the aid of "experts", and is aimed at effecting particular changes in the "commercial behavior", one may say, or the "buying habits" of the "customers", viewed as a somewhat alien group of people, and perhaps even reified and considered as not "fully human".

Another illustration of the ability of shift of context within a form of life to change language games is provided in D.Z. Phillips's "Wittgenstein's Full Stop" (Phillips (1983)). He is concerned there to argue against a conception of Wittgensteinian philosophy which would make it a deadening influence on religious dialogue. The specific complaint is expressed by Copleston (1974), quoted by Phillips: "The idea of autonomous language games, each of which can be understood only from within, by those who actually play the game in question, and which is therefore immune to all external criticism, seems to me open to objection...If it is carried to a point at which any fruitful dialogue between religious belief and critical philosophy is excluded, theology retreats into a kind of ghetto, cut off from the cultural life of which philosophy is one expression" (cited by Phillips (1983, p.181)). Hence one type of move among philosophers sharing Copleston's view has been to claim that relative to a form of life there are different uses of language, and there are also different purposes, or perhaps orientations, such as the religious purpose or orientation. Then "when 'religious' is used with 'language' it should draw our attention to the fact that certain concepts are being used

for religious purposes, and not that some kind of semantic or substantive shift has been made to a new type of discourse" (Bell (1969, p.6); cited in Phillips (1983 p.183)). Phillips's answer to Bell refers to Wittgenstien's discussion in his "Lecture on Ethics" (Wittgenstein (1965b)) of prudential "ought" vis-a-vis moral "ought": "You ought to keep your matches dry" versus "You ought to treat her decently". Here Wittgenstein differentiates between "absolute" and "relative" uses of "ought". We are not doing the same thing when we appropriately utter the two sentences cited just above-- we are not "commending" in both cases, say. "The moral context makes a difference to what the commending comes to" (Phillips (1983, p. 183)). So what Bell has alleged about the collocation of "religious" with "language" cannot, Phillips tells us, be said of the moral use of "ought": "One cannot say that the concept is used for moral purposes, since it is these so-called purposes which give the concept its distictive grammatical status" (loc. cit.). Phillips then shows that Bell's article actually contains an apparently unwitting admission of the distinct grammatical status of the religious realm: " ' "Asking, thanking, cursing, greeting, praying": In theological and religious behaviour these language-games usually have liturgical functions which only partially parallel their ordinary use'" (Bell, op. cit. p.13; cited in Phillips, op. cit. p. 183).

Generalizing over the two examples just cited, we can say that within the all-encompassing framework of a form of life, there seem to be less basic systems of reference, or perhaps of self-definition. Like forms of life, such systems function to define sense and coherence within their

own ambit. The "systems", in this sense, of the advertiser, of the philosopher, and of serious practitioners of particular religions, are the examples we have cited.

It seems we need a name, at least for our purposes, for the type of context that occasions distinctions between "absolute" and "relative" word uses. Above, the phrase "religious realm" was used for one instance of this sort of context; Bell (above) refers to "theological and religious behaviour". Perhaps a fruitful move would be to adopt the term of Goodman (1978) and call these contexts "worlds". Thus we would have the theolgical and religious world, the world of philosophical argumentation, the advertising world, and so forth.

A number of comments must be made concerning what the concept of "world" is to Goodman, and exactly how we propose to adapt it for our own purposes. First of all, we immediately wish to sever the idea from its connection with Goodman's nominalism. This, of course, involves a major reinterpretation, from a nominalist world, as it were, to a Wittgensteinian one. But, once this reinterpretation is agreed upon, it would seem that Goodman's concept is just what is required. Let us see how he uses it: "In what sense are there many worlds?" (Goodman (op. cit. p. 1)). "We are not speaking in terms of multiple possible alternatives to a single actual world but of multiple actual worlds" (op. cit. p. 2). (Among these worlds) "we have no neat set of frames of reference, no ready rules for transforming physics, biology, and psychology into one another, and no way of transforming any of these into Van Gogh's vision,

or Van Gogh's into Canaletto's" (op. cit. p.3). "Identity or constancy in a world is identity with respect to what is within that world as organized" (op. cit. p.8).

Goodman says that: "Several relevant kinds of the one world, rather than being absent from the other, are present as irrelevant kinds; some differences among worlds are not so much in entities comprised as in emphasis or accent, and these differences are no less consequential" (op. cit. p 11). This raises the issue, within a Wittgensteinian framework and for the extension we are making of it, as to whether worlds are "complete" in the sense that forms of life seem to be: a form of life is a self-contained complex of behaviors, ways of thinking, and norms of language use. Here again, Goodman can help us: "What counts as emphasis, of course, is departure from the relative prominence accorded the several features in the current world of our everyday seeing" (op. cit. p. 11). This suggests, for our purposes and within our framework, that where an area of the basic, background form of life is not appropriated for particularistic use by a world, it is nonetheless a part of that world, albeit a quiet, seldom-used part. Moreover, and importantly, such an element of the background form of life would have a meaning in the host world that is particular to this realm. For instance, it is conceivable that concepts of dental hygiene might enter into the religious world, say within a sermon; but it is not to be expected, it is not "emphasized", in Goodman's terms, and it will have its own meaning particular to the religious world. In the world we will be examining, that of the Members of the Canadian Parliament, language games such as wishing others a Merry

Christmas and a Happy New Year are certainly not emphasized, but they are present--and they certainly have a meaning special to this world (3). So there seems to be a sense in which worlds are open-ended, able to absorb elements of the larger form of life. And for this reason, we probably want to say that worlds are not complete--only forms of life, and imaginary language games, are (4).

Finally, it is to be observed that there is no end to worldmaking, to use Goodman's (1978) term. An important source of worlds, as Goodman suggests in the passage cited two paragraphs above, is the artist. The following comments of Ullmann (1962, pp.49-50) provide an illustration of one such world: "<I>n stylistic criticism...it has often been found that the complete significance of an important term can be grasped only in the light of the work as a whole. When one begins to read Camus's novel La Peste, the word 'peste'<-->plague<-->seems at first to refer to a specific disease which devastated the town of Oran in the 1940s. As one reads on one gradually realizes that the term also has several superimposed layers of symbolic significance: it is an allegory of the German occupation of France and, in a wider sense, of evil in all its metaphysical and moral aspects, and these implications continue to be broadened and deepened until the final sentence of the book."

With these preliminaries out of the way, we can turn to a discussion of the theoretical foundations of the research presented here. The basis of our strategy of word prediction is simply that learning the normal use of a word--that is, learning the word --is the same as coming to know the

conditions under which the word is normally used (Cf. Section I1 above.) Then, to the extent that it makes sense, for specific practical purposes (Cf. Section I4 above), to restrict the means of determining conditions of use to an examination of the line of text in which the word occurs, and to the extent that this examination in fact makes evident these conditions, learning a word means discovering the normal contents of this line of text.

But it is not words but usages of a word that we wish to predict, so actually "usage-of-word" should be read for "word" in the last paragraph. Now, in fact there are no fixed boundaries between the usages of a word (Cf. Section I3 above), so that a second instance of imposing limits where they do not naturally exist, for special purposes (Cf. Section I4 above), consists in setting up rules for the definite categorization of contexts of occurrence.

The world (Cf. Section W above) with reference to which our investigation is being conducted is "the Hansard world", the world of the Members of the Canadian House of Commons of the mid-to-late 1970s (5). The degree of applicability to other worlds of predictions derived in this process is an open question. Since there exists a family resemblance (Cf. Section I5 above) among worlds within the form of life (Cf. Section I2 above) of, say, English-speaking civilization, and since boundaries among them are fluid (Cf. Section I3 above), there is a sense in which a given language game (Cf. Section I2) played within one such world is the same game as the corresponding one played within another such world; but there

is also a sense in which they are not the same. Now if the games are identical, then usages within them ought to be sames as well. If the games are not identical, then it is just not clear how much carryover to expect in passing to a new world. So the degree to which the sets of defining conditions predicted here will serve for other related worlds is unknown.

Whether it does make sense to define evidence for conditions of use as, essentially, the five words just before and the seven words just after the word being learned (6), is not an a priori question but one for empirical investigation. Moreover, there are two separate applications of the question to our enterprise. First, can humans do it? That is, can someone familiar with a particular world inspect a concordance for some word w as used in that world, where a concordance line is of the approximate form "w plus-7-minus-5 words", and consistently assign one of a small number of usage labels to each line (more accurately, to each occurrence of w as node word)? And can he or she do it right--that is, are the resulting groups of concordance lines satisfying to the person doing the labelling, and, even better, to one or more other people familiar with the same world? Second, can computers do it? This second application of our question depends on a positive answer to the first application, at least under the particular statistical approach employed here, for the answer to this second application of the question is no more or less than the correlation between the human expert's labelling of a number of concordances in the manner just described, and a computer's labelling of the same set of concordances. If the answer to both applications of the question is in the affirmative, then it makes sense, with reference to

our investigation, to define evidence for conditions of use as concordance lines, node +7-5.

Assuming that application 1 of the above question gets a "yes" answer, the present research is designed to offer an answer to application 2. In fact there are at least three reasons to believe we can assign an affirmative answer to application 1. First is the author's own experience in labelling concordances drawn from the Hansard database. Even at the start of the labelling process, when he had no very close acquaintance with the Hansard world, the author found that for a large number of words, labels could be assigned with very little hesitation to almost all lines, once about a page (60 lines) had been scanned to determine the universe of usages present. Of course, becoming familiar with a world by examining concordances is a cumulative process, and the labels assigned in the earliest stages of the process are often judged incorrect later on, simply because the breakdown of usages did not accurately reflect the world. So the more significant observation is that once a good degree of familiarity with the Hansard world was achieved, there was the same facility in labelling noted above for the early stages, and that the resulting labels were now satisfying to one with a close knowledge of the world. Also significant is the large size of the corpus from which the concordances were drawn, viz. 20-25 million words, and the large number of concordances examined: 100 of them were given extremely close scrutiny.

A second reason to believe in application 1 is the report of Stock (1984) concerning lexicographers working at the huge University of

Birmingham (England) lexicography project. This project employs on the order of 20 professional lexicographers (J. Sinclair, personal communication) who work from concordances over 7.3 million words of heterogeneous English: it includes 1.5 million words of spoken English, the rest being written English; it is 70% British English, 30% other varieties; most of the 20 genres of the Brown Corpus (Kucera and Francis (1967)) are represented; and there is a conscious effort to produce "variety and completeness" of range (Renouf (1985); also see Renouf (1986)). Concordance lines are 98 bytes long (Stock (1984, p.133)). Stock tells us: "Lexicographers (i.e. those of the Birmingham project, EB) are able to read through a large number of usages and determine distinct senses in a significant proportion of cases" (Stock (1984, p.132)). Stock goes on to conduct an extremely illuminating discussion of the possible consequences of this phenomenon for lexicography and to make a number of highly interesting theoretical points about meaning. However, our concern right now is to address the fact that in only "a significant proportion of cases", and not in, say, almost all cases, were the Birmingham lexicographers able to pick out the right usage. (We will not discuss the question of speed, since no data are provided; however, the fact that Stock bases a suggested approach to word sense discrimination for working lexicographers on this procedure suggests that recognition is relatively rapid.) There are two countervailing influences at work here, it would seem. On the one hand, the relatively great length of the Birmingham concordance line (roughly node +9-6, on average) should certainly increase performance over, say, the 80-byte, approximately node +7-5 concordance-based labelling efforts of the author, other things being

equal. On the other hand, the extremely heterogeneous nature of the Birmingham Corpus should work to slow performance down.

This is an important point. Before we discuss it, we note that, of course, this is in no way a criticism of the Birmingham project, since they need to have as broad-based a corpus as possible, in view of their aim to construct general-English dictionaries. The philosophical question as to whether this is a coherent aim, and the question, relative to the theory of predictivity in text, as to how accurate the predictions of an extremely heterogeneous corpus might be for large corpora specific to a particular "form of life", are inappropriate given that the Birmingham approach is defined by a practical imperative, that of producing dictionaries with a general coverage.

The point, then, is that the form of life corresponding to the entire English-speaking civilization serves as context for the lexicographer attempting to reach a quick and accurate labelling decision on a line of Birmingham concordance. The difficulty here is that, first, it is inconceivable, or at least highly improbable, that this lexicographer's "expert status" with reference to *all* of the *different* worlds within this form of life equals that of someone familiar with a much more restricted world with reference to it. Hence the categorizations of concordance lines according to particular language games should be both slower and less accurate in the more comprehensive situation, due to hesitancy and lack of sure-footedness. Second, even assuming a coherent partition of linguistic behavior into language games can be carried out for such a broad

domain, the number of these would be so large as to represent a new source of confusion and slowness of response. In sum, then, it appears that other things are by no means equal with reference to the advantage accruing to the Birmingham concordance-labeller on the basis of the 98-byte length of a line of concordance. In particular, the negative influence of the difficult-to-manage set of worlds over which his or her corpus ranges appears signficiantly to outweigh the considerable help provided by increased concordance line length (7).

Notwithstanding the difference in coverage just noted, the report of Stock is worth highlighting, because of the large number of investigators it concerns--it seems fair to assume that this ability is not the privileged province of a minority of the 20 or so Birmingham professionals--and because of the impressive number of observations on which Stock's generalization is based: the 20 were not perusing a 50,000-word corpus, say, yielding puny concordances of 15 or 25 instances of even "core" words, but rather a mammoth 7.3 million-word source, yielding hundreds and thousands of instances of large numbers of words.

We now come to the third reason to credit application 1 of the question as to whether it makes sense to define evidence for conditions of use as concordance lines. In addition to providing additional strong grounds for believing that this sort of definition does indeed make sense, the third reason seems to confirm the analysis made just above of the contrast in reported accuracy of concordance labelling between the work at Birmingham and that of the author. (Recall that, above, we pointed

out that we were able, even early on, to assign labels with little hesitation to almost all concordance lines of Hansard text. On the other hand, the testimony of Stock, cited above, suggests that this level of efficiency was not quite attained at Birmingham.)

Kelly and Stone (1975) led a team of eight (?) individuals, including themselves, none of whom were professional lexicographers, and some of whom were present for only part of the concordance-labelling project (Kelly and Stone (1975, p.ix). Their goal was the improvement of a pre-existing computational data analysis package for performing content analysis of corpora in the behavioral sciences. This improvement was to be achieved by the automatic tagging, preliminary to statistical analysis, of each "target word" in a corpus to be analyzed, as representing one or another of the usages ("senses") previously agreed upon by the concordance-labellers. What is of importance right now is that this team labelled a concordance of 510,976 words created as a sample of a roughly 6 million-word behavioral science corpus. Fifty-six different sources were drawn from, "representing nine basic content areas (conversational material, personal documents, dreams, survey responses, TAT stories, literature, speeches, editorials, and folktales)" (Kelly and Stone (1975, p.5). What is clear about both the original corpus and about the sample used by Kelly and Stone, is that it relates to a coherent "world": "we wanted both the specification of the dictionary domain and the construction of entries to reflect English as the behavioral scientist is likely to observe it" (Kelly and Stone (1975, p.5)). Their concordance lines are 79 bytes long, but they discard word fragments at the beginning

and end of a line, unlike either Birmingham or the author. On the other hand, unlike the author but like Birmingham, they do not restrict each concordance line to a single sentence, but rather include in a line whatever words make up the 79 bytes, including material from contiguous sentences. Span would seem to be node+-7 (Kelly and Stone (1975, p.5)). This means that for quantity of evidence for labelling concordance lines, disregarding for the moment the all-important question of the "world" represented by the corpus, the three concordance varieties under consideration rank as follows, in descending order: Birmingham, Kelly and Stone, the author. Against this background, now consider the following statement: "Our first observation was that the amount of context supplied by the KWIC line is ALMOST ALWAYS SUFFICIENT to determine the sense in which a token is used" (emphasis mine, EB; Stone and Kelly (1975, p.10) (8).

The bearing of this statement on our current concerns is clear. First of all, application 1 of our question seems to be on a rather solid footing by now. Secondly, we see that, using less evidence, concordance-labellers who were not professional lexicographers (a number, in fact, being college undergraduates) were able to label lines "almost always", as opposed to professional lexicographers with more evidence who managed to label lines "in a significant proportion of cases" only. The difference seems to be the degree of amplitude and coherence of the world or set of worlds involved (9).

There were two notions which, we argued above, needed to make sense before we could claim that, for our special purposes, learning a word

means discovering the normal contents, relative to some world, of concordance lines in which the word occurs. The first was that it must be reasonable for us to restrict the means of determining conditions of use to an examination of such concordance lines. This is the point we have just finished arguing for. The second was that our computational examination must in fact make evident the conditions of use in question. Our experimentation, reported in Chapter Three, is devoted to verifying just this claim. That is, we wish to demonstrate significant agreement between the sense labellings, based on concordance-line context, supplied by humans, and those arrived at by a computational procedure previously "fed" with many correctly-labelled sample lines.

But not just any manner of arriving at such agreement between "people results" and "computer results" will do. For example, we need real, palpable procedures, input and output--not "on-paper" algorithms. So we do not wish to be told that, "in the limit", the problem of word-sense identification from labelled concordance lines is the trivial one of "asking every possible question" of the text. Every conceivable permutation of words, or even sub-word entities, could certainly be made to count as a candidate defining context for a particular sense-number. And then, "in the limit", the problem of using a decision tree to accurately label new text would be solved. But the problem is that the quantity of data required to allow the valid performance of a decision tree defined over the enormous number of tests which such an "in-the-limit" approach would yield, would be many orders of magnitude greater than, say, all the

text in possession of all corpus researchers known to the author (10), (11).

As we saw in Chapter One, there are a number of approaches to automatic word sense discrimination which have been suggested, and even in some cases implemented. In attempting to evolve the best possible procedure for identifying word senses, we are of course well advised to incorporate whatever results of past research seem likely to prove useful. Let us examine each of the methods presented in Chapter One, with the new purpose of determining what principles we would profit by including in a new system. Recall that those who have offered methods are Debili, Amsler, Kelly and Stone, Amsler and Walker, Weiss, Gross, and Sinclair. We have also just noted the "in-the-limit" approach.

Now the morphologically-based methods of Debili were shown in Chapter One to have such serious difficulties that it would not appear that they have much to offer us at the present stage of their development. For a different reason, the dictionary-based conceptual approach of Amsler (as opposed to Amsler and Walker; see Chapter One) cannot be of use to us either. This is because actual implementation of this method of disambiguation, as opposed to paper-and-pencil illustration of its possibilities, such as Amsler provides in his thesis, presupposes the full disambiguation of Webster Pocket Dictionary definitions as proposed in Amsler (1980), as well, perhaps, as the derivation of a considerable amount of information from larger editions of Webster (See Amsler (1980).)

