RAOUL N. SMITH - EDWARD MAXWELL

AN ENGLISH DICTIONARY FOR COMPUTERIZED SYNTACTIC AND SEMANTIC PROCESSING SYSTEMS

1. INTRODUCTION

R. F. SIMMONS (1970) and M. PAČAK and A. W. PRATT (1971) point out that no computerized system using natural language either as part of the processor or as the object processed and having a syntactico-semantic component has a lexicon of more than a few hundred items (except for the SNOP' s medical lexicon). It is obvious from the lack of success of large-scale computerized systems using natural language data that better solutions will be reached if these systems have a large lexicon as an integral component. Our purpose is to build a large scale dictionary[1] of English which will incorporate important recent research into language structure and which will have the potential of being used either as part of a computerized natural language-using system or as a large data base, itself a source for further syntactico-semantic studies.

There are a number of specific problems that anyone who constructs a large-scale computerized dictionary must resolve. First, as discussed in R. N. SMITH (1972) and P. B. GOVE (1972), a computerized dictionary must incorporate additional types of data than is available in standard dictionaries. Since standard dictionaries and some of their computerized counterparts define words in terms of other words, they are of necessity circular. In addition, the efficiency of any system will depend on the size and form of the dictionary. Any usable large-scale dictionary of English probably would have to contain at least 200,000 entries (including inflected forms).

---

[1] We distinguish lexicon and dictionary by considering an entry in the latter as being the same information as the corresponding entry in the former but with added definition.

If each entry is defined as in a standard dictionary with, say, 20 words used in the definition then there must be storage for 4,000,000 words. In addition if, as has been proposed in N. CHOMSKY (1965), each entry has syntactico-semantic features attached we will encounter a similar problem: entries probably need on the average 20 features to specify them. Finally, when words are arbitrarily stored in computer systems, with pointers directing the search from word to word (cf. M. R. QUILLIAN, 1968), the search algorithm can be long.

With all of these problems in mind, we have defined a theoretical model which we expect will eliminate or substantially reduce these very real limitations of computerized dictionaries discussed above. The purpose of our research is to implement the scheme so that it may be used in artificial intelligence systems; as a data base for computer assisted instruction systems (e.g. PLATO), and as a tool for lexical testing (cf. J. OLNEY, D. RAMSEY, 1972) and information retrieval (e.g. cf. C. SALTON, 1971; W. A. WOODS, 1972).

2. LEXICAL INFORMATION

Until recently much of the interesting research in lexicology has been carried out in the Soviet Union. The Soviets have long been concerned with automated language processing and attribute the lack of success at this task to the lack of a sophisticated lexical theory. Dictionaries are quite inadequate in giving us insight into the nature of words. There is no way, for example, that one could learn a language using a dictionary. In addition, definitions in dictionaries are circular: every word is defined in terms of every other word (actually approximately 50 % of the vocabulary appears in the definitions (JOHN OLNEY, personal communication)).

Some of the most innovative research in Soviet lexicology has been carried out by Žolkovsky, Mel'čuk and Apresjan (cf. A. K. ŽOLKOVSKY and I. A. MEL'ČUK, 1970, YU. D. APRESJAN, 1967, and YU. D. APRESJAN, I. A. MEL'ČUK and A. D. ŽOLKOVSKY, 1969). Initially, they felt that the detailed syntactic properties of a word composed its meaning – in a structural rather than substantive sense. Their approach was first to classify words using grammatical criteria; for example, Apresjan classified verbs as being able to undergo the passive transformation or as being able or unable to take a complementary infinitive, accusative objects or locative adverbial phrases.

Their theory is essentially a structuralist one. In one of their studies they proposed a revision of the notion " word field ". That is, their lexico-structural analysis begins with an enumeration of the phrase types of a language, revealed by syntactic analysis; an indication of the frequency of each of the structural patterns; and finally an enumeration of the word meanings found in each pattern.

In YU. D. APRESJAN, I. A. MEL'ČUK and A. D. ŽOLKOVSKY (1969) and elsewhere they propose that a dictionary which displays "the process of text generation as an integral succession of steps" be constructed. They state that the dictionary should be based on the following principle:

> ... it must be fully sufficient for a smooth, idiomatic and flexible expression of a given meaning; that is to say, it must display in an explicit and logical form whatever information may be necessary for the correct choice and usage of words and phrases to convey a given idea in a speech context.

The proposed dictionary is " combinatory " because " it is primarily intended to display the combinatorial properties of words. " It is " explanatory " because the syntactic government patterns are semantically interpreted with the goal of providing idiomatic expression of any given meaning.

