ANNUAL REPORT:
AUTOMATIC INDEXING
AND ABSTRACTING

PART I

M-21-66-1 March 1966

Annual Progress Report
Office of Naval Research
Contract Nonr 4440(00)

Electronic Sciences Laboratory
LOCKHEED MISSILES & SPACE COMPANY
A Group Division of Lockheed Aircraft Corporation
Palo Alto, California

PRÉCIS

RESEARCH PROGRESS REPORT

Title: "Annual Report: Automatic Indexing and Abstracting," Annual Progress Report, Part I, Office of Naval Research, Contract Nonr 4440(00).

Background: This investigation is concerned with the development of automatic indexing, abstracting, and extracting systems. Basic investigations in English morphology, phonetics, and syntax are pursued as necessary means to this end.

Condensed Report Contents: The second annual report on automatic indexing and extracting consists of 8 papers summarizing progress in three areas of investigation:

(1) Application of English word morphology to automatic indexing and extracting

(2) Use of combined syntactic and entropy selection criteria in automatic indexing

(3) Studies in phonetic English

The first four papers are concerned with the relationship between the part of speech of words and their graphic form. An operational definition of affixes is given, the usefulness of affixes in the automatic determination of parts of speech is discussed, and an algorithm is outlined for determining parts of speech with a dictionary look-up of less than 200 affixes and less than 800 words. The inflection of adjectives is also discussed, anticipating the need for future refinement of the part-of-speech algorithm, which at present identifies 11 part-of-speech categories. For some objectives these categories may be inadequate, necessitating further breakdown, for example adjectives might be further distinguished as relative, comparative, etc.

The fifth paper is a progress report on the development of a method for automatic indexing without reference to any pre-prepared dictionary, thesaurus, etc. It shows the current results on five text excerpts.

The final three papers are concerned with the relationship between English phonetics and English morphology. One of the papers is concerned with homonyms, which represent a problem area in transformation from phonetic to graphic English. Another discusses a function for mapping written English into spoken English, and the third describes a computerized study of transcribed English phonetics as given by different dictionaries.

For Further Information: The complete report is available in the major Navy technical libraries and can be obtained from the Defense Documentation Center. A few copies are available for distribution by the author.

FOREWORD

The issue of this report marks the completion of the second year in which the Office of Naval Research has contributed support to the program of research in the information sciences at the Palo Alto Laboratories of the Lockheed Missiles & Space Company.

It is convenient to think of the work reported here as dealing with a data base. During the first year of the program, a major part of the effort went into establishment of the data base. Illustration of its nature and use are provided by the five volumes of The English Word Speculum which was distributed to ONR program participants during the last year. In this report, examples of exploration and application of the data base to problems in linguistics and information analysis are given. Special attention should be given to the report by R. P. Mitchell which shows how the research methods developed for written English can be used in an approach to the problems of synthesis and recognition of spoken English.

One part of the year's work is not reported here. This deals with development of a technique for obtaining index phrases in English from untranslated Russian technical texts. This work is described in a separate report.

The group at Lockheed takes this opportunity to express its thanks for the continued support and encouragement given by the members of the Information Sciences Branch of the Office of Naval Research.

B. D. Rudin
Principal Investigator

ABSTRACTS

1. STRUCTURAL DEFINITION OF AFFIXES FROM MULTISYLLABLE WORDS

In July 1964, H. L. Resnikoff and J. L. Dolby presented a paper at the Bloomington meeting of the AMTCL entitled "The Nature of Affixing in Written English." In that paper, an algorithm for the structural definition of affixes was developed and applied to data consisting of all the words of the form CVCVC in the Shorter Oxford English Dictionary. Fourteen strong prefixes and twelve strong suffixes, seven weak prefixes and forty weak suffixes were defined, but it was noted that all the affixes could not be expected to show up in two-vowel-string words. This paper summarizes the results of applying a modified form of the operational definition to data consisting of all the four-, five-, six-, and seven-vowel-string words in Webster's Third New International Dictionary of the English Language. Thirteen additional weak suffixes, nineteen weak prefixes, seventeen strong prefixes, one strong suffix, and twelve possible suffix-compounding elements were found.

2. PART-OF-SPEECH IMPLICATIONS OF AFFIXES

This paper describes a systematic investigation of the extent to which the part-of-speech of words can be identified from their prefixes and suffixes. The results indicate that it is possible to determine, with 95 percent accuracy, the inclusive part of speech of an affixed word from a consideration of its prefixes, suffixes, and length.

By "inclusive" parts-of-speech we mean a string which will include all of the parts-of-speech assigned by both dictionaries considered, but which may include one or two extraneous parts-of-speech. The extra parts-of-speech will differ according to the class of words, as adjectives may have an extra part-of-speech "noun" or "adverb," while nouns may have an extra part-of-speech "verb." The part-of-speech implications of 72 prefixes and of 87 suffixes are given.

3. ON THE INFLECTION OF WRITTEN ENGLISH ADJECTIVES

The inflection of adjectives in English is investigated on the basis of the number of admissible vowel strings contained in a given word. Two types of comparison are distinguished: the terminational and the analytic. The investigation has determined a direct relationship between the inflection of adjectives (a given set of adjectives of a certain graphemically defined type) and an easily observed structural property, and the following claim is made:

1. The standard adjectives in W (the set of one-vowel-string words which do not end with the sequence consonant -le is denoted by W) which are not standard adverbs are inflected analytically, i.e., by using more and most.
2. The standard adjectives in W which are also standard adverbs are inflected terminationally, i.e., by using the suffixes -er and -est.

In view of the discussion of the relation of the traditional parts-of-speech classes to structural properties of English, the asserted claim takes on a special importance. It asserts that the set of adjectives of a certain graphemically defined type (namely, those that bel) can be partitioned into two classes, one containing the

analytically inflected adjectives, and the other containing the terminationally inflected adjectives. This partition can be determined solely from the parts-of-speech classes to which the adjectives belong.

4. AUTOMATIC DETERMINATION OF PARTS OF SPEECH OF ENGLISH WORDS

A procedure for automatically assigning part-of-speech characteristics to English words is discussed in this paper. The development of the algorithm is traced, and the algorithm itself is described in terms of three basic graphemic rules, which are used in conjunction with a group of affixes (less than 200) and a list of exception words (less than 800) whose part of speech must be stored in the computer for direct look-up. The results of a test of the algorithm on a 500-word random sample from a 73,582 word dictionary are given. Ninety-five percent of the samples are assigned the correct "inclusive" part-of-speech string, where the inclusive string is defined as including all parts of speech given in the dictionary, but which may include one, or rarely two, extra parts of speech.

5. A SYNTACTIC-STATISTICAL METHOD FOR AUTOMATIC INDEXING

The objectives and method of an algorithm developed for automatic indexing are briefly presented. The reduction level achieved by the algorithm is indicated and the results of tests on five text excerpts are shown.

6. STATISTICS OF OPERATIONALLY DEFINED HOMONYMS OF ELEMENTARY WORDS

This computerized study of the homonyms of elementary words (roughly equivalent to monosyllabic words) has allowed the compilation of exhaustive lists of homonym sets, using phonetic transcriptions from five different dictionaries. Of the 5,706 elementary words, 3,000 were involved in at least one homonym set, indicating that homonyms will present a significant problem in mechanized word recognition. The effects on the homonym sets of changing from the phonetic transcription of one dictionary to another were tabulated, as were the effects of removing dialectal pronunciations. Since the effects of dialectal variations turned out to be relatively small, it was possible to categorize and list for study the actual words whose dialectal pronunciations caused homonym-type confusion with other words.

7. ACOUSTIC PHONETIC TRANSCRIPTION OF WRITTEN ENGLISH

The function that maps the words of written English onto the corresponding words of spoken English is described. The simplest hypothesis is that the function F, defined on the symbols forming the letters of the alphabet, maps each letter onto a sound and maps the sequence of letters L_1L_2 as $F(L_1L_2) \quad F(L_1)F(L_2)$. This hypothesis is false, since F is not always well defined in the sense that its values are not always unique and the equation does not always persist. On the basis of an exhaustive dictionary search, we have shown that it is possible to define F in a context-dependent manner such that its restriction to consonant strings is uniquely defined.

with this new definition, the equation holds for consonant strings of the grammatically homogeneous one-vowel-string words of written English; the consonant strings in these words coincide with those uninfluenced by the rules of euphonic combination.

8. COMPUTER STUDY OF TRANSCRIBED ENGLISH PHONETICS: A PROGRESS REPORT

A summary of a computer-oriented study of the relations between orthographic and transcribed phonetic forms of elementary English words is presented. The principles used to generate transcribed phonetic data from the orthography of elementary words are described. The computer programs which embody these principles were used to accurately obtain phonetic data contained in five authoritative sources, yielding phonetic transcriptions for several recognized dialects of English.

These data were analyzed and several important results were obtained. A brief summary of these results is presented, among them (a) the significance of homonyms, (b) predictable dependencies of vowel values upon consonant values, and (c) the isolation of phonetic segments which are independent of transcriptions in which they may occur.

CONTENTS

ILLUSTRATIONS

TABLES

INTRODUCTION

During the past year, experiments in automatic indexing have proceeded in four areas, three of which are covered in this report:

(1) Application of English-word morphology to automatic indexing and extracting

(2) Automatic indexing using combined syntactic and entropy selection criteria

(3) Studies in phonetic English

Each section comprises papers describing specific efforts within that area. The first phase of the fourth area of research, direct Russian-to-English indexing, has been completed and is documented in a separate volume of this report, Part II.

The first section of this report is a continuation of the last annual report. As that report indicated, only 66 percent of the 500 words were assigned a correct part-of-speech string by the part-of-speech algorithm, whereas 95 percent accuracy is desired for the indexing experiment. Therefore a more complete study of the two crucial factors in the algorithm was undertaken:

- The determination of affix sequences and of their part-of-speech implications (Papers 1 and 2)
- Exceptions to the basic theoretical premise that words with one-syllable kernels have parts of speech noun, adjective, and verb, while those with multisyllable kernels have parts of speech noun and adjective (Paper 3)

These studies indicated that 95 percent accuracy is not obtainable from considerations of vowel-string and word affixing. Accordingly, the goal of 95 percent accuracy was dropped, at least temporarily, in favor of a more obtainable goal of 95 percent "inclusive accuracy," wherein the string assigned to a word includes all parts of speech given by the dictionaries, but which may include one, or rarely two, extraneous parts of speech. This decision was based on the judgment that in most utilizations of

parts-of-speech information, it is easier to eliminate or compensate for extra parts of speech than to infer the existence of, or compensate for, missing parts of speech. The goal of 95 percent inclusive accuracy has been reached on the 500-word random sample from the dictionary. The modified part-of-speech algorithm can now be recoded for the IBM 360 Model 30 now available to the research laboratory, and coding of other programs necessary to the compilation of a sentence dictionary can proceed. The sentence dictionary will be used to investigate the relationship between syntax and "indexible" sentences, as described in the first annual report. It is also expected to play a more general role in the study of syntax, analogous to that played by word dictionaries in the study of morphology. It will provide an ordered list of observed sentence constructions which will be useful both in the derivation and testing of syntactic algorithms.

The second section gives the results to date of an evolving algorithm for combining syntactic and entropy criteria in the automatic selection of index items and phrases. In this experiment, a parsing program is used to select the syntactic units upon which the entropy criteria is then imposed. Because of a change in computers, it has been possible to complete tests on only five text fragments. Further refinements and tests must await the recoding of the programs for the IBM 360 Model 30 now in use by the Research Laboratory.

The third section describes progress in the investigation of relationships between the phonetic and graphemic forms of English words. Such studies are expected to make it practicable to use human speech as input to and output from a computer programmed for any desired automatic processing of language.

I

APPLICATION OF ENGLISH WORD MORPHOLOGY TO AUTOMATIC INDEXING AND EXTRACTING

1. STRUCTURAL DEFINITION OF AFFIXES FROM MULTISYLLABLE WORDS*

L. L. Earl

In this paper the goal is to define affixes from structural criteria alone. The problem of when an affix sequence is genuinely acting as an affix (as re may be considered a prefix in react but not in read) will not be considered, although the categorization into strong and weak affixes is intended to anticipate this problem. The validity of the defined affixes will be indicated only by comparison with existent affix lists. A more utilitarian evaluation of affix validity can be made after the syntactic and phonetic implications of the defined affixes have been investigated.

The definitions for affixes given in this paper are essentially unchanged from the definitions given by Dolby and Resnikoff,[1] but are extended to include both one- and two-syllable affixes. The data set to which these definitions are applied is the four-, five-, six-, and seven-vowel-string words, a set of about 11, 250 words. From this set the one-vowel-string affixes which did not occur in the two-vowel-string data set (used in Reference 1) will be defined, along with the two-vowel-string affixes which could not have occurred in the two-vowel-string data.

The extended definition for strong prefixes can be summarized as follows (consonant strings referred to in the definition are given in Table 1-1): Given a word of the form $C_1V_1C_2V_2C_3V_3 \ldots$, if either C_2 or C_3 is an inadmissible consonant string, there is a mandatory syllable break within the string, and everything preceding that break is defined as a prefix possibility. A prefix possibility is defined as a prefix probability if in the data there are at least four words with the same prefix possibility

*This work was supported by the Office of Naval Research and by the Independent Research Program of Lockheed Missiles & Space Company.

Table 1-1

INITIAL AND TERMINAL STRINGS

ADMISSIBLE INITIAL CONSONANT STRINGS OF CVC WORDS

B	N	BL	GL	SH	TR	SCH
C	P	BR	GN	SK	TW	SCR
D	Q	CH	GR	SL	WH	SHR
F	R	CL	KN	SM	WR	SPH
G	S	CR	KR	SN		SPL
H	T	DR	PH	SP		SPR
J	V	DW	PL	SQ		STR
K	W	FL	PR	ST		THR
L	Z	FR	RH	SW		THW
M		GH	SC	TH		

ADMISSIBLE FINAL CONSONANT STRINGS OF CVC WORDS NOT ENDING WITH E

B	BB	MP	SH	GHT
C	CH	ND	SK	LCH
D	CK	NG	SM	LPH
F	CT	NK	SP	LTH
G	DD	NN	SS	MPH
H	FF	NT	ST	MPT
K	FT	NX	TH	NCH
L	GG	PH	TT	NTH
M	GH	PT	WD	NTZ
N	GN	RB	WK	RCH
P	LD	RC	WL	RSH
R	LF	RD	WN	RST
T	LK	RF	XT	RTH
W	LL	RK	ZZ	SCH
X	LM	RL		TCH
Z	LP	RM		
	LT	RN		
	MB	RP		
	MM	RR		
	MN	RT		

arising from the same consonant string. A prefix probability becomes a strong prefix if the same prefix probability arises from two or more inadmissible consonant strings. The definition for strong suffixes is analogous, proceeding from the other end of the word. Thus, given a word of the form ... $V_3C_3V_2C_2V_1C_1$, if either C_2 or C_3 is an inadmissible string, there is a mandatory syllabic break within the string, and everything following that break is defined to be a suffix possibility. Then the definition for suffix probability and for strong suffix is the same as for prefixes; the word suffix can be substituted for the word prefix wherever it occurs. The consonant string C_1 may be blank in either case. The criterion of four or more words in establishing an affix probability, and the criterion of two or more consonant strings in defining an affix from a probability, were established in Reference 1. These criteria were established heuristically, and have been retained here not only for the sake of consistency but also because they were proven effective.

The definition for weak affixes has also been extended to include two-syllable affixes. Weak affixes are so classified because their definition is based on a probable syllabic break rather than a mandatory break. Because such probable breaks are not interior to a consonant string, weak prefixes end with a vowel and weak suffixes begin with a vowel. For prefixes, given a word of the form $C_1V_1C_2V_2C_3V_3$ ..., if either C_2 or C_3 is an admissible initial string but not an admissible final string, everything preceding that consonant string is a prefix possibility. For suffixes, given a word of the form ... $V_3C_3V_2C_2V_1C_1$, if either C_2 or C_3 is an admissible final string but not an admissible initial string, everything following that consonant string is a suffix possibility. The criteria by which an affix possibility becomes an affix are the same as for strong affixes. Note that these definitions exclude admissible final strings from C_2 or C_3 for prefixes, and admissible initial strings from C_2 or C_3 for suffixes, to

increase the reliability of the definition by reducing the probability of postulating a break before (for prefixes) or after (for suffixes) C_2 or C_3 where a break does not exist. Consider the prefix case first. If C_2 or C_3 is an admissible initial string, and also an admissible ending string, the syllabic break could be logically either before or after the string. The string CH is such a string, as the following words illustrate.

enrich/ment ta/chometer

poach/er re/christen

By eliminating such doubtful strings we should increase somewhat the reliability of the definition of our prefix possibilities, but we do not completely eliminate chance for error, because even with initial strings that are not also final strings, a break may occur internal to a multiletter string or after a single letter string. The strings BR and GR are such multiletter strings, as the following words illustrate.

sub/routine ag/riculture

re/broadcast de/gree

The chances of this happening in two multiletter strings with the same prefix possibility is judged small enough to be discounted, since here we are simply defining prefix sequences. The chances of error due to a break after a single letter seems greater, as with the letter S.

re/sidual

res/ident

However, since there are only three single consonants which are beginning but not ending strings (J, S, V), and since again it takes two consonant strings to cause a sequence to be defined as an affix, this problem, too, can be discounted

It is suspected that the situation for suffixes is more difficult in that the set of terminal consonant strings left after removing initial strings has more members which show a tendency to break internally. For example, breaks in the following strings are common.

c/t	as in	lac/tate	m/b	as in	am/bition
r/t	as in	fer/tile	m/p	as in	am/pere
p/t	as in	ap/titude	r/l	as in	pur/loin
r/b	as in	ar/bor	n/d	as in	ban/dit

Therefore, more difficulty in determining when a defined weak suffix is actually acting as a suffix in a given word could reasonably be anticipated. It would be interesting to subject each of the weak suffixes to a qualifying test, namely that in the two-syllable data set there not be two sets of illegal strings preceding the suffix, where each had at least four members. When this test was applied to the five suffixes a, age, ah, ent, ock, two of the suffixes, a and ock, failed the test. But, both a and ock obviously sometimes act as suffixes (they are both listed in the dictionaries as such), so it is unwise to eliminate them at this point in the research. What is indicated, perhaps, is the structural classification of the weak suffixes by degree of weakness, as a means of approaching the suffix-in-context problem.