We do not know whether this has been accomplished, but even if it has, we do not have access to this body of data.

As to Kelly and Stone, we feel there is much of value in their approach, and we will follow them in a number of respects. A basic principle which emerges from the practice of Kelly and Stone, is that of working from the inside of a domain. First, decisions were made as to what constituted a sense of a word, as a function of the textual domain under analysis. Second, their set of content categories, or "semantic" categories, is highly biased by their need to make those distinctions which count in the behavioral sciences, in their professional judgment. Notice especially that these content categories serve, in conjunction with syntactic or "structural" categories, to define the contextual environments in which words occur which are to be disambiguated. Further, to the extent that these categories do not make the distinctions made by those who originally spoke or wrote the words transformed into computational form, we can expect that they will fail to support the differentiation of individual words into senses in a manner that is either coherent with the text or of consequence to the speakers and hearers involved. So what we will take from Kelly and Stone is the notion that a portion of the burden of sense prediction is to be carried by content categories "tailor made", as it were, to the domain from which the text originates.

However, we must note the limitations of Kelly and Stone's work as well as its strengths. They were able to consider only about 1800 words,

and this took them some seven years' work (Kelly and Stone (1975, p.vii))! This is, of course, because they had to do all their work by hand and further, because they needed to consider each word de novo. That is, there was no carryover of the analysis of one word to the understanding of another.

These two major stumbling blocks to progress in this area, we propose to eliminate in two steps. The fact that they needed to do all work by hand is of course the fundamental problem, but it is just too difficult at the present time to attempt to leap this barrier at a single bound. We can only set as a goal the complete automation of the process of word sense discrimination, and hope that our efforts to handle more tractable difficulties will make automation of the process an easier task than it has been so far.

The second stumbling block, however, we will indeed try to remove. Recall that this is the fact that Kelly and Stone were constrained to examine each word afresh, with no transfer of results from one word to another. One of the methods we will implement and test in Chapter Three relies principally on a content analysis of the Hansard world, just as Kelly and Stone relied to a large extent upon their expert knowledge of the concepts and distinctions peculiar to the domain of the behavioral sciences. There will be a major difference, however, between our procedure and that of Kelly and Stone. This is that instead of examining each of the tens of thousands of words which occur multiple times in our 20-25-million- word corpus, we will sample 100 of them only. A very close

examination of all the contexts in which these 100 words occur, along with the classification into content categories of the words and expressions which are to be found in their environments, will enable us, we claim, to project the results of our analysis of a sample onto completely new cases, i.e. words not belonging to the 100-item sample. This will be done by means of statistical techniques of which we will give a brief account in the course of our presentation of our experimentation, in Chapter Three.

Turning to the work of Amsler and Walker, what we can take from them is the notion that domain-general content categories may be of use in the discrimination of word senses. We must stress that we do not propose to evaluate the actual application that they suggest making of such categories--more precisely, of the subject categories contained in the on-line version of the Longman Dictionary of Contemporary English. Recall that their proposal was to determine the category which occurred most frequently in a paragraph, and to let words take on senses labelled with this category whenever this was a choice. They also stated quite plainly, as we note in Chapter One, that they realize that there is a great need to find a way to increase the number of such categories, if this method is to be made useful. They mentioned that they were actively looking for such additional categories. With this caveat, we will in fact test the degree to which the Longman subject categories can be useful in separating out word senses within the Hansard corpus, when it is their occurrence in the same concordance line as the node word, and not in the same paragraph, which is the criterion of classification.

What we will take from the work of Weiss and of Sinclair is their emphasis on structural factors--better, collocational factors--as determinants of word sense selection. In Jones and Sinclair (1974), results are presented which argue for the windows node +-1 and node+-4 as optimal for capturing the more "structurally" oriented facts of collocation, and the more "content" oriented such facts, respectively. We incorporate this breakdown in one of the approaches we will test in Chapter Three, namely the content-analytic approach referred to just above in our discussion of Kelly and Stone. That is, even though 61 of the 81 categories we will employ in this categorization scheme, are content-analytic in character, the remaining 20 are "structural", and in fact consist of the 20 most frequent word pairs, n+-1, for the concordance being treated.

Weiss, on the other hand, while in effect also making the distinction between "local" and "sentential" collocation, sets the window size of the former at node+-2. (Recall that Weiss's terms are "templates" and "contextual rules".) This difference of node+-2 vis-a-vis node+-1 may appear to be a trivial difference with Sinclair, but in fact we will see that there is some plausibility in arguing that the difference can have very consequential effect indeed. In any case, in our experimentation, "sentential" context is never changed. It is always the 80 bytes--roughly node-5+7--of our concordance line which constitute the environment for non-"local" categories.

What we will not take from Weiss--and it is necessary to stress this point--is his manner of using "templates" and "content rules" to differ-

entiate senses. There is no need for us to point out once more the serious flaws of his method which are noted above. We will use the same statistical approach to the implementation and evaluation of all three methods put to the test in Chapter Three. Frankly, the idea of confining oneself to the words actually present in the text as the basis for sense identification is a fairly obvious one. We consider that the data processing aspects of Weiss's proposal constitute its meat, and that since we find these ideas unusable, all we are carrying over from Weiss is the definition of window size at w+-2, and perhaps the stricture that only single words, and not multiple-word units, are to be counted as predictors.

This leaves the work of Gross to consider. We will not be able to test Gross's ideas in Chapter Three, for the simple reason that we do not have access to a lexicon-grammar of English. Nor is it apparent that one exists of any dimensions that would be adequate to the analysis of large corpora. In fact, Gross (1985, pp. 10-11) calls precisely for the development of such tools as prerequisites to the statistical analysis of text. Since we will not be testing Gross's ideas experimentally, we will devote a fair amount of space in the remainder of this chapter to evaluating his proposals more informally, from the point of view of the Hansard data base.

We feel there is a very close relation between Gross's ideas on sense discrimination and those of Sinclair. In particular, it would seem that Gross's ideas are if anything more elaborated than those of Sinclair, as far as the use of syntax for disambiguation is concerned. That is, Gross

brings in the notion of transformation to relate what appear to be different constructions which nonetheless "share the same sense", although he perhaps would not put it this way (Cf. Chapter One). For this reason, we consider the conclusions to be reached as to Gross's system to be more or less applicable to the proposals of Sinclair. We will now examine Gross's methods in some detail. We think it is fair to say that the conclusions we reach with regard to his work carry over to the proposals of Sinclair, which, like those of Gross, are syntax-based.

In Gross (1985), the only document in which ideas specific to word sense identification are advanced by him, to our knowledge, dictionary entries intended to subserve disambiguation efforts are presented for two words, namely "office" and "waiting". As it happens by coincidence that we also have performed a rather detailed analysis of the former word, as it occurs in our own data base, we focus on "office". (Recall that our data base is the 20-25 million-word Hansard corpus, which comprises the official proceedings of the Canadian House of Commons for a period in the late 1970s.)

We wish to argue that Gross's criteria for sense differentiation are either precise and often wrong, or open-ended and no different from a content-analytic approach such as, say, Kelly and Stone's, except that the actual membership of such content categories is not even given. First, then, let us set out his rules for disambiguating "office". We will need to quote in extenso:

We give here an outline of the description of the two word forms: "wait" and "office". This outline is intended to show how a systematic separation of entries can be achieved on SYNTACTIC grounds. The list of entries will have to be refined further...A few examples of compound nouns have been given for various word forms. The list should be completed and each compound should be represented with its support verb(s)...

The entry "office"

NOUNS

1) N0 have an office =: Bob has an office
<A physical description of an office can be given in terms of elementary sentences such as the following ones:.
There is a desk in Bob's office etc.
Bob's office has two windows, etc.
The office is 10 feet wide
The office is whitewashed>

2) N0 (be, work) in an office =: Bob (is, works) in an office
<Here the meaning of "office' is abstract as opposed to the concrete meaning of entry 1). The difference is easily perceived in the sentences:
Bob locked his office
Bob shut down his office
Semantic specifications of this entry correspond to such sentences as:
Bob's office hired ten clerks
Bob's office makes money
Since the specifications given above for the concrete meaning are incompatible with these specifications for the abstract meaning 2) of "office", interdictions such as the following one can be used to reduce potential ambiguities:
* Bob's whitewashed office deals with cash sales>

3) The (office, whole office) celebrated Bob's promotion
<This entry may be derived from entries 1) or 2) by means of a general rule which states that a non human container noun of humans can be used (metonymically) for the contained humans>

4) N0 (hold, take) an office =: Bob holds the office of treasurer

N0 (be in, come into, go out of) office =: Max (is in, went out) of office

<These sentences contain support verbs that are synonymous to within aspectual differences, which is one of their regular features. The sentence:
Max will suppress Bob's office of treasurer
is a complex sentence in which the verb has for its object a reduction of a sentence with support verb.
Offices can be described by means of analytical or definitory sentences such as:
Treasurer is an office
Dealing with protests is an office>

5) N0 propose his offices to N1 =:
The lawyer proposed his good offices to Bob
Max used the lawyer's offices

<This entry may, in fact, be a compound noun with obligatory plural: "good offices">

6) N0 performed an office =: The priest performed an office for the dead

COMPOUND NOUNS
lawyer's office, doctor's office, etc.
box office, post office, etc.
divine office, etc.
office boy, office building, office hours, office job, etc.
(Gross 1985 pp. 18-20).

There seem to be two sorts of conditions which are presented here, which one might call strictly syntactic and selectional. To the extent that precise claims are being made, these would appear to be that certain syntactic constructions mark, or perhaps are necessary characteristics of, particular senses. But on this construction, the claims are quite often incorrect. So, it is not the case that sense 1)--physical office--is always in a construction derived from "N0 have an office", viz. Bob's

office is big, this office belongs to Bob(?). Moreover, Gross's own later examples use the phrase "Bob's office" in senses 2) and 4). A few examples of the use of this "office" in the physical sense, from the Hansard data base, which do not conform to these specifications, are:

ustry, a factory or any	office	and asked the people responsible for
e of the Edmonton south	office	could prove inconvenient to a signif
firm that won a federal	office	cleaning contract.
, was laid off from her	office	cleaning job.
erely to one storefront	office	but will include the mobile teams as
t myself--setting up an	office	beside H. & R. Block, the tax consult
fficials, with the head	office	being located in Ottawa.

On the other hand, there are, in this case, a good number of instances of this sense which do in fact exhibit the characteristics Gross points to:

ial government into his	office	and ask to see its books?
gh, I can go back to my	office	and come back with many more.
had to come back to my	office	and find someone who was bilingual on
I went back to my	office	and found about three and a half feet
A garbageman came to my	office	and said that he had something in com
Mr. Trudeau) sat in his	office	and sent out under his own letterhead
him to come to the UIC	office	and study it the way I did, and they

In the case of sense 2), however, it is not even very often that the "abstract" sense of "office" occurs in the context "(be, work) in an office":

n Ottawa man called the office of our parliamentary leader yesterda
eral phone calls to the office of the Minister of Transport (Mr. Ma
ceive a letter from the office of the Minister of Indian Affairs an
o were in touch with my office this morning, that according to my u
e it has not reached my office.
register at a Manpower Office and immediately collect benefits.
since the New Brunswick office handles both provinces.
The district office of the Department of Transport now r

Counter-examples for the "political" sense of "office"--

sense 4)--include:

make election to public office a realistic goal for the many rather
his appointment to that office almost one year ago.
minister of energy from office and give him the job of House leader
seek separation from my office and go to work in other fields.

upon your attainment of office and on the way you conduct that impo
candidate to stand for office and to serve a term here knowing tha
reality of the abuse of office by a party in power is a subject on
with the tenure of that office by the hon.

At the same time, like sense 1), but unlike sense 2), there are also

a good number of occurrences of "office" in this sense which do fit

Gross's criteria:

d are appointed to hold office during "good behaviour for a peri
government has been in office during the new committee regime a
is good lady, on taking office and wish them well.
m, I want to keep it in office because of money problems.
overnment since it took office had built transportation faciliti
ne since I have been in office has been perfect.

Senses 6) and 7) are not represented at all in the Hansards.

Now the performance of the strictly syntactic sense identification criteria is fair--neither particularly good nor particularly bad. It seems that the selectional criteria would provide the needed

supplementation to render Gross's approach more useful. One problem is that these criteria are merely sketched in. But more generally, we may observe that this sort of criterion amounts to nothing more or less than a content-analytic approach a la Kelly and Stone, say--if it is done with attention to the peculiarities of each domain for which disambiguation is to take place. Otherwise, it will simply get swamped. For instance, the two "types of office" cited in sense 4) are those of treasurer and of "person who deals with protests". But the class of offices, in this sense, is open-ended, and, importantly for our purposes, varies with the domain. For instance, compare the offices, in this sense, which are referred to in the Hansards with those that crop up in another corpus of roughly 25 million words, and one which ranges over multiple domains and genres, including novels, sports, marriage manuals, literature, politics, biography, art, and history; issues of magazines such as Fortune, Datamation, and Readers Digest are also included in this data base. Hansard "offices" include:

minister of energy from	office	and give him the job of House leader
Liberal government took	office	and placed a freeze on the interest
will mark his tenure in	office	as our first commoner.
decided to seek public	office	at the federal level, about giving u
nt to run for political	office	at the federal, provincial or munici
rnor in Council to hold	office	during pleasure.
hen a party has been in	office	for a long time, it will continue to
swers whatsoever in his	office	as Minister of Consumer and Corporat
The only way the NDP in	office	has changed bad practice is to make
ent Prime Minister took	office	have not, I think, worked out as wel
nce 1953, I ran for the	office	of mayor of Napanee on a platform of

Offices in the more general corpus include:

s Secretary of War he had found that office to be the most congenial h
n of a new government and its taking office . Neither the President no
l Address in March would require the office of president to give it ef
any of the bishops whom she found in office . To offset any papal lean
e , and his brethren stripped him of office and threw him out at long
ymouth . Nor did the church fill the office of prophet , doubtless bec
h in the church (being sometimes by office a deacon of the church in
atter's resignation to accept church office in 1644 , and for five yea
lose the day the sheriff went out of office , turning to meet his one
epublican congressmen seeking higher office is Delaware's William Roth

In the same way, the set of "compound nouns" is open, and no list, especially a supposedly domain-general list, can hope to approach completeness. From the Hansards we have, to cite but a few: storefront office, office equipment, national office, office desk, office expenditures, provincial office, branch office, law office, federal office, office cleaning, ticket office, and office work. In the more general data base, a sampling includes: church office, office building elevator, Dean's office, post office, tourist office, Registrar's office, police office, barge office, Provost Marshall's office, and office arm-chair. In both cases, this is a minuscule sampling.

The conclusion we can reach about the selectional criteria set forth by Gross is that statements such as "The list of entries will have to be refined further" and "A few examples of compound nouns have been given for various word forms. The lists should be completed..." simply will not do, in the sense that they unwittingly conceal a difficulty in principle in Gross's approach. This is that a large part of the work seems to need to be done selectionally, and that to meet this challenge, either

he tries to be domain-general and gets swamped in data, or else he focuses on the particular domain he is dealing with, and then becomes more or less indistinguishable from someone pursuing a content-analytic approach. As stated above, we feel that there is not sufficient difference between Gross's proposals and those of Sinclair, where they are made precise, to allow us to suppose that the conclusion just enunciated fails to apply to Sinclair's work as well as Gross's (12).

While the syntactic approaches of Gross and Sinclair have weaknesses, as just argued, it is important to recall that in almost all of the senses of our sample word "office" there were in fact syntactic criteria of identification which, if not by any means sufficient from a practical point of view to predict sense labels, nonetheless were often on target. Therefore the fact that the selectional criteria also presented needed more thought and adaptation to specific domains should not obscure the real contribution of the syntax-based method of disambiguation. Comparing the Hansard and domain-general corpora as to their adherence to these syntactic guidelines, we found that there does seem to be a syntactic substrate across domains, at least for the word we looked at, which rules such as those of Gross, or Sinclair, can capture with fair accuracy. Therefore, we can conclude that sense discrimination algorithms ignore such generalizations at their peril, and can formulate as a desideratum for full-fledged sense identification routines that they somehow take account of what might be called "middle-range" syntactic structure, in addition to the extremely local syntax--say, node +-1 or

+-2-- which has characterized a number of approaches to date, in particular those of Weiss and of the author.

To summarize the present chapter, we have attempted brief explications of the Wittgensteinian notions of meaning as use, forms of life and language games, vagueness, stipulative definitions, and family resemblance. In addition we introduced the conception of "worlds", following Goodman to some extent, in order most profitably to adapt these ideas of Wittgenstein to our interests within the realm of language study. We then applied this perspective to the question in focus here, which is the computational identification of word senses in context. First, we sought to demonstrate that humans are capable of accurately and rapidly labelling senses given simply KWIC concordance lines of roughly node +-6 or 7. In addition, we argued that superior accuracy seems to be achieved when a single, coherent "world" or domain is the textual source of the concordance which is inspected. Second, we noted that experimentation to be presented in Chapter Three will address the issue of whether computers can effect this same sort of sense labelling. The last portion of this chapter was then devoted to a discussion of a few computationally-oriented, or at least algorithmically-oriented, approaches to this problem, approaches whose best insights we hope to incorporate into our own models. In particular, the strengths and weaknesses of the proposals of Debili, Amsler, Weiss, Amsler and Walker, Kelly and Stone, Sinclair, and Gross were examined.

The following chapter concentrates on experimentation using, on the one hand, domain-general, dictionary-derived categories of words and expressions, and, on the other hand, domain-particular, content-analytic such categories, for the purpose of identifying word senses in concordances drawn from a very large written corpus of English. The results obtained illustrate the usefulness within linguistic study of the philosophical notions discussed in the present chapter.

FOOTNOTES TO CHAPTER TWO

1. (Wittgenstein 1961) is the Tractatus.

2. The preceding remarks, as well as much of what follows, generally follow Baker and Hacker (1985a, 1985b) in the interpretation of Wittgenstein's philosophy. While Wittgenstein scholarship is marked by numerous serious differences among commentators, the scope and seriousness of the two-volume Baker and Hacker work would seem to give its analyses as much claim to authoritativeness as any others in the literature. We have also relied on a number of the articles in Block (ed.) (1983), which are cited where appropriate. Other secondary sources were consulted whose views we do not follow; among these are Pitcher (1964), Finch (1977) and Ayer (1985). See the last-named for an interesting and readable critique of Wittgenstein's post-1929 ideas from the perspective of modern analytic philosophy.