The typical entry in their dictionary would have the following format:

a) Entry word
b) Morphological information
c) Definition
d) Syntactic potential of word
e) Regular lexical functions
f) Non-regular lexical functions
g) The " lexical universe " of entry
h) Examples
i) Phraseology (idiomatic expressions)
j) Discrimination of synonyms and near-synonyms.

Concerning the definition (c), they specify that they *not* be circular. They state that " if this requirement is met, all the definitions will in the long run be reduced to a small number of indefinable units of meaning (elementary meanings). " (This is the same goal as the UCLA lexicography project.) An example of this can be found in McCawley's

work with lexical atoms (J. McCawley, 1968). That is, *redden* has the semantic atoms " cause to come to be red. " It is important to note that the definition of a word should be an exact paraphrase of the word using these semantic atoms.

The notion of lexical functions is the principle innovation of their dictionary. Lexical function involves establishing relationships between words. Examples from Yu. D. Apresjan, I. A. Mel'čuk and A. D. Žolkovsky (1969) are the following:

(1) Syn-synonym
Syn (*to help*) – *to aid*
Syn (*to call*) – *to name*

(2) Conv-conversive
Conv (*to build*) – *to be built* (*by someone*)
Conv (*to contain*) – *to be contained* (*by something*)
Conv (*A precedes B*) – *B follows A*
Conv (*A beat B*) – *B loses to A*
Conv (*A sold B to C*) – *C bought B from A*

(3) Anti-antonym
Anti (*beautiful*) – *plain, ugly*
Anti (*before*) – *after*

(4) Gen – genus
Gen (*liquid*) – *substance*
Gen (*blue*) – *color*
Gen (*crawl*) – *move*

(5) S_o-noun coinciding with the verb
S_o (*to move*) – *movement*
S_o (*to be white*) – *whiteness*

(6) A_o-adjectives coinciding with the verb
A_o (*sun*) – *solar*
A_o (*time*) – *temporal*

(12) Sinstr-noun denoting instrument of word
Sinstr (*think*) – *brain*
Sinstr (*clap*) – *hands*

(15) Mult-noun denoting aggregate
Mult (*flowers*) – *bunch*
Mult (*sheep*) – *flock*

(18) Figur-standard figurative designation
Figur (*passion*) – *flame*
Figur (*misery*) – *abyss*

(31) Oper-verb connecting name of participant with action
oper (*support*) – *to lend*
oper (*defeat*) – *to suffer*

oper (*recession*) – *to experience*
(39) Fin-verb meaning " to cease ", " to stop "
Fin (*sound*) – *to die away*
Fin (*patience*) – *to lose*
(46) Son-verb denoting typical sound
Son (*lion*) – *to roar*
Son (*snake*) – *to hiss*

What they mean by information about the lexical universe of a word is " an informal description of a sufficiently broad *piece* of reality including the given situation as a constituent element. " For example, the lexical universe of *student* would include such lexical items as *books, classes, college, instructor, study, exam* and so on.

Finally, description of near synonyms would involve a listing of all words connected to a lexical item by connotations. Connotations involve, of course, literary and emotional overtones of words. A *terrorist* is, for example, a *guerrilla* whose cause we have emotional disagreement with.

Their notion of syntactic potential corresponds somewhat to Fillmore's case frames. That is, for Fillmore, a dictionary must specify the case potential of words. For example, in the sentences

(1) a) *John hit the ball with a bat.*
b) *The bat hit the ball.*
c) *John hit the window with the ball.*
d) **John hit the window with the ball with the bat.*
e) **The window hit.*
(2) a) *John broke the window with the ball.*
b) *The ball broke the window.*
c) *The window broke.*
d) *The ball broke the window.*

both *hit* and *broke* can have agents as subject. Notice also that in the case of *hit* the object always remains after the verb, but *broke* allows the object to be the subject. Both verbs allow the instrument to be the subject. And all of this information comes under " syntactic potential. "

Fillmore's current (C. J. FILLMORE, 1970) " cases " are agent, experiencer, instrument object, source, goal, place, time, and extent. The syntactic potential of a word (in the sense of APRESJAN *et al.*, 1969) determines the case of a lexical item (and the case frame of a verb). For example, in the sentences

*

(3) a) **Personally, I'm sixty-five.*
b) *Personally, I'm happy.*

the reason for the non-bizarreness of (b) is that the subject of *be happy* must be an " experiencer ". On the other hand the verbal *be warm* can have an experiencer, object, instrument, place, or time as its subject:

(4) a) *Algernon is warm.*
b) *The rock is warm.*
c) *The coat is warm.*
d) *Texas is warm.*
e) *Summers are warm.*