Table 1-2 reviews the prefixes and suffixes defined in Reference 1, using the two-vowel-string words as the data set. Table 1-3 shows the new suffixes defined using four-, five-, six-, and seven-vowel-string words, with the preceding letter strings and occurrence counts which established them as suffixes. Surprisingly, there is only one that can be considered a strong suffix, and that actually turned up as the weak suffix ation. Since all of the preceding letter strings turned out to be of the form Ct (where C = c, l, n, or r), and since phonetic breaks were consistently before the t,

Table 1-2

AFFIXES FROM TWO-VOWEL STRING WORDS

Strong Prefixes		Strong Suffixes	
ac	in	ful	ly
ad	mis	land	lock
al	out	ler	man
con	sub	less	ment
dis	sun	let	mess
en	trans	ling	ward
ex	us		

Weak Prefixes
a
be
cy
de
e
i
re

Weak Suffixes		
a	in	con
age	ine	ue
ah	ing	er
al	ion	um
an	is	et
ant	ish	ure
ar	ite	ic
ard	ive	us
at	o	ie
ed	ock	ier
ee	on	ile
el	or	
en	ot	
ent	ow	

Table 1-3

SUFFIXES FROM MULTISYLLABLE WORDS

Suffix	Preceding Letter Strings	No. of Occurrences of Suffix Following Given Letter String
(t)ation	c(t)	5
	l(t)	6
	n(t)	36
	r(t)	5
able	ll	8
	nt	4
ial	nn	8
	nt	37
ate	ll	6
	nn	5
ist	ll	4
	nt	12
	pt	4
ism	ll	4
	nt	5
ian	ll	4
	nn	4
ium	ng	5
	rd	4
ia	ps	12
	rd	4
	nt	5
y	rg	4
	ps	4
	rm	4
	rr	36
	st	19
	x	13
ous	ll	6
	rm	6
	rp	11
ide	x	9
	lf	7
is	ps	6
	x	28

(as in plantation), it seemed reasonable to consider tation a strong suffix. Of the thirteen newly defined suffixes, able, ial, ate, ist, ism, y, ous, ian, ium, ia, and ide, are all commonly recognized as such, while only tation or ation and is are not.

It was expected that more than one two-vowel-string suffix would materialize. Instead, a number of sequences were observed which appear to act as inner suffixes, or suffix compounding elements, which occur frequently in combination with one-syllable suffixes. Thus, the sequence tic is frequently encountered followed by al, ize, or ide to form tical, ticism, ticize, ticide as in elliptical, aspticism, didacticism, ascepticize, romanticize, and infanticide. Such interior sequences which meet the occurrence criteria set up for suffixes are listed in Table 1-4. It is expected that these sequences will have little syntactic meaning but may be helpful in word hyphenation techniques.

Table 1-5 shows the prefixes defined using four-, five-, six-, and seven-vowel-string words, with the following letter strings and occurrence counts which established them as prefixes. The three newly defined strong two-syllable prefixes circum, inter, and hyper are well known. Three other common prefixes, over, under, and super, were encountered with a good many letter strings, but always failed to meet the requirement of more than three occurrences with a given letter string.

Of the strong one-syllable prefixes defined, ab, at, ap, com, an, em, im, and ec are recognized by dictionaries, while vul is not. Of the weak two-syllable prefixes, auto, demo, iso, photo, epi, and tele, are commonly recognized, but ana, apo, deni, and irre* are not. None of the one-syllable weak prefixes (au, ca, hy, ma, mi, lu, pro, sa, su, vi) are familiar as meaningful prefixes except for pro. Therefore, the next step, in which the part of speech implications of the structurally defined affixes

*irre is no doubt a combination of the recognized prefixes i and re.

Table 1-4

ELEMENTS COMBINING WITH SUFFIXES

Suffix Compounding Element	Terminal Letter String Associated	No. of Occurrences
-cat-	rc	9
	nc	12
-mat-	rm	22
	mm	18
-pos-	mp	8
	rp	6
-pat-	lp	6
	rp	6
-sit-	ns	8
	rs	5
	ss	5
-sat-	ns	12
	rs	5
	ss	16
-tat-	lt	16
	nt	46
	rt	11
-tur-	ct	6
	et	19
	nt	8
-tic-	ct	13
	nt	7
	pt	13
-tor-	ct	33
	nt	6
-ter-	ct	8
	et	9
	nt	8
	pt	44
-tin-	nt	6
	rt	6

Table 1-5

PREFIXES FROM MULTISYLLABLE WORDS

Weak Prefixes		
Prefix	Following Letter String	No. of Occurrences Of Prefix Preceding Given Letter String
ana	cl	4
	gl	6
apo	cr	4
	str	4
auto	cr	4
	gr	4
	tr	4
deni	gr	4
	tr	8
demo	cr	12
	gr	6
epi	gr	12
	sc	7
	cl	4
irre	fr	5
	pr	7
	tr	6
iso	cr	4
	tr	4
photo	gr	7
	tr	4
tele	gr	8
	sc	6
au	sc	6
	str	5
ca	j	4
	pr	4
	sc	5
hy	dr	85
	gl	5
ma	cr	18
	j	9
	tr	15

Table 1-5 (Cont.)

Prefix	Following Letter String	No of Occurrences of Prefix Preceding Given Letter String
mi	cr	69
	thr	4
pro	gl	6
	pr	4
sa	cr	8
	pr	5
su	bl	6
	pr	11
	sc	5
vi	br	8
	sc	5
	tr	4

Strong Prefixes		
Prefix	Defining Letter String	No. of Occurrences of Prefix With Given Letter String
at	tm	15
	ttr	11
ap	ppl	15
	ppr	46
an	ndr	18
	ngl	9
	nh	6
	nth	20
	nthr	35
em	mbl	12
	mbr	20
im	mbr	5
	mpl	21
	mpr	66
com	mpl	28
	mpr	13
vul	lc	6
	ln	4

Table 1-5 (Cont.)

Prefix	Defining Letter String	No. of Occurrences of Prefix With Given Letter String
cc	cc	4
	ccl	10
	cst	4
ob	bj	19
	bs	21
	bsc	9
	bst	6
	bstr	5
	bt	6
ab	bd	4
	bn	7
	bs	19
circum	mc	5
	mf	7
	mr	5
	mscr	6
	mst	10
	mv	10
inter	rcl	5
	rj	6
	rpr	9
	rsp	6
hyper	rcr	4
	rpl	4
	rtr	5

are investigated, will be especially interesting for this group. It is, in fact, the next step, in which the various applications and implications of the structurally defined affixes are investigated, that the utility and therefore the validity of these structural definitions will be tested.

ACKNOWLEDGMENT: The author wishes to thank Dan L. Smith for writing many of the computer programs used in deriving the affixes.

REFERENCE

1. J. L. Dobly and H. L. Resnikoff, "The Nature of Affixing in Written English," presented at AMTCL Meeting, Jul 1964; published in Mechanical Translation, Vol. 8, Oct 1965.

2. PART-OF-SPEECH IMPLICATIONS OF AFFIXES*

L. L. Earl

In a highly inflected language, the structure of a word is indicative of its syntactic role. A relationship between form and part-of-speech might also be expected in English, a language not highly inflected but closely related to more inflected languages. Such a relationship was noted by J. Dolby and H. Resnikoff[1] who show that a high percentage of a set of words called "elementary words" (roughly equivalent to the set of one-syllable words) can be used as nouns, adjectives, or verbs, while a high percentage of the remaining multisyllable words can be used only as nouns or adjectives. If this relationship can be regarded as a general rule, and if subrules can be developed to cover the considerable number of exceptions to the general rule, it will be possible to identify part-of-speech by algorithm. Intuitively, it would be expected that prefixes and suffixes are key structural elements; this expectation is reinforced by the structure of the European languages whose beginnings and endings indicate the grammatical properties of words.

A logical step in an effort to classify words from their structure is to examine the relationship between the affixes of words and their part-of-speech possibilities as listed in a dictionary. The part-of-speech information from The Shorter Oxford Dictionary[2] and from the Merriam Webster New International Dictionary[3] was recorded on magnetic tape. A computer was used to correlate the affixes of words with their part-of-speech possibilities. A total of 73,582 words was recorded, but, of course, not all of these words contain affixes.

---

*This work was supported in part by the Office of Naval Research; the computer time was supported by the Independent Research Program of Lockheed Missiles & Space Company.

The first problem encountered is that of selecting a list of affixes. Two sets of affixes have been selected, the first being the operationally defined affixes derived from dictionaries solely on graphemic evidence,[4,5] and the second being all "beginnings or endings" listed in A Dictionary of Modern English Usage[6] which were not already on the first list. Both lists are given in Table 2-1. The inflectional suffixes ed and ing and the adverbial ly were not considered in this study because they have well recognized implications. It is believed that the number of words ending in ed, ing, or ly whose parts-of-speech differ from the expected is small enough so that such words can be listed as exceptions.

The second problem encountered is that of determining when an affixing unit is acting as an affix in a given word, as re is a prefix in react but not in read. This problem is complicated by an uncertainty as to what the words prefix and suffix signify. It is difficult to determine from the definitions currently in use to what unit an affix is expected to attach (word, stem, or syllable), to what extent the function of an affix is semantic, and to what extent the affix should indicate phonetic syllabic boundaries (as pre indicates syllabic boundaries in prefix but not in preface). Since we hope to use affixes in determining part-of-speech from form alone, we will use a formal definition. For purposes of this study, an affix will be recognized as an affix under only two formal and reproducible conditions. First, the unit to which any affix attaches must contain one or more vowel strings. Second, the unit to which any prefix attaches must begin with an admissible initial consonant string, and the unit to which any suffix attaches must end with an admissible terminal consonant string. The admissible initial and terminal strings, whose derivation is given in Reference 1, are listed in Table 2-2. Refinements of these rules are possible, to produce a closer correspondence with any given definition, but these criteria seem adequate for our purposes.

Table 2-1

AFFIXES SELECTED FOR CORRELATION

Affixes Set I

Prefixes			Suffixes		
a	dis	ob	a	ia	lock
ab	e	out	ah	ic	man
ac	ec	photo	al	ie	ment
ad	em	pro	an	in	ness
al	en	re	ar	is	o
an	epi	sa	at	ial	on
ap	ex	sac	age	ier	or
at	hy	sub	ant	ile	ot
ana	hyper	sun	ard	ine	ow
ape	i	tele	ate	ion	ock
auto	im	trans	able	ish	tation
be	in	un	ee	ism	ue
ca	incon	uncon	el	ist	um
circum	inex	vi	en	ite	us
com	inter	vul	er	ium	ure
con	irre		et	ive	ward
cy	lu		ey	ler	y
de	ma		ent	let	
demo	mi		eon	land	
deni	mis		ful	less	

Affixes Set II

Prefixes		Suffixes		
air	for	ac	ise	ty
aero	fore	al	ist	ular
bi	hecto	as	ity	valent
by	homo	cy	ize	ways
bye	non	ex	ible	worthy
brain	para	eer	iana	
co	self	ern	lily	
conti	semi	est	logy	
deca	super	ette	latry	
deci	vice	genic	phile	
demi	yester	ix	th	

Table 2-2

INITIAL AND TERMINAL STRINGS

ADMISSIBLE INITIAL CONSONANT STRINGS OF CVC WORDS

B	N	BL	GL	SH	TR	SCH
C	P	BR	GN	SK	TW	SCR
D	Q	CH	GR	SL	WH	SHR
F	R	CL	KN	SM	WR	SPH
G	S	CR	KR	SN		SPL
H	T	DR	PH	SP		SPR
J	V	DW	PL	SQ		STR
K	W	FL	PR	ST		THR
L	Z	FR	RH	SW		THW
M		GH	SC	TH		

ADMISSIBLE FINAL CONSONANT STRINGS OF CVC WORDS NOT ENDING WITH E

B	BB	MP	SH	GHT
C	CH	ND	SK	LCH
D	CK	NG	SM	LPH
F	CT	NK	SP	LTH
G	DD	NN	SS	MPH
H	FF	NT	ST	MPT
K	FT	NX	TH	NCH
L	GG	PH	TT	NTH
M	GH	PT	WD	NTZ
N	GN	RB	WK	RCH
P	LD	RC	WL	RSH
R	LF	RD	WN	RST
T	LK	RF	XT	RTH
W	LL	RK	ZZ	SCH
X	LM	RL		TCH
Z	LP	RM		
	LT	RN		
	MB	RP		
	MM	RR		
	MN	RT		

To correlate the affixes in Table 2-1 with parts of speech, a computer program was written to examine all double standard* words with two or more vowel strings. It sorted out all words that had an affix, that is, a beginning or ending that matched a member of the affix list and met the established criteria. Each of these words had a part-of-speech** string given for it, that is, the list of parts-of-speech possible for that word. Since the dictionaries do not always agree, the string is taken as the parts-of-speech that are associated with standard meanings of the word in either dictionary. The program associated the part-of-speech string of a given word with that word's prefix or suffix. Up to nine different strings could be associated with an affix. For each affix, a count of the number of words with that affix was made for each encountered part-of-speech string, with the counts divided according to the number of syllables in the words. The following example will help to clarify.

The result for the prefix inter is shown in Fig. 2-1. A 1 indicates presence in the dictionary of the part-of-speech identified by the abbreviation at the head of the column. Thus, the first line of Fig. 2-1 indicates that the first part-of-speech string encountered in the words prefixed with inter was noun and verb, and that there were 23 total words with this part-of-speech string, one of them a two-vowel-string word and 22 of them three-vowel-string words. The next line shows that there were three total words with the string noun, adjective, and verb, one of them a two-vowel-string word and two of them three-vowel-string words. This continues until the tenth line, which indicates that more than nine part-of-speech strings had been encountered, at

*To avoid the complication of considering archaic or little-used words, only words having a standard meaning in both dictionaries were used.

**The parts of speech recorded on tape are as follows: noun (N), adjective (AJ), verb (V), adverb (AV), preposition (PR), conjunction (CJ), pronoun (PN), interjection (IJ), past verb (PV). The category other (OT) was used whenever the dictionary gave some part of speech other than the nine listed; OT comprises mainly participles and collective nouns.

INTER

N	AJ	V	AV	PR	CJ	PN	IJ	PA	OT	NO. OF OCCURRENCES VS. VOWEL STRINGS 2VS	3VS	4VS	5VS	6VS	7VS	TOT
1	0	1	0	0	0	0	0	0	0	1	22	0	0	0	0	23
1	1	1	0	0	0	0	0	0	0	1	2	0	0	0	0	3
1	0	0	0	0	0	0	0	0	0	1	7	27	22	4	3	54
0	0	1	0	0	0	0	0	0	0	0	25	10	6	3	0	44
0	1	0	0	0	0	0	0	0	0	0	1	23	15	7	0	46
0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1
1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1
1	1	0	0	0	0	0	0	0	0	0	2	7	4	3	0	16
0	1	0	0	0	0	0	0	0	1	0	1	7	4	1	0	13

MORE THAN 9 UNIQUE POS. PATTERNS ENCOUNTERED

ABBREVIATION KEY

N	noun	PR	preposition	PA	past
AJ	adjective	CJ	conjunction	OT	other
V	verb	PN	pronoun	VS	vowel string
AV	adverb	IJ	interjection	TOT	total sample

VS columns show the part-of-speech distribution for 2, 3, 4, 5, 6 and 7 vowel-string words. The TOT column shows the part-of-speech distribution in the total sample.

Fig 2-1 Example of Affix Statistics Output by the Computer Program

which point the program terminated the examination of this affix. Note that the column headed TOT shows the distribution according to part-of-speech of all words prefixed with inter and the columns headed Nvs show the distribution according to part-of-speech of words with N vowel strings. The distribution according to vowel strings was obtained because it had been noted that there was a general tendency for the percentage of noun-adjective words to increase with the number of syllables.

Study of the part-of-speech distributions of the words with affixes in Set I shows that the words with a given affix have an average of eight or more part-of-speech combinations associated with them, and, in general, there is wide distribution of the words among the different part-of-speech strings. In fact, the results indicate that it will be impossible to assign a 100 percent unique part-of-speech string to a word on the basis of its affixes. What should be possible is to establish an algorithm which will be 95 percent correct in assigning an "inclusive" part-of-speech string, by which we mean a string which will include all of the dictionary-assigned parts-of-speech, but which may include some extraneous parts-of-speech.

Since, as already noted, the majority of multisyllable words can be used only as nouns or adjectives, this will be the point of departure in deriving a part-of-speech algorithm. All words which do not behave as nouns, or adjectives, or nouns and adjectives only are to be considered exceptional, to be listed or to be identified as exceptional by examination of their affixes. The algorithm will be constructed to identify the exceptions, leaving the rest to be given the basic assignment of noun-adjective for multisyllable words or noun-adjective-verb for one-syllable words.