3. Neither VOTERS nor WORKERS are wished a Merry Christmas and a Happy New Year--this formula is reserved for POWERFUL_PEOPLE. The three capitalized expressions are the names of three categories forming part of our content analysis of the Hansard world. What links them is that they are the only categories in this set which directly have to do with people. In other words, we claim that the most predictive partition of the words and expressions of the Hansard data base can be made if these three categories together cover all language directly naming humans or bearing on "personal" matters such as health. Notice that one could conceive of the Members joining together, say, to wish all American Indians, or all fishermen, or all Canadians, a Merry Christmas. But there is no record of such events in the Hansards. From the point of view of prediction, one can be reasonably sure of finding an exponent of the category POWERFUL_PEOPLE in the neighborhood of, say, "wish" and "Merry Christmas". In the society at large, the greeting is not so restricted.

4. Actually, it is probable that only imaginary language games are complete. There is interpenetration of cultures, so that forms of life are not hermetic. It is therefore no surprise that the less inclusive "worlds" are not either. Wittgenstein's philosophy leads us to expect such a state of affairs, in fact. The only way we should be able to "pin down" a world is by defining it for some special purpose. This has to do, in turn, with the lack of clear separation between language and its setting. Rules, linguistic or otherwise--if such a dichotomy makes sense--need to be applied when used. This is therefore true of rules for conducting or expressing oneself within the context of some "world". It is not a matter of concern that at no precise point can we stop and say, "Here we are definitely in world w in its pure state". To expect otherwise would be to imagine that worlds have a static nature, unlike forms of life, language and human rule systems in general.

5. "Hansard" is used in England to refer to the official published reports of the debates and proceedings of the British Parliament, and in Canada, similarly, for the Canadian House of Commons. One Luke Hansard (1752-1828) and his descendants compiled the British reports until 1889.

6. As is now common in collocational and related studies, we adopt the following elements of the vocabulary of Sinclair (1966): "We may use the term NODE to refer to an item whose collocations we are studying, and we may then define a SPAN as the number of lexical items on each side of a node that we consider relevant to that node. Items in the environment set by the span we will call COLLOCATES" (p. 414). Unless otherwise noted, we will always take "item" or "lexical item" to refer to typographical words, i.e. blank-delimited character strings (or number strings, or mixed character and number strings). From here forward, we will use the expressions "node+-x" or "node+x-y" to mean, respectively, a span of x words in either direction from the node, and a span of x words to the left of the node and y words to its right. We will use expressions such as "w+-x", "w+x-y", "n+-x", and "n+x-y" in the identical way, where "w" or "n" stand for "word", and are identical with "node".

7. Again, we are talking about accuracy here, not speed. It is obvious that it takes longer to read through 90-byte lines than 80-byte lines.

8. A KWIC line is simply a line of concordance in which the node word is located more or less at the center of the line, flanked perhaps by extra blank spaces, and the preceding and following context in the source text is positioned to the left and right, respectively, of the node.

9. It is true that a possible confounding factor may be degree of delicacy of definitions. However, this would seem to be something of a red herring. That is, let us assume for the sake of argument that the Birmingham lexicographers in fact had a larger number of senses to choose among, in making a labelling decision, than did, say, the Kelly and Stone team. (Actually, we do not know whether or not this is the case, although perhaps one may suppose it to be so.) Now, we suggest that if there are more choices for the Birmingham lexicographers, this is simply, like the relative frequency of indeterminate cases, an effect of the extremely broad domain of coverage, and not, as this objection would have it, a cause of this same indeterminacy. In Chapter One, we noted that Kelly and Stone were concerned first and foremost with providing to the community of behavioral scientists a tool which is practical and effective. They wished to draw whatever distinctions a behavioral scientist would in their judgment find useful. From the point of view of their purposes, they might as well not have undertaken to build their system if they were not going to make all the sense distinctions which were capable of doing any work within their elected domain. If it turns out that the number of sense distinctions required to

effectively cover the behavioral science field is less than that needed to set up a general-purpose dictionary of English, this is emphatically not a case of grossness of meaning distinctions, of failure to make detailed and precise distinctions within a semantic space. Rather, it is the effect of a huge disparity in the amplitude of the two domains which are being mapped.

10. R.L. Mercer, personal communication. The author wishes to thank Dr. Mercer and Dr. P. DeSouza for enlightening discussions of this point.

11. This "in-the-limit" approach is a generalization of the method of Weiss (1973), which is not to claim that Weiss himself would advance such a position. Rather, this is simply a possible point of view that occurred to the author in the course of his consideration of Weiss's procedures.

12. The difference between the Harrisian notion of selection employed by Gross, and that of the content category, in the work of Kelly and Stone, or of the author (cf. Chapter Three) is that the former concept is defined syntactically, while the latter is defined in terms that might perhaps be identified as experiential. A word or construction related by selection to another such, must by definition stand to it in some "grammatical relation". Now, of course, "grammatical relations" are objects which are defined only with reference to particular theories of grammar. It is then conceivable that a theory could be put forth such that in effect any two words or constructions would be selectionally related. In this case, much of the formal difference between selection and content categorization would disappear, in the sense that they both would reduce to co-presence within some window, when looked at from a strictly formal viewpoint. However, the great difference of motivation in the definition of the two concepts would remain. The basis of content-analytic categories--the grounding, one wishes to say--in forms of life or worlds, is crucial. On the other hand, Gross takes a strictly formal approach to sense differentiation, perhaps partially because of his skepticism as to the "correctness" or possibly usefulness, of commercial dictionary definitions (Gross (1985, p.3 fn.3)).

CHAPTER THREE: AN EXPERIMENT IN WORD SENSE DISCRIMINATION VIA COMPUTER

This chapter presents the experiment in word sense discrimination which we have carried out. We begin with a statement of its general aim, and then pass to considerations of the materials employed, of the procedure effectuated, and of the results obtained. At the chapter's conclusion is a discussion of these results, and an indication of how they might be extended and, eventually, applied.

Our aim in conducting the experiment to be reported is to determine the relative power of three distinct approaches to the task of identifying English word senses via computer. The first method, which we will call the domain-general method, or DG, classifies the linguistic environment of a word token which is to be assigned one of several stipulated senses, according to the combinations of dictionary-derived categories which, roughly speaking, tend to co-occur in concordance with a particular sense of the word under study. Actually, categories can contribute to the identification of a given sense by their habitual absence in its neighborhood as well as by their frequent presence. The second method, one we will label the domain-specific, exclusively structural method (DS1), uses as its criterion of word classification, grosso modo, frequency of occurrence in the source text as a whole. However, this notion of frequency is applied in two different ways. "Function words" (1) are counted only within the window node+-2, while "content words" are counted wherever

they may occur in an 80-byte (about-13-word) concordance line. The third method--the domain-specific, content-analytic/ structural method (DS2) is like DS1 in employing two separate classification schemes for words and, in the present case, multi-word items. The first of the schemes is frequency-based and small-span, as is DS1's. In particular, it takes as its basis for categorization frequency in the environment node+-1, and ranges over all words, "function" or "content". The second of the schemes is quite different from the DS1 approach, and depends on intensive preliminary work of a taxonomic nature by a human who is familiar with the domain or world which the source text represents. We are assuming, in fact, that the corpus does exemplify a single, coherent world, and make no claims for generalizability of our findings to corpora which are essentially heterogeneous in character. To restate our aim, then it is to compare DG, DS1 and DS2 according to their ability correctly to predict word sense variation as a function of context, and specifically as a function of the contents of the concordance line in which the word appears whose sense is to be selected.

We will now describe in some detail the materials used in this experiment. In each case--DG, DS1, DS2--these materials consist of sets of computer files. Here we discuss the generation of these files. Below we indicate the manner in which the files were processed statistically in order to yield experimental results. In addition, we describe briefly in the immediately following text the initial preparation of the source corpus.

Our source document consists, then, of some 20-25 million English words of text, representing the official transcripts of the proceedings of the House of Commons of Canada, for a period situated roughly in the mid-to-late 1970s. We received the text after it had been processed by SCRIPT, IBM's document formatter. A few types of SCRIPT descriptors, as well as some unexcised typographical instructions whose typical length was two characters, crept into the lines of the concordances whose generation we are about to describe. It is likely that this noise does account for some slight diminution in accuracy of predictions, since the presence of two characters is sufficient to displace the beginning or end of a word or expression which may be criterial for a particular sense selection. For instance, if a search is undertaken for the character string "Solicitor General of Canada", and what is found is "*boSolicitor General of Can", a null result is obtained. Nonetheless, this noise probably damaged results by no more than a percent or so, and we will therefore ignore it. In future applications, however, we intend to eliminate this problem.

Our first step in data preparation was to submit portions of corpus to the TUPLES (2) text analysis package, designed by R. Byrd of the IBM T.J. Watson Research Center. We began by obtaining a frequency count of some 6 million words of the text. From this list we selected about 1000 words as a pool from which items could be chosen for analytic and experimental purposes. The criteria of selection were, first, to ensure an adequate representation of words with a large number of occurrences in the corpus, since the statistical processing to be undertaken later would

require very many instances of use for each word investigated. An additional criterion, somewhat at odds with the first, was the desire to include in the 1000-type sampling words with a broad range of functions in the corpus and with a considerable variety of different frequencies. The two criteria combined to yield the following decision on word selection: About 200 items were chosen among the very most frequent items on the list, roughly those with frequencies anywhere from 1200 to 130,000. Then some 100 each were chosen with around 900, 800, 600, 500, 300, 200, 100, and 50 occurrences in the 6-million-token corpus. Five words were then selected from this pool of roughly 1000, as the foci of experimental processing. That is, the ability of DG, DS1 and DS2 to differentiate the senses of these five words would constitute the basis of our evaluation of the three methods.

The selection of the five test words proceeded as follows: A program was written to generate random numbers between 1 and 1026, the actual number of words in the pool. Given a random number n, we considered the nth word in the pool, and applied to it a number of criteria of selection. If a word met all of the criteria, it was taken as a test word. The process of selection was continued until all five words had been chosen. The criteria were that the word must occur at least 2000 times in the whole 20-25-million-word corpus; that it must have at lease three clear senses within a single part of speech (3); that none of the senses be "too easy", i.e. that there not exist a single telltale cue which always, or almost always, accompanies some sense, and never, or practically never,

occurs with other senses. The five words selected were: interest, point, power, state, and terms (4,5).

Once the terms were decided upon, complete concordances over the 20-25-million-word Hansard data base were generated for them, again using the TUPLES package. A close analysis was made of each resulting concordance, with the aim of defining three or four senses within a single part of speech. Once this was accomplished, lines were discarded in which the node occurred qua parts of speech other than the one selected. Then each sense was assigned a number between 1 and 3, or 1 and 4, and each of the 2000 lines of each concordance was assigned the sense label 1, 2, 3, or 4. Then a program was written and run which partitioned each approximately-2000-word concordance into a file of about 1500 records of "training data" and one, of exactly 500 records, of "test data".

We now turn to a description of the other materials used, namely the category files essential to methods DG, DS1 and DS2, respectively. DG was generated in the following manner: First, a frequency count for single words (as opposed to multi-word units) was performed on each test-word concordance. Then, roughly the 500 most frequent words on the list were selected as input to the program generating DG files. (Taking the top 500 words was equivalent to taking those words occurring anywhere from 5 times or more to 10 times or more in the 2000 lines, or about 26,000 words, of a test concordance.) The program which yielded DG files worked in a straightforward manner. It looped through the list of 500 words, looked up each word in the on-line version of the Longman Dictionary of

Contemporary English, and determined which "subject categories" appear in the word's definition. (As reported in Chapter One, Longman "subject categories" are labels designed to inform the dictionary user when a word or sense is associated particularly closely with some special area of life, such as: Hairdressing, Numismatics, Plumbing, Spiritualism, Tobacco, Transport, or Winter Sport. There are 124 "major" subject categories and 251 "minor" subject categories, which are subdivisions of the major ones; hence there are 375 possible classifications.) Once a word's subject category listings were determined, the program simply added the word to a separate file for each category. (Obviously, in case a word was the first one on the list to have a particular subject category in its definition, a new file was created, and the word was its first and, so far, only entry.) The final product for each of the five test words, then, was the following: a set of files, each of which contained all words among the 500 most frequent of the test word's concordance that include some particular subject category within the text of their definition. So, to take an imaginary example, for the test word "watch", one might end up with a file, among others, named "ENTERTAINMENT", which indeed is the name of one of the Longman subject categories, and this file might contain the words "movies", "play", "television" and "performance". This would mean that all and only these four words had in their definitions the symbol denoting the "ENTERTAINMENT" subject class.

In most cases, several fewer subject category files were produced in this manner than the 81 being sought. (See below as to why exactly 81 files were required.) In these instances, it was necessary to add a small

number of null or "dummy" files to those obtained automatically. In one or two instances, more files were produced than necessary. In this case, a small number of files containing but one word were dropped.

The thrust of DS1 is to attempt to disambiguate nodes in the test data by using as cues to sense identity the content words and node-local structures associated with each sense in the training data. DS1 was produced as follows: For a given test-word concordance, the 1500-line "training data" file was subjected to a single-word frequency count. The 41 most frequent "content words" (1) were extracted from the sorted list, and each was placed in a file bearing its name. These were the "content" categories. Next a sort was performed which counted only words in the window node +-2. The 40 highest-scoring words from this sorted list were selected. Then the "test data" file was run through by another program which listed every occurrence within the window node +-2 of each word among the 40 just mentioned. The result of this processing was that every test-data word string of the form xwn, wxn, nxw, or nwx was located, where n = the node word, w = any of the 40 words referred to above, and x = any word at all. The complete such list for any w constitutes an exhaustive enumeration of all occurrences in the test concordance of w in window node+-2. Each such list was placed in a file whose name is that of w. There were now 81 "entirely structural" categories: 41 consisting of the most frequent content words in the training concordance, and 40 more comprising the most frequently occurring words in window node+-2 of the training data, listed in extenso in the precise environments in which they occur in the test concordance.

It remains for us to consider how the category files of DS2 were generated. There are two types of DS2 files; the first are "structural", the second "content-analytic". The structural files were produced in a manner not unlike that by which DS1's were created. The training concordance was sorted by frequency using the window node+-1, and the 20 most frequent word pairs thus obtained were each assigned to a category of their own. That is, each became the sole record of a file bearing its name.

Now the "content-analytic" files of DS2 were generated in a manner completely unlike any of the procedures we have seen so far. Recall that 1026 types were chosen from the type list for 6 million words of Hansard text. A chance selection was made from these 1026, of 100 types, for which concordances were then produced, via TUPLES as usual, over the full 20-25-million-word corpus. Then each such concordance was analyzed quite closely from what might be called a content-analytic point of view, and many of the words and expressions which were uncovered in the course of this analysis were loaded into one or another of the 61 "content-analytic" DS2 files.

Actually, the loading of the 61 files took place through a bootstrapping process which started with no "content-analytic" files at all and, of course, no words or expressions entered in such files; and which ended with 61 fully-stocked files. The procedure by which this result was obtained was the following: The 100 concordances were examined seriatim. In the case of any given concordance, this examination began

with a first reading whose aim was the stipulation of a set number of node word meanings relative to the corpus (6). Each sense was assigned a label consisting of a number. There followed a second reading of the concordance, in which all words and expressions which occurred were listed and partitioned according to the node sense number of the line in which they were found (7). Next, an attempt was made to partition along thematic lines each list of sense-particular words and expressions. In the course of processing the first ten or fifteen concordances, any theme was entertained. Some of these early themes which did not survive were: Advertising, Capital Goods, Industry, Obligative, and Try. However, beyond this watershed it started to become clear what the useful thematic categories were for this "world". (For the notion "world", see Chapter Two, Section W.) A theme was considered useful and adopted if the presence of one of its exponents in a concordance line--either alone or in conjunction with exponents of a small number of other themes--sufficed in a large percentage of cases to determine the sense number of the line's node. The proliferation of themes was, in an informal sense, asymptotic, and fell off dramatically after about forty concordances had been considered. In a final reading, conducted after the processing of all 100 concordances was complete, those concordances taken up at the beginning of the process were reanalyzed in terms of the final set of thematic categories, so that uniformity of classification was ensured (8).

We now present a number of examples of thematic categories. In each case, illustration is provided of the manner in which the category functions. This illustration always takes the same form: We display, in

concordance, one or more of the 100 types examined, for which the category appears criterial--either alone or in combination with a small number of other categories--in separating one node sense from others. In each instance, a sampling of the entries of the category is provided (9).

BODY_OF_WATER

This category lists names of types of body of water (e.g. river) as well as associated words and expressions (maritime, seaworthy).

32	If you were to	ship	a car of potatoes to Newfoundlan
18	erefore an incentive to	ship	goods from the prairies, and thu
328	raw material, and would	ship	it to the United States where it
446	er from a platform or a	ship	in deep WATER.
240	MARITIME Command, in a	ship	or other VESSEL, or in a repair u
604	In the	ship	HARBOUR case, there was agreement

316	.SP The psychological	damage	which had been done to this you
198	nvestment that it would	damage	Canadian economic prospects.
24	lly in crimes involving	damage	and danger to the state, that t
143	ause SHORELINE property	damage	and create the risk of possible
498	de, waiting to see what	damage	the INFLOWING WATERS cause befo
163	SHORELINE FLOODING and	damage	next summer when the HIGH WATER

296	one sense, but rather a	deep	analysis; an evaluation of a ver
389	the seconder in a very	deep	and thoughtful speech reminded u
236	Animosities	deep	and dangerous have been given wa
88	We do not do so without	deep	anxiety and a recognition that t
108	l punishment issue is a	deep	clash of values between various
197	le of receiving the USN	Deep	SUBMERGENCE Rescue Vehicle.
770	r, the WATERS are quite	deep	in that area, and if the BREAKWA
987	The site of the	deep	SEA PORT that has been talked ab

Sample members of this class are:

cove, fish, fishery, inshore, maritime, merchant fleet, offshore, off...shore, port, river, sea, submergence, supertanker, tanker, water, waters

(X...Y indicates that X and Y must both appear in the concordance line, and that X must precede Y.)