In particular, C. J. FILLMORE (1969, p. 109) feels that the lexicon must make accessible to the user

(i) the nature of the deep-structure syntactic environments into which the item may be inserted;
(ii) the properties of the item to which the rules of grammar are sensitive;
(iii) for an item that can be used as a " predicate ", the number of " arguments " that it conceptually requires;
(iv) the role(s) which each argument plays in the situation which the item, as predicate, can be used to indicate;
(v) the presuppositions or " happiness conditions " for the use of the item, the conditions which must be satisfied in order for the item to be used " aptly ";
(vi) the nature of the conceptual or morphological relatedness of the item to other items in the lexicon;
(vii) its meaning; and
(viii) the phonological or orthographic shapes which the item assumes under given grammatical conditions.

Although dictionaries are the most popular way to define words, there are other ways than dictionaries for specifying the meanings[2] of a word within a certain lexical system. For example, U. WEINREICH (1963), in his review of Soviet semantic research, speaks of three ways of specifying word-meanings:

---

[2] It should be clear that we are not using the term ' meaning ' and ' definition ' synonomously.

1) by lexicographic definition (like the dictionary);
2) by locating the lexical item in a synonym system;
3) by establishing the syntactic properties of the lexical items.

Point (1) has been discussed above. As for point (2) M. MINSKY (1968) has a few interesting comments on the possibility of constructing a thesaurus-like dictionary (which would be, in effect, a synonym dictionary):

> My thesis is simply that we must not try to evade the 'thesaurus problem' just because we (rightly) can never be satisfied with any particular thesaurus. We must still learn how to build them, and find ways to make machines first to use them, then to modify them, and eventually to build for themselves new and better ones (p. 27).

There has been much recent research in current linguistic theory with respect to Weinreich's third way of analyzing a terminological system, by syntactic characterization of words. The J. FRIEDMAN (1971) computerized lexicon included information about the types of transformations that a word can undergo as well as some rudimentary semantic information (in the form of features). Other information that has not been included in computerized systems to any great extent are such notions as "factivity" (as defined by P. KIPARSKY and C. KIPARSKY, 1970) and notions of "genericity" and "specificity" (as discussed in R. JACKENDOFF, 1973). Another important syntactic development that has found its way into lexical systems is "case structure" as mentioned earlier and as elaborated in C. J. FILLMORE (1968, 1969, 1971), R. P. STOCKWELL *et al.* (1973). Most of these interesting and important facts of language have not been incorporated into computerized or standard dictionaries.

An additional type of information to be included in a lexicon should be the non-discrete syntactic and semantic features proposed by Ross and by Lakoff. Both linguists, working in syntax and semantics, respectively, have discovered variable acceptability of syntactic and semantic features within a given structure. Lakoff proposes to account for this variable strength probabilistically, basing his research on results from the theory of fuzzy sets. In our work on interactive lexicon construction, we have found a variation in responses due, we presumed, to regional, social, psychological and perhaps chronological differences. This probabilistic information, measured in response time, should also be included in a lexicon as information pertinent to utterance understanding and production.

3. CONTENTS OF THE LEXICON

The purpose of this section is to describe in specific detail what our dictionary will look like and how we plan to incorporate the data discussed in the previous section.

First, we propose to tag the following syntactico-semantic information on nouns, verbs, adjectives, and adverbs which we assume to be crucial: for every lexical entry in each part of speech we will record:

1) Entry word.
2) Part of speech.
3) Semantic field.
4) Dictionary definition.
5) Irregular inflectional morphology.
6) Derivational morphology. (Prefixed forms are relatively easily retrievable from the hyphenated form of the word in the dictionary with a table of prefixes. Suffixed forms can be retrieved for productive suffixes by checking the ending against a list of suffixes including the combining forms recorded in *Webster's*. The purpose of this will in part be to be able to relate lexical entries from the same root.)
7) Synonyms including synonymous cross-references (available from *NIH* research group) plus annotations from synonym paragraphs in the *Webster's Dictionary*. Suffixed forms are retrievable in part from run-on entries with notation as to source and target parts of speech.
8) Antonyms when available.
(1, 2, 3, 5 and 7 are available directly from *Webster's*: 4 and 6 are available in part from the derived data sets from the Lexicography Project users group.)
9) Example of use for each definition under a traditional main entry available from the *Brown English Corpus*.
10) Response time for sentences by informant and averaged by sentence over all informants.
11) Informant data, available from informant, including region, class, sex, age, race and economic status.