Because they are manageably few, all adverbs not ending in ly, and all prepositions, conjunctions, interjections, and irregular past tense verbs can be removed and put in a special exception list. This leaves combinations of noun, adjective, verb, and "other"

to deal with, where "other" comprises participial forms and collective nouns. Regular forms of participles can be recognized by the inflectional endings ing or ed and irregular forms of participles and collective nouns are few enough so that they can be added to the exception list. (So also can all words which end in ing or ed but are not participial forms.) Seven possible part-of-speech combinations remain:

(1)	noun	N
(2)	adjective	AJ
(3)	noun and adjective	N-AJ
(4)	verb	VB
(5)	noun and verb	N-VB
(6)	adjective and verb	AJ-VB
(7)	noun, adjective, and verb	N-AJ-VB

Since most nouns can be used as adjectives, and since the AJ-VB combination is uncommon except for participles, which are already taken care of, the seven combinations can be reduced to four by merging 3 with 1, and 5 and 6 and 7, to give:

(1)	noun and adjective	NA
(2)	adjective	AJ
(3)	verb	VB
(4)	verb and (noun and/or adjective)	NAVB

To put it another way, there are two large classes of multisyllable words, NA and NAVB, which must be distinguished. In addition, the class AJ must be distinguished from the NA and the class VB from the NAVB. Whenever these distinctions cannot be made with 95 percent accuracy, assignments will be made to the inclusive set.

The construction of the algorithm thus becomes quite simple, a matter of studying the distribution of the part-of-speech strings for each affix, ignoring any part-of-speech other than noun, adjective, or verb. In accordance with the 95 percent criteria, an affix for which 95 percent of the words with that affix have a single part of speech, either

AJ or VB, will be classified as "adjectival" or "verbal," respectively, and the algorithm will simply assign words containing such an affix to the AJ or the VB class instead of to the basic NA class. Affixes for which 95 percent of the words are nouns and/or adjectives, but not verbs, may be considered as "neutral," since words containing them behave as nouns and/or adjectives in accordance with the general rule. An affix, however, for which 5 percent of the words (and more than 5 words) have a verb usage will be classified "noun-verbal," and words containing such an affix will be assigned to the NAVB class. As already indicated, all words which do not contain an affix and which are not in an exception list are classified as NA if multisyllable and NAVB if one syllable.

It must be realized that a good many ambiguities will be introduced by this algorithm. For example, for words prefixed with inter, 71 of the 211 words in our data set have a verbal usage, with further breakdown as follows:

noun and verb	23		
noun adjective and verb	3	NAVB	27
	or		
adjective and verb	1		
verb	44	VB	44

Accordingly, words beginning with inter will be assigned to the NAVB class, obtaining the correct inclusive part-of-speech for 71 words at the cost of introducing the extraneous part-of-speech VB to the 140 well-behaved NA words. The situation is worse in the ambiguity between the AJ and the NA classes. For example, although about 8 percent of words ending in the suffix ful are adjectives, 34 out of the total 169 have a noun usage, so rather than take a 20 percent error of omission, ful is regarded as a neutral suffix and an extra part-of-speech has been introduced in 80 percent of the words. By stretching a point, the suffix less can be considered adjectival, since it is 94 percent adjectival, but many other adjective-tending affixes encountered cannot (ic, 54 percent; able, 79 percent, ish, 70 percent; ial, 61 percent; us, 57 percent; mis, 61 percent).

A part-of-speech implication of either NAVB, VB, AJ, or neutral (i.e., NA) has been determined for all of the affixes. These implications are listed in Table 2-3. When there were fewer than five words with a given affix, no assignment was made. The implications of the operational affixes and of the Dictionary of Modern English Usage affixes break down statistically as follows;

	Operational	English Usage
Neutral	33	20
NAVB	77	17
AJ	1	1
VB	0	1

In Table 2-3, some of the affixes have star superscripts. These are affixes with a NAVB implication which in words of four or more syllables may be regarded as neutral, since in the dictionary there were fewer than three 4- to 8-vowel-string words with these affixes which possessed verbal usages. NAVB affixes which are neutral for 5- to 8-vowel-string words were not considered because there are only about 1,250 of these, while there are about 11,250 4- to 8-vowel-string words.

There are some words, of course, which have both prefix(es) and suffix(es). As the part-of-speech tabulations for suffixes were independent of prefixes, and vice versa, there was a possibility of a particularly influential and common affix introducing an extra part-of-speech into the part-of-speech counts of other affixes. For example, suppose that all the words with the prefix "trans" were always nouns except those which ended in verbal suffixes such as er or ate as "transfer" and "translate." Then "trans" would be assigned the implication NAVB when it should have been neutral. To test this possibility, the Set I prefix counts were repeated with all words having nonneutral suffixes omitted from the data set. However, the part-of-speech implication of all prefixes remained the same. Since none of the part-of-speech implications of the prefixes

Table 2-3

PART-OF-SPEECH IMPLICATIONS OF AFFIXES

Affixes Set I

Prefixes	POS Code	Prefixes	POS Code	Suffixes	POS Code	Suffixes	POS Code
a	NAVB	hyper	Nuetral	a	Neutral	ion	NAVB
ab*	NAVB	i*	NAVB	ah	Neutral	ish*	NAVB
ac	NAVB	im	NAVB	al	Neutral	ism	Neutral
ad*	NAVB	in	NAVB	an	Neutral	ist	Neutral
al*	NAVB	inex	Neutral	ar*	NAVB	ite*	NAVB
an*	NAVB	irre	Neutral	at*	NAVB	ium	Neutral
ap	NAVB	incon	Neutral	age*	NAVB	ive	Neutral
at*	NAVB	inter	NAVB	ant*	NAVB	ler	Neutral
ana	NAVB	lu	Neutral	ard	NAVB	let*	NAVB
apo	Neutral	ma	NAVB	ate	NAVB	land	Neutral
auto	Neutral	mi	NAVB	able	Neutral	less	AJ
be*	NAVB	mis	NAVB	ee*	NAVB	lock*	NAVB
ca	NAVB	ob	NAVB	el*	NAVB	man	Neutral
cy	Neutral	out*	NAVB	en	NAVB	ment	NAVB
com*	NAVB	pro	NAVB	er	NAVB	ness	Neutral
con	NAVB	photo*	NAVB	et*	NAVB	o	NAVB
circum	NAVB	re	NAVB	ey*	NAVB	on	NAVB
de	NAVB	sa*	NAVB	ent	NAVB	or*	NAVB
dis	NAVB	sac	Neutral	eon*	NAVB	ot*	NAVB
demo	NAVB	sub	NAVB	ful	Neutral	ow*	NAVB
deni	Neutral	sun	Neutral	ia	Neutral	ock*	NAVB
e	NAVB	tele*	NAVB	ic	Neutral	tation	Neutral
ec	Neutral	trans	NAVB	ie*	NAVB	ue	NAVB
em	NAVB	un	NAVB	in*	NAVB	um*	NAVB
en	NAVB	uncon	Neutral	is*	NAVB	us	Neutral
ex	NAVB	vi*	NAVB	ial	Neutral	ure	NAVB
epi	NAVB	vul	Neutral	ier*	NAVB	ward	Neutral
hy	NAVB			ile*	NAVB	y	NAVB
				ine	NAVB		

Table 2-3 (Cont.)

Affixes Set II

Prefixes	POS Code
air	Neutral
aero	Neutral
bi*	NAVB
by	Neutral
bye	
brain	
co*	NAVB
centi	Neutral
deca	NAVB
deci	NAVB
demi	NAVB
for*	NAVB
fore	NAVB
hecto	
homo	Neutral
non	Neutral
para	NAVB
self	Neutral
semi	Neutral
super	NAVB
vice	
yester	

Suffixes	POS Code
ae	
al	Neutral
as	Neutral
cy	Neutral
ex	NAVB
eer	NAVB
ern	NAVB
est	NAVB
ette*	NAVB
genic	AJ
ix	NAVB
ise	NAVB
ist	Neutral
ity	Neutral
ize	VB
ible	Neutral
iana	
lily	Neutral
logy	Neutral
latry	Neutral
phile	
th	NAVB
ty	Neutral
ular	Neutral
valent	
ways	
worthy	Neutral

Affixes Set III

Suffixes	POS Code
tion	Neutral
sion*	NAVB
tial	Neutral
sial	AJ
tive	Neutral
sive	Neutral
tious	AJ
ous	AJ

changed, it was decided that it was unnecessary to test suffixes on a set from which prefixed words had been removed.

Prefixes were chosen for the test because the suffixes seem to have a stronger influence than prefixes in multi-affixed words, as for example the neutral ism wins over the NAVB ex in "exorcism," and the verbal ize wins over the neutral vul in "vulcanize." Suffixes would thus cause much more of a problem in the prefix counts than prefixes in the suffix counts. The one easily noted exception to the rule of suffix ascendancy is for such words as "automation" and "vulcanization," in which the neutral auto and vul seem to be ascendent over the NAVB ion. However, a consideration of other words in which both prefix and suffix are NAVB, as in "demolition," construction, accession, etc., indicate that there is a group of important suffixes beginning with t or s which failed to show up in the operational definition of affixes. To test this hypothesis, these possible suffixes were subjected to the part-of-speech tests for affixes with the following results:

Suffix	POS Implication
tion	Neutral
sion*	NAVB
tial	Neutral
sial	AJ
tive	Neutral
sive	Neutral
tious	AJ

Examination of the suffix tious led to examination of the weak suffix possibility ous, which, like tious, turned out to have strongly adjectival implications. Undoubtedly, these suffixes do exist and have strong part-of-speech connotations. For the sake of completeness, they have been added to Table 2-3 as Set III.

Whether or not the use of the part-of-speech implications reported in this paper will be adequate to produce 95 percent accurate part-of-speech by algorithmic assignment remains to be seen. They are, of course, guaranteed to produce 95 percent inclusive accuracy on words with listed affixes. It is not yet known how many non-affixed words there are, nor how well they fit the general rules. Before comprehensive testing can take place, it may be necessary to develop more definitive rules for determining when an affix is acting as an affix in a given word.

ACKNOWLEDGMENT: The author wishes to thank Dan L. Smith who wrote the computer program referred to in this paper.

REFERENCES

1. J. Dolby and H. Resnikoff, "On the Structure of Written English Words," Language, 40, 2, April-June 1964
2. The Shorter Oxford English Dictionary on Historical Principles, 3rd ed., revised with addenda, Oxford at the Clarendon Press, 1959
3. Webster's Third New International Dictionary of the English Language, Springfield, Mass., G. C. Merriam Company, Publishers, 1961
4. H. Resnikoff and J. Dolby, "The Nature of Affixing in Written English," Mechanical Translation, 8, June and October 1965
5. L. L. Earl, "Structural Definition of Affixes in Multisyllable Words," (in manuscript)
6. H. W. Fowler, A Dictionary of Modern English Usage, 2nd ed., revised and edited by Sir Ernest Gowers, New York and Oxford, Oxford University Press, 1965

3. ON THE INFLECTION OF WRITTEN ENGLISH ADJECTIVES

H. L. Resnikoff and J. L. Dolby

The part-of-speech algorithm under development is predicated on the assumption that it is possible to determine the parts of speech of English words without the use of extensive dictionaries. But it is by no means evident that the eight traditional parts-of-speech classes* are meaningful reflections of the structural properties of the English language, and it must be supposed that they have relevance to English only insofar as English bears a genetic relationship to Latin. However, the two languages are vastly different in important respects, and there is, therefore, no real reason to believe that the Latin norms are meaningful in the description of English.

The traditional definitions of the English parts of speech do not help to allay the suspicion that the parts-of-speech classes are the product of the desire of the early English grammarians to fit English to the Latin mold. Gleason has written,[1]**

> English grammar is traditionally described in terms of eight parts of speech .... These eight classes are of quite diverse character and validity. The familiar definitions overlap and conflict, or are so vague as to be nearly inapplicable. Some parts of speech gather together a number of not very obviously related types of words. In other cases, the line of demarcation between parts of speech is rather arbitrary.

These views contrast sharply with the basic premise of the Indexing Project. The Project is attempting to index texts by using a sentence dictionary, that is, a collection of the distinct parts-of-speech sequences occurring in English sentences, based on the traditional parts-of-speech classifications with only minor modifications. If, indeed, these classes are meaningless, or if the assignment of English words to these classes

*Noun, pronoun, adjective, verb, adverb, preposition, conjunction, and interjection.
**Page 92.

is capricious, then it is not possible for the sentence dictionary to have much utility in the solution of the indexing problem.

For this reason it is important to show that the traditional parts-of-speech classes do correspond closely to structural properties of English words. In fact, if a close correspondence can be discovered, then it can be used to provide a structural definition of the parts-of-speech classes, and this will have the virtue of essential agreement with existing sources of data, e.g., dictionaries.

There are several distinct ways of illustrating structural properties of parts-of-speech classes. One way is to construct an algorithm that will generate the parts-of-speech class of a given word from the graphemic shape of the word (together with certain other structural information which is independent of the particular word under examination, and without the use of comprehensive dictionaries). It is not yet known to what extent this is possible, although certain progress has been made. For instance, the multivowel-string words ending with a are very uniformly nouns. The authors[2*] showed that the set of one-vowel-string words depleted by the "structure words" and the -le suffixed words form a part-of-speech category; that is, almost all such words belong to the category noun-adjective-verb. Results reported in the first annual report[3] show that it is possible to construct a reasonably straightforward algorithm which will correctly determine the parts-of-speech class of a random sample drawn from a dictionary with an accuracy of between 70 and 80 percent on the standard words. This is not very good in terms of an algorithm that can be used reliably as a component in a functioning, utilitarian, English text processing system. However, it is strong evidence that the traditional parts-of-speech classifications must indeed bear a close relationship to the structural properties of English.

---

*Section 7 and footnote 22.

It might be true that the algorithm reflects the structured assignment of parts of speech but that parts of speech have nothing, or at best little, to do with the structure of English. In other words, it might happen (although the authors believe this to be farfetched) that the traditional classification is orderly but that the order is one imposed by the early grammarians in some complicated way not really related to the direct properties of the language. If this possibility is admitted, it becomes of interest to find some relationship between the parts-of-speech assignments and some clearly significant structural property of English. In this paper we will describe such a relation.

The traditional grammarian, George Curme, distinguishes two types of comparison, i.e., inflection, of adjectives in English.[4*]

> There are two quite different types of inflection employed in comparing English adjectives – the terminational and the analytic
>
> 1. Termination type of comparison. In this type we add to the positive -er to form the comparative and -est to form the superlative: strong, stronger, strongest. This way of comparing adjectives was universal in Old English, but it is now confined to words of one syllable and a large number of words of two syllables, especially these in -er, -le, -y, -ow, -some. . . .
>
> 2. Analytic type of comparison. Here we put more before the comparative and most before the superlative: beautiful, more beautiful, most beautiful. Adjectives and participles with more than two syllables regularly follow this type, also many words with two syllables. . . .[5*]

Gleason defines adjectives as those words which are inflected using the terminational type of comparison described by Curme; words occurring in the environments in which adjectives are found (but which compare) use the analytic type of comparison he calls adjectives.[1***] Both types will be referred to as adjectives in this paper.

---

*See Reference 4 for an extensive discussion of English verb inflection.
**Page 220, 104.B.
***Pages 92–93. There are also a small number of irregularly inflected adjectives.

In Curme's description quoted above, the number of syllables contained in the adjective under examination is important in determining to which type of inflectional paradigm it belongs. In the study of written English the notion of syllable, which is phonological, is not present. It must be replaced by the number of admissible vowel strings contained in the word, according to the method developed in Reference 2. For the present it will be enough if we approximate to that definition by counting the final e in a word as a consonant, and then counting the number of remaining vowel strings (i.e., the number of connected sequences of vowels) in the word. Then Curme's description states that terminational comparison of adjectives is reserved primarily for one-vowel-string words and certain two-vowel-string words containing selected suffixes, whereas analytic comparison occurs for the remaining adjectives.

Of particular interest are the one-vowel-string adjectives. Contrary to Curme's description, there are large numbers of one-vowel-string adjectives which inflect analytically. It has already been remarked that most one-vowel-string words are noun-adjective-verbs, and, hence, in particular they are adjectives. Almost any one of these words provides an illustration of analytic comparison for one-vowel-string words. Thus:

charm , bloat , squint , ring , bound , flash , etc.

That these words are compared analytically is not due to any hypothetical inability to carry the comparative terminational suffixes. Each of the words given in illustration has a corresponding noun form with the suffix -er appended, but in no case is this form the comparative of the word. Thus it would appear that Curme's description does not agree with the facts in any significant way, although his description is traditional.

The traditional description of comparison for one-vowel-string words is in general disagreement with the facts. Nevertheless, it does contain a hidden kernel of truth

which leads to a rather startling structural relationship between certain classes of words. It must be exactly this relationship to which the inadequately phrased traditional description is attempting to draw attention.

Suppose that the set of one-vowel-string words which do not end with the sequence consonant -le is denoted by W .* If a word has a standard usage as a traditional part of speech, in Merriam-Webster's New International Dictionary, third edition, hereinafter abbreviated "MW3," then it will be called a standard noun, standard adjective, standard verb, etc.

CLAIM: (1) The standard adjectives in W which are not standard adverbs are inflected analytically, i.e., by using more and most.

(2) The standard adjectives in W which are also standard adverbs are inflected terminationally, i.e., by using the suffixes -er and -est.

In the following paragraphs we will substantiate this Claim. First, some remarks are in order as to the meaning of the Claim, if indeed it is true. In view of the discussion of the relation of the traditional parts-of-speech classes to structural properties of English, the assertion takes on a special importance. The assertion is that the set of adjectives of a certain graphemically defined type (namely, those that belong to W) can be partitioned into two classes – one containing the analytically inflected adjectives and the other containing the terminationally inflected adjectives – and that this partition can be determined solely from the knowledge of the parts-of-speech classes to which the adjectives belong. Thus, a direct relationship between the traditional parts-of-speech classes and an easily observed structural property is asserted. This lends weight to the traditional classification in a very impressive way.