EXPRESS_FEELING

"Thanks to" may be said to function as a subordinating conjunction in a large number of its occurrences. On the other hand, Members (POWERFUL_PEOPLE) EXPRESS thanks to other POWERFUL_PEOPLE. The factors which come together to distinguish the two uses are complex and involve, often, precedence by a comma:

372	.SP Now, however,	thanks	to the energy crisis we have s
46	for every other child,	thanks	tc the influx of federal monie
391	y after day we realize,	thanks	to the statistics supplied to
243	f the locked-in system,	thanks	to the Minister of Energy, Min
683	things were happening,	thanks	to the NDP, and that the outst
262	lso finally of Toronto,	thanks	to the new Canadians who made

and/or being immediately followed by "the":

```
  230 tern world, and this is  thanks   to the continuing efforts of
  228 ied during this session  thanks   to the initiative of the Mi
  312 are of today's problems  thanks   to the information media.
  159 deral government exists  thanks   to the provinces.
  424  House and it is mostly  thanks   to the provisions of this b
```

But note:

```
  244 ps today we should GIVE  THANKS   to the good Lord for the mo
  511 ould like to EXPRESS my  thanks   to the hon.
  474 tice, in time, will SAY  THANKS   to the hon.
```

Another factor is the possibility of a NEGATIVE_CONNOTATION to the right of the node expression:

```
  399 es twice as much money,  thanks   to the INCREASE in the COST
```

Normally, however, there are GOOD factors to the immediate right of "thanks to" in this use:

```
  572 the west has the power,  thanks   to federal SUPPORT, and this i
  133 e work of its citizens,  thanks   to its nearly BOUNDLESS natura
  126 employment and poverty,  thanks   to our GREAT NATIONAL PRODUCT
  559                          Thanks   to the FORESIGHT of my grandfa
```

Withal, one of the factors identifying the FEEL use of "thanks" or "thanks to" is EXPRESS_FEELING:

```
  268 aper--wishes to EXPRESS  thanks   and APPRECIATION to various ex
  173 SP *roCity editors GIVE  THANKS   for news of shipwreck and prom
  247  not only to EXTEND the  THANKS   of this House to His Excellenc
   60  think we OWE a DEBT OF  THANKS   to my hon.
  258 embers in EXPRESSING my  THANKS   to two very likeable Canadians
  560 ave to OFFER our humble  THANKS   to Your Excellency for the gra
```

GOODS

These are the names of specific items for sale, or generic terms such as "farm products" or "commodities".

GOODS patterns with "accept order(s)" in opposition to, e.g. accepting (agreeing with) a principle, idea or policy, or to (legally) accepting a report of a commission or an official application for government contracts, say, or for employment:

534	ervative like myself to	accept	a principle of greater governm
659	So we must	accept	as a principle that work itsel
127	I	accept	his first idea, that of giving
339	If so, is he willing to	accept	it or will another proposal, b
638	ces Commission power to	accept	a report of any government-app
288	ght be in a position to	accept	an application in the spring.
100	at the director may not	accept	any application for a new esta
511	not the cabinet should	accept	the report of the committee wh
283	tlantic, of refusing to	accept	orders for delivery of SUGAR to
282	ENCE oil REFINERIES TO	ACCEPT	ORDERS FOR FUTURE DELIVERIES TO

While "ship" as a physical object is associated with BODY_OF_WATER, "ship" meaning roughly "to transport" is connected in the Hansards with the GOODS category:

446	er from a platform or a	ship	in deep water.
240	Maritime Command, in a	ship	or other vessel, or in a ship rep
604	In the	ship	harbour case, there was agreement

32	If you were to	ship	a car of POTATOES to Newfoundland
18	erefore an incentive to	ship	GOODS from the prairies, and thus
328	RAW MATERIAL, and would	ship	it to the United States where it

Both the means of transportation and the vocabulary of finance and buying and selling cluster with GOODS in conjunction with "deliver", in contrast to some of its other uses: deliver a baby, deliver an opinion, deliver an address.

200	delec, in my riding, to	deliver	a political speech or distrib
232	stand in this House and	deliver	a speech in French.
327	had to whip himself to	deliver	a speech in which I recognize
173	ancouver Island, and to	deliver	an address to the John Howard

198	.SP "To	deliver	an opinion is the right of al

728	hern nurses have had to	deliver	babies and relieve haemorrhag

197	decided arbitrarily to	deliver	that OIL to some other count
107	ailways are not able to	deliver	the GOODS.
706	decided arbitrarily to	deliver	the OIL to some other countr
317	t would be difficult to	deliver	the PRODUCT to Montreal wher
488	Eagle Canada Limited to	deliver	the required FUEL to the peo
348	the Arctic Islands and	deliver	the RESOURCES to be found th
359	another telling them to	deliver	their CROPS to the market, a
259	n that he either cannot	deliver	the GRAIN or the price has d
373	GRAIN, but they cannot	deliver	the GRAIN they now have and

GREATLY

These are adverbial expressions and single words signalling something like "to a great degree".

GREATLY picks out one of the numerous uses of "due": "due to" or "because", as opposed to "due (or proper) respect", "due (owing) and payable", and "due (scheduled) to take effect", among other uses.

25	to Canadian refineries,	due	allowance being made for the diffe
388	nt Service did not give	due	and timely notice of such an impen
13	I say this with all	due	awareness that there is a regional
267	turned men its just and	due	appreciation of the inestimable va
324	House without giving it	due	consideration at the second readin
249	I trust they will take	due	note of it.
400	at if equalization were	due	as a share of export tax revenues,
519	orm indicating a rebate	due	can, if the right to the receipt o
622	nion, and all sums then	due	or payable for such lands, mines,

487	reports, and another is	due	soon.
78	e government paid or is	due	to pay for that oil covered by a c
146	ate is construction (i)	due	to start (ii) due for completion (
252	oil, which increase is	due	to take effect as early as this co
678	which have been too low	due	IN LARGE MEASURE to government act
545	t by small business are	due	IN NO SMALL WAY to the attitude of
17	rce have remained open,	due	LARGELY to the efficiency of our a
61	ut declined in Ontario,	due	MAINLY to rising costs and the dif
324	has occurred in Canada	due	PRIMARILY to the direct and indire
552	n the cost of living is	due	MAINLY to the increase in the exor
551	.9 per cent in February	due	MAINLY to higher mortgage interest
378	tween 1971 and 1972 was	due	PRIMARILY to increased cost of su
588	ve at present PRIMARILY	due	to heating fuel costs, but I see
15	man consumption LARGELY	due	to inadequate chilling between th

PLACE

This test includes place names, such as "Ontario", words referring to locations, such as "area" and "region", and terms meaning location qua land : "land", "grounds", "greenbelt".

For the word "ship", the usage as "transport" correlates both with GOODS, as has already been noted, and with PLACE: GOODS are shipped to a PLACE. When the word refers to a seagoing vessel, it co-occurs with BODY_OF_WATER expressions.

32	If you were to	ship	a car of potatoes to NEWFOUNDLAND
18	erefore an incentive to	ship	goods from THE PRAIRIES, and thus
328	raw material, and would	ship	it to THE UNITED STATES where it
446	er from a platform or a	ship	in deep water.
240	Maritime Command, in a	ship	or other vessel, or in a ship rep
604	In the	ship	harbour case, there was agreement

Note that the "transport"/"vessel" usage split coincides with the verb/noun difference for the word "ship". One might think that part-of-speech labelling would be sufficient here to demarcate usages. This might be so for "ship", providing no verbal usage creeps in such as "to ship out", i.e. to depart on a ship, especially as part of the crew. However, the same cannot be said of "shipping", where the same correlations hold to a considerable extent:

98	57, to amend the Canada	Shipping	Act (coasting trade).
186	utional jurisdiction of	shipping	and navigation.
498	the Atlantic so far as	shipping	and ship repairs are concer
139	lity for navigation and	shipping	and the harbours and ports
33	international merchant	shipping	and the threat this shortag
555	of the Canadian coastal	shipping	trade has already been rais
268	would refer them to the	Shipping	and Shipbuilding Directory

354	e SOVIET UNION had been	shipping	armaments to the ARAB nation
233	ked out the business of	shipping	oil by tanker from WEST COAST
315	SAUDI ARABIA on Aramco	shipping	oil to CANADA?
726	e same, yet the cost of	shipping	petroleum to LYNN LAKE is 30
488	stern CANADA, let alone	shipping	to EUROPE.
282	not understand that in	shipping	to PORTLAND, MAINE, the oil t
352	in that the U.S.S.R. is	shipping	tremendous amounts of munitio

PLACE tends to separate physical damage--due to flooding, storms, or crop failure--from more intangible, if no less serious, damage, such as that to national security, commerce, the free flow of trade, or damage to an entire industry (e.g. the wheat trade):

24	lly in crimes involving	damage	and danger to the state, that t
16	this point without any	damage	being done to the current expor
255	ake hours to detail the	damage	done by the government's lack o
656	Irreparable	damage	could be done to a system which
180	This practice can	damage	competition if it unfairly prev
511	yet quite overcome the	damage	caused by these abortive attemp
111	lation rate, but it may	damage	business in many respects.
926	he has done irreparable	damage	and harm to the wheat producer.

569	ents suffered extensive	damage	both to their RESIDENCES and to
134	sing shoreline PROPERTY	damage	and creating the risk of possib
35	s of Canada, to prevent	damage	to shorelines, HOMES, LANDSand
569	of the fact that flood	damage	in the city of MOOSE JAW is est
587	member regarding flood	damage	in ALBERTA.
290	oncerning serious storm	damage	in PRINCE EDWARD ISLAND.
441	ic storm and high water	damage	on the GREAT LAKES?
426	SP .CE OFF .CE ON STORM	DAMAGE	IN Newfoundland LAST NOVEMBER--
438	S*ro .SP .CE OFF .CE ON	DAMAGE	ON Great Lakes--ALLEGED DELAY I
260	SP .CE OFF .CE ON FLOOD	DAMAGE	TO CROPS IN AREAS OF Manitoba--

An important aspect of the bootstrapping process referred to above is that as more and more concordances were read with care, the author's familiarity with the Hansard "world" increased considerably. Only when the point had been reached where to some extent he could "think like" a Member of the House of Commons, or at least understand to some degree "how they think"--how they categorize people, events, ideas; what they consider positive and negative; what activities they engage in professionally; what their goals are, for themselves personally, for their party, for their nation--only when this point had been attained could he lay down thematic categories capable of projection onto new, yet unseen types of the corpus. This is because there is a constancy, a sameness, a "world" (Cf. Chapter Two, Section W) underlying the Hansard text, and

it is the "world" of the Members. The correctness of this perception is demonstrated in the results obtained, displayed below, when the 61 thematic categories derived through the bootstrapping process we have discussed are used as criteria of node sense differentiation in the five test word concordances (along with the 20 "structural" categories of DS2). Remember that these five words did not form part of the 100-type sample on whose basis the thematic categories were developed. Therefore the experiment involving these five types amounts to an attempt to project into new lexical realms the categorization of the Hansard world established through consideration of quite a small sampling of types.

Having provided an account of the goals of our experiment and of the materials used, we address ourselves now to the procedure followed in applying these materials to the experimental situation. Recall that for each method--DG, DS1 and DS2--we generated 81 categories (81 files) for use in connection with each of the five test words. In the case of DS2, the categories were substantially constant over the five tests. That is, they remained the same except for the 20 structural categories, which were, of course, different for each test word, being derived from the item-specific training data. As to DG and DS1, any overlap from item to item was more or less coincidental. Nonetheless, an important distinction must be made between all DG categories, on the one hand, and all DS1 categories and the 20 structural DS2 categories, on the other hand. This distinction is the following: the former are text-derived only in a superficial sense, while the latter are genuinely text-derived. That is,

it is only for the sake of convenience that the DG categories employed in our experiment are not simply the 375 subject classes of the Longman Dictionary of Contemporary English, each consisting of all words in the dictionary which include the class in their definition. Actually, an 81-member subset of this set would be selected, but we have seen that this would be no problem, since in most cases fewer than 81 files are produced by our own method of generation. So the point is that the DG categories are in reality not at all text-derived but rather, fully dictionary-derived. The DS1 and structural DS2 classes, however, are truly text-derived; simple reflection on the process by which they are created reveals this. We draw attention to this fact here because it will be commented upon later as we evaluate our experimental results.

Two further points are in order before we offer a brief description of our experimental procedure. The first concerns the manner in which the training and test concordances were processed by computer in preparation for their statistical analysis along lines to be set out just below. As mentioned above, single words and phrases were far from the only units permissible as criteria of category membership on the part of a concordance line. We now wish to make this observation precise. The following "questions" could be "asked" of a concordance line in order to determine whether or not a given content category characterized the line:
1. Is word w in the line (or, more generally, is "word" x in the line, where x is any blank-delimited character sequence with no internal blanks)? (Henceforth, by "w" we will understand "w or x", i.e. what we informally call a word can always in fact turn out to be a blank-delimited

character sequence of any sort, again with no internal blanks.) 2. Is phrase p in the line (p being equivalent to w1, w2, wn, viz. an uninterrupted sequence of words or "words")? 3. Is c, i.e. any exponent of category C, in the line? 4. Is z- in the line (z- essentially a prefix)? 5. Is -z in the line (-z essentially a suffix)? 6. Is -z- in the line (-z- essentially an infix)? 7. Is there in the line any sequence q...r, where q, r range over w, p, c, z-, -z, -z-, and where q...r means that units q and r are in the line, with q preceding r at any distance? 8. Is there in the line any pair of elements q&r, with the range of q and r as in 7, and where q&r means that both q and r are in the line, with no restrictions on order? 9. If we designate w, p, c, z-, -z, and -z- as "terms", and state that the negation of any term is, for the purposes of the current rule only, also a term, where negation is interpreted as absence from the concordance line being queried, then is there in the line any pair of type 7 or 8 above, in which exactly one term is negative? Somewhat less formally, the types of tests which could be applied to a concordance line in order to determine its category membership included the presence in the line of a particular word, phrase, prefix, suffix, or infix, or of any exponent of another category (10); the co-presence of any two of the above types of element, or of two exponents of the same element type, in both cases either in ordered or unordered fashion; and pairs of items, of the sort just mentioned, but where one of the members of the pair is actually a stipulation that the element it represents not appear in the line.

The second point to be considered before we describe our procedure has to do with how the rules of each category--rules of the form just discussed--were actually applied to each of the training and test concordances, and what the final form of these data was. A suite of programs was written by M. Scott (11) which carried out every test of every category upon each line of a concordance, and produced as output a bit vector of length 81 (actually, of length equal to the number of input questions, whatever that turned out to be in a given run). This bit vector was a translation of the results of the concordance-line queries into the form required by the decision-tree program described below. It featured a 1 in the appropriate position within the vector when a test was passed, and a 0 when one was failed.

To proceed, then, without further preliminary to our experimental procedure, it was statistical, consisting in the automatic construction of a decision tree (cf. Meisel (1972)) which functioned to induce a particular manner of organization on an input 81-member category set, i.e. on the DG, DS1, or DS2 categories, relative to some test word t . We will henceforth refer to these "relativized" category sets as, respectively, DG_t , $DS1_t$, and $DS2_t$. The manner of organization in question was the following: First, that category among the 81 was discovered which, speaking intuitively, was the most helpful in separating out the concordance lines of the training data in terms of the sense labels of their nodes. (It is

important to bear in mind that all 2000+ lines of a test-word concordance, including the 1500+ lines of training data, and the 500 lines of test data, had been hand-labelled by the author according to the sense number (1-3 or 1-4) of the node of each line. Further, the program generating the decision tree had access to the label number of each line, and in fact this information was crucial to the tree's construction.) A bit more specifically, a binary division of the complete data set was induced by means of this first decision. If the category c_1 chosen as the label of the tree's root was present in the list of categories which applied to a given concordance line, then that line was among those represented by the node which was "right son" of the root; and if not, the line was in the class which the "left son" node stood for. Next, a category was chosen from the remaining 80 classes such that it was "most helpful" in the binary partition of the set of concordance lines which did not feature the category which labeled the root; and one was chosen which best partitioned the complementary set. In particular, category c_2 was that which served to label the "left son" node of the root, and was also the one which was "most helpful" in cleaving the set of all those concordance lines in which c_1 did not appear. Here the division effected was into one class of concordance lines of which c_2 but not c_1 was in fact a descriptor--and this class was represented at the "right son" node of c_2 --and another whose

descriptors included neither c_2 nor c_1 --and this class corresponded to the "left son" node of c_2. The remaining nodes of the decision tree can be accounted for along the same lines. The elucidation of our intuitive notion "most helpful" is in terms of mathematical information theory (12). For a presentation both of the relevant information-theoretic notions and of the particular decision-tree algorithm used in the present experiment, see Lucassen (1983) (13).

Once a decision tree was obtained for each DG_t, $DS1_t$, $DS2_t$, utilizing the training concordance for each t, the Scott concordance-analysis program was run again, but this time on the test concordance for each t, once for each method, for a total of fifteen runs. A program was written by the author to apply an information-theoretic measure of goodness, called the perplexity measure (14), to each of the fifteen files representing the output of the test data analyses. Again in an intuitive sense, what this goodness measure consisted of was the quantification of the predictive power of the decision tree derived, for the training data, on the basis of each DG_t, $DS1_t$, $DS2_t$, when such decision tree was applied to the test data. That is, three different methods--three different category sets--were used to analyze

each of the five test words. Characteristics of the environments of each sense of each test word were identified according to three different schemes, based on training data. Now, by applying these same classifcations to new data--to the test concordances--we were able to see how effective each of the methods was at achieving the goal of predicting the node sense numbers of concordances over the five test words.

Perplexity is 2 raised to the power corresponding to the entropy, in an information-theoretic sense, of a set of data. Entropy, in turn, can be thought of as the average of the logarithms to the base 2 of a set of probabilities. Now the probabilities with which we were concerned in judging the performance of our three methods were obtained as follows: Every possible combination of "plus" and "minus" values for a set of labels of a decision tree--in the present case, every possible combination of values for the presence and absence of categories in a concordance line--corresponds to one or another of the leaves of the tree. Since a leaf of a tree is a node which is just like other nodes except that it has no branches, the leaf corresponds to a set of items being classified, and therefore in this case to a set of concordance lines. Because some of these (training data) concordance lines bear, say, sense label 1, others label 2, still others label 3, and so on, the probability can be calculated, with reference, of course, to the training data, of arriving at a given leaf in the course of analyzing a single concordance line via the tree, and finding that the line bears label 1, or 2, or 3, etc. This is exactly what the evaluation procedure amounts to.