In addition we will record information peculiar to each part of speech:

For nouns (to be derived from defining formula whenever possible, otherwise interactively and by hand):

1) The following syntactico-semantic features:

$\pm$human, $\pm$animate, $\pm$count, $\pm$concrete, $\pm$male, $\pm$female.

Also, the following non-binary features which could be treated as a property list or as a set of functions in the sense of S. MARX (1972): used as an instrument, indication of quantity or degree, movable, prolonged, separable, color, and shape. These were posited on the basis of the defining formulae in *Webster's*.

2) Case markings.

3) Metaphorical extension. (We may find that this category, as well as others, are probably derivable from other information, but at the moment it isn't clear and so this information is being listed separately.)

4) Sociolinguistic restrictions on use of the entry.

For verbs:

1) Complementizers.

2) Subcategorization.

3) Defining verb, that is, the verb, if present, used in defining the entry, e.g. *be*, *become*, *come*, *have*, *make*, etc. These may be relatable to McCawley's interpretation of *kill* as " to cause to become not alive, " and to our notion of semantic field discussed below.

4) Selectional features related to noun features such as animate subject.

5) Presuppositions and their differences from synonyms of the entry.

6) Case structure (number and type of arguments.)

For adverbs:

1) Type: time, manner, location, direction, degree.
(Much of this can be gotten from the defining formulae.)

2) Position sensitivity: subject-oriented, speaker-oriented, verb-oriented, or sentence-oriented.

For adjectives:

1) The kind of noun it can or must modify, e.g. animate, concrete, count; and the manner in which it modifies (e.g. *warm stove*, *warm coat*) and whether it is a relative term (*hot/cold*) or absolute (*black/white*).

2) Semantic properties/functions: color, time, location, size, and quality. (These are disjunct sets.)

Examples:

Sample Noun Entry

1) Entry word: *man*
2) Semantic field: *Person*
3) Part of Speech: *Noun*
4) Dictionary Definition: *an adult male*
5) Irregular Inflectional Morphology: *men*
6) Derivational Morphology: *manly*
   manish
   manliness
7) Synonyms: *fellow*
   chap
8) Antonyms: *woman*
   boy
9) Example of Use (*from Brown English Corpus*): *The man killed the lion.*
10) Response Time for the Acceptability of the Sentence:
    A man is a male over 13 years of age: 5 seconds, negative response.
11) Informant Data: female student, age 19, Midwest.
12) Syntactico-Semantic Features:
    + *Concrete*
    + *Animate*
    + *Human*
    + *Male*
    + *Count*
13) Case Markings:

Agent	*The man killed the lion.*
Experiencer	*The lion killed the man.*
Source	*Only a man could make such a statement.*
Goal	*Give the book to the man.*

14) Sociolinguistic Restrictions on Use: (old) man = *husband* (*youth*); man = *boss* (*black*)

Sample Verb Entry

1) Entry word: *feel*
2) Part of Speech: *verb*
3) Dictionary Definition: *to touch in order to have a tactile sensation.*
4) Irregular Inflectional Morphology: *felt*
5) Derivational Morphology: *feeler*
6) Synonyms: *touch*
7) Antonyms: *to be numb*
8) Example of Use: *John felt the surface of the table.*

9) Response Time for Acceptability of the Sentence:
I am feeling the table: 3 seconds, negative response.
10) Informant Data: male student, age 24, northeast.
11) Complementizers: none of the regular complementizers can be used with the verb *to feel* under the above definition. Notice that if the *that* complementizer is used this indicates a change of definition:
John felt that the treatments were too painful.
12) Subcategorization: + *Transitive*
± *Stative* (— *stative* when aware of texture)
13) Defining Verb: none (implication: word defines semantic field).
14) Selectional Restrictions: + *Human Subject.*
15) Presuppositions: Instrument is part of Agent's body.
16) Case Structure: [*A*, *O*, (*I*)] v [*E*, *O*, *LOC*]

4. METHODOLOGY

The plan for the dictionary is to produce a core English lexicon consisting of the 20,000 most frequent words listed in H. KUČERA and W. N. FRANCIS (1967). The reason for choosing these is that in theory they account for 98 % of the words in running text.

As described in section 3 we have a very good idea of what to include in the lexicon, although this must obviously be left open-ended. There are problems of division of labor, however: that is, how can we most efficiently capture the information that we want to include. We have narrowed the various possible ways down to three:

1) by hand (including a real time text editing scheme)
2) interactively
3) by automated processing of a standard dictionary.