---

*A more accurate restriction is this: W denotes the set of elementary words, as defined in Reference 2. In particular, almost all of the elementary words are one-syllable words in our dialects, and conversely.

The Claim must be generously interpreted. It would be false to assert that it has no exceptions; what is really meant is that the proportion of exceptions, and even the particular properties of the exceptions, show them to constitute a maverick and rare set of words, which either belong to the nucleus of words with so many meanings or such frequent usage that it is almost impossible to modify or destroy them, or that they belong to the fringes of the current language and can be expected to fade out with time.

Current English is in a state of rapid change. Many people object to many of the changes which, they contend, debase the language. In particular, there has been increasing use if not acceptance of such phrases as drive slow, run quick, fresh cut, etc. The words slow, quick, fresh, etc., as adjectives, have the terminational comparison:

slower , slowest , quicker , quickest , fresher , freshest , etc.

According to the Claim, the words should also be adverbs. If the Claim represents a productive property of English, then such words as slow, quick, fresh, etc., must either lose the terminational inflection as adjectives, or take on the part-of-speech adverb in addition to their other parts of speech. Evidently the latter is just what occurs. But, in reality, these words are not assuming adverbial usage as a current novelty; each of them has adverb meanings in older unabridged dictionaries such as the Merriam-Webster 2nd edition.

We will now turn to the data. To check the hypothesis that the inflection of one-vowel-string adjectives is a function of the adverb part-of-speech class, a random sample was drawn from English Word Speculum, Vol. I.*

In a sample of 11,200 words, randomly distributed, there were 110 one-vowel-string words which had at least the parts-of-speech noun-adjective-verb and were standard with respect to each of these classes. Since 111 is about 0.98 percent of 11,200, and since the Speculum I contains about 75,000 words, one can expect to find about 750 words with these properties in a medium-size dictionary.

Of the 111 words, 95 had no adverbial usage, 13 were standard adverbs, and 3 were nonstandard adverbs. Thus, about 12 percent of the 111 words were standard noun-adjective-verb-adverbs, and one would expect to find a total of about 90 such in a medium-size dictionary.

Of the 95 words which did not have any adverbial usage, only 2 inflected the adjectival form using terminational inflection, i.e., about 2 percent. This supports the first part of the Claim, that standard adjectives which are not standard adverbs are inflected analytically.

Of the 3 words that had nonstandard adverbial usage, 2 had obsolete adverbial usage, and 1 had dialectical adverbial usage. The obsolete words follow the analytic inflection, while the dialectical word follows the terminational inflection. This is not surprising, first because the obsolete forms may be already discarded from the current language, and second because dialectal forms may be quite contemporary and popular,

---

*Reference 6 is the English Word Speculum, whose several volumes are referred to as Speculum I, Speculum II, etc. Speculum I contains more than 73,000 distinct words [the word list of the Shorter Oxford Dictionary (SOX)] together with part-of-speech and status classes from both the SOX and the MW3, ordered in a statistically random fashion. Speculum II contains an extracted word list from Speculum I together with parts-of-speech and status information, organized so that all words with a fixed number of vowel strings are brought together, and within each of these classes, the words are forward alphabetized.

and thus reflect the productive forms of the language. In this instance, the dialectal adverb was black, with inflection blacker, blackest.

Of the 13 standard adverbs in the collection, 5 inflect analytically. This is about 38 percent of the total, and does not verify the second part of the Claim in any significant way. But a sample of 13 words is too small to have any statistical significance. Furthermore, in attempting to analyze the adverb-adjectives that inflect analytically we encounter a lexicographical problem which may prove to be decisive for the limited collection of words which must be examined. Dictionaries typically indicate the terminational inflections of adjectives explicitly; when a terminational inflection is indicated for an adjective, we may be quite certain that it does in fact exist in text samples. However, if a terminational inflection is not explicitly indicated, this may be due to one of several causes: the adjective is inflected analytically; the lexicographer did not work enough on the particular word; or there were a number of terminational inflections for the adjective that appeared in the corpus, but this number was small and therefore discounted. In the last case, one must worry about the smallness relative to the usage of the word, which presents further complications. Therefore, in general, one can be confident of the information explicitly given in dictionaries, but must be wary of information which can only be inferred from the absence of explicit statements. For example, we cannot be certain that the standard adjective-adverb dang does not have the inflection danger, dangest, although these forms are not attested in MW3. But the comparative form is quite unlikely, both because it coincides with a more common word with very different meaning, and also because it is difficult to assign a comparative to a word such as dang for semantic reasons, although the superlative presents neither of these problems.

This last example illustrates yet another difficulty associated with the determination of the analytically inflecting adjectives. There are certain adjectives which do

not occur in the comparative or superlative. For these adjectives, the absence of explicit information about terminational inflections does not necessarily imply the existence of analytic inflections; it may well be that these adjectives cannot support inflected forms for semantic or other reasons, or it may simply be that their frequency of usage is so low that the inflected forms have not yet been observed. The latter is probably true of dang, while the former seems to be a reasonable explanation for the lack of terminational inflection for the adjective-adverb last; for the analytic forms more last and most last do not appear likely.

The 13 standard adverbs are listed in two columns. The left-hand column contains those words with terminational inflection of the adjective; the right-hand column contains those for which the inflections are not terminational.*

stiff	pat
near	dang
keen	south
light	last
dear	snap
cool	
fine	
dry	

In the right-hand column, the words dang and last have already been discussed; the geographical directions north, south, east, and west, all are exceptions to the Claim (as are the ordinal numbers). It may be that terminational forms of pat remain to be uncovered. If all these factors are taken into account, the second part of the Claim may not be in great difficulty after all. But the sample is much too small to be of guidance.

To study the second part of the Claim, we must have a larger collection of adjective-adverbs belonging to the set W . To this end the standard one-vowel-string

*Once again we warn the reader that this does not imply that there are analytic inflections for these words; there may be no observed inflections whatsoever.

noun-adjective-verb-adverb words not ending with consonant -le have been collected from Speculum II. These words, 97 in number, are listed in Table 3-1. Also listed in Table 3-1 are the 5 words of this category that did end in consonant -le to give some indication of just what we are omitting from our collection.

Of the 97 words in the collection, only 60 use the terminational inflection of adjectives; 37 have no such indication in the Merrian-Webster 2nd edition (these two classes are given explicitly in Table 3-1). This represents only about 60 percent agreement with the second part of the Claim, reflecting almost exactly the proportion indicated by the small listed sample of 13 standard adverbs. But now that this substantially larger and complete collection is available, it will be possible to analyze it in a more detailed fashion.

We have partitioned the set of 37 nonterminational words into two parts: the set of words which are standard adjective-adverbs in both the SOX and the MW3, and the set of words that one of these sources indicates a nonstandard adjective or adverb usage. Table 3-2 shows this classification. The notation following the words in the second column indicates the nonstandard usage according to the following conventions. the letters s , c , r , d , and o occurring inside of parentheses refer to standard, colloquial, rare, dialectical, and obsolete usage, respectively. The four positions within the parentheses refer, reading from left to right, to noun, adjective, verb, and adverb usage, respectively. A period (.) in one of the positions indicates that the corresponding usage is not given in the source under consideration. Each parentheses is followed by either the letter x , denoting SOX, or the letter w , denoting MW3.

Of the 13 words in the top part of Table 3-2 the word bias contains an inadmissible vowel string, and really should not appear in the corpus; however, it is the only such word, and it may be simpler for the reader if it is included along with the other words

Table 3-1

STANDARD ONE-VOWEL-STRING NOUN-ADJECTIVE-VERB-ADVERBS FROM SPECULUM II[a]

Consonant -le Words

double	single	tickle	treble	triple

Nonconsonant -le Words With Terminational Adjectival Inflection

blind	flat	mean	shrill	spruce
chance	flush	pat	slack	square
clean	foul	prime	sleek	steep
clear	fresh	prompt	slick	stern
close	full	pure	slight	stiff
cold	glib	queer	slow	straight
cool	grave	quiet	small	sweet
dry	just	right	smart	thin
faint	keen	rough	smooth	tough
fair	lax	sharp	snug	trim
fine	loose	sheer	sour	true
firm	low	short	spare	warm

Nonconsonant -le Words Without Terminational Adjectival Inflection[b]

back	dutch	north	side	squab
bias	east	part	slant	stick
bone	flounce	pi	smash	stump
chock	front	plumb	snap	third
dab	home	rear	snell	west
damn	jam	rush	sole	
darn	last	scale	splash	
dog	mock	shoal	splay	

(a) The various parts of speech of these words are standard in at least the SOX or MW3 and have no other parts of speech.

(b) Note that this does not imply the existence of analytic adjectival inflection.

Table 3-2

NONTERMINATIONALLY INFLECTED WORDS FROM TABLE 3-1

Standard Adjective-Adverbs
in Both SOX and MW3

back	north	squab
bias	part	third
east	plumb	west
home	shoal	
last	slant	

Nonstandard Adjective-Adverb Usage
in Either SOX or MW3

bone	(sss.)x	rear	(ssso)x
chock	(s.ss)w	rush	(s.s.)x
dab	(s.s.)w	scale	(s.s.)x
damn	(s.s.)x	side	(sdso)x
darn	(s.s.)x	smash	(s.s.)x
dog	(sss.)x	snap	(s.s.)x
dutch	(sss.)x and (cess)w	snell	(sssd)w
flounce	(s.s.)x	sole	(sss.)x
front	(sss.)x	splash	(s.ss)x and (sss.)w
jam	(sss.)x and (s.ss)w	splay	(sss.)w
mock	(rss.)x	stick	(s.s.)x
pi	(sd..)x	stump	(sss.)w

at this stage of the argument. It may be worth remarking that it is the only two-syllable word in the collection (in our dialects).

The bottom part of Table 3-2 shows that there is considerable disagreement between the SOX and the MW3 with respect to the classification of adverbs, and to a lesser extent, of adjectives. It is evident that the SOX is much more conservative, i.e., has a higher frequency threshold for the admission of adverbial usage than does MW3. But it is also evident that the SOX principles are in close accord with the second part of our Claim.

If we agree that dictionaries are most reliable when several of them agree, then we will be urged to discard the words in the right-hand column of Table 3-2 when

examining the agreement with the Claim. If this is done,* then there remain 73 words in the collection (including bias), of which only 13 do not have terminational inflection. That is, the second part of the Claim is true for 82 percent of the words.

An examination of these 13 words which do not agree with the second part of the Claim is fruitful. These are the words in the top part of Table 3-2. The three geographic directions east, north, and west appear,** and the ordinal third also appears. It is evidently impossible for ordinals to have comparative or superlative inflections for semantic reasons: the most third is no better than the thirdest. Similar remarks apply, but based on more personal evaluations, for the words home and last, and perhaps for some of the others as well.

Thus it may be that, after semantic considerations have been accounted for, the agreement with the second part of the Claim will be in the 90 percent range. Due to the difficulties inherent in obtaining adequate and complete information to test the Claim, such a level of agreement would be impressive. For the present, however, we will have to satisfy ourselves with the weaker 82 percent agreement.

The relationship between terminationally inflected adjectives and adverbs can be used in the determination of the parts of speech of certain two-vowel-string words. Both the comparative and the superlative of such adjectives are two-vowel-string words (because the words discussed above were all, with the exception of the excludable word bias, one-vowel-string words). But the comparative suffix, -er, coincides with a suffix with a quite different structural role, and therefore can be confused with

---

*We really should eliminate those words in Table 3-1 which do have terminational adjectival inflection but are not standard adjectives and adverbs in both SOX and NW3; but we have not actually done this. It seems that the results would not be much different, although the expenditure of effort would be considerable.

**The reader will recall that Speculum II contains only those words whose parts of speech are included among noun, adjective, verb, and adverb. The word south has other parts of speech, and therefore does not appear in Table 3-1 or Table 3-2; the same is true of dang, which appeared in the random sample discussed.

the latter. It is clear that the comparatives of adjectives are themselves adjectives, and dictionaries often take this fact for granted and do not explicitly indicate that a given word is the comparative form of an adjective. For example, the words cooler and fuller are both listed as nouns but not as adjectives in both SOX and MW3. Clearly, the dictionary user is supposed to recognize that these are, in addition to noun (and perhaps still other) usage words, comparatives of adjectives. This being the case, it is necessary for a parts-of-speech predicting algorithm to distinguish those -er forms which are not comparatives from those that are.

This can be achieved in the following way. Only those one-vowel-string words that are adverbs as well as adjectives compare using terminational inflection; we will assume, in agreement with the second part of the Claim, that all such words do compare in this way. Then a two-vowel-string word ending with -er can be expected to be the comparative of an adjective, say A , if the word is of the form Aer and if A is both an adjective and an adverb. As we have seen, the collection of all one-vowel-string adjective-adverbs is not large;* hence, these can be stored in a dictionary in a parts-of-speech-predicting algorithm.

In illustration, consider the forms cooler and fuller discussed above. They are of the form Aer with A standing for cool and full respectively, both of which are adjective-adverbs in the one-vowel-string word class. Hence, both cooler and fuller are comparatives of adjectives (hence are adjectives) in addition to any other parts of speech properties they may have.

---

*Note that Table 3-1 does not contain all one-vowel-string adjective-adverbs, but rather only those that are also noun-verbs, and such that all four of these parts-of-speech categories are standard for the words involved, and such that no other parts-of-speech classes occur. But these restrictions do not diminish the size of the class by a large factor due to the fact that the one-vowel-string words essentially form a parts-of-speech category, namely, noun-adjective-verb-adverb.

In the general application of the procedure just outlined, it will sometimes be necessary to take into account algorithmic spelling changes. For example, the adjective-adverb dry compares as drier and driest, the y changing to i. Similar consistent changes are described in Reference 4, and will not be further discussed here.

REFERENCES

1. H. A. Gleason, An Introduction to Descriptive Linguistics, New York, Holt, Rinehart, and Winston, 1961
2. J. L. Dolby and H. L. Resnikoff, "On the Structure of Written English Words," Language, Vol. 40, 1964, pp. 167–196
3. J. L. Dolby, L. L. Earl, and H. L. Resnikoff, The Application of English-Word Morphology to Automatic Indexing and Extracting, M-21-65-1, Lockheed Missiles & Space Company, Palo Alto, Calif., April 1965
4. H. L. Resnikoff and J. L. Dolby, "The Inflection of Written English Verbs" (to be published)
5. George O. Curme, English Grammar, New York, Barnes & Noble, 1947
6. J. L. Dolby and H. L. Resnikoff, The English Word Speculum, Sunnyvale, California, Lockheed Missiles & Space Company, 1964

4. AUTOMATIC DETERMINATION OF PARTS OF SPEECH OF ENGLISH WORDS

L. L. Earl

INTRODUCTION

This paper describes the development and details of a procedure for automatically assigning part-of-speech characteristics to English words, largely from graphemic considerations. The development of the algorithm began with the observation of Dolby and Resnikoff[1] that the parts-of-speech associated with one-syllable words are frequently noun (or noun and adjective) and verb, while the parts of speech associated with multi-syllable words are usually noun and adjective only. Development of a working part-of-speech algorithm required the study of exceptions to this general rule so that analytical subrules and exception lists sufficient to automatically identify all such exceptions could be derived. Two avenues for the isolation and study of exceptions were utilized:

(1) Exhaustive sorts of a 73,582 word dictionary on magnetic tape were used to separate and classify words consistent with the general rule from those that were not.

(2) Analysis of possible part-of-speech implications of affixes was carried out, by computer, on the same dictionary.

The resulting algorithm developed utilizes a prepared dictionary of less than 800 words and an affix list of less than 200 entries.

PARTS OF SPEECH USED, AND THEIR ABBREVIATIONS

The tape dictionary used for both analyses contained 73,582 words, with part-of-speech and word status information from The Shorter Oxford Dictionary[2] and the

Merriam Webster New International Dictionary.[3] The tape dictionary is reliable in most respects, since it was made from punched cards transcribed directly from the dictionaries, verified by different personnel, and spot checked periodically during the process. Nevertheless, errors did occur, particularly in the recording of part-of-speech information which was not always understood by the keypunchers. The parts of speech recorded are as follows:

noun	N	adverb	AV	pronoun	PN
adjective	AJ	preposition	PR	interjection	IJ
verb	VB	conjunction	CJ	past verb	PV

In addition, the category "other" (OT) was used whenever the dictionary gave some part of speech other than the nine listed above. OT comprises mainly participles, numerals, articles, and collective nouns. The algorithm was designed to assign these same nine parts of speech (excluding OT) with the addition of four more which were unfortunately subsumed under OT: present participle (PA), past participle (PP), auxiliary verb (AX), and plural or collective noun (NP). The category noun was changed to the category noun-adjective (NA) on the grounds that nearly all nouns can act as adjectives under some circumstances; therefore, although we will try to distinguish AJ from NA, we will not try to distinguish N from NA. Collective nouns will be assigned the string NA and NP to show possible use with either singular or plural verbs.[4] Although a dictionary may show additional or fewer parts of speech for participial forms, their use (or lack of use) as nouns, adjectives, or verbs will be considered here as implicit in the participle assignment, and no attempt will be made to distinguish them. Thus, present participles will implicitly be possible nouns, adjectives, or in a verb phrase, and past participles will implicitly be adjectives, past verbs, or in a verb phrase. An attempt will be made to identify participles which have any other special usages, and to identify irregular past tense and past participial forms.