The evaluation program written by the author takes up each of the 500 lines of the test concordance, one by one, determines on the basis of the output of the Scott program which categories are present, and which absent, in the line, and then simply finds the leaf of the decision tree that also bears that pattern of category presences and absences, or pluses and minuses, or bit-vector 1s and bit-vector 0s. Once the proper leaf has been located, the program notes the actual sense number of the test line, and looks up the probability of a line's simultaneously being at that leaf and bearing the sense label it does. This is the probability of the particular line of test data. Its base 2 logarithm is calculated, and this value is added to a tally of such figures for all 500 lines of test data. When the last of the 500 lines is in fact reached, the tally is divided by 500, yielding the average of the base 2 logarithms of the probabilities, in the sense just provided, of the 500 lines of test data. That is, if we recall the definition of entropy given above, we now have the entropy of our data. (Actually, the entropy is this figure multiplied by -1.) To get an intuitive feel for what this statistic might amount to in a given case, consider that if we had four possible senses for a node, and if on the average our decision tree allowed us to correctly predict the sense number exactly 50% of the time, then other things being equal (these other things including notably the distribution of the samples, i.e. the number of lines with senses 1, 2, 3, etc.) the entropy of these data would be 1.00. If we were right 25% of the time--that is, again other things being equal, if our method were no better than chance--the entropy would be 2.00. As a last example, if we had a perfect method, one that always predicted the correct sense label, then the

entropy would be 0. So some idea of the space being explored has hopefully just been provided. Finally, note that perplexity is merely a transformation of entropy; in particular, it is 2 raised to the power corresponding to the entropy of some set of data. So in the first case discussed just above, the perplexity would be 2, in the second case 4, and in the last case 1.

We have so far considered the goal of our experiment, its materials, and its procedure. We are therefore ready to state the results obtained, and then to proceed to discuss them. The results of the fifteen analyses of test data conducted on the basis of the decision trees obtained from the fifteen corresponding analyses of training data can be summarized as follows:

TEST WORD	PERPLEXITY(DG)	PERPLEXITY(DS1)	PERPLEXITY(DS2)
interest	2.77	1.78	2.35
point	2.78	1.69	1.56
power	2.83	1.67	1.70
state	3.09	1.92	1.60
terms	2.61	2.31	1.99
AVERAGES	2.82	1.88	1.84

A further result is that when, for two of the above words selected at random, the 20 structural categories of DS2 were replaced by the 20 DS1 categories that appear highest in the DS1 decision tree, the result was nearly identical in each case with the minimum of DS1 and DS2. Specifically, this combination of DS1 and DS2 for the word "point" resulted in a perplexity of 1.56, which is exactly identical with min(DS1, DS2) for this word. The same combination applied to the word "state" yielded a perplexity of 1.58, which is within 0.02 of min(DS1, DS2) for "state" (in particular, it is better by that amount). Projecting this result onto the entire set of five test words by averaging the minima of DS1 and DS2, we obtain an estimated perplexity of 1.72 using the particular method of combining the two DS approaches which is described just above.

Some initial comments are in order with reference to the performance of DG. Reference was made earlier in this chapter to the need to take into account, in evaluating statistics of the type with which we are dealing here, the distribution of the test samples themselves. Recall that we said, for instance, that the entropy of a four-sense test-word concordance would be an even 2.00 if a particular method of sense prediction proved so poor as to be correct only 25% of the time, AND if there were exactly the same number of samples of each of the four senses. Now, some idea of how very poorly method DG performed may be gained by noting that, if we take into account the entropy of the test data for the word "point", for example--a word which happens to exhibit quite a skewed sample distribution in its test concordance, at 329, 128 and 43 respec-

tively for the three senses--then DG did considerably worse than chance. That is, by chance selection of sense numbers, one would attain a perplexity of 2.3; but we can see that the DG perplexity was considerably worse than that, at 2.78. And the DG score for "point" is not unusually high (i.e. unusually bad) compared with its score for the other words; in fact there is only one of its perplexity values which is lower than 2.78, viz. that for "terms".

An analysis of the DG trees themselves, and of the contents of the DG categories, helps us to understand why this method fails. For instance, those categories which rise to the top of even a mildly useful decision tree, in the present experimental environment, are connected with the thematic or structural functioning of the node word in a way that is intuitively obvious. So, the category "PLACE" appears at the top of the DS2 decision tree for "point", since almost every sample in which the word is used to mean "geographical location" has some exponent of this category, and not many other lines do. Hence, "PLACE" is quite "helpful" (informative) with reference to the disambiguation of the nodes of this concordance. Similarly, in the DS1 tree for "terms", the categories "IN", "THE" and "OF" rise to positions at or near the root, and the reason for this is intuitively clear. It has to do with the predictive power of such frequent expressions as "in terms of" and "the terms of", each of which characterizes two of the four sense of "terms" (although not the same two) to the practically complete exclusion of the remaining meanings. But this phenomenon fails to occur in the DG trees, and this failure takes on at least two forms. First, there are cases where the categories that exist

in the list of 375 Longman subject classes--categories one might choose on an intuitive basis as "relevant"--do not even attract a single one of the 500 most frequent words of a test item, and therefore simply do not get a chance to appear near the root, or for that matter anywhere else in the tree. But there are also instances where the intuitively likely categories are in fact represented in the tree, and then there is no discernible tendency for them to inhabit the upper reaches of the tree. Rather, their distribution does not seem to follow any particular pattern. For instance, in the DG tree for "interest", classes such "COMMERCE", "BANKING" and "FINANCIAL" do occur in the decision tree, but they are nowhere near the top, and actually serve to categorize only quite small numbers of samples. The point is that this is counter-intuitive, and suggests that the DG method is just failing to reflect what is going on thematically, or in any other way for that matter, in the concordance.

A look into the DG category files themselves furthers our understanding of what goes wrong when this method is utilized. On the one hand, even the most common "function words" are given very complete descriptions in terms of the subject classes, so that, for example, every occurrence of "on" or ' ff" is categorized as indicating that there is some sense in which the topic "ELECTRICITY" is being discussed. It is doubtful that this is a shortcoming in a dictionary, but it would certainly appear to be one in a tool for word sense discrimination, at least via the method under discussion. On the other hand, one might think, on the basis of what was just said, that DG performance would improve substantially if function words were removed from the input to the Longman

categories, and only content words were therefore retained. Unfortunately, we also find a bad situation in this sphere. What typically happened is that words which were "content" items but still high in frequency, such as "point" when it occurred in concordances other than its own, were placed in a rather large number of categories, including, for example, "MATHEMATICS" and "LAW". In most cases, these categorizations have nothing whatever to do with the topics characterizing the concordances, and, if this sort of "misclassification" were either an isolated or a random sort of event, presumably no harm would eventuate. However, because of the consistently comprehensive nature of the Longman classifications--comprehensive, of course, from the point of view of their own subject classes--frequent content words start to accumulate in classes irrelevant to the topics marking the test concordance, and powerful categories emerge which are at cross purposes with the aims of the procedure. As an example, "make", "times", "power", "number" and "sum" might easily join "point" under "MATHEMATICS". If function words were allowed, "and" , "by" and "into" might also appear. Now, in the Hansards, for example, "power" has to do with the DS2 themes ENERGY (e.g. hydro power), POWER (e.g. thirst for power), RESPONSIBILITY (e.g. The Board has the power to revoke that ruling), and a few others. "Times" would either fall under the "o'clock" meaning, in which case it would be associated in DS2 with the TIME_WHEN theme, or under the "newspaper" meaning. In the latter eventuality, it would come into contact with, and possibly even be a member of, the INFORMATION and/or DOCUMENT classes of DS2. In the case of DS1, where "power", say, and "times" would fall would depend to a great extent upon the node word and its lexical patterns. However,

it is possible to cite some fairly clear examples of the probable functioning of these terms given certain nodes. Where the node is "party", for instance, "power" should be both a content category and a structure class; in the first case, expressions such as "came to power" and "thirst for power" would pull "power" outside the node+-2 range and make it a candidate for content category status if it is sufficiently frequent. In the second case, there are locutions such as "party in power" which occur quite often and which should suffice to include "power" within the structural categories. To cite two obvious cases, where the nodes were "London" or "York", "times" would quite possibly be structural categories within DS1, but in DG, the word would still face the prospect of being dragged into an association with the other high-frequency words cited above, say under the "MATHEMATICS" rubric.

When the results presented above are considered globally, at least two observations stand out. One is that it seems that well-chosen exclusively structural categories did as well as a combination of about 3/4 thematic and 1/4 structural classes. The other is that DG had neither accurate thematic categories nor any structural ones at all. We will discuss the first of these points at some length, and then return to consider the second point in the light of these comments.

Let us begin by accepting the DS1/DS2 results at face value. On the one hand, the improvement achieved by combining the two methods points to tasks for future research, such as, first, varying the number and character of DS1 categories entering into combination with the DS2 host;

and second, attempting to discern the source or sources of DS1's power by teasing apart its elements and investigating their performance in different combinations, and in combination with the DS2 structural classes. On the other hand, the near-equivalence of the DS1 and DS2 scores demands consideration on its own terms.

A brief but careful review of the structurally-oriented method of Gross and an examination, as promised, of some of Sinclair's (1985) specific lexical analyses might prove helpful at this point. Regarding Gross, recall the dilemma posed by his approach. If he toed a strictly structure-oriented line, his results would be interesting but not particularly impressive. But if he did a conscientious job on what he calls selection, it was not clear how he could avoid adopting, to some degree at least, a content-analytic approach such as that of DS2. Although there are certainly relatively long-range structural phenomena which Gross's approach but not DS1 would capture, there is also a fairly extensive common ground between the two methods. Certainly the analysis by Gross of the word "office", cited at length in Chapter Two, can to some extent be recast in DS1 terms. It goes without saying that all nine compound nouns enumerated there fall into the DS1 net if they are frequent in the source text. If they are not frequent, there may still be a way of handling them, perhaps within Gross's scheme, perhaps within the framework of Mel'chuk, to whom Gross pays homage in Gross (1985), as we saw in Chapter Two. We return below to Mel'chuk for some brief comments. Unfortunately, in either of these cases, extensive work by hand would be involved. Debili (1982), as we saw earlier, claims to have automated the

discovery of instances of some of the Mel'chuk categories (cf. Mel'chuk (1984), Evens et. al. (1983)), but the only ones he in fact uses are those equivalent to the traditional "grammatical relations", and not those expressing "lexical relations", which would be the useful ones in cases such as, say, "box office" vis-a-vis "divine office", each of which had one or two occurrences in the Hansard data base.

Continuing with our enumeration of points of contact between the Gross method and DS1, we may note that his senses 1 - 3 share the environments Human's office (e.g. Bob's office) and possessive pronoun office (e.g. my office), unfortunately not to the complete exclusion of the remaining senses, but at least the two examples he cites of these constructions with other senses sound awkward and may be infrequent: Max will suppress Bob's office of treasurer (his sense 4); Max used the lawyer's offices (meaning the lawyer's good offices--his sense 5). Again, with the window set at node+-2, sense 6 (perform an office), sense 5 (good office(s)), sense 4 (hold, take, in, into, out of office), and possibly some other senses, fall within DS1's purview. What we make of all this is that to the (apparently considerable) extent that there is overlap between the Gross approach and DS1, we can apply to the latter the evaluation given of the former. DS1 should be good but not great--and it is. Further, it should be in need of supplementation via thematic categories; and this is what is indicated at least preliminarily by the immediate improvement obtained on the very first attempt at combining DS1 and DS2.

Redeeming another promise made in earlier chapters, we now look at some of Sinclair's analyses of the lemma YIELD, as rendered in Sinclair (1985). The relevance of this undertaking to our present interests--those of coming to understand our results, and especially, for the moment, the DS1/DS2 near-equivalence--is that we will see that Sinclair's analyses simultaneously reveal the strengths and weaknesses of the structural approach to word sense discrimination. Thus on the one hand, they suggest fruitful means by which structural factors might be put to greater use in future algorithms, and on the other hand, they reinforce our analysis of Chapter Two in which the notion "world" features prominently.

First recall Section I4 of Chapter Two, and the emphasis throughout this study on the relativity of rules to one's purposes. If, returning to the theme of our Introduction, we reiterate that our framework of investigation is the quest for predictivity in text, then we are able to sharpen our appreciation of the pros and cons of Sinclair's structural method. Any of the sets of examples he presents for a "single sense" would illustrate our point, so we focus on the very first such set he adduces. To perhaps overstate the case Sinclair wishes to make regarding the following set of examples, he argues that YIELD in its intransitive-verb avatar always has the meaning "give way" as opposed to "produce", to "lead to", and to a number of "minor meanings" (Sinclair (1985, p. 5)). The examples in question (p. 3) are these:

> But we did not yield then and we shall not yield now. "On the wall of his office, Professor M... (24)
>
> ...Watergate conversation, which Mr. Nixon agreed to yield to the courts, "did not exist". As it happe... (35)
>
> ...modernize their precepts of Judaic orthodoxy, and yield to the demands of Western culture as well as... (42)
>
> ...means to prevent it. Ovid recommends: "Love yields to business, be employed, you're safe", and... (122)
>
> ...as once the masculine province of the area but is yielding and its glamour has been somewhat diminis... (84)
>
> ...to Denmark and Sweden. In Sweden the authorities yielded at once to the threats which swiftly follo... (56)
>
> (Sinclair (1985, p. 3)).

The important point here is this: Let us grant, for the sake of argument, Sinclair's claim for the falling together of intransitivity in YIELD and of the meaning "give way". Now, after harvesting whatever predictive force accrues to the presence in a clause of an intransitive verb as opposed to other constructions capable of supplanting it, and after making some slight play of the (probably rather general) association between this sense of YIELD and "demands" (42) / "threats" (56), what have we gained in terms of predictivity? There seem to be no generalizations in sight as to typical agents or recipients of this sort of "yielding", no hint as to prior, concomitant, subsequent or otherwise related activities to "yielding". (There is the presence, four times of six, of the word "to" either just after YIELD or a few positions to the right. All methods we have seen (except DG) would probably catch this. But note that, since "to" is such a frequent word--often the most frequent of a

text sample--many errors would be made unless it were required to immediately follow YIELD, and/or a good job of parsing were done at least on the few words to the right of YIELD. Even then there would be a good number of misfires. Sinclair's examples include quite a few with "to" in the right neighborhood but which are not assigned the sense "give way" or a variation thereof.)

What we conclude from this excursus is first, that we must allow a strong language-general foundation to discourse--but that it is "structural" in the narrow sense, and of interesting but ultimately limited power. The successes, even partial, of Gross's analysis of "office" on the Hansards; the plausibility, to an extent, of Sinclair's sense/structure association as argued through the example of YIELD; and most immediately, the equivalence of the DS1 and the DS2 scores, impel us to adopt this viewpoint. Our second conclusion, however, is just as basic. This is that thematic predictivity, which ultimately is founded on a familiarity with a "world" or domain, is flagrantly lacking from the sharply structural orientations of Gross and Sinclair, and is only supplied to an extent in DS1 where a particular word carries a theme with singular frequency. One of the most adamantine obstacles facing DS1 would seem to be the situation where a theme is expressed with regularity in the neighborhood of nodes bearing a given sense, but where there is quite considerable variation in the words and expressions which represent the theme. One possible explanation for the egregious--and anomalous--score of DS1 for the word "terms" is that two node senses, between them accounting for a bit less than half the samples, were related in the fol-

lowing fashion: First, they shared both left and right immediate context with great regularity. Second, no particular word or words seem to predominate as exponents of the key themes differentiating them. Yet, their difference is quite easily appreciated, is enshrined in the two college dictionaries consulted on the question, and is apparently somewhat more easily captured by the thematic approach of DS2 than by DS1's structural rules. The two senses are: terms of a law, a ruling, an indictment (its articles, its strictures) and terms of a sale, a labor agreement, a loan (we would maintain that, basically, the terms are the sums or the benefits slated for transfer or award); they share the environment "the terms of". So a major aim of continuing research in this area should be to hit upon ways of fruitfully combining the strengths of the thematic and structural approaches.

While on the topic of progress which may take place in the future, we wish to take brief note, consequent on our expressed intention to do so, of the lexicological work of Mel'chuk and his associates (Mel'chuk (1984), Evens et. al. (1983)). Mel'chuk (1984) is an extremely detailed application to 50 French words of principles which have apparently been undergoing constant development for decades now. A large battery of "lexical functions", relations mapping an argument word into its lexical image along one or another semantic axis, have been enumerated. They can enter into combinations, yielding an extremely rich set of lexical descriptors. (For an account of some of these possibilities, see the two references cited just above.) The problem is that, at least for French (no work of which we are aware is going on in this tradition in English),

while the analyses of the 50 words in Mel'chuk (1984) are satisfying and quite likely useful as an extension of language-general facts into the realm of detailed and specific lexical affinities, large-scale treatment of the lexicon along these lines would seem to be a rather futuristic goal.

We wish now to redeem a final promise made in previous chapters and to compare the definitions we arrived at for our five test words, with those utilized for the same words by Kelly and Stone. Recall from Chapter Two that their expressed aim in carrying out their sense- and part-of-speech-identification research was that "we wanted both the specification of the dictionary domain and the construction of entries to reflect English as the behavioral scientist is likely to observe it" (Kelly and Stone (1975, p. 5)). With this goal in mind, they adopted as source text "conversational material, personal documents, dreams, survey responses, TAT stories, literature, speeches, editorials, and folktales" (ibid.). Finally, as we saw in Chapter One, Kelly and Stone state: "the set of senses is itself relative to our pragmatic aims--i.e. we ask which of 'the senses' of this entry seem useful, or worth discriminating. To this end we were aided by our accumulated knowledge of what kinds of distinctions are important to content analysis work" (op. cit., p.10). As was pointed out in Chapter One, they presumably mean content analysis work in the behavioral sciences when they refer simply to "content analysis work".

One reminder we will derive from this comparison will be this: It was noted long ago by Chomsky (15) that traditional grammars, while gen-

erally offering highly cogent linguistic analyses, rely for their understanding upon the application made of their principles by the "intelligent reader"--normally, by the "native speaker" of the language described. If we change "traditional grammar" to "traditional dictionary" and, in what is no more than a shift of emphasis, replace "native speaker" by "participant in a form of life or in a 'world'", then we will have a convenient guiding notion as we juxtapose these two sets of definitions. We return to this idea below.