Method (1) is obvious. As for method (2) Olney (J. OLNEY, D. RAMSEY, 1973, p. 16) says, " what better source than the disambiguated parsed [= formatted] transcripts of *W* 7 and *MPD* [The *Merriam Pocket Dictionary*, which is also on tape] is there likely to be in the near future for obtaining semantic data pertaining to the English vocabulary as a whole? ". We feel that there is a better source, at least for the kinds of information that we are interested in, and that is the native speaker of English. R. N. SMITH (1972) describes a way of obtaining this data

interactively (in a system which has been described by R. L. WIDMANN (1972, p. 9) as " one of the most successful projects currently under way ") and the reader should consult that work for details.

As to method (3) we have been influenced by the work of one of the largest groups and one of the most potentially successful groups involved in automating the process of lexicon construction from standard dictionaries, viz., the user's group emanating from the Lexicographic Project headed by JOHN OLNEY of the Institute of Library Research at the University of California, Los Angeles, in collaboration with Systems Development Corporation. This project began in July 1966 with the initiation of transcribing *Merriam-Webster's Seventh Collegiate Dictionary* in computer processable form. Since then collaboration with over 30 researchers at various institutions has led to the creation of approximately 50 data sets derived from the dictionary transcript. A few of the data sets have been used in disambiguating the entries in the dictionary – the principal first goal of this philosophically, rather than linguistically, oriented project. Some of this has been relatively successful but based on the scope and the methods used, it is clear that still a great deal more time and effort will have to be expended.

Some of the already existent derived data sets are useful. The group at *SDC* has formatted the original transcript of *Webster's Seventh* so that the main entry, the etymology, the pronunciation, etc. are all put into a fixed format of card image records where the first character of each record specifies the type of information recorded, e.g. whether the record is one of the words used in the definition of a main entry. All of the subsequent data sets have been derived from this formatted version. One of these is an alphabetized list of the first 86 characters of all definitions separately and by part of speech. In addition all synonomous cross-references have been extracted, alphabetized on the main entry form and on the word referred to. Also, there are various suffixal data sets used in aiding to correlate suffixes with definitions.

Samples of print-out for sorted definitions within part of speech and end-alphabetized within part of speech are appended. The former has been especially productive by giving us quite a good deal of insight into so-called defining formulae and these defining formulae have in turn allowed us to posit certain features which can be extracted directly from the definitions. These defining formulae will be used in extracting some of the features from *Webster's*. (Some features such as " + human " cannot be extracted automatically, except by listing, by the interactive scheme described above or, by some inferential scheme.) We have also

constructed a KWIC concordance for a portion of the data on non-function words in the definitions which will lead to short-cuts for syntactic-semantic tagging.

5. STRUCTURING THE DATA

The innovation that we propose to implement in this computerized dictionary that will allow us to structure and store all of the information discussed above efficiently and accurately is that of the " semantic field. " The theory of semantic fields is not new; what is new is the use of this concept to structure semantic information. Its most appealing characteristic is that it eliminates the need for redundant information (the problem with the feature approach which is widely used) and it makes retrieval much more efficient. First we will discuss the motivation for such a system as a model for semantic structure.

Some of the most interesting empirical evidence for semantic fields has been in work done by Marshall and Newcombe in psycholinguistics and by Whitaker, Kehoe, Schnitzer and others in neuro-linguistics. H. A. WHITAKER (1971) has described the remarkable correspondence of the distinct cellular arrays in the cortex of the brain to the classical divisions of the language system: the semantic/syntactic component, the lexicon, and the phonological component.

For example, it has been found that the lexicon has an existence apart from the syntactic-semantic (or logical) aspects of language. A case study reported by H. A. WHITAKER (1971), described a woman who was unable " to initiate conversation or to demonstrate general cognitive skills – in brief, the semantic and syntactic aspects of language were totally lost. She was however, able to repeat verbal material well, ... " (p. 190). Whitaker has postulated that the lexicon is a separate neural component, perhaps biochemically coded in nerve cells. That the lexicon, a separate component, is organized in some sort of semantic field arrangement was pointed out again and again by Whitaker. In work done by E. WEIGL and M. BIERWISCH (1970), they described errors which were the results of substitutions of words for other words from the same semantic fields; e.g., *trousers* for *blouse*, *tie* for *cuff*, *bodice* for *cardigan*, *sandals* for *socks*, *peaches* for *oranges*, *bananas* for *figs*, *potatoes* for *vegetables*. Of particular note is that the substitutions usually occur at the same taxonomic level, that is, the substitution is rarely an

item for the name of the field containing the item (e.g., *peaches* for *fruit*).

In another study, by H. GOODGLASS, B. KLEIN, P. CAREY and K. JONES (1966), the investigators chose words which came within the categories of objects, forms, letters, actions, numbers, colors, and body parts. They found that the patients had an easier time understanding object names than producing them, but producing letters was easier for them than understanding them.