DESIGN PLAN

In the design of a part-of-speech algorithm, a goal of 95 percent accuracy was set. To begin with, three basic rules were postulated:

Rule A: The part-of-speech string associated with a word containing only one vowel string in its kernel will be NA – VB, where a kernel will be defined as a word stripped of its affixes. Similarly, the part-of-speech string associated with words with multivowel string kernels will be NA.

Rule B: The part-of-speech string associated with a word ending in ed will be PP, and with a word ending in ing will be PA. All PP will also be considered PV. A NA classification will be changed to NP for all words ending in single s.

Rule C: The part-of-speech string associated with a word ending in ly will be AJ – AV.

Rule A is basically a refinement of the original Dolby-Resnikoff hypothesis and depends on the Dolby-Resnikoff definition of a legal vowel string. It also depends on the existence of an operational definition of affixes.[5,6] Rules B and C are a recognition of the most consistently used and meaningful suffixes of English.

Design of the algorithm was conceived of as requiring three steps:

Task 1: Tabulation of the exceptions to Rules B and C

Task 2: Tabulation of special-purpose words, with part-of-speech PR, CJ, PN, or IJ, which are not covered by Rules A, B, or C.

Task 3: Modification of Rule A as much as necessary to achieve 95 percent accuracy, using a study of affixes, or tabulation of exceptions, or both, as a means to this end.

The first two tasks will be discussed first, and then the considerably more involved Task 3 will be summarized. The first two tasks could be accomplished by sorting the dictionary on magnetic tape, as mentioned in the introduction, although it may be of interest that not all of the data handling necessary could be accomplished with a generalized sort routine. 7094 SORT was used, but special-purpose routines were also developed.

DICTIONARY STUDIES

Task 1: Exceptions to Rules B and C

For Tasks 1 and 2 the tape dictionary entries were divided into 2 categories, those with parts of speech (POS) limited to NA, AJ, VB or AV and those having at least one part of speech other than NA, AJ, VB or AV. To find the exceptions to Rule B, the entries in the second category were separated into two lists.

List 1: Words ending in ed, ing, or single s.

List 2: Words not ending in ed, ing, or single s.

According to Rule B, all words in List 1 should be categorized as OT and all those in List 2 should not be. Exceptions to Rule B arising from List 1 are in Table 4-1 and those arising from List 2 are in Table 4-2. Only words in standard usage are shown in any of the tables. There were only 18 words in the exceptions arising from List 1, and these are all shown in Table 4-1. This list of 18 words does not comprise all the words ending in ed, ing, or s which are not categorized as OT, as there are many more of these in the NA, AJ, VB category, also. Fortunately most such category 1 words need not be considered. Words ending in ing need not be considered because their actual parts of speech (usually NA, as for pudding) are subsumed under the participle heading; classifying them as present participles will be correct from the

Table 4-1

EXCEPTION WORDS ENDING IN s, ed, OR ing

From Category I		From Category II	
aliped	NA	across	AJ AV PR
atlas	NA VB	alas	IJ
biped	NA	anything	N(A) AV PN
bonus	NA VB	besides	CJ PN
callus	NA VB	bring	N(A) VB IJ
canvas	NA VB	cross	NA VB AV PR
caucus	NA VB	during	PR
census	NA VB	hoicks	VB IJ
childbed	NA VB	minus	NA AV PR
chorus	NA VB	nothing	N(A) AV PN
circus	NA VB	plus	NA VB AV PR
debarras	VB	something	NA VB PN AV
debus	VB	theirs	PN
disfoliaged	NA	this	N VB AJ PN
embarras	VB	unless	N(A) PR CJ
embed	NA VB	various	AJ PN
embus	VB	whereas	N(A) CJ
exceed	VB	whing	N(A) VB IJ
fissiped	NA		
focus	NA VB		
gossiped	NA		
hocus	NA VB		
hotbed	NA		
hundred	NA		
interbed	NA VB		
lobsided	NA		
milliped	NA		
misdeed	NA		
mohammed	NA		
monied	NA		
[illegible]	NA		
palmiped	NA		
pinniped	NA		
quadruped	NA		
rebus	NA VB		
sacred	AJ		
soliped	NA		
thoroughbred	NA		
vertabed	NA		
watershed	NA		
worsted	NA		

Table 4-2

EXCEPTION WORDS DERIVED FROM LIST II

Irregular Participle and Past Tense Verbs

bet	NA VB PP
beaten	PP
begotten	PV PP
bidden	PV PP
bitten	PP
blown	NA PP
blent	PP
blest	PP
bound	NA VB PP
bode	NA VB PP
borne	AJ PV PP
born	AJ PV PP
bought	AJ VB PP
boundc:2	PP
broke	NA VB PP
brought	PV PP
brant	NA PP
brabien	PP
bracken	NA PP
broken	NA PP
built	NA VB PP
burst	NA VB PV PP
came	NA VB PP
caught	AJ VB PV PP
chosen	NA PP
clad	AJ VB PP
clove	PV PP
clung	PP
cleft	NA PV PP
cloven	PV PP
could	AX
crept	PP
cut	NA VB PP
did	PV PP
done	NA PV PP
drove	NA PV PP
drunk	NA PP
drawn	PP
drew	PV PP
drunken	PP
driven	PP
felt	NA VB PV PP
flown	ABSENT
flew	NA PP
fought	PP
fraught	AJ PP
frozen	PP
gilt	NA PP
given	PP
gone	NA PP
got	PP
ground	NA VB PV PP
grit	NA VB PP
grew	PV PP
graven	PV PP
had	PV PP
held	PP
hewn	PP
hidden	PP
hung	VB PV PP
knit	NA VB PP
known	NA PP
lay	NA VB PP
let	NA VB PP
left	NA VB PP
lent	PP
made	AJ PP
met	NA PP
meant	PP
might	NA PV PP
misgotten	PV PP
mown	PV PP
molten	PP
ought	NA VB AV PX PP
paid	AJ PP

Table 4-2 (Cont.)

Irregular Participle and Past Tense Verbs

pent	NA	PP			split	NA	VB	PP	
put	NA	VB	PV	PP	spent	VB	PP		
quit	NA	VB	PP		spoken	PP			
rang	VP				stole	NA	VB	PV	PP
read	NA	VB	VP	PP	strung	PP			
reft	NA	PP			stung	PP			
rent	NA	VB	PP		stricken	PV	PP		
rung	NA	PP			stolen	NA	PP		
run	NA	VB	PP		sung	PP			
said	AJ	PP			sunk	AJ	PP		
saw	NA	VB	PP		sunken	PP			
sewn	PV	PP			swam	PP			
sent	NA	PP			sworn	PP			
should	AJ	AX			swollen	PP			
shod	AV	PP			taught	AJ	PV	PP	
shone	PV	PP			thrown	AJ	PP		
shrunk	PP				thought	PV	PP		
shook	NA	VB	PP		threw	PP			
shorn	NA	PV	PP		thrust	NA	VB	PV	PP
shot	NA	VB	PP		told	PP			
shaken	PP				torn	PV	PP		
shapen	PP				trodden	PP			
shotten	PP				went	PP			
shaven	PP				were	PP			
riven	PP				wet	NA	VB	AV	PP
slunk	NA	PP			widen	PV	PP		
slit	NA	VB	PP		woke	PV	PP		
slew	NA	VB	PP		worn	PP			
smelt	NA	VB	PP		would	NA	AX		
sought	PP				wound	NA	VB	PV	PP
soaden	AJ	VB	PV	PP	wove	NA	PP		
spoke	NA	VB	PP		sewn	PV	PP		
spread	NA	VB	PV	PP	written	PV	PP		
sprung	PP				wrought	AJ	PP		
spun	PP				wrung	PP			

Irregular Plural or Collective Nowns

apache	NA NP
cattle	NA VB NP
carp	NA VB NP
caribou	NA NP
chinook	NP
cherubim	NA NP
dice	NA VB NP
couple	NA VB NP
crane	NA VB NP
crustacea	NP
cutlery	NA NP
data	NP
dicta	NP
fish	NA VB NP
foe	NA NP
fulcra	NA NP
game	NA VB NP
geese	NP
genera	NP
grouse	NA VB NP
help	NA NP IJ VB
hosiery	NA NP
ice	NA VB NP
ingesta	NP
irish	NA AJ NP
japanese	NA NP
lice	NP
like	NA VB AV PR CJ NP
lynx	NA NP
marabou	NA NP
maxima	NP
mice	NP
milanese	NA NP
men	NP
pence	NP
people	NA VB NP
perch	NA VB NP
pike	NA VB NP
poultry	NA NP
regalia	NA NP
rice	NA VB NP
roe	NA NP
secreta	NA NP
seraphim	NA NP
sheep	NA VB NP
snipe	NA VB NP
sperm	NA NP
spawn	NA VB NP
spoor	NA VB NP
squid	NA VB NP
steer	NA VB NP
strata	NP
starfish	NA NP
swine	NP
tripe	NA NP
tuna	NA NP
viscua	NP
young	NA NP

point of view of an "inclusive" part-of-speech. By an "inclusive" part-of-speech string is meant that string which is sure to contain all the parts of speech attributed to the word by either dictionary, but which may also contain one more or, rarely, two more parts of speech. Since use of inclusive part of speech becomes necessary in Task 3, its justification will be discussed when Task 3 is taken up. Words ending in ed which are not OT but are either AJ or VP will similarly be correct from an inclusive part-of-speech viewpoint. However, some non-past-participles ending in ed are NA. Some of these can be identified by the use of suffixes, to be discussed later. All others are given in Table 4-1. Most words ending in single s will have the correct inclusive part-of-speech assigned by the Rule B – Rule A combination; all exceptions are also given in Table 4-1. Table 4-1 thus contains all the necessary exception words ending in s, ing, or ed.

Table 4-2 shows participles, past tense verbs, and plural or collective nouns which cannot be recognized from s ing, or ed endings. It is a subjective list derived from the 1,380 or so entries in List II which had OT designations. To make automatic determination of part of speech substantially faster than dictionary lookup, the exception lists were kept as small as possible. The 1,380 entries in List II with OT designations include numerals, obscure collective nouns (e.g., herb, scrub), words which become collective only when s is added (e.g., geriatric), and some errors in judgement by the keypuncher as well. It is believed that this list can safely be reduced to the words shown in Table 4-2 without dropping below the goal of 95 percent accuracy. All of the irregular participles and past tense verbs have been retained, but only a partial list of collective nouns has been included.

Exceptions to Rule C were found by extracting from the entire dictionary all words which, though ending in ly, were not adverbs, or conversely, though not ending in ly,

were adverbs. Contrary to expectations, there were a large number of such words (slightly over 1,500). Many of these words were judged rare, or rare in the usage in question (e.g., dog-fly as NA, or dash, pi, rife, smell, thistle, as AV); others could be predicted by an extension of the affix lists, to be discussed later. In accordance with the philosophy of maintaining a relatively short exception list without sacrificing too much accuracy, this list of 1,500 words has been reduced to a list of 357 of the common words which are exceptions to Rule C, as shown in Table 4-3.

Task 2: Tabulation of Special-Purpose Words Which are not Covered by Rules A, B, or C.

For Task 2, List II was again used. To review, List II contains all the words which

- Have at least one standard meaning corresponding to a part of speech other than NA, VB, AJ, or AV (the parts of speech assigned by Rules A, B, C)
- Have all "irregular" entries removed (fragments, etc.)
- Have all words ending in ed, ing, or s removed (the suffixes covered by Rule B)

By extracting from List II all words with standard meaning corresponding to a part of speech PR, CJ, IJ, or PN we should get an exhaustive list of those structural, special-purpose words which are so important in a mechanized handling of English.

Table 4-4 shows the 249 function words so extracted. Note that Fig. 4-1 lists the 18 function words ending in s or ing. Because of a difficulty in sorting, certain OF words (27) which are irregular adverbs and collective nouns are included in this group, although they should appear in Fig. 4-3 instead. Because of a misunderstanding by keypunchers in the original creation of the dictionary, some important pronouns were not so classified in the Merriam Webster part-of-speech designations and are therefore missing from the list (I, your, his, we them, our, us, their they). The word as has

Table 4-3

COMMON EXCEPTION WORDS TO RULE C

Word	POS	Word	POS
backhand	N AJ V AV	broad	N AJ AV
bare-backed	AJ AV	cheap	N AJ AV
bare-headed	AJ AV	clean	N AJ V AV
between-whiles	N AV	damn	N AJ V AV
co-ally	N	double	N AJ V AV
cock-sure	AJ AV	east	N AJ V AV
counter-clockwise	AJ AV	faint	N AJ V AV
counter-current	N AJ AV	fair	N AJ V AV
criss-cross	N AJ V AV	false	AJ AV
cross-country	N AJ AV	fine	N AJ V AV
cross-grained	AJ AV	flat	N AJ V AV
double-quick	N AJ V AV	flush	N AJ V AV
free-hand	N AJ AV	forte	N AJ AV
god-damn	N AJ V AV	foul	N AJ V AV
half-and-half	N AJ AV	free	AJ V AV
half-way	N AJ AV	fresh	N AJ V AV
happy-go-lucky	N AJ AV	front	N AJ V AV
harum-scarum	N AJ AV	full	N AJ V AV
now-a-days	N AJ AV	hard	N AJ AV
off-hand	AJ AV	hence	AV
oft-times	AV	here	N AJ AV
old-fashioned	N AJ AV	heynne	AV
over-hard	AJ AV	hone	N AJ V AV
over-long	AJ AV	ill	N AJ AV
over-supply	N V	just	N AJ V AV
point-blank	N AJ AV	keen	N AJ V AV
post-haste	N AJ AV	large	N AJ AV
pot-belly	N	last	N AJ V AV
right-handed	AJ AV	late	AJ AV
rough-and-tumble	N AJ AV	lax	N AJ V AV
second-class	N AJ AV	least	N AJ AV
side-saddle	N V AV	long	N AJ V AV
single-handed	AJ AV	loose	N AJ V AV
sky-high	AJ AV	loud	AJ AV
so-and-so	N AJ AV	low	N AJ V AV
topsy-turvy	N AJ V AV	maybe	N AJ AV
under-arm	AJ AV	mean	N AJ V AV
up-country	N AJ AV	much	N AJ AV
up-grade	N V AV	needs	AV
up-stream	AJ AV	new	N AJ AV
up-wind	N AJ V AV	nope	N AV
aft	N AJ AV	north	N AJ V AV
back	N AJ V AV	odd	N AJ AV
bad	N AJ AV	oft	AV
blind	N AJ V AV	old	N AJ AV

Table 4-3 (Cont.)

Word	POS
part	N AJ V AV
pat	N AJ V AV
prompt	N AJ V AV
queer	N AJ V AV
quick	N AJ V AV
quite	N AV
real	N AJ AV
right	N AJ V AV
sic	N V AV
snug	N AJ V AV
soon	AJ AV
sour	N AJ V AV
square	N AJ V AV
straight	N AJ V AV
thence	AV
twice	N AJ AV
west	N AJ V AV
worse	N AJ AV
wrong	N AJ V AV
yea	N AV
yep	AV
yes	N V AV
ablaze	AJ AV
adrift	AJ AV
afield	AJ AV
aground	AJ AV
ajar	AJ AV
akin	AJ AV
alias	N AV
alike	AJ AV
alive	AJ AV
almost	AJ AV
alone	AJ AV
aloud	AV
always	AV
amuk	N AJ AV
andante	N AJ AV
apart	AJ V AV
apiece	AV
aright	AV
askew	AJ AV
astray	AJ AV
away	AJ AV
awful	AJ AV
awhile	AV
broadcast	N AJ V AV
broadside	N AJ V AV
broadway	N AJ AV
complete	AJ V AV
costly	AJ
counter	N AJ V AV
curly	AJ
direct	N AJ V AV
dirty	AJ V AV
doily	N
doubtless	AJ AV
earthly	AJ
even	N AJ V AV
ever	AV
farther	AJ V AV
farthest	AJ AV
further	AJ V AV
furthest	AJ AV
galore	N AJ AV
gratis	AJ AV
gully	N V
hearty	N AJ AV
heaven	N AJ V AV
herein	AV
hereof	AV
higher	N AJ AV
highest	N AJ AV
hilly	AJ
holly	N AJ
holy	N AJ
imply	V
indeed	AV
indoor	AJ AV
indoors	AV
jelly	N V
july	N
largo	N AJ AV
later	N AJ AV
latest	N AJ AV
lengthways	AV
lento	AJ AV
lesser	AJ AV
lily	N AJ
longways	N AV
lower	N AJ V AV

Table 4-3 (Cont.)

Word	POS	Word	POS
lowest	N AJ AV	thereat	AV
matchless	AJ AV	thereof	AV
measly	AJ	threefold	AJ AV
merry	N AJ AV	tidy	N AJ V AV
middling	N AJ AV	topside	N AJ AV
midstream	N AV	twofold	N AJ AV
mighty	N AJ AV	upright	N AJ V AV
molly	N	very	N AJ AV
never	AV	vivace	AJ AV
nohow	AV	weary	AJ V AV
noways	AV	wellnigh	AV
offshore	AJ AV	whereat	AV
offside	N AJ AV	wherein	AV
often	AV	whereof	AV
open	N AJ V AV	whereon	AV
outboard	N AJ AV	wily	AJ
outright	AJ AV	abundant	AJ AV
perchance	AV	adagio	N AJ AV
perforce	N AJ AV	aflutter	AJ AV
perhaps	N AV	afterward	N AV
piano	N AJ AV	afterwards	N AV
plenty	N AJ AV	aglitter	AJ AV
pronto	AV	akimbo	AJ AV
proper	N AJ AV	alibi	N V AV
rally	N V	alongshore	N AJ AV
ready	N AJ V AV	already	AV
reckless	AJ AV	amidships	AJ AV
reply	N V	anywhere	N AV
restless	N AJ AV	apriori	N AJ AV
reverse	N AJ V AV	bareback	AJ AV
sally	N V	barefoot	AJ AV
scaly	AJ	butterfly	N AJ V
seldom	AJ AV	careless	N AJ AV
sheepish	AJ AV	cowardly	AJ V
slantways	AV	crescendo	N AJ V AV
slantwise	AJ AV	elsewhere	AV
smelly	AJ	evermore	AV
sooner	N AV	extempore	AJ AV
speedy	AJ AV	falsetto	N AJ AV
starboard	N AJ V AV	family	N AJ
steadfast	AJ AV	forehand	N AJ AV
steady	N AJ V AV	foremost	AJ AV
sudden	N AJ AV	forever	N AV
sully	V	forzando	AJ AV
tally	N V	furthermore	AV

Table 4-3 (Cont.)