We will proceed alphabetically through the five test words. Kelly and Stone (1975, p. 185) provide two nominal senses for "interest". They are: 1. curiosity, concern, involvement, stake; 2. the cost of borrowing money. In analyzing the Hansards, we were led to distinguish four senses. The first corresponds to Kelly and Stone's second meaning, but it would probably be phrased "return on investment". Although perhaps trivial in appearance, this difference in perspective with regard to this sense of "interest" seems to us to be symptomatic of a difference in world view which will become more apparent as we proceed. Two further Hansard meanings for "interest" which will concern us (we will not dwell on the fourth, the locution "in the interest of", used as the rough equivalent of "for the purpose of", "in order to") are: Hansard sense two--"concern", curiosity; and Hansard sense three--stake, potential gain. What is interesting about Hansard sense two is that neither disinterested curiosity nor personal concern is normally denoted when this sense is employed. Rather, what is expressed seems to be a sort of self-interest whose pursuit is taken for granted:

e will watch with great	interest	the uses to which the railway land
ecall it, related to an	interest	by Soquip in buying shares in Pana
I await with	interest	the response of the government whe
t the period of keenest	interest	exists in the last few days of the
off because there is no	interest	by this government.
the minister show some	interest	in the fisheries, back up the prom
and we shall look with	interest	on the kind of arrangement which c
ey, too, know the great	interest	that municipalities have in the ur
.SP I followed with	interest	the address by the mover of the mo
Why this sudden	interest	by the government leader who, duri

For this reason, sense two enjoys a rather fluid relationship with Hansard sense three, which, curiously, actually acquires a more disinterested aspect than is associated with sense two, since it is frequently used in such locutions as "national interest" and "public interest":

enthusiasm and with the	interest	of the farmers very much at heart.
on, bearing in mind our	interest	over the long run in maintaining p
Canada has no greater	interest	than in the maintenance of a stabl
u all have at heart the	interest	and the well-being of veterans of
ultimately in the best	interest	of western Canada and of B.C. in p
ecessarily, against the	interest	of the people of this country, by
hink it would be in the	interest	of everybody in Canada today, part
ntention to protect the	interest	of the consumer and deter restrict
question of conflict of	interest	of cabinet ministers.
lve with the legitimate	interest	of each sector of the economy bein
o find it in the public	interest	to control the price they must pay

Because of this fluidity, there are a good number of borderline cases, some of which are:

o draw attention to the	interest	of some groups in metropolitan Toro
) .SP .CE ON GOVERNMENT	INTEREST	IN PINETREE RADAR SITE .SP .CE OFF
he public interest, the	interest	that we as parliamentarians I think
and indicate the large	interest	that group has in Ontario.

Contrast Hansard sense two with Kelly and Stone sense 1, which might be said to telescope Hansard meanings two and three, but which in fact seems to have running through it the notion of actual personal involvement, either with other individuals or with ideas.

Turning to the noun "point", we have three Kelly and Stone senses: 1. temporal or spatial (much less frequent) location, level, degree, stage; a dot, sharp end (infrequent); 2. observation, argument, main idea, purpose, thrust, gist, meaning; and 3. distinguishing feature or characteristic, usually of a person-- "good points". We also have three Hansard senses: 1. stage or degree; 2. geographical location; and 3. topic of discussion, idea. The notion of linguistic communication is at least superficially a locus of contact for the two domains, so that Kelly and Stone meaning 2 is in some sense parallel to Hansard meaning 3. Note, however, that Hansard 2 and 3 are compressed into Kelly and Stone 1, which even bears the note that what would be Hansard 2 is "much less frequent". A characteristic of the Hansard world is that it has extensive national and worldwide "concerns" having to do with places.

For the noun "power", Kelly and Stone list only two noun uses, and one is in fact labelled "noun-adj". This fact is suggestive in and of itself. The two Kelly and Stone senses are: 1. control, influence, strength, ability, faculty; physical energy; and 2. nation, group or individual having power. The Hansard senses are: 1. control over others; 2. legally delegated authority; 3. mechanical energy; and 4. nation-state. It is actually rather surprising that Kelly and Stone 2 includes

"nation...having power", which is rather similar to Hansard 4. Kelly and Stone 1 features the rather positive terms "strength", "ability" and "faculty"; Hansard 1, on the other hand, appears regularly in the environments "thirst for power", "clutch at power", "grasp for power", "take power", and so forth, which have pronounced negative connotations. There is some representation in the Hansards of the "positive" Kelly and Stone sense 1 sub-senses, but it is quite scant.

Kelly and Stone's "state", rather unexpectedly, includes two nominal idioms which feature prominently in the Hansards, namely "state of affairs" and "ship of state". In addition, there are two noun senses: 1. body politic--area of government; and 2. condition, often relating to consciousness or perception. The Hansards feature: 1. condition; 2. American administrative unit; 3. the government; and 4. the conduct of official government policy, especially foreign policy. Two remarks are in order here. First, it is of course to be expected that in the Canadian House of Parliament, the "political" notion of "state" should be more elaborated than in a behavioral science environment. Second, there is no great tendency for Hansard 1 to relate to consciousness or perception, as does Kelly and Stone 2:

ate and report upon the	state	and management of the business, or a
in a natural and decent	state	after the project is completed?
f its relatively feudal	state	and brought it up to, perhaps 10 or 1
areas in their natural	state	and it will cost more and more money.
economy is in a fragile	state	and must be handled with considerable
onsiders to be a crisis	state	and the powers come into effect for a
ed full employment is a	state	in which every man and woman of this
ediate reference on the	state	of our economy to the Standing Commit
we are discussing a sad	state	of affairs.
kept in an undesirable	state.	
he confused and chaotic	state	of land titles-- .SP I do not need to
specially in the frozen	state,	in order to protect the jobs of Cana
o are in a transitional	state	following confinement to a penal inst
rea, the extremely poor	state	of the roads and facilities there and
rations, the health and	state	of our waters, the supply of our fish
finds their dilapidated	state	just too costly to keep in repair.

This suggests, when taken along with other senses reviewed above, somewhat more of a preoccupation with things (vapor state, liquid state, sad state, sorry state) than with people.

Turning lastly to "terms", both sources feature the locution "in terms of", which we will not consider. Kelly and Stone also have four noun senses: 1. specific word or expression; 2. a period of time; 3. conditions or stipulations; and 4. relations or standing with another. The Hansards have: 1. words, manner of expression; 2. point of view (in economic, dollar, percentage terms); 3. provisions (of a law, order, etc); and 4. conditions of sale or other monetary transaction. What should be noted is the "collapsing" of Hansard 3 and 4 into Kelly and Stone 3, and the more or less total absence in the Hansards of Kelly and Stone 4. A Hansard analogue of Kelly and Stone 2 exists, but is relatively infrequent and is about evenly divided between locutions amounting to "prison terms" and "periods during which an elected official remains in office".

It would appear noteworthy that in none of these five cases which, after all, were selected randomly, was there identity between the two sets of definitions. The differences seem to fluctuate between obvious and deep differences of basic orientation and Weltanshauung, and what, at least on the surface, look like mere omissions on the part of one or another source. In any case, what does emerge from this brief exercise is a clear picture of two different arrangements of meaning, of life. In the case of the Hansards, the arrangement seems coherent enough to call it a "world". The Kelly and Stone source text seems somewhat mixed, and in any case we know that its authors do not share the sort of tight-knit association that the Members of Parliament do; rather, one might say that they share a general orientation. Certainly all these speakers-- Members, social scientists, clients, raconteurs--share a form of life.

To pass once over lightly the idea put forward just before we began comparing the Kelly and Stone with the Hansard senses, we might use the word "interest" in its sense "concern" to illustrate once more the Wittgensteinian point (cf. Chapter Two, Sections I3 and I4) that him in no sense a distortion. A thoughtful consideration of the status of this sense of "interest", and of the likely relations of this term with hundreds of other key words, phrases and constructions of the Hansard vocabulary, may conduce to an appreciation of the value of thematic or content analysis as an independent tool of text analysis when the latter is pursued with the goal of maximizing predictions over textual events.

At this point, it might be well to comment on the status of the Hansard world from the viewpoint of natural language processing in general. So far, successful computational analysis of text has been undertaken on extemely-restricted-domain text types, and contrapositively, whatever work has been done on unrestricted text has been, frankly, speculative. Under the first rubric we may count systems for the treatment of "sublanguage" texts, in the sense of Kittredge and Lehrberger ((eds.) 1982). Examples of research productive of working, usable systems, and focusing strictly on highly restricted and narrowly defined varieties of language, include the work of N. Sager (Sager 1982, Hirschman and Sager 1982) and of Kittredge (Chevalier et. al. 1978) and their associates. The "sublanguages" dealt with in the research just cited consist of the language of journal articles in pharmacology (Sager (1982)); that of the "hospital discharge summary"--the portion of a hospital patient's record on which the attending physician makes notes by way of a final indication of the patient's status at the time of discharge from the institution--(Hirschman and Sager (1982)); and the language of weather reports in Canada (Chevalier et. al. (1978)). An early "sublanguage" grammar was that of Damerau (1971); its domain was that portion of U.S. Patent applications, in the field of electronics, in which petitioners described their innovations. Yet another sublanguage study is Gross and Tremblay (1985), which bears on the terminology of nuclear power. Some of the main characteristics of domain-limited language analysis, and some of its important limitations as well, are touched on in the following passage from this report:

> Il est certain que pour atteindre un tel raffinement dans la description des elements d'un lexique, la delimitation du domaine

> a etudier et la collaboration d'un specialiste du domaine constitueront des conditions essentielles de reussite. La delimitation du champ d'etude nous obligera a deployer notre activite sur un nombre limite d'elements correspondant a un reseau notionnel restreint. Il faudra s'assurer une connaissance approfondie du domaine etudie pour decider du type de rapports entre les termes qu'il y a lieu de faire ressortir compte tenu de l'application envisagee. (Gross and Tremblay (1985 p. 67)).

Note that the domain is so restricted that the aid of a specialist is said to be indispensable to insure accuracy in the grouping of content terms. (It may be that this requirement is obviated where, as in Hirschman et. al. (1975), automatic means are applied to this task.) A second characteristic of "sublanguage" studies which is pointed to here is the highly reduced vocabulary, and the severely bounded repertoire of concepts which one can expect to see expressed. Third, it is observed that considerable familiarity with the domain is required if terms are to be appropriately related via whatever formalism is adopted. We will return to the interesting and revealing remarks of Gross and Tremblay in short order.

Examples of the other end of the natural language processing spectrum--wide-ranging, language-general strategies--are ubiquitous. Certainly approaches such as Amsler's for instance, which rely on general-purpose dictionaries as the sole source of information concerning English words, seem to bespeak an ambition to be able to "handle" any text type whatever. Pursuing the example of Amsler's thesis, we find that he does in fact speculate along just these lines: "It...is reasonable to expect that a fully disambiguated dictionary could form the basis for a disambiguation procedure useful on any text" (Amsler 1980 p. 125) (16). Similarly, Gross would seem to view his lexicon-dictionary as capable of serving as the only information source for a computational system of word

sense discrimination in any environment except for such narrowly circumscribed "sublanguages" as are alluded to above. In Gross and Tremblay (1985), for instance, a pre-existent terminology bank in the field of nuclear power is analyzed using what one may broadly call Gross's methods, but a lexicon-dictionary is in effect constructed entirely de novo for the sublanguage. This includes new transformations whose range of applicability is apparently limited to the realm of texts concerning atomic power. It would appear that a more or less neat separation is intended between such specialized linguistic varieties as that which forms the subject of Gross and Tremblay (1985), and the French which is classified in the mainstream lexicon-dictionary. In fact, the language codified there is subject to the following idealization: "a first approximation of a description should not include socio-linguistic data in the grammatical representations. Thus, we make no provision for notions such as idiolect, multiple grammars for a single speaker or separate grammars for separate groups of speakers. Nor do we distinguish between styles, such as literate speech and slang, for example" (Gross (1984, p.380)).

Let us try to place the Hansard world with reference to the two poles, "sublanguage" domains and whatever textual realm it is that corresponds to a "language-general" perspective of English. Returning to the three hallmarks of sublanguage domains enumerated in the citation above from Gross and Tremblay (1985), we may start by noting that the Hansard vocabulary is not restricted, unlike that of a sublanguage. A sort on 6 million words of the Hansard data base shows 43,020 different types--hardly an impoverished lexicon. A second sublanguage character-

istic alluded to by Gross and Tremblay is the need for expert assistance in classifying the content words and expressions of the corpus. Again we find a difference vis-a-vis the Hansards. While a typical clause in the medical text with which Sager and her associates are dealing, is apparently "The abdomen shows no organomegaly", or "The cerebrospinal fluid shows a WBC of 55", or "The chest shows slight intercostal retractions" (Hirschman and Sager (1982 passim)); and while a sampling of the terms to which Gross and Tremblay devoted their attention looks like this: "neutrons lents", "puissance massique", "circuit secondaire de refroidissement", "equipment de transfert incline", "anneau de pulverisation basse pression" (Gross and Tremblay (1985 passim)); some typical concordance lines from the Hansard data base are:

Speaker, I wish to	direct	a question either to the Minister o
B., of a city hall that	cost	$1,200,000 to the builder, whereas th
I have not seen the	thing	happen either inside industries.
a letter on their desks	stating	that they had been laid off effect
is the first time this	particular	kind of situation has arisen.
er reinstituting it and	making	an announcement about it within the
Hydro for the month of	January	1969, one can read this: .SP s*i
.SP Mr. Speaker, I	hope	the government really means what it s
committee, on which the	fishermen	were represented, recommended ag
Republic with a view to	establishing	diplomatic relations and deve
is House, despite three	days	of time wasted arguing against them.

This language is hardly recherche. In fact, one is tempted to say that there is nothing special about this language at all, and that it is simply "regular English", the type of text that a "general-purpose" version of English can accommodate. Actually, however, as we have just seen, the Hansard world imposes its own interpretation on its vocabulary, a meaning

discernible to greater and greater extent as one reads Hansards and becomes familiar with their world.

The upshot of the immediately preceding discussion for the placement of the Hansards with reference to the "sublanguage"/"general English" spectrum is that the second criterion of sublanguage domains enunciated by Gross and Tremblay--the need for expert assistance in analysis--does not apply to the House of Commons proceedings. Rather, we have seen that the non-technical nature of the Hansards allows the linguist himself or herself to become an "expert" in the domain after a good deal of careful reading and preliminary content analysis of the source text. In fact, in domains much more obviously value-laden and marked by human interaction than weather reports or pharmacology articles, it is easy to imagine the influence of the so-called "expert" upon the linguist as inhibiting and even obfuscating. For instance, we reached the conclusion that, in the Hansard world, human beings are classed as either "Powerful People", including the Members themselves, or "Voters", or "Workers". One doubts that such a trichotomy would have recommended itself spontaneously to an "expert" Parliamentarian as the most perspicuous. So, while Gross and Tremblay's second typical property of sublanguages is not characteristic of the Hansards, the third one--the need for a close familiarity with the domain (on the part of the analyst)--does indeed apply.

In sum, we see that the "Hansard world" fits neither the picture which experience has drawn of the "sublanguage", nor its complement, so to speak. We may speculate that to the extent that a large text source

is "natural", i.e. not a compilation for study purposes of sub-corpora of divergent types--and, of course, to the extent it is not a sublanguage domain (17)--to this extent, this same "limbo" between sublanguage and "general-English" status will always be obtained. That is, the third of Gross and Tremblay's criteria, the need for intimate acquaintance with the domain or "world", will never fail to apply where real-world text sources are concerned. So it seems to us. Only actual attempts at sense discrimination, or at related types of text processing, on other large corpora which are not simple compendia, can indicate the degree of correctness of this speculation. One last point here is this: Since the question addressed by the speculation is an open one, one which we lack sufficient knowledge to resolve, let us face the fact that we have no reason whatever at present to speak, as we did just above, of a "limbo" between sublanguages and "general English". We know what a sublanguage source text looks like; we have numerous examples of them. But what is a "general English" source text--one for which, for instance, the Longman categories would be effective in identifying word senses? It is important to realize that, given the state of our understanding of text, it is at least as valid to regard the notion "general English source text" as an artificial or even dubious one as it is to say that the Hansard data base inhabits some sort of "limbo" between sublanguages and "general English".

We will glance now at future research goals and directions, and then conclude. The underlying purpose of the experimentation presented here has been to test the efficacy of variant basic orientations in the dis-

crimination of English word senses. None of the approaches employed could ever be used, exactly as is, to actually process large volumes of text with the aim of automatically differentiating word senses. This is because of the necessity of hand labelling of concordances, and further, of the grand scale on which this labelling is required. In the background lies another difficulty, but one whose elimination is most probably a somewhat distant objective. This difficulty is that the content analysis itself had to be carried out by a human hand. While automation of this latter task does not seem a forlorn hope, obviating the need for hand labelling does appear the greatly more tractable task of the two at present.

What has to be said in this regard is simply that the replacement for such hand labelling is cluster analysis using those categories found to be optimal in experimentation of the sort carried out in the present study. The problem here is that the number of clustering algorithms is vast, and there does not seem to be a royal road through this terrain. Nonetheless, research directions in prosecution of this study are clear. Most immediately, the various combinations of DS1 and DS2 need to be investigated. Hopefully a local optimum, as it were, can be reached with reference to the perplexity scores attainable by these two methods as currently set up.

Once such an optimal category set is obtained, the proper type of cluster analysis, once identified, should permit the automatic labelling of concordance lines by word sense. That is, we assume that lexical de-

finitions are stipulative (Cf. Section I4 of Chapter Two). Then, granting the validity of our category set with reference to the corpus under analysis, it would appear preferable on several counts to allow a machine, rather than a human, to "stipulate" senses. The obvious advantage is that the time required for humans to label thousands of concordances, each with thousands of lines, is prohibitive. But a second advantage is that the mechanical labelling will be optimal with reference to the category set, whereas the human will make mistakes.

A third advantage is that phenomena such as multiple categorization of a single token and typicality of a token qua exponent of one or another sense, can be captured via cluster analysis. A token midway between two clusters, for instance, might be considered to represent two senses simultaneously. And distance form the center of a cluster might be used as a metric of typicality. It could be that predictions based on sense categorization are reliable in proportion to the typicality enjoyed by a token as a representative of its sense.

In the medium and long ranges, two tasks which suggest themselves are the application to new "worlds", or at least new domains, of the sort of thematic analysis which was undertaken in creating the DS2 category set. One question which looms is the extent of overlap among content-analytic category sets for different domains, where each set is developed with the sole purpose of maximizing predictivity over the particular text type involved. A large degree of commonality among such sets would be

surprising, but it would also indicate a basis for the automatization of thematic analysis mentioned earlier as a rather remote objective.