J. C. MARSHALL and F. NEWCOMBE (1966) reported errors such as the following: their patient read *liberty* as *freedom*, *canary* as *parrot*, *abroad* as *overseas*, *entertain* as *entertainment*, *political* as *politician* and *beg* as *beggar*. Later studies of the same patient showed that the patient had twice as much difficulty with verbs than with nouns and that adjectives were harder than nouns but easier than verbs. One of the problems encountered was the patient's tendency to read verbs as the corresponding derived nominal and to read nominals derived from adjectives as the original base form of the adjective. Words like *uncle*, *priest* and *poet* were harder than *horse*, *lion*, and *insect*. *Large* was read as *long*, *short* as *small*, *tall* as *long*, *little*, as *short*.

H. A. WHITAKER (1971) reports patients who read verbs as their corresponding derived nominal form: *decide* is read as *decision*, *conceal* as *concealment*, *nominate* as *nomination*, *portray* as *portrait*, *bathe* as *bath*, *speak* as *discussion*, *remember* as *memory*. Whitaker also reports that the opposite phenomenon has been found where derived forms are read as their base forms: *refusal* was read as *refuse*, *darkness* as *dark*, *whiteness* as *white*, *amazement* as *amaze*.

Psycholinguistic and anthropological data therefore point to the reality of organization into semantic fields and success of information retrieval schemes has often been tied into a division of the semantic universe into fields. It would seem not only an obvious desideratum but a *sine qua non* in a dictionary to include information of semantic field.

Once the data has been recorded so that all words are completely defined we will eliminate redundant information so that storing of the lexicon can be accomplished most economically. The elimination of redundancy will be done by means of structuring the data in a specific way. This method has been discussed in E. MAXWELL (1973).

In effect what happens is this: the head of a semantic field (call it *L*) is defined in a certain way; the members of that semantic field (λ_1, λ_2 ...λ_n) are defined in relation to *L*. All the information that need be specified to define, λ_1, etc. is that information that is unique to them.

For example, there is the semantic field (described in C. J. FILLMORE, 1971) made up of the verbs: *judge, accuse, blame, scold, forgive,* etc. All of the verbs are verbs of *judging* (which is the name of the semantic fied.) They are uniquely defined in terms of their presuppositions (i.e. *accuse* presupposes that the action done is bad). Therefore, by defining *judge* and by saying that *accuse,* etc. are kinds of *judging* except for their presuppositions all redundant information can be deleted and the specific definitions can be derived with inferential schemata.

An example of how the information would be stored is the following (using the word *boil*):

[(* * *)	(* * *)	XX	(* * *)	(∧)	(* * *)	(∧)	(* * *)]
↓	↓		↓		↓		↓
SF	AGENT		OBJECT		PLACE		INSTR
(' COOK ')	(HUMAN)		(EDIBLE/POTABLE)		(HEATED)		(WATER)

The partial description of the word *boil* gives the following information: that it is a member of the semantic field " cook "; that the agent must be a member of the semantic field " human "; that the thing boiled must be edible or potable; that the place the boiling is done must be heated (actually this information is redundant since the place for cooking must also be heated); and the instrument in which the boiling is done must be water. The symbol xx means that the object can be subject if no agent is stated:

Alice boiled the eggs.
The eggs boiled quickly.

The parentheses around the operators mean that the choice of place and instrument is optional.

Using this model we can state relationships between derivational morphemes and nominalizations that have not as yet been stated in computerized lexicons. (*Reliable* is passively related to *rely*: " able to be relied on "; while *comfortable* is actively related to *comfort*: " able to comfort ").

SUMMARY. Our purpose is to construct a 20,000 word core dictionary of English to be used in computerized natural language using systems. It is to include as much syntactico-semantic information as necessary to be used in most current theoretical frameworks both in

sentence recognition and production as well as for linguistic studies of English syntax and semantics.

We eventually would like to parse the definitions so that this information can be put in some formal notation and used for further dictionary organization but we feel at the moment that our core-English dictionary must be pre-requisite to any such definition parsing (cf. O. WERNER, 1972 for a model to account for taxonomic relations derivable from definitions).