Word	POS	Word	POS
henceforth	AV	therefore	N AV
hereabout	AV	thereto	AV
hereafter	N AV	thereupon	AV
hereby	AV	thousandfold	N AJ AV
hitherto	AJ AV	twelvefold	AJ AV
homily	N	unaware	AJ AV
however	AV	underground	N AJ AV
howsoever	AV	underhand	N AJ V AV
hundredfold	N AJ AV	ungodly	AJ
impromptu	N AJ V AV	unholy	N AJ
inasmuch	AV	unruly	AJ
innuendo	N V AV	unsightly	AJ
insomuch	AV	unworldly	AJ
legato	N AJ AV	uppermost	AJ AV
lentamente	AJ AV	upriver	AJ AV
lifelong	AJ AV	verbatim	N AJ AV
manyways	AV	whereabout	N AV
miserly	AJ	whereby	AV
nevermore	AV	wherefore	N AV
ninefold	N AJ AV	whereupon	AV
outermost	AJ AV	wholesale	N AJ V AV
overboard	AV	yesterday	N AJ AV
overhand	N AJ V AV	altogether	N AJ AV
overhead	N AJ AV	beforehand	AJ AV
overland	N AJ AV	contrariwise	AJ AV
overnight	N AJ V AV	everyway	AV
overtime	N AJ V AV	everywhere	N AV
piecemeal	N AJ V AV	fortissimo	N AJ AV
sevenfold	AJ AV	henceforward	AV
sforzando	N AJ AV	heretofore	N AJ AV
sforzato	AJ AV	incognito	N AJ AV
sideway	N AJ AV	malapropos	N AJ AV
sideways	AJ AV	melancholy	N AJ
sixtyfold	N AJ AV	moderato	AJ AV
somehow	AV	monopoly	N
sometime	AJ AV	nevertheless	AV
someway	AV	oftentimes	AV
somewhere	N AV	pianissimo	N AJ AV
staccato	N AJ V AV	pizzicato	N AJ AV
straightaway	N AJ AV	prestissimo	N AJ AV
thenceforth	AV	sometimes	AV
thereabout	AV	thenceforward	AV
thereabouts	AV	unawares	AV
thereafter	AV	underhanded	AJ AV
thereby	AV		

Table 4-4

SPECIAL-FUNCTION WORDS

Word	POS
he	N AJ V PN IJ
she	N AJ PN OT
the	AJ AV OT
me	N PN
a	N AJ V PR OT
dead	N AJ AV OT
mid	N AJ AV PR
bold	N AJ AV OT
and	N AV CJ
beyond	N AV PR
round	N AJ V AV PR
thud	N V AV IJ
whence	N AV CJ
since	AV PR CJ
once	N AJ AV CJ
bounce	N V AV IJ
jee	N AV IJ
strange	AJ AV IJ
like	N AJ V AV PR CJ OT
while	N V CJ
vile	AJ AV OT
same	N AJ AV PN
some	AJ AV PN OT
thine	PN OT
mine	N AJ V PN
one	N AJ V PN OT
none	N AV PN OT
prone	N AJ AV OT
woe	N AV IJ OT
ere	N AV PR CJ
there	N AJ AV PN
where	N AV CJ PN
maugre	AV PR
fore	N AJ AV PR IJ
more	N AJ AV PN
wise	N AJ V AV OT
whose	AJ PN OT
ante	N V AV PR
save	N V PR CJ
bove	AV PR
aye	N AV IJ
off	N AJ V AV PR

Word	POS
if	N CJ
self	N AJ V PN
of	N PR OT
stag	N AJ V AV OT
dang	N AJ V AV OT
whang	N V AV OT
each	AJ AV PN
which	AJ PN
rich	N AJ AV
such	AJ AV PN
nigh	N AJ V AV PR
though	AV CJ
through	N AJ AV PR
plash	N V AV IJ
swash	N AJ V AV IJ
swish	N AJ V AV IJ
with	N AV PR
both	N AJ AV CJ PN
south	N AJ V AV PR
crack	N AJ V AV IJ
stock	N AJ V AV OT
rank	N AJ V AV OT
plunk	N V AV IJ
whisk	N V AV IJ
all	N AJ AV PN OT
fell	N V AV PV
well	N AJ V AV OT
till	N V PR CJ
still	N AJ V AV CJ
him	N PN
whom	PN
from	PR
cum	AJ AV PR
than	PR CJ
been	PV
then	N AJ AV CJ
when	N AV CJ PN
in	N AJ V AV PR
lain	PV
on	N AJ AV PR
con	N AJ V AV PR
down	N AJ V AV PR

Table 4-4 (Cont.)

Word	POS
who	N PN IJ
no	N AJ AV
pro	N AJ AV PR
so	N AJ AV CJ PN
to	N AV PR
sweep	N V AV
plop	N V AV IJ
pop	N AJ V AV IJ
up	N AJ V AV PR
bar	N AJ V PR
dear	N AJ V AV IJ
near	N AJ V AV PR
her	N AJ PN
per	AJ PR
or	N AJ PR CJ
for	N PR CJ
nor	CJ
whirr	N V AV IJ
at	N AV PR CJ PN
neat	N AJ AV OT
great	N AJ AV OT
that	N AJ AV CJ PN
what	N AJ AV CJ PN IJ OT
wet	N AJ V AV OT
yet	AJ AV CJ
left	N AJ AV OT
light	N AJ V AV OT
aught	N AV PN
caught	AJ V PA OT
ought	N V AV PN OT
it	N PN
not	N AV PR
spot	N AJ V AV OT
fast	N AJ V AV IJ
joist	N AJ AV PR OT
midst	N AV PR
best	N AJ V AV OT
lest	CJ
most	N AJ AV PN OT
but	N AJ AV PR CJ PN
out	N AJ V AV PR IJ
bout	N AJ V AV PR
next	N AJ AV PR
you	N PN OT

Word	POS
how	N AV CJ IJ
now	N AJ AV CJ
ay	N AV IJ
by	N AJ V AV PR
why	N AV IJ
whizz	N V AV IJ
.....	N AV
supra	AV PR
contra	N AV PR
instead	AV OT
abroad	AJ AV PR
amid	PR
inland	N AJ AV OT
behind	N AJ AV PR
around	AJ AV PR
aboard	AV PR
toward	AJ AV PR
astride	AV PR
aside	N AV PR
beside	AV PR
inside	N AJ AV PR
outside	N AJ AV PR
unlike	N AJ AV PR CJ
before	N AJ AV PR CJ OT
because	AV CJ
despite	N PR
above	N AJ AV PR
himself	N PN
herself	PN
ourself	PN
yourself	PN
itself	PN
myself	PN
along	AV PR
endlong	AV PR
among	PR
anigh	AV PR
although	CJ
enough	N AJ AV IJ
awash	AJ AV OT
beneath	AJ AV PR
argal	N AV CJ
until	PR CJ
bidden	PV

Table 4-4 (Cont.)

Word	POS
between	N AV PR
amen	N V AV IJ
certain	N AJ AV OT
within	N AJ AV PR
upon	PR
ago	AJ AV
into	PR
presto	N AJ V AV IJ
asleep	AJ AV OT
atop	AJ AV PR
anear	AV PR
yonder	AJ AV PN OT
under	N AJ AV PR
rather	AV OT
whether	N AV CJ
either	AJ AV CJ PN
neither	AJ AV CJ PN
whither	AV CJ
other	N AJ AV PN
after	N AJ AV PR CJ
better	N AJ V AV OT
whoever	PN
over	N AJ V AV PR
atour	AV PR
abaft	AV PR
outwrought	PA OT
albeit	CJ
howbeit	AV CJ
aslant	AJ AV PR
except	V PR CJ
athwart	AV PR
amort	AV OT
amidst	N AV PR
amongst	PR
against	PR
midmost	N AJ AV PR
anust	N AJ AV OT
about	AJ AV PR
throughout	AJ AV PR

Word	POS
without	N AV PR
betwixt	AV PR
adieu	N AV IJ
below	N AJ AV PR
midway	N AJ AV PR
bully	N AJ V AV IJ
only	AJ AV PR CJ
any	AJ AV PN OT
alongside	AV PR
opposite	N AJ AV PR
oneself	PN
sidelong	AJ AV PR
underneath	N AJ AV PR
wherewith	N AV PN
unequal	N AJ AV OT
overall	N AJ AV OT
unbeknown	N AJ AV OT
another	AJ PN
whichever	AJ PN
whomever	PN
whenever	AV CJ
whensoever	AV CJ
whosoever	AJ PN
wherever	AV CJ
whatever	AJ AV PN
somewhat	N AV PN
unbethought	AV OT
amidmost	AV PR
undermost	AJ AV OT
anyhow	AV OT
anyway	AV OT
bimonthly	N AJ AV OT
instantly	AV CJ OT
exenterate	AJ V AV OT
wherewithal	N AV PN
anybody	N PN
everybody	PN
immediately	AV CJ

been lost in the sorting process. No other significant omissions have been noted, but are of course possible since checking of the tape dictionaries was not exhaustive. The parts of speech given in Tables 4-1 through 4-4 were taken from the ape dictionary and have been verified in the dictionaries themselves.

Task 3: Modification of Rule A Using a Study of Affixes

Rule A is based upon a general observation. The business of Task 3 is to discover it is possible, by considering prefixes and suffixes (which might well be expected to be key structural elements indicative of syntactic roles), to convert a general rule evidently effective in a majority of cases to an exhaustive rule effective for 95 percent of the words. It was first necessary to develop a formal and reproducible definition of prefixes and suffixes, as is described in The Nature of Affixing in Written English[5] and Structural Definition of Affixes in Multisyllable Words.[6] It was then necessary to investigate the extent of the correlation between affixes and part-of-speech, as described in Part-of-Speech Implications of Affixes.[7] The results of the correlation can be briefly described here.

All words with part of speech AV, PR, PX, NP, IJ, PA, PP, VP, and CJ can be automatically assigned part of speech by reference to the word lists in Tables 4-1 through 4-4, followed by application of Rules B and C for words not in these lists. Part-of-Speech Implications of Affixes was therefore concerned only with words whose part of speech string contained the elements NA, AJ, and VB, which allows the five possible combinations VB, NA, AJ, NA-VB, AJ-VB, NA-AJ being considered equivalent to NA. Attempts to establish a 95 percent correlation between the part of speech string of a word and its affixes failed. However, it was noted that the correlation was closer for four- to seven-syllable words than for two- to three-syllable words, and that a very

good correlation could be obtained for all words between an "inclusive" part-of-speech string and the affixes. Thus in some cases affix-vowel-string considerations enable an absolute identification of the part of speech of a word, but in other cases identification is to a more inclusive set. For example, a NA or a VB may be classified as NA-VB or an AJ may be classified as a NA. Such a classification is justifiable on the following grounds:

- It is the natural task of a syntactic analysis program to choose among several possible parts of speech, and it is easier to do so than to supply a missing part of speech.
- Dictionaries are very reliable in the information explicitly given, but implications inferred from the absence of information are less reliable. Thus the inclusive part of speech string assigned by the algorithm may in some cases be more correct than the more limited one assigned by a particular dictionary. In our experience with the SOX and MW3 dictionaries we found many instances of nonagreement; usually one was more inclusive than the other.

In Part-of-Speech Implications of Affixes, the results of the correlation study are given for 72 prefixes and 87 suffixes. Implications are of the form NA, or NA-VB, or VB or AJ. For 41 of the affixes, the part-of-speech implication changes with the length of the word, from NA-VB for two- and three-syllable words to NA for four- to eight-syllable words.

Later a correlation was made for the affixes, previously mentioned, which seemed to be likely candidates for reducing the exception lists by aiding in the identification of adverbs or in the identification of words ending in ed which are not past participles. Though not operationally defined these affixes are of practical importance and are therefore listed below, with their part-of-speech implications.

Prefixes	POS	Suffixes	POS
north	NA AV	seed	NA
south	NA AV	weed	NA
west	NA AV	like	NA AV
a-	AJ AV	wise	AJ AV
		ward	NA AV
		wards	NA AV
		-fly	NA

TESTING AND EVALUATION

Rules A, B, and C, the exception lists, and the prefix and suffix implications reported in Reference 7 were incorporated into a computer program for testing the algorithm. In the program the exception lists were checked first, then the word was separated into kernel and affix parts, then rules B and C and the other affix rules were executed, and finally rule A was applied to all words still without a part-of-speech assignment. There are some complications involved in some of these steps, particularly in separating a word into kernel and affix parts, and in assigning parts of speech on the basis of affixes. The logic used by the program for these steps is given in Fig. 4-1.

To summarize briefly, the criteria by which an affix sequence was accepted as an affix in a given word was the same as that given in Reference 7. Prefixes were given priority in the stripping of affixes from the kernel, but suffixes were given priority in assigning the parts of speech of the word (as is also explained in Reference 7).

To test the algorithm, 500 words were chosen at random from the tape dictionary[2,3] and the parts of speech assigned by the algorithm were compared with those given