In a word, the ultimate aim of a program of investigation such as the one indicated by our study is the automatization of content analysis. Related goals include the discrimination by computer of domains or "worlds", and hopefully of more delicate thematic units.

It goes without saying that at every stage of the research program just sketched out, attempts should be made to apply knowledge obtained to real-world tasks involving natural language processing. It is hard to think of a sub-discipline within this area of study which would not stand to benefit from a truly workable system of English word sense discrimination. Related fields such as information retrieval and character recognition might also find such a procedure of use. Whether a system of this description can in fact be realized, however, is a question to which we do not yet have the answer.

Summarizing this study, in Chapter One we reviewed efforts to date aimed at computationally discriminating the senses of English words. We found that a morphological and dependency-relational approach to the task had been developed by Debili (1982), dictionary-based methods suggested by Amsler (1980) and by Amsler and Walker (1985), syntax-driven efforts at disambiguation conceived by Gross (1985) and by Sinclair (1985), and

a content-analytic system of English word sense differentiation originated and instantiated by Kelly and Stone (1975).

In Chapter Two, theoretical background was presented in the form, first, of a discussion of points of Wittgenstein's later philosophy which we consider to explain and justify the sort of research reported here; the notions of meaning as use, forms of life and language games, vagueness, stipulative definitions, and family resemblance were examined. A modest elaboration of the "form of life" concept was undertaken and was informed by Goodman's (1978) conception of "worlds". Second, the content-analytic approach to text analysis was explicated in the Wittgensteinian terms introduced earlier. Third, an evaluation was carried out of the syntax-based models of Gross and Sinclair; they were found to be potentially useful, but to require supplementation via content analysis.

The current chapter presented the aim, materials, procedure, and results of an experiment contrasting three different approaches to the identification of English word senses. The aim was to investigate the differential performance of a dictionary-based model (DG) of automatic word sense discrimination; of a model whose categories are derived by directly partitioning the corpus submitted for analysis (DS1); and of a partially content-analytic, partially local-collocational method (DS2). The materials included a 20-25-million-word source text consisting of official proceedings of the House of Commons of Canada; a roughly 2000-line concordance derived from this source for each of five randomly

selected test words; and, for each of the three experimental methods under examination, five sets of categories of words and expressions. Each such set represented a (potential) classification of the graphemic environments, with respect to their concordances, of the five test words. A random partition was induced on each test-word concordance, separating it into files of sizes 500 and about 1500 respectively. The procedure followed in applying these materials to each of the fifteen experimental tasks consisted, first, in the automatic construction, with respect to the 1500-line training concordance, of a decision tree over the 81 categories. Second, the resulting tree was employed to categorize the appropriate 500-line test-data file. Third, a value in terms of the information-theoretic measure "perplexity" (Bahl et. al. (1983)) was output for each test data file. Our results showed an extremely poor performance for DG, and better, nearly equivalent scores for DS1 and DS2. The combination of the DS1 and DS2 categories, carried out on two of five test data sets, yielded results nearly exactly equal, in each case, to the minimum of DS1, DS2.

The results were discussed and accounted for in terms, first, of the complementary strengths of the local-structure-oriented DS1 and of the mostly content-analytic DS2, and second, of the lack of either of these advantages on the part of DG. The value of domain-general classification systems to sense discrimination efforts applied to corpora embodying a single "world" was argued to be confined to strictly syntax-oriented or "lexical-functional" approaches. However, even these approaches were considered to require supplementation via "world"- or domain-particular

thematic categories derived, at least for the present, through the careful content analysis of complete concordances for a number of sample word types on the order of 100 for a multi-million-token corpus.

Future research was reported to be oriented, in the short term, toward further exploration of the possible combinations of DS1 and DS2; in the medium term, toward replacing hand labelling of test and training data with induction of sense categories by means of cluster analysis; and in the long term, toward achieving the automation of content-analytic category construction.

FOOTNOTES TO CHAPTER THREE

1. We make no attempt to define this nebulous but highly useful concept. As good a listing as any, for illustrative purposes, may be found in the appendices to Bradley (1983), which, incidentally, is itself a very good example of the utility within language study of the notion "function word". As to why exactly 41 content-word categories were chosen, and not some other number of them, we wished to obtain a total of 81 categories, in order to achieve parity with method DS2, to be described below. Since method DS1 had two types of category, it was decided that each of the two types would make up half of the DS1 category set. Hence there were 41 of the type currently being described, and 40 of the kind to be accounted for below.

2. See Byrd (1986). The author wishes to express his gratitude to Mr. R. Byrd for his unfailing generosity and patience in helping him to make optimal use of the TUPLES package. The capabilities provided by TUPLES were of great value in our effort to process millions of words of text.

3. Of course, the word may have additional senses within other parts of speech, but these will be discarded. So a better formulation of the first two criteria is that there must be 2000 instances of use of a word within the same part of speech, ranging over at least three different senses. Also note that the reason we require that the senses fall within a single part of speech is that we are not in the business of part-of-speech labelling, but rather of sense selection. (See footnote 15 of Chapter One.)

4. We were restricted to five words by the highly labor-intensive nature of the experimental task. Notably, processing the five concordances entailed hand-labelling 10,000 lines of concordance as to which of a set of senses was exemplified by the node word of each line.

5. All five test words were considered in their nominal forms. Nonetheless, each has a verbal form, and the question arises as to why none of these was used. The answer is simply that in each case, the verbal use failed one or more of the criteria for selection enumerated above. On a related point, it is not surprising that no word which is exclusively a verb figures among the five test items. Professor M. Chodorow reports (personal communication) that in Webster's New Collegiate Dictionary, there are about 40,000 noun listings and around 9,500 verb listings. If we assume, as appears reasonable, that there is a good deal of overlap between nouns and verbs--i.e. that it is often the case that a single spelling is listed both as a noun and as a verb--then it is to be expected that a random list of five words would fail to include any word that is uniquely a verb. While it is of course possible that the average "spread" of verb senses is narrower in the Hansards than that of the average noun,

we cannot take the results of this random selection of test words as evidence for this claim.

6. We stress the word "stipulation". We found, as was to be expected given Chapter Two, Section I4, that no clear boundaries could be established between senses; that one could augment or diminish the number of senses at will by crediting or overlooking particular distinctions; and that on numerous occasions a single occurrence of the node word could be argued plausibly to exhibit several senses at once.

7. Actually, in addition to words and expressions, quite a number of other items were sought and listed. For details, see below.

8. To redeem a debt incurred earlier in this chapter, we see why DG and DS1 needed to have 81 categories each. It is because it had simply turned out that 61 "content-analytic" categories were deemed adequate for the mapping out of the 100 sample types, and the decision was made to limit to 20 the number of structural categories of DS2, whose total number of categories thus became 81. Therefore the requirement of parity in the number of categories over the three methods dictated that DG and DS1 also have 81 classes.

9. A complete listing of the 61 "content-analytic" or "thematic" categories, with illustrative entries for each category, is available from the author on request, at IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, N.Y. 10598.

10. The program instantiating these conditions was set up in such a manner that no self-reference was possible, either in a single step or indirectly; in this way the danger of infinite recursion, as an unwanted byproduct of the ability of one category to reference another, was eliminated.

11. See Scott (1986). The author wishes to express his deep gratitude to Mr. M. Scott of Columbia University for his extremely valuable assistance during his Summer 1986 stay at the IBM T.J. Watson Research Center. Not only did Mr. Scott write and meticulously document the software package referred to above, and the programs mentioned in footnote 12 below, but he gave of himself unstintingly, in discussions with the author concerning the mathematical and computational aspects of the present research, and related topics, all the while maintaining an admirable sense of balance and humor.

12. An exposition of information theory which presupposes no calculus is Abramson (1963). One which is more popular, and requires much less mathematical background, is Pierce (1961). For a classic presentation which is addressed to the mathematician see Jelinek (1968).

13. The algorithm itself was originated by Dr. R.L. Mercer of the IBM T.J. Watson Research Center, and was instantiated in a program written by Mr. P. Brown of the same institution, and adapted for the author's use by Mr. M. Scott of Columbia University, during his Summer 1986

stay at the Watson Research Center. The author wishes to thank these three individuals for their extremely generous help in connection with the utilization of the decision-tree algorithm and program, without which the experiment reported here could not have been carried out in its present form.

14. On the perplexity measure, see Bahl et. al. (1983).

15. E.g. Chomsky (1965, p. 5).

16. Notice that there is an "escape hatch" in this citation, corresponding to the words "form the basis for". That is, it might be the case that the claim being advanced is nothing more than that a fully disambiguated dictionary could form one part of a system meant to disambiguate any sort of text--a particularly important ("basic") part. But to the extent that such accompanying portions of the putative system are not specified, or even suggested--and they are not--this "escape hatch" interpretation of the excerpted sentence flirts with vacuity. Certainly there would be no such thing as testing the claim on this construction. If we ignore this possibility, however, and take the statement at face value--that is, where "form the basis for" means something like "constitutes"--then a testable, and in fact quite interesting proposal is being put forward. It is to be hoped that one day we will be able to judge the merits of an instantitated system along these lines.

17. There is no particular correlation between paucity of available text and sublanguage status of the language employed within such text. Damerau (1971) had access to 7 million words of patent text, of which he processed some 5 million (Damerau (1965), p. 71). Nonetheless, he counted only 8510 types in the corpus (Senko and Damerau (1966, p. 2)), and was able to come up with a rather compact set of grammatical rules to encompass the syntax of this highly stereotyped language source.

APPENDIX I: LIST OF SAMPLE WORDS

The following is a list of the 100 words whose concordances over the 20-25-million-word Hansard corpus were examined closely as a basis for the generation of the content categories of method DS2.

access
accept
acting
accident
address
afternoon
agency
anticipate
appreciation
April
bargaining
bear
billion
caucus
cheaper
Christmas
collect
collected
co-operate
constitute
contingency
convince
cope
correctly
damage
dated
dates
debt
decrease
deep
define
definite
deliver
depending
depends
deserves
distribution
doubt
due
either
exercise
facing
fall
federal-provincial
feet
financing
force
fund
hand
hour
knows
marketing
movement
person
redistribution
refusal
regional
rejected
rely
research
resources
sat
satisfy
scale
scrutiny
seriousness
sets
ship
shipping
showing
sign
sincere
spokesmen
stations
suddenly
suffered
sufficiently
suggested
supplied
supplier
surrounding
sympathy
tabling
taxes
tells
tendency
test
text
thanks
threaten
time
transfer
tribute
UN
unacceptable
uncertainty
urged
uses
variety
voice

APPENDIX II: DG, DS1, DS2 CATEGORIES FOR WORD "INTEREST"

METHOD DG: CATEGORIES FOR THE WORD "INTEREST"

The DG categories for "interest" are listed below.

AGRICULT	COMMUNIC	ELECTRIC	GEOMETRY
ANATOMY	COOKERY	ELECTRON	GOLF
ARCHITEC	CRICKET	ENGINEER	HAIRDRES
ART	CALENDAR	ENGLISH	HISTORY
ASTRONOM	CARPENTR	ENTERTAI	HOROLOGY
AUTOMOTI	CHESS	EPITHET	HORSERAC
BANKING	CLOTHING	EQUESTRI	HOUSEHOL
BEVERAGE	COLOR	FIREARM	HUNTING
BIOLOGY	COMMERCE	FISHING	INSTITUT
BOTANY	CURRENCI	FOOD	INSURANC
BOXING	DANCE	FUR	JEWELRY
BREED	DEGREE/A	GAMBLING	LABOR
BRICKLAY	DRAMA	GAME	LAW
BUILDING	ECONOMIC	GENERALI	LEXICOGR
BUSINESS	EDUCATIO	GEOLOGY	LINGUIST
MARKETIN		NUMBER	
MATHEMAT		NUMISMAT	
MEASURE/		OBSTETRI	
MEDICINE		OCCULT	
METEOROL		OCEANOGR	
MILITARY		PATHOLOG	
MORPHOLO		PHONOLOG	
MUSIC		POKER	
NAUTICAL			
NAVY			

METHOD DS1: CATEGORIES FOR WORD "INTEREST"

1. CONTENT WORD CATEGORIES (WINDOW = ENTIRE CONCORDANCE LINE)

DEBT
ECONOMY
ORDER
SHOWN
SPEAKER
HIGH
CANADIAN
PAYMENTS
SAFETY
COUNTRY
MR
BEST
GOOD
INCOME
MADE
THINK
CONCERN
TAXES
THROUGH
LOANS
MEMBERS
BEEN
PAY
MATTER
GETTING
PAID
MORTGAGE
CANADIAN
MONEY
RATES
HOUSE
HON
RATE
CENT
PEOPLE

2. FUNCTION AND CONTENT WORDS (I.E. UNRESTRICTED)--WINDOW N+-2

PUBLIC
GOVERNMENT
PER
THEIR
MINISTER
ALL
GREAT
MORE
AN
WHICH
AT
CANADA
ARE
BY
HAS
WE
AS
NATIONAL
CONFLICT
HIS
WILL
FOR
HAVE
WOULD
OR
NOT
PUBLIC
THIS
I
WITH
IT
ON
BE
IS
A
THAT
TO
AND
IN
OF
THE

METHOD DS2: CONTENT CATEGORIES FOR WORD "INTEREST"

The content categories of method DS2 do not change from word to word, but rather remain constant throughout the analysis of all items of the Hansard corpus vocabulary.

APPROX
ARGUDROP
BAD
BEGIN/END
BODY_OF_WATER
CAPIDROP
CHANGE
CO_OPERA
CONTROVERSY
CRIMDUMM
DECIDE
DECRDUMM
DOCUMENT
DOCUVERB
EDUCDUMM
ENERGY
EVENT
EXPRESS_FEELING
FEEL
GETDUMM
GOOD
GOODS
GOVT_BODY
GREATLY
HAPPEN
HELP
INCRDUMM
INFORMATION
INSTITUTIONS
ISSUE
MATEDROP
MILITARY/FORCE
MONEY
MONEYBAD
MONEYGOOD
MUCH/GREAT
NEG_CONNOTA
NON_NEGS
NUMBER
PARLIA_MOVES
PLACE
POLICY
POS,DUMM
POS.DUMM
POS_CONN
POWER
POWERFUL_PEOPLE
PRE_TIME
PRICE
RESPONSIBILITY
SUFFDUMM
THOUGHT/BELIEFS
TIME_PER
TIME_UNIT
TIME_WHEN
TRAFDROP
TRANSPORT
TWO,DUMM
UNITS
VOTERS
WORKERS

METHOD DS2: STRUCTURAL CATEGORIES FOR WORD "INTEREST"

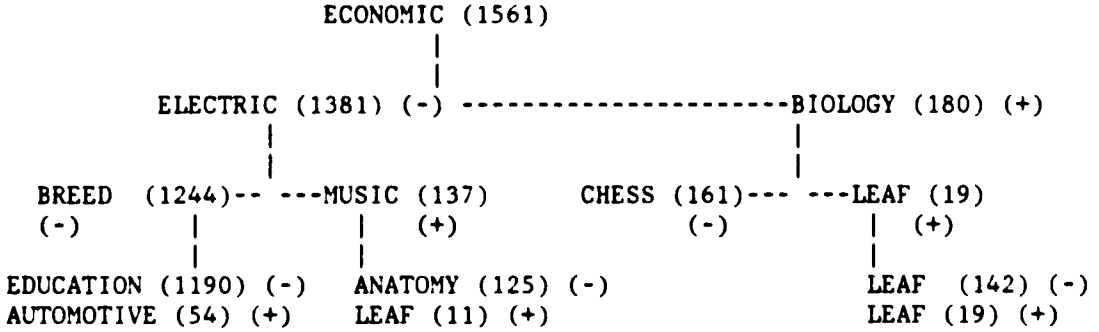
W_R refers to the bigram "interest W"; R_W refers to "W interest"; example--FOR_R indicates "interest for"

Note that trigrams of form W1 node W2 are often captured by bigram categories such as these. For instance, the expression "in the interest of" is the combination of THE_L and OF_R.

FOR_R
BEST_L
HIS_L
MORTGAGE_L
RATES_R
OR_R
THE_R
AND_L
GREAT_L
NATIONAL_L
WITH_L
THAT_R
IN_R
PUBLIC_L
TO_R
ON_R
AND_R
OF_L
THE_L
OF_R

APPENDIX III: SAMPLE DECISION TREES FOR DG, DS1, DS2

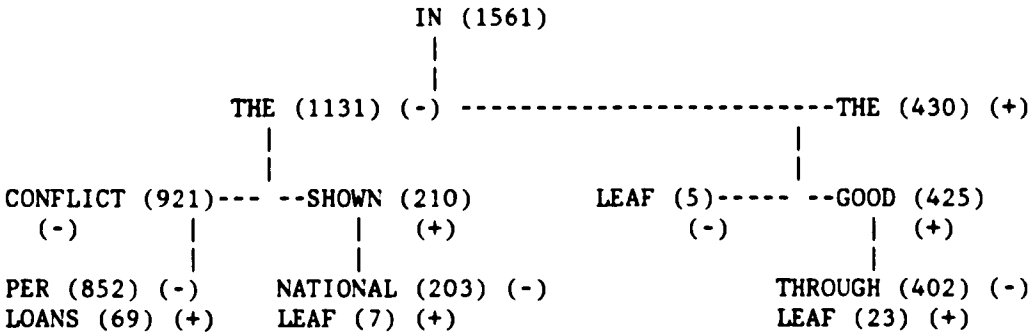
TOP PORTION OF DG DECISION TREE FOR TEST WORD "INTEREST"



It is hard to see what any of the categories above, with the exception of "ECONOMIC", which is at the root, have to do with "interest" as "curiosity, concern" (Sense 1), as "stake" (Sense 2), as "purpose--'in the interest of'" (Sense 3), or as "payment for borrowing money" (Sense 4). Moreover, from the perspective of these senses, it seems unlikely that the words which make up each of the categories above form coherent groupings. (Again, the class "ECONOMIC" constitutes the exception to this rule.) Examples are: 1. "ELECTRIC" consists of "current", "off", "on", and "point"; 2. "BIOLOGY" includes the single word "order"; 3. "BREED" is just the word "by"; 4. "MUSIC" comprises "air", "low", "note", "number", "part", "set", "subject", and "time".

Finally, just as for DS1 and DS2, only the uppermost portion of the tree is shown; the full tree has 19 nodes.