APPENDIX I

OF , RELATING TO , OR SUITABLE FOR A FEAST OR FESTIVAL	AJ FESTIVE
OF , RELATING TO , OR SUITABLE TO A LETTER	AJ EPISTOLARY
OF , RELATING TO , OR SUITED TO AN EPICURE	AJ EPICUREAN
OF , RELATING TO , OR SUPPORTED BY CHARITY	AJ ELEEMOSYNARY
OF , RELATING TO , OR TEACHING THE BASIC SUBJECTS OF EDUCA	AJ ELEMENTARY
OF , RELATING TO , OR TENDING TO CAUSE DEGENERATION A [DI]	AJ DEGENERATIVE
OF , RELATING TO , OR TENDING TO PRODUCE AN ELECTRIC CURREN	AJ ELECTROMOTIVE
OF , RELATING TO , OR USING THE METHODS OF GEOCHEMISTRY	AJ GEOCHEMICAL
OF , RELATING TO , OR UTILIZING DEVICES CONSTRUCTED OR WORD	AJ ELECTRONIC
OF , RELATING TO , OR WRITTEN IN A SIMPLIFIED FORM OF THE	AJ DEMOTIC
OF , RESEMBLING , OR COMPOSED OF FILM	AJ FILMY
OF , RESEMBLING , OR PRODUCING A DISK $\leq$ AS	AJ DISCOIDAL
OF , USED FOR , OR ASSOCIATED WITH BURIAL A PHARACH$\neq$S [CHA	AJ FUNERARY
OF , USING , OR INVOLVING EQUATION OR EQUATIONS	AJ EQUATIONAL
OF A DULL BROWNISH YELLOW TAWNY	AJ FULVOUS
OF A FAVORABLE CHARACTER OR TENDENCY [NEWS BOUNTIFUL FERT]	AJ GOOD
OF A HIGH DEGREE OF EXCELLENCE SUPERB	AJ GOLDEN
OF A KIND GROWN IN THE OPEN AS DISTINGUISHED FROM ONE MORE	AJ GARDEN
OF A KIND RELATED TO OR RESEMBLING ANOTHER KIND THAT IS USU	AJ FALSE
OF A LIGHT BLUISH GRAY OR BLUISH WHITE COLOR	AJ GLAUCOUS
OF A LIGHT YELLOWISH BROWN	AJ FALLOW
OF A MIXED EUROPEAN AND ASIATIC ORIGIN	AJ EURASIAN
OF A PALE YELLOW GREEN COLOR	AJ GLAUCOUS
OF A PARTICULAR SORT SPECIFIC	AJ EXPRESS
OF A PLEASANT CHEERFUL DISPOSITION	AJ GOOD-NATURED
OF A RUDDY HEALTHY COLOR	AJ FLUSH

APPENDIX II

HAVING AN EMBRYO	AJ	EMBRYONATED
HARD AND DENSE LIKE IVORY	AJ	EBURNATED
GIVEN TO OR MARKED BY DISSIPATION DISSOLUTE	AJ	DISSIPATED
PROVIDED WITH OR CHARACTERIZED BY WINDOWS	AJ	FENESTRATED
HAVING ONE OR MORE OPENINGS OR TRANSPARENT SPOTS	AJ	FENESTRATED
RETICULATE [LEAVES	AJ	FENESTRATED
RAISED ESP. ABOVE THE GROUND OR OTHER SURFACE [HIGHWAY	AJ	ELEVATED
MORALLY OR INTELLECTUALLY ON A HIGH PLANE [MIND	AJ	ELEVATED
FORMAL DIGNIFIED [DICTION	AJ	ELEVATED
EXHILARATED	AJ	ELEVATED
BROKEN	AJ	FRACTED
CAST DOWN IN SPIRITS DEPRESSED	AJ	DEJECTED
DOWNCAST	AJ	DEJECTED
THROWN DOWN	AJ	DEJECTED
LOWERED IN RANK OR CONDITION	AJ	DEJECTED
NOT CONNECTED INCOHERENT	AJ	DISCONNECTED
HAVING A POSITIVE OR NEGATIVE SENSE [LINE SEGMENT	AJ	DIRECTED
CUT DEEPLY INTO FINE LOBES A [LEAF	AJ	DISSECTED
HAVING GREAT NATURAL ABILITY TALENTED [CHILDREN	AJ	GIFTED
REVEALING A SPECIAL GIFT [VOICES	AJ	GIFTED
DELIGHTFUL	AJ	DELIGHTED
HIGHLY PLEASED	AJ	DELIGHTED
SEEING OR ABLE TO SEE TO A GREAT DISTANCE	AJ	FARSIGHTED
HAVING FORESIGHT OR GOOD JUDGMENT SAGACIOUS	AJ	FARSIGHTED
HYPEROPIC	AJ	FARSIGHTED

REFERENCES

YU. D. APRESJAN, *The experimental study of the semantics of the Russian verb* (in Russian), Moscow, 1967.