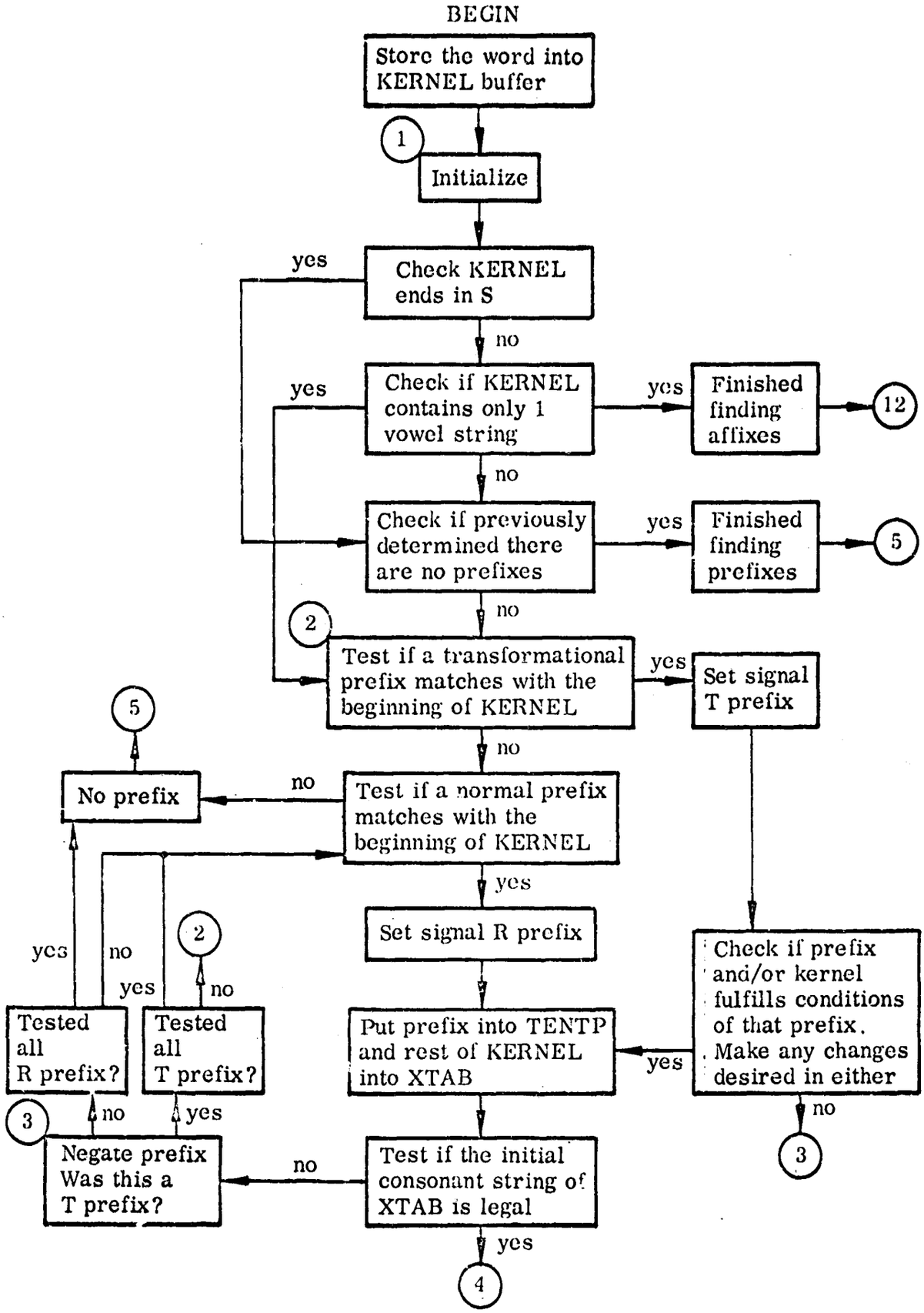

Fig. 4-1 Search-for-Affixes Flow Diagram

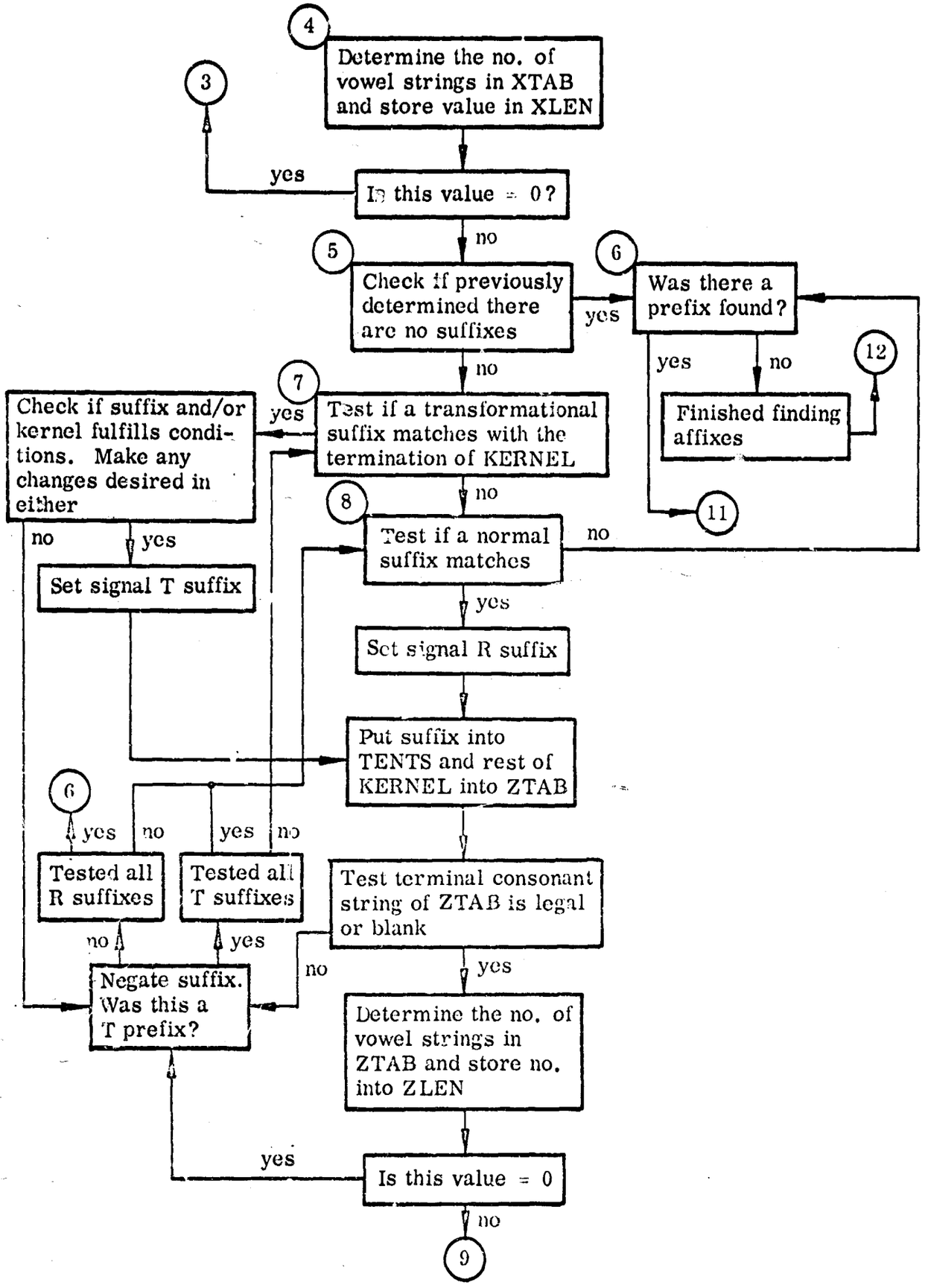

Fig. 4-1 Search-for-Affixes Flow Diagram (Cont.)

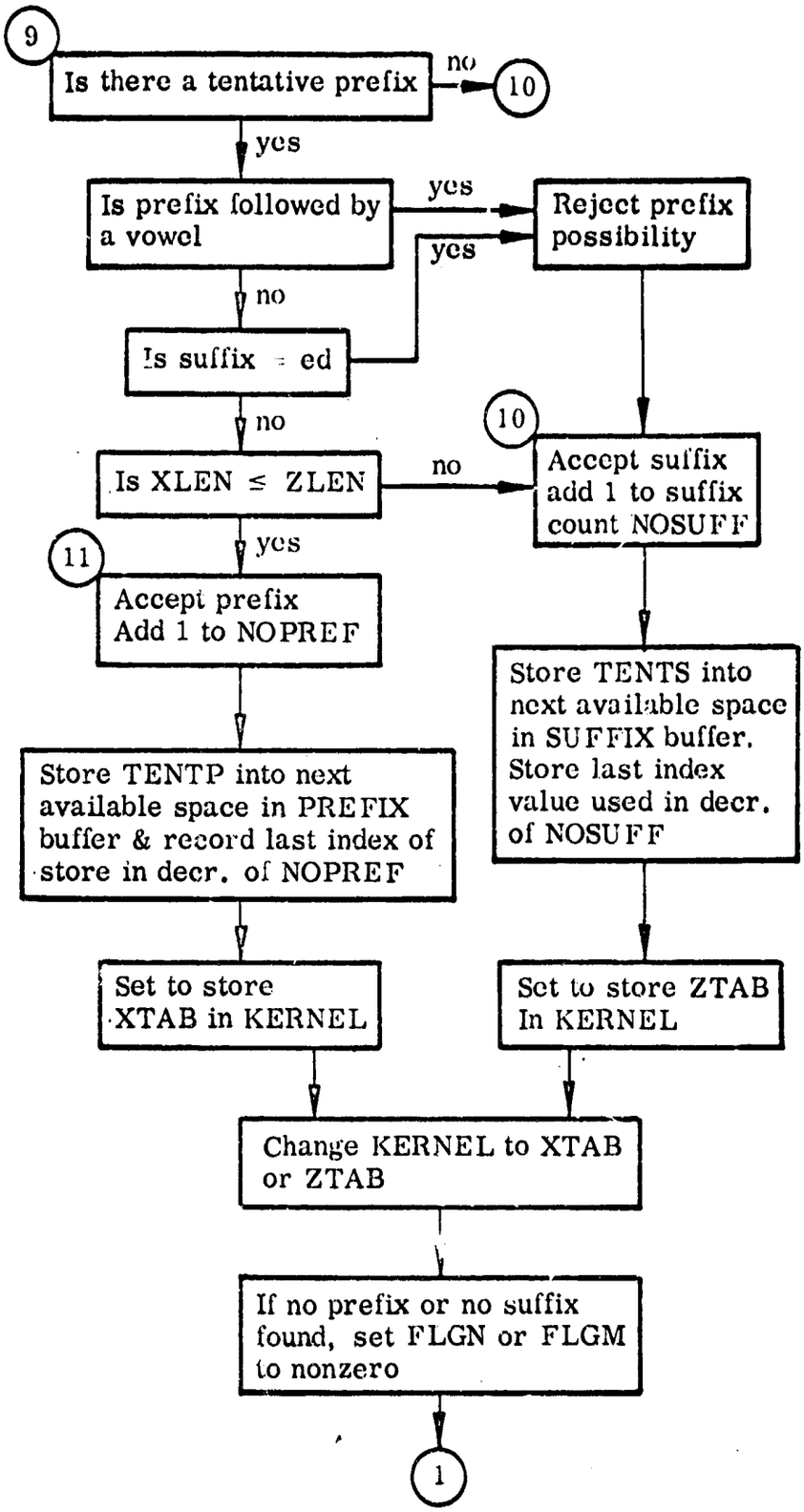

Fig. 4-1 Search-for-Affixes Flow Diagram (Cont.)

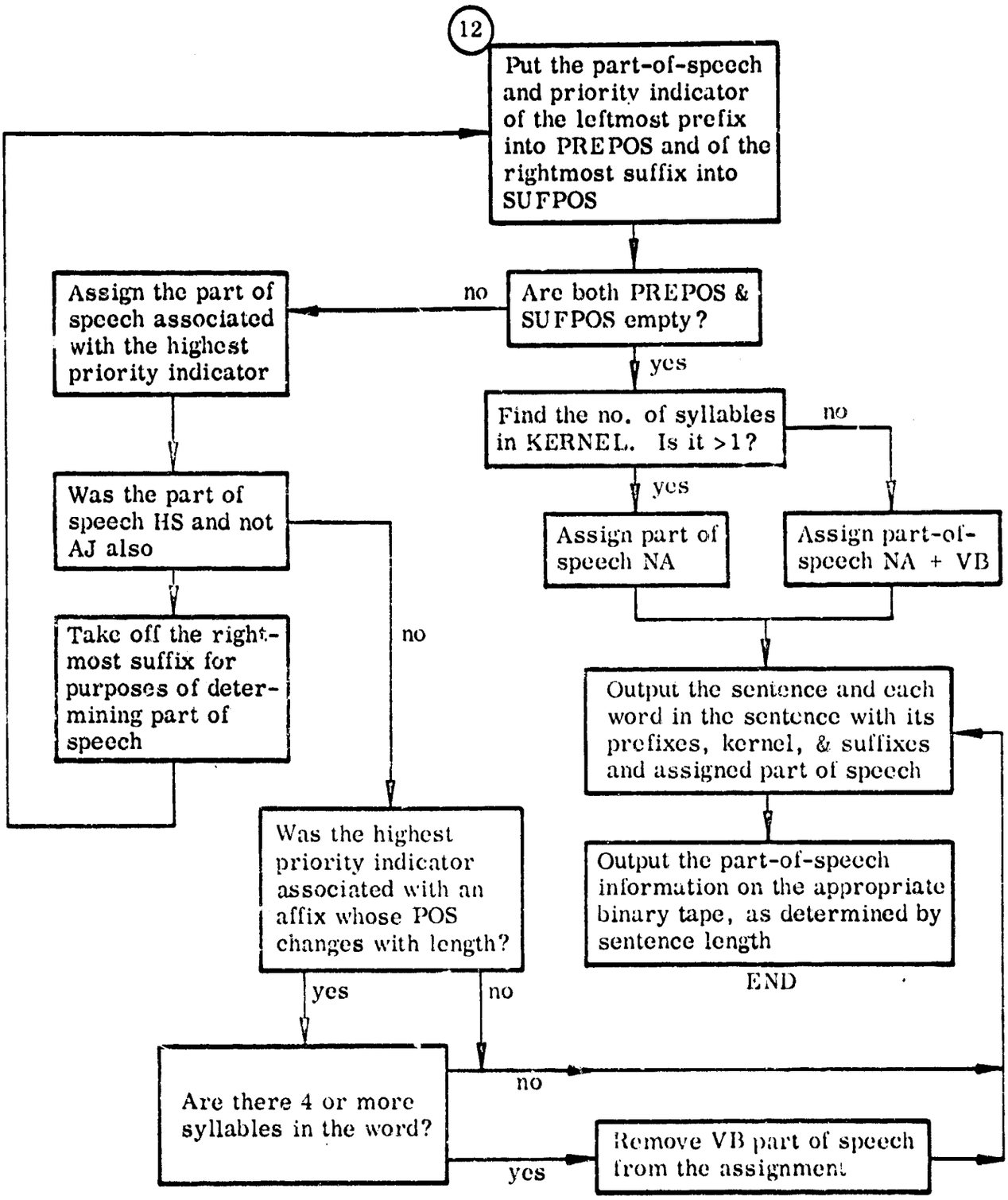

Fig. 4-1 Search-for-Affixes Flow Diagram (Cont.)

in the dictionary. If dialectal, obsolete, archaic, and rare words causing errors are removed, and if program errors are corrected, results are as follows:

Category	Number of words in category
Assigned POS matches dictionary POS	271
Extra POS assigned	196
Missing POS	16
POS does not match at all - Error	8
Total sample	491

This shows that 95.1 percent of the words were assigned the correct inclusive part of speech and 55.2 percent were assigned parts of speech exactly coinciding with those assigned by the dictionary. Thus, the goal of 95 percent has just been achieved.

It is interesting to consider how little the affix implications have improved the results for this sample. Taking the first 192 of the 500 alphabetized words and applying the original Rules A, B, and C only, 20 words are shifted into the exact match category and 25 words from the exact match category for a net loss of 5 words, where 2 of these go into the error category. Six words are added to the words with missing part of speech while two words are taken out of the category. Thus the total loss is 4 more words into the missing category and 2 more words into the error category, or about a 3 percent loss from the point of view of inclusive part of speech. Rule A, it will be remembered, requires the removal of affixes from the kernel of the word. If this kernelizing of the word is omitted, there is about a 13 percent loss from the point of view of inclusive part of speech, indicating that the fact that a word is affixed is more important in predicting part of speech than what the affix is (the affixes ing, ed, and ly excepted). Nevertheless, using the implications of affixes is a refinement in an area where refinement is sorely needed.

It might be interesting at this point to evaluate the two original premises, that elementary words are largely noun-verb and all other words are largely noun only.[1] To test the first premise, the standard one-vowel-string words in the tape dictionary were divided into two sections, those which were NA-VB (and only NA-VB) and those which were not. (The OT category was ignored.) There were 2,520 words in the NA-VB category and 1,925 words with more or less parts of speech than NA-VB. The 1,925 word list includes the 132 one-syllable members of the word-class with parts of speech PR, CJ, IJ, PN, and PV listed in Table 4-4. Discounting these 132 function words then, the first premise is true for 2,520 out of 4,313 cases, or about 58 percent. To get 95 percent of the one-syllable words assigned as in the dictionary, most of the 1,793 non NA-VB words would have to be in an exception dictionary. However, since most of these are NA, from the point of view of inclusive part of speech, the NA-VB rule for elementary words is quite good, giving results very close to those obtained in the 500-word random sample of all words (55 percent exactly matching dictionary, 95 percent giving correct inclusive part of speech).

The second premise has not been directly tested, but may be inferred from the 500-word random sample, since we have just proven that the one-syllable words (there are 46 in the sample) do not affect the results substantially. In its general form the second premise is true about 70 percent of the time, as is reported in Reference 1. In its modified form as stated in Rule A, and tested by our 500-word sample, it is true for only about 55 to 60 percent of the cases, but is good for about 90 to 95 percent of the cases from the point of view of inclusive part of speech, with something less than 5 percent variation, depending on whether or not part of speech implications of affixes are used.

SUMMARY

The net result of the part-of-speech studies is an algorithm which, used in conjunction with a dictionary of less than 800 words and an affix list of less than 200, gives a correct "inclusive" part of speech for 95 percent of a 500-word random sample, and which should do better on textual material. The dictionary is derived from an exhaustive compilation of words which the algorithm is not capable of handling. Such words are either adverbs, function words, participles, or collective nouns not recognized by the program, or conversely, words so classified which should not be. The number of words in the exhaustive list is 3,163, of which only 754 were selected for the dictionary. However, as explained in the body of the text, all of the 267 function words with parts-of-speech other than NA, AJ, VB, or AV have been included, as have all of the irregular past verbs and past participles and the more commonly used adverbs and collective nouns. The 2,409 words omitted are mainly less common adverbs and collective nouns and they comprise only about 3 percent of the total 73,582 word dictionary.

REFERENCES

1. J. Dolby and H. Resnikoff, "On the Structure of Written English Words," Language, 40, 2, April–June 1964
2. The Shorter Oxford English Dictionary on Historical Principles, 3rd ed., revised with addenda, Oxford at the Clarendon Press, 1959
3. Webster's Third New International Dictionary of the English Language, Springfield, Mass., G. C. Merriam Company, Publishers, 1961
4. H. W. Fowler, A Dictionary of Modern English Usage, 2nd ed., revised and edited by Sir Ernest Gowers, New York and Oxford, Oxford University Press, 1965

5. H. Resnikoff and J. Dolby, "The Nature of Affixing in Written English," Mechanical Translation, 8, June and October 1965

6. L. L. Earl, "Structural Definition of Affixes in Multisyllable Words" (in manuscript)

7. L. L. Earl, "Part-of-Speech Implications of Affixes" (see paper 2 of this report)

II

AUTOMATIC INDEXING USING COMBINED SYNTACTIC AND ENTROPY SELECTION CRITERIA

5. PROGRESS REPORT ON A SYNTACTIC-STATISTICAL METHOD FOR AUTOMATIC INDEXING

L. L. Earl

A method for automatic indexing has been developed with two basic aims in mind. The first aim has been to provide a two-level index:

Level (1): Index terms which are descriptive of the subject matter yet represent a drastic reduction of volume

Level (2): Index phrases, arranged under the terms, which consist of selected phrases from the text containing the terms and which give a more complete picture of the subject matter

These two levels would allow selective storage or retrieval depending on the capacities of the system or the needs of the individual information seeker. With the present algorithm the reduction on the first level is to somewhere between 0.05 and 0.25 percent of the original volume, and on the second level to somewhere between 0.5 and 3 percent. The wide range of reduction on both levels has to do with the second aim, which is to adjust the density of terms and phrases to correspond with the information content of the text.

In the method developed, the text is first reduced by selecting a group of phrases on the basis of their syntactic function; the noun phrases selected are the subjects of verbs, and the objects or predicate complements of verbs or infinitives, together with any modifying genitive phrases. Frequency counts are taken of the remaining words, then index terms are taken from the words with highest frequency counts, using entropy criteria to determine the number of index terms chosen. Deriving suitable criteria to meet the second basic aim, adjustment of the density of index phrases according to the

level of information content, has been a major concern. Five texts have been used as samples for experimental purposes:

Text I – An excerpt from an electronics textbook, concerned with amplifiers

Text II – An excerpt from a text on information processing, concerned with the encoding of speech and pictures

Text III – A talk given by a college president on the inauguration of a department of synnoetics

Text IV – A chapter from a book about the personal history of the scientists who developed the atomic bomb (titled Brighter Than a Thousand Suns)

Text V – A journal article on a technique in processing of lists by computer (titled "Multiword List Items")

The indexes produced using the current algorithm are shown in Figs. 5-1 through 5-5; each of the five texts has been reproduced in Figs. 5-6 through 5-10. The first six letters of each index term appear on the left. The index phrases containing the index terms are listed underneath the terms, indented 7 spaces. In the opinion of the author, results are best for the textbook excerpts (texts I and II), satisfactory for the light and low content (texts III and IV), and poorest on the more specialized and abstract journal article.

```
AMPLIF
      USE OF MULTIPLE AMPLIFIER STAGES
      POWER AMPLIFIERS
      GROUNDED GRID AMPLIFIER
      ENVELOPE RESPONSE CHARACTERISTICS OF BAND PASS AMPLIFIERS
      OUTPUT OF ONE AMPLIFIER
      PRACTICAL PUSH'PULL CLASS B AMPLIFIERS
      TWO AMPLIFIERS
      ENVELOPE RESPONSE OF BAND PASS AMPLIFIER
      VIDEO AMPLIFIER
      PUSH'PULL AMPLIFIERS
      THUS GROUNDED GRID AMPLIFIERS
      MOST MULTISTAGE AMPLIFIERS
      TRANSIENT RESPONSE CHARACTERISTICS OF BAND PASS AMPLIFIERS
      TUNED AMPLIFIERS
      TRANSIENT RESPONSE OF EQUIVALENT VIDEO AMPLIFIER
      PC THAN CLASS AB AMPLIFIERS
      RESPONSE OF BAND PASS AMPLIFIER
      DISTRIBUTED AMPLIFIER
      ENVELOPE RESPONSE OF AMPLIFIER
      CIRCUIT DIAGRAM OF AN AMPLIFIER
      PERFORMANCE OF BAND PASS AMPLIFIER
      MORE STAGES OF VIDEO AMPLIFICATION
CIRCUI
      CIRCUIT DIAGRAM OF AN AMPLIFIER
      DESIGN OF OUTPUT COUPLING CIRCUITS
      GROUNDED CATHODE CIRCUIT
      EQUIVALENT PLATE CIRCUIT OF FIGURE
      ANALYSIS OF CIRCUITS
      CIRCUIT CONNECTION
      SIMPLE CIRCUITS
      COMMON CAUSE OF CIRCUIT CONDITION
RESPON
      RESPONSE OF BAND PASS AMPLIFIER
      ENVELOPE RESPONSE OF AMPLIFIER
      ENVELOPE RESPONSE OF BAND PASS AMPLIFIER
      SAME TRANSIENT RESPONSE CHARACTERISTICS
      ENVELOPE RESPONSE CHARACTERISTICS OF BAND PASS AMPLIFIERS
      TRANSIENT RESPONSE CHARACTERISTICS OF BAND PASS AMPLIFIERS
      POSSIBLE RESPONSE
      DETERMINATION OF ENVELOPE RESPONSE
      TRANSIENT RESPONSE OF EQUIVALENT VIDEO AMPLIFIER
```

Fig. 5-1 Index for Text 1

```
PICTUR
        COMPLETE PICTURE
        NEW PICTURE
        PICTURE SIGNAL
        VOLTAGE OF PICTURE SIGNAL
        NEXT PICTURE
        SORTS OF PICTURES
        ACTUAL PICTURE*TRANSMISS SYSTEMS
        BRIGHT OF COLOR TV PICTURE
SIGNAL
        STRENGTH OF SPEECH SIGNAL
        PICTURE SIGNAL
        VOLTAGE OF PICTURE SIGNAL
        ORIGINAL SPEECH SIGNAL
        SLOW VARIATIONS OF SIGNAL
        CONSIDERABLE STRETCH OF SIGNAL
        OF SIGNAL
        DESCRIPTION OF SIGNAL
        LIMITATIONS OF SOURCE SIGNAL
        ANALOG SIGNAL
        SPEECH SIGNAL SOUNDS
SPEECH
        ELECTRICAL REPLICA OF SPEECH
        SPEECH QUALITY
        SPEECH TRANSMISSION
        STRENGTH OF SPEECH SIGNAL
        MOST COMPLICATED SPEECH SOUNDS
        SPEECH SIGNAL SOUNDS
        INTELLIGIBLE SPEECH
        FINE TEMPORAL STRUCTURE OF SPEECH WAVE
        ORIGINAL SPEECH SIGNAL
        SPEECH TRANSMISSION PROBLEM
        SPEECH SOUNDS
        ENTROPY OF SPEECH
VOCODE
        UNNATURAL SOUND OF CHANNEL VOCODE
        EVEN IMPERFECT VOCODE
        TRANSMITTING ANALYZER AND RECEIVING SYNTHESIZER UNITS OF VOCODE
        CHANNEL VOCODE OF FIGURE VII-4 NEEDS
        SORT OF VOCODE DESCRIBED SENDS INFORMATION
        MOST VOCODE
        FORMANT TRACKING VOCODE
        VOCODE QUALITY
        COST OF VOCODE EQUIPMENT
        VOCODE WAY OF
        VOICE-EXCITED VOCODE
```

Fig. 5-2 Index for Text II

Fig. 5-3 Index for Text III

ATOMIC
- CERTAIN ATOMIC
- MINDS OF ATOMIC PHYSICISTS
- ATOMIC BOMB
- ATOMIC ARMAMENT
- FUTURE OF ATOMIC
- ATOMIC SCIENTISTS
- ONE OF ATOMIC EXPERTS
- ATOMIC AIR RAID
- ATOMIC PHYSICIST
- ATOMIC WEAPON
- ATOMIC LABORATORIES
- RAIN OF ATOMIC BOMBS
- ATOMIC TEST EXPLOSION
- OF ATOMIC PHYSICISTS
- NO ATOMIC SECRET
- DIRECTION OF NEW ATOMIC INDUSTRY
- FEDERATION OF ATOMIC SCIENTISTS

NEW
- NEW LEGISLATIVE PROPOSAL
- NEW LEGISLATION
- DIRECTION OF NEW ATOMIC INDUSTRY
- MEMBERS OF NEW SCIENTIFIC
- DANGEROUS EFFECTS OF NEW POWER
- OF NEW KIND OF WEAPONS DEVELOPMENT
- NEW FRIENDS
- NEW POWER
- SINGLE ONE OF NEW BOMBS
- SCIENTISTS* VISION OF NEW WORLD

PHYSIC
- MINDS OF ATOMIC PHYSICISTS
- OF ATOMIC PHYSICISTS
- YOUNG AMERICAN PHYSICIST
- LEARNED PHYSICIST
- ATOMIC PHYSICIST
- ONE OF AMERICAN PHYSICISTS
- THREE PHYSICISTS
- YOUNG PHYSICIST

SCIENT
- GROWING TENDENCY OF SCIENTISTS
- ATOMIC SCIENTISTS
- YOUNG SCIENTISTS
- SCIENTISTS* BLOOD
- OF SCIENTISTS
- DOZEN OF YOUNGER SCIENTISTS
- INDIRECT METHOD OF SCIENTISTS
- SCIENTISTS* VISION OF NEW WORLD
- MEMBERS OF NEW SCIENTIFIC
- FEDERATION OF ATOMIC SCIENTISTS

Fig. 5-4 Index for Text IV

Fig. 5-5 Index for Text V

Fig. 5-5 Index for Text V (Cont.)

PUSH•PULL AMPLIFIERS CAN BE USED TO CONSIDERABLE ADVANTAGE WHEN THE TUBES OPERATE IN THE SWITCHING MODE , BECAUSE THE CIRCUIT CONNECTION CAN BE USED TO REDUCE THE EVEN HARMONIC DISTORTION , AND THE SWITCHING MODE CAN BE USED TO INCREASE THE EFFICIENCY.

THUS TWO TUBES OPERATED IN PUSHPULL CLASS AB OR B WILL PROVIDE A GREATER POWER OUTPUT THAN THE SAME TUBES OPERATED IN CLASS AB PUSH•PULL OR PARALLEL.

THE ADVANTAGE OF THE PUSH•PULL CONNECTION FOR CLASS B OPERATION IS CLEAR FROM FIGURE (12.4) OR (12.5).

FIGURE (12P5) SHOWS THAT THE THIRD HARMONIC IS ZERO IN CLASS B.

THE SECOND HARMONIC HAS A VALUE OF ABOUT 4250/0 , BUT THIS CAN BE REDUCED TO A SMALL FIGURE BY THE PUSH•PULL CONNECTION.

MOREOVER , THE CIRCUIT NOW HAS A MAXIMUM THEORETICAL EFFICIENCY OF 78.5 PC , A CONSIDERABLE IMPROVEMENT OVER THE 50 PC POSSIBLE IN CLASS AB OPERATION.

FURTHERMORE , PRACTICAL PUSH•PULL CLASS B AMPLIFIERS COME CLOSER TO 78P5 PC THAN CLASS AB AMPLIFIERS COME TO 50 PC FOR THE SAME AMOUNT OF HARMONIC DISTORTION.

A PARASITIC IS AN UNWANTED OR SPURIOUS OSCILLATION IN AN ELECTRONIC CIRCUIT.

THEY OCCUR FREQUENTLY IN TUNED POWER AMPLIFIERS.

AS NOTED IN CHAPTER 10 , THE MOST COMMON CAUSE OF PARASITICS IS INADVERTENT FORMATION OF A TUNED GRID , TUNED PLATE OSCILLATOR OPERATING AT SOME FREQUENCY OTHER THAN THAT FOR WHICH THE CIRCUIT WAS DESIGNED.

A COMMON CAUSE OF THIS CIRCUIT CONDITION IS THE USE OF SHUNT FEED BOTH THE PLATE AND GRID CIRCUITS.

Fig. 5-6 Text I

THIS SHOULD BE AVOIDED IF POSSIBLE , BUT IF IT MUST BE DONE , THE CHOKES SHOULD BE AS DISSIMILAR AS POSSIBLE.

THE RATIO OF CHOKE INDUCTANCES SHOULD BE ABOUT 100.

INTERTUBE PARASITICS OFTEN RESULT WHEN TUBES ARE OPERATED IN PARALLEL.

THE USE OF PARASITIC SUPPRESSING RESISTORS OF 10 TO 50 OHMS IN SERIES WITH THE GRID OF EACH TUBE , OR THE USE OF CHOKES IN THE PLATE LEADS , WILL OFTEN REMOVE THIS DIFFICULTY.

PARASITICS OFTEN RESULT THROUGH THE USE OF UNGROUNDED RADIO FREQUENCY TUNING CAPACITORS , EXCESSIVELY LONG LEADS TO THE NEUTRALIZING CONDENSER , OR THROUGH THE USE OF MULTIPLE RADIO FREQUENCY GROUNDS.

SPURIOUS OSCILLATIONS ALSO RESULT FROM COMPLEX CIRCUITS FORMED WHEN TAPS ARE PLACED ON THE TANK COIL FOR THE PURPOSES OF LOADING OR TUNING.

LONG LEADS FROM TUBE CONNECTIONS TO TANK CIRCUITS WILL OCCASIONALLY CAUSE U.F PARASITICS.

IT WAS NOTED IN SECTION (12P11) THAT TUNED AMPLIFIERS WILL OFTEN OSCILLATE BECAUSE OF FEEDBACK PRODUCED THROUGH CGP.

IF THIS FEEDBACK IS DELIBERATELY ENCOURAGED , THE POWER AMPLIFIER IS CONVERTED INTO A POWER OSCILLATOR.

THUS POWER OSCILLATORS ARE DESIGNED AS HIGH EFFICIENCY TUNED AMPLIFIERS OPERATING IN THE SWITCHING MODE , USING THE DESIGN PROCEDURE GIVEN IN SECTION (12.10).

WHEN THE DESIGN IS COMPLETE , STEPS ARE TAKEN TO DETERMINE THE AMOUNT OF FEEDBACK REQUIRED TO PRODUCE THE NECESSARY GRID EXCITATION , DRIVING POWER , AND SO ON.

NEARLY ALL THE CONSIDERATIONS AFFECTING POWER AMPLIFIERS ALSO APPLY TO POWER OSCILLATORS.

Fig. 5-6 Text I (Cont.)

THE ONE NOTABLE EXCEPTION IS THAT FIXED BIAS CANNOT BE USED WITH CLASS C OSCILLATORS BECAUSE THE TUBE WOULD ALWAYS BE CUT OFF , PLATE CURRENT WOULD NEVER FLOW , AND THE OSCILLATION WOULD NEVER START.

HENCE GRID LEAK BIAS IS NEARLY ALWAYS USED IN POWER OSCILLATORS.

ANOTHER SUPERFICIAL DIFFERENCE BETWEEN POWER AMPLIFIERS AND OSCILLATORS IS THAT CRYSTALS ARE OFTEN USED IN PLACE OF THE GRID TUNED CIRCUIT TO PROVIDE FREQUENCY STABILIZATION.

A FEW MISCELLANEOUS TOPICS SHOULD BE MENTIONED BEFORE CLOSING THE DISCUSSION.

BECAUSE OF THE HIGH HARMONIC CONTENT IN THE OUTPUT OF CLASS C AMPLIFIERS , THESE CIRCUITS ARE FREQUENTLY USED AS DOUBLERS AND TRIPLERS.

IN THESE CASES THE PLATE TANK CIRCUIT IS TUNED TO THE DESIRED HARMONIC FREQUENCY.

BECAUSE OF THE SHORT CONDUCTION ANGLES REQUIRED FOR EFFICIENT OPERATION AS FREQUENCY MULTIPLIERS , TUBES OPERATED IN THIS MANNER REQUIRE LARGE GRID BIAS VOLTAGES AND LARGE SIGNAL VOLTAGES.

THIS DIFFICULTY CAN BE PARTIALLY OVERCOME BY USING HIGH U TRIODES AND BEAM POWER TUBES AND PENTODES.

IN CONNECTION WITH NEUTRALIZING PROBLEMS , IT SHOULD BE REMEMBERED FROM CHAPTER 3 THAT THE GROUNDED GRID AMPLIFIER IS LESS SUSCEPTIBLE TO OSCILLATION THAN THE GROUNDED CATHODE CIRCUIT.

THUS GROUNDED GRID AMPLIFIERS ARE WIDELY USED IN CLASS C POWER AMPLIFIERS TO MINIMIZE NEUTRALIZATION PROBLEMS.

THE GENERAL OPERATION AND ANALYSIS OF SUCH CIRCUITS PROCEEDS ALONG THE SAME LINES AS THOSE ILLUSTRATED FOR THE GROUNDED CATHODE CIRCUIT.

Fig. 5-6 Text I (Cont.)

THERE ARE SOME DIFFERENCES IN THE COMPUTATION OF THE POWER OUTPUT.

THE DESIGN OF THE OUTPUT COUPLING CIRCUITS IS NOT COVERED HERE BECAUSE SUCH MATTERS ARE GENERALLY COVERED IN DETAIL IN STANDARD BOOKS OF CIRCUIT THEORY.

FOR EXAMPLE , IN A GREAT MANY CASES , AMPLIFIERS ARE REQUIRED TO HAVE A CERTAIN VOLTAGE GAIN VS. FREQUENCY CHARACTERISTIC OVER A SPECIFIED FREQUENCY RANGE.

THE CIRCUIT DIAGRAM OF SUCH AN AMPLIFIER IS SHOWN IN FIGURE (4.18).

THE EQUIVALENT PLATE CIRCUIT OF FIGURE 4 IS OBTAINED FROM FIGURE FIVE , IF THE OPERATING FREQUENCY IS HIGH ENOUGH SO THAT THE REACTANCE OF THE COUPLING CAPACITATOR IS PRACTICALLY ZERO.

THE CIRCUIT CAN BE FURTHER SIMPLIFIED AS IN THE FIGURE BY COMBINING RP AND RG INTO AN EQUIVALENT RESISTANCE R2.

TWENTY OR MORE STAGES OF VIDEO AMPLIFICATION MAY BE REQUIRED IN TELEVISION STUDIO INSTALLATIONS.

EVEN THE ORDINARY BROADCAST FREQUENCY SUPERHETERODYNE RECEIVER USES SEVERAL STAGES OF AMPLIFICATION.

CONSEQUENTLY , IT SHOULD BE CLEAR THAT THE USE OF MULTIPLE AMPLIFIER STAGES IS COMMON.

IF THE OUTPUT OF ONE AMPLIFIER IS CONNECTED TO THE INPUT OF ANOTHER , THE TWO AMPLIFIERS ARE SAID TO BE CONNECTED IN CASCADE.

A REPRESENTATION OF THE CASCADE CONNECTION IS SHOWN IN FIGURE (5.1).

THIS IS THE MOST COMMON METHOD OF USING MULTIPLE STAGES OF VOLTAGE AMPLIFICATION , SO THAT MOST MULTISTAGE AMPLIFIERS ARE OF THIS TYPE.

Fig. 5-6 Text 1 (Cont.)

A SIGNIFICANT EXCEPTION IS THE DISTRIBUTED AMPLIFIER , WHICH IS DISCUSSED TOWARD THE END OF THE CHAPTER.

THIS CHAPTER IS CONCERNED WITH THE STEADY STATE RESPONSE OF MULTISTAGE AMPLIFIERS.

THE RESPONSE TO PULSE INPUTS IS DISCUSSED IN THE NEXT CHAPTER.

IN MANY ELECTRONIC SYSTEMS USED FOR PULSE TRANSMISSION , BOTH LOW PASS AND BAND PASS AMPLIFICATION