TOP PORTION OF DS1 DECISION TREE FOR TEST WORD "INTEREST"

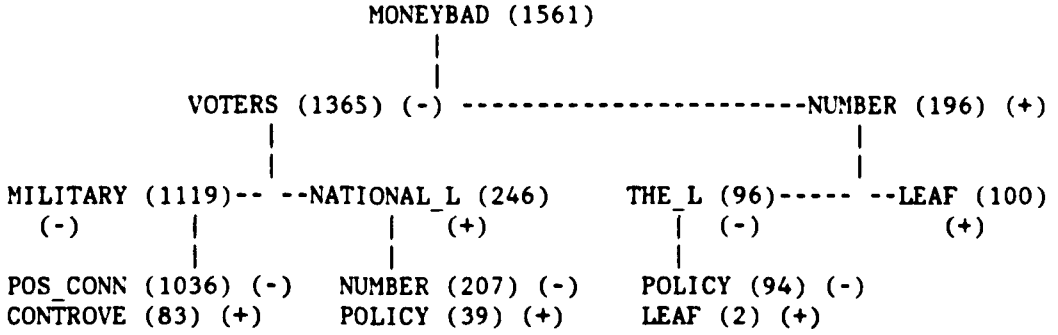


Of the node labels above, some must occur in the window node+-2, and others may occur anywhere in the concordance line. The former include: "IN", "THE", "CONFLICT", and "NATIONAL"; the latter are: "SHOWN", "GOOD", "PER", "THROUGH", and "LOANS". "-".

Interpreting some of the paths through the tree, the path -IN, -THE, +LOANS expresses the predictivity of the collocation of "interest" and "loans", in the absence of "in" and "the". It turns out that the 69 concordance lines with this profile display the sense "interest on money" 97% of the time. To take a second sample path, +IN, +THE, +GOOD ends in a leaf whose 23 concordance lines all represent the sense, "in the interest of = for the purpose of". Expressions such as "in the interest of the maintenance of good order" and "in the interest of good order" account for this categorization. Finally, the path -IN, +THE, +SHOWN ends in a leaf with 7 concordance lines, all of which exemplify the sense "interest as curiosity or concern". In the expression "to show and interest in something", it is the latter sense which is normally used.

As in the DS2 tree displayed above, we show only the very top part of a large tree--in this case, one with 36 nodes.

TOP PORTION OF DS2 DECISION TREE FOR TEST WORD "INTEREST"



Each node of the above tree is associated with a category, a number of concordance lines (in parentheses), and a probability distribution over sense numbers (not shown). Left children are rated "-" for the parent's category, right children rated "+". That is, for instance, the left child of the root node corresponds to the 1365 lines of the training concordance of the word "interest" which DO NOT feature an exponent of the category "MONEYBAD". Similarly, the right child of the root node symbolizes all those lines which DO include a word or expression of the "MONEYBAD" category.

Lexical events labelled "MONEYBAD" are associated with financial circumstances considered unwelcome within the Hansard "world". Examples are an increase in the cost of living, a decrease in gross national product, and bankruptcy. "VOTERS" are all those among the Members' constituents who are not notably wealthy or powerful; this group contrasts with "POWERFUL_PEOPLE". The notation X_L, Y_R indicates that the words X, Y are to the left or right, respectively, of the node. So "NATIONAL_L" indicates the expression "national interest".

"POS_CONN" is the class of words and expressions possessing positive connotations with respect to the Hansard world. Instances are: "amiable", "down to earth", and "operating at full capacity". "CONTROVE" lists the vocabulary of debate: "ambiguous", "approbation", "contend". "POLICY" covers locutions employed in the context of establishing the government's course of action on particular issues: "affairs of state", "controls", "guideline". War- and police-related matters are discussed in terms of "MILITARY" language, and "NUMBER" consists of numerical expressions.

Since this tree contains 43 nodes, only a small portion of it is presented, for purposes of illustration.

APPENDIX IV: SAMPLE CONTINGENCY TABLES--THE WORD "TERMS"

Contingency tables are provided below for the word "terms", as an indication of the way our data behave when analyzed via this tool of data presentation.

A table is to be interpreted as follows: The correct sense labels are to be found along the vertical axis. By following any such label number--say label 1--horizontally across its row, one learns how many concordance lines were considered MOST LIKELY to be each of the four different senses, and how many lines were not categorized at all.

Contingency Table For Method DG, Word "Terms"

		Labels Assigned by Method DG				
		1	2	3	4	No Label
Actual Labels	1	92 (57%)	18 (11%)	7 (4%)	7 (4%)	28 (24%)
	2	44 (37%)	39 (33%)	5 (4%)	7 (6%)	24 (20%)
	3	43 (28%)	8 (6%)	93 (60%)	2 (1%)	8 (15%)
	4	23 (35%)	11 (16%)	14 (21%)	9 (14%)	9 (14%)

Contingency Table For Method DS1, Word "Terms"

Actual Labels	Labels Assigned by Method DS1									
		1		2		3		4		No Label	
	1	66	(41%)	68	(42%)	19	(12%)	0	(0%)	8	(5%)
	2	15	(13%)	100	(84%)	4	(13%)	0	(0%)	0	(0%)
	3	5	(3%)	8	(5%)	140	(91%)	1	(1%)	0	(0%)
	4	8	(12%)	12	(18%)	34	(52%)	6	(9%)	6	(9%)

Contingency Table For Method DS1, Word "Terms"

Actual Labels	Labels Assigned by Method DS1									
		1		2		3		4		No Label	
	1	128	(80%)	20	(12%)	1	(1%)	12	(7%)	0	(0%)
	2	33	(37%)	73	(61%)	30	(3%)	9	(8%)	1	(1%)
	3	19	(12%)	1	(1%)	113	(73%)	21	(14%)	0	(0%)
	4	17	(26%)	5	(7%)	6	(9%)	38	(58%)	0	(0%)

BIBLIOGRAPHY

Aarts, J. and W. Meijs (eds). 1984. Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research. Amsterdam: Rodopi.

Abramson, N. 1963. Information Theory and Coding. New York: McGraw-Hill.

Ager, D.E., F.E. Knowles and J. Smith (eds). 1979. Advances in Literary and Linguistic Research. Birmingham, England: AMLC.

Amsler, R.A. 1980. The Structure of the Mirriam-Webster Pocket Dictionary. Ph.D. Dissertation, University of Texas at Austin. TR-164.

Amsler, R.A. and D. Walker. 1985. Knowledge resource tools. Presentation at IBM Corporate Technical Institute, Thornwood,NY. December, 1985.

Atwell, E.S. 1983. Constituent-likelihood grammar. Newsletter of the International Computer Archive of Modern English (ICAME News) 7. Pp. 34-66.

Ayer, A.J. 1985. Wittgenstein. Chicago: University of Chicago Press.

Bahl, L., F. Jelinek, and R.L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. IEEE Transactions

on Pattern Analysis and Machine Intelligence. Vol. PAMI-5.2. Pp. 179-190.

Baker, G.P. and P.M.S. Hacker. 1985a. An Analytical Commentary on Wittgenstein's 'Philosophical Investigations'. Volume 1. Chicago: University of Chicago Press.

Baker, G.P. and P.M.S. Hacker. 1985b. Wittgenstein: Meaning and Understanding. Volume 1. Chicago: University of Chicago Press.

Baldi, P. (ed.) 1984. Papers form the XIIth Linguistic Symposium on Romance Languages, University Park, April 1-3, 1982. Amsterdam: John Benjamins.

Bar-Hillel, J. 1964. Language and Information. Reading: Addison-Wesley.

Bazell, C.E., J.C. Catford, M.A.K. Halliday, and R.H. Robins. 1966. In Memory of J.R. Firth. London: Longman.

Beale, A.D. 1985a. Grammatical analysis by computer of the Lancaster-Oslo/Bergen (LOB) corpus of British English texts. Ms., U. of Lancaster, England.

Beale, A.D. 1985b. A probabilistic approach to grammatical analysis of written English by computer. Ms., U. of Lancaster, England.

Bell, R.H. 1969. Wittgenstein and descriptive theory. Religious Studies. 5.

Block, I. (ed.) 1983. Perspectives on the Philosophy of Wittgenstein. Cambridge, Mass.: MIT Press.

Botha, R. P. 1982. On 'the Galilean style' of linguistic inquiry.

Lingua. 58. Pp. 1-50.

Bradley, D. 1983. Computational distinctions of vocabulary type. Bloomington: Indiana University Linguistics Club.

Budd, R.W., R.K. Thorp, L. Donohew. 1967. Content Analysis of Communications. New York: Macmillan.

Byrd, R. 1986. The TUPLES text analysis system: description and users' guide. IBM Research Report. Yorktown Heights, N.Y.: IBM T.J. Watson Research Center.

Byrd, R.J., J.L. Klavans, M. Aranoff, F. Anshen. 1986. Computer methods for morphological analysis. In Proceedings of the 24th Meeting of the Association for Computational Linguistics, Columbia University. Pp. 120-127.

Chodorow, M.S., R.J. Byrd, G.E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In Proceedings of the 23rd Meeting of the Association for Computational Linguistics, University of Chicago. Pp. 299-304.

Chomsky, N. 1965. Aspects of the Theory of Syntax. Cambridge: MIT Press.

Chomsky, N. 1980. Rules and Representations. New York: Columbia U. Press.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20.1. Pp. 37-46.

Copleston, F.C. 1974. Religion and Philosophy. Dublin: Gill and

Macmillian.

Damerau, F.J. 1971. Markov Models and Linguistic Theory: An Experimental Study of a Model for English. The Hague: Mouton.

de Saussure, F. 1969. Cours de linguistique generale. Paris: Payot.

Debili, F. 1982. Analyse syntaxico-semantique fondee sur une acquisition automatique de relations lexicales-semantiques. These de Doctorat d'Etat, Universite Paris XI. No. d'ordre 2541.

Evens, M.W., B.E. Litowitz, J.A. Markowitz, R.N. Smith, O. Werner. 1983. Lexical-Semantic Relations: A Comparative Survey. Edmonton, Canada: Linguistic Research, Inc.

Fiengo, R. 1980. Surface Structure. Cambridge, Mass.: Harvard U. Press.

Finch, H.L. 1977. Wittgenstein--The Later Philosophy: An Exposition of the 'Philosophical Investigations'. Atlantic Highlands, N.J.: Humanities Press.

Firth, J.R. 1957. Papers in Linguistics 1934-1951. London: Oxford University Press.

Firth, J.R. 1966. The Tongues of Men, and Speech. London: Oxford University Press.

Forney, G.D. Jr. 1973. The Viterbi algorithm. Proceedings of the IEEE. Volume LXI. Pp. 268-278.

Garside, R. and F. Leech. A probabilistic parser. Ms., U. of Lancaster, England.

Gentner, D. 1975. Evidence for the psychological reality of semantic components: the verbs of possession. In Norman and

Rumelhart 1975, pp. 211-246.

Goodman, N. 1978. Words, works, worlds. In Ways of Worldmaking. Indianapolis: Hackett Publishing Company

Granger, R.H., Jr. 1977. FOUL-UP: A Program that figures out meanings of words from context. IJCAI-77. PP. 172-178.

Gross, M. 1979. On the failure of generative grammar. Language. 55. Pp. 859-885.

Gross, M. 1984. A Linguistic environment for comparative Romance syntax. In Baldi (ed.) 1984. Pp. 373-446.

Gross, M. 1985. Projecting the lexicon-grammar on texts. Paper delivered at the 1st International Roman Jakobson Conference, New York University. August 1985.

Gross, M. and D. Tremblay. 1985. Etude du Contenu d'une Banque Terminologique. Technical Report. Paris: Laboratoire d'Automatique Documentaire et Linguistique.

Hartmann, R.R.K. (ed.) 1984 LEXeter '83 Proceedings: Pepers from the International Coference on Lexicography at Exeter, 9-12 September 1983. Tubingen: Max Niemeyer Verlag

Hayes, P.J. 1975. A process to implement some word-sense disambiguations. Ms. U. of Rochester.

Hirschman, L., R. Grishman and N. Sager. 1975. Grammatically-based automatic word class formation. Information Processing and Retrieval. 11. Pp. 39-57.

Jelinek, F. 1968. Probabilistic Information Theory. New York: McGraw-Hill.

Jones, S. and J. McH. Sinclair. 1974. English lexical collocations:

a study in computational linguistics. Cahiers de lexicologie. 24.1. Pp. 15-61.

Katz, J.J. 1972. Semantic Theory. New York: Harper and Row.

Kelly, E.F. and P.J. Stone. 1975. Computer Recognition of English Word Senses. Amsterdam:North-Holland.

King, M. (ed.) 1983. Parsing Natural Language. London: Academic Press Hartmann (ed.) 1984 pp. 101-108

Kucera, H. and W.N. Francis. 1967 Computational Analysis of Present-Day American English. Providence: Brown University Press.

Landau, B. and Gleitman, L.R. 1985. Language and Experience: Evidence from the Blind Child. Cambridge, Mass.: Harvard University Press.

Langendoen, D.T. 1968. The London School of Linguistics: A Study of the Linguistic Theories of B. Malinowski and J.R. Firth. Research Monograph No. 46. Cambridge: MIT Press.

Longman Group Limited (eds.). 1978. Longman Dictionary of Contemporary English. Burnt Mill, England: Longman.

Lucassen, J.M. 1983. Discovering phonemic base forms automatically: an information theoretic approach. IBM Research Report RC 9833. Yorktown Heights, N.Y.: IBM T.J. Watson Research Center.

Lyons, J. 1963. Structural Semantics: An Analysis of Part of the Vocabulary of Plato. Publications of the Philological Society. XX. Oxford: Basil Blackwell.

Lyons, J. 1977. Semantics. Volume 2. Cambridge: Cambridge University

Press.

Meisel, W.S. 1972. Computer-Oriented Approaches to Pattern Recognition. New York: Academic Press.

Mel'chuk, I.A. 1984. Dictionnaire explicatif et combinatoire du francais contemporain: recherches lexico-semantiques I. Montreal: Les Presses de l'Universite de Montreal.

Merriam, G. and C. (eds.). 1963. Webster's Seventh New Collegiate Dictionary. Springfield, Massachusetts.

Merriam, G. and C. (eds.). 1971. The New Merriam-Webster Pocket Dictionary. New York: Pocket Books.

Mitchell, T.F. 1971. Linguistic "goings on": collocations and other lexical matters arising on the syntagmatic record. Archivum Linguisticum. 2. Pp. 35-69.

Moyne, J.A. 1980. Language use: a performance model. International Journal of Computer and Information Sciences. 9.6. Pp. 459-481.

Moyne, J.A. 1982. Human Information Processing: Language Comprehension--A Report to the National Science Foundation. Flushing,NY: Queens College Press.

Moyne, J.A. 1985. Understanding Language: Man or Machine. New York: Plenum Press

Munroe,A. 1975. Linguistic theory and the LNR representation. In Norman and Rumelhart 1975, pp. 88-113.

Norman, D.A. and D.E. Rumelhart. 1975. Explorations in Cognition. San Francisco: W.H. Freeman.

Palmer, F.R. 1968. Selected Papers of J.R. Firth 1952-1959.

Bloomington: Indiana University Press.

Phillips, D.Z. 1983. Wittgenstein's full stop. In Block (ed.) 1983.

Pierce, J.R. 1961. Symbols, Signals and Noise: The Nature and Process of Communication. New York: Harper and Row.

Pitcher, G. 1964. The Philosophy of Wittgenstein. Englewood Cliffs: Prentice-Hall.

Reed, A. 1977. CLOC: A collocation package. Association for Literary and Linguistic Computation Bulletin. 5. Pp. 168-173.

Reed, A. and J.L. Schonfelder. 1979. CLOC: A General-purpose concordance and collocations generator. In Ager et. al. (eds.) 1979. Pp. 59-72.

Reed, A. 1984. Anatomy of a text analysis package. Comput. Lang. 9.2 Pp. 89-96.

Renouf, A. 1984. Corpus development at Birmingham University. In Aarts and Meijs (eds.) 1984.

Renouf, A. 1985b. A Corpus of spoken interaction. Newsletter of the International Computer Archive of Modern English (ICAME). 9. Pp. 12-13.

Ritchie, G. 1983. Semantics in parsing. In King (ed.) 1983, pp. 199-218.

Rosengren, K.E. (ed.). 1981. Advances in Content Analysis. Volume 9, Sage Annual Reviews of Communication Research. Beverly Hills: Sage Publications.

Rumelhart, D.E. and J.A. Levin. 1975. A Language comprehension system. In Norman and Rumelhart 1975, pp. 179-208.

Scott, M. 1986. The MKQTN/QTREE concordance-processing system. Ms.,

IBM T.J. Watson Research Center. August, 1986.

Seal, B.D. 1981. In Search of Significant Collocation. M.A. Thesis, UCLA.

Senko, M.E. and F.J. Damerau. 1966. Full text search terms for transistor and nonlinear conductor systems. IBM Research Paper RC-1692. Yorktown Heights, N.Y.: IBM T.J. Watson Research Center.

Sinclair, J. McH. 1966. Beginning the study of lexis. In Bazell et. al. (eds.) 1966. Pp. 410-430.

Sinclair, J. McH. 1985a. Sense and structure in lexis. Ms., U. of Birmingham, England.

Sinclair, J. McH. 1985b. Basic computer processing of long texts. Ms., U. of Birmingham, England.

Small, S. 1980. Word Expert Parsing: A Theory of Distributed Word-Based Natural Language Understanding. Ph.D. Dissertation, University of Maryland. TR-954.

Small, S. 1983. Parsing as Co-operative Inference. Understanding through Memory Interactions. In King (ed.) 1983, pp. 247-276.

Stock, P. 1984. Polysemy. In Hartmann 1984, pp.131-140.

Ullmann, S. 1962. Semantics: An Introduction to the Science of Meaning. New York: Barnes and Noble.

Weiss, S.F. 1973. Learning to disambiguate. Information Storage and Retrieval. 9. Pp. 33-41.

Wilks, Y.A. 1972. Grammar, Meaning and the Machine Analysis of Language. London: Routledge and Kegan Paul.

Wittgenstein, L. 1961. Tractatus Logico-Philosophicus. Translated by

D.F. Pears and B.F. McGuinness. London: Routledge and Kegan Paul.

Wittgenstein, L. 1965a. The Blue and Brown Books: Preliminary Studies for the 'Philosophical Investigations'. New York: Harper and Row.

Wittgenstein, L. 1965b. Wittgenstein's lecture on ethics. Philosophical Review. 74.

Wittgenstein, L. 1968. Philosophical Investigations. Translated by G.E.M. Anscombe. Oxford: Basil Blackwell.