YU. D. APRESJAN, I. A. MEL'ČUK, A. D. ŽOLKOVSKY, *Semantics and lexicography: towards a new type of unilingual dictionary*, in F. KIEFER (ed.), *Studies in syntax and semantics*, 1969, pp. 1-33.

C. J. FILLMORE, *The case for case*, in E. BACH, R. T. HARMS (eds.), *Universals in linguistic theory*, New York, 1968.

C. J. FILLMORE, *Types of lexical information*, in F. KIEFER (ed.), *Studies in syntax and semantics*, 1969, pp. 109-37.

C. J. FILLMORE, *Verbs of Judging: an exercise in semantic description*, in C. J. FILLMORE, D. T. LANGENDOEN (eds.), *Studies in linguistic semantics*, New York, 1971.

J. FRIEDMAN, *A computer model of transformational grammar*, New York, 1971.

H. GOODGLASS, B. KLEIN, P. CAREY, K. JONES, *Specific semantic word categories in aphasia*, in « Cortex », II (1966), pp. 74-89.

P. B. GOVE, *English dictionaries of the future*, in H. D. WEINBROT (ed.), *New aspects of lexicography*, Carbondale, 1972.

R. JACKENDOFF, *Semantic interpretation and generative grammar*, Cambridge (Mass.), 1973.

P. KIPARSKY, C. KIPARSKY, *Fact*, in M. BIERWISCH, K. E. HEIDOLPH (eds.), *Progress in linguistics*, The Hague, 1970, pp. 143-173.

H. KUČERA, W. N. FRANCIS, *A computational analysis of present-day American English*, Providence, 1967.

G. LAKOFF, *Hedges: a study in meaning criteria and the logic of fuzzy concepts*, ms.

J. C. MARSHALL, F. NEWCOMBE, *Syntactic and semantic errors in Paralexia*, in « Neuropsychologia », IV (1966), pp. 169-176.

M. MARSHALL, F. NEWCOMBE, J. C. MARSHALL, *The microstructure of word-finding difficulties in a dysphasic subject*, in G. B. FLORES D'ARCAIS, W. LEVELT (eds.), *Advances in Psycholinguistics*, Amsterdam, 1971.

S. MARX, *Deductive question-answering with natural language inputs*, diss., 1972.

E. MAXWELL, *Graphical representation of semantic fields*, Paper read at Conference of the Association for Computational Linguistics, Ann Arbor (Mich.), 1973.

J. MCCAWLEY, *The role of semantics in a grammar*, in E. BACH, R. T. HARMS. (eds), *Universals of linguistic theory*, New York, 1968.

M. MINSKY (ed.), *Semantic information Processing*, Cambridge (Mass.), 1968.

J. OLNEY, D. RAMSEY, *From machine-readable dictionaries to a lexicon tester: progress, plans, and an offer*, in « Computer studies in the humanities and verbal behavior », III (1972) 2, pp. 213-220.

M. PAČAK, A. W. PRATT, *The function of semantics in automated language processing*, reprint from *Proceedings of the Symposium on information storage and retrieval, College Park (Md.), April 1-2, 1971.*

M. R. QUILLIAN, *The teachable language comprehender*, in M. MINSKY. (ed.), *Semantic information processing*, Cambridge (Mass.), 1968.

C. SALTON, *The performance of interactive information retrieval*, in « Information processing letter », I (1971).

R. F. SIMMONS, *Natural language question answering systems*, in « Communications of the ACM », XIII (1970), pp. 15-30.

R. N. SMITH, *Interactive lexicon updating*, in « Computers and the Humanities », VI (1972) 3.

R. P. STOCKWELL, P. SCHACHTER, B. H. PARTEE, *The major syntactic structures of English*, New York, 1973.

E. WEIGL, M. BIERWISCH, *Neuropsychology and linguistics: topics of common research*, in « Foundations of Language », 1970.

U. WEINREICH, *Lexicology*, in T. SEBEOK (ed.), *Current trends in linguistics*, Vol. I, The Hague, 1963.

O. WERNER, *Ethnoscience*, mimeographed, Northwestern University, 1972.

H. A. WHITAKER, *Neurolinguistics*, in W. O. DINGWALL (ed.), *A survey of linguistic science*, College Park (Maryland), 1971.

R. L. WIDMANN, *Recent scholarship in literary and linguistic scholarship*, in « Computers and the Humanities », VII (1972) 1, pp. 3-27.

W. A. WOODS, *The lunar sciences natural language information system*, Cambridge (Mass.), 1972.

A. K. ŽOLKOVSKY, I. A. MEL'ČUK, *Semantic synthesis*, in « Systems Theory Research », XIX (1970), pp. 170-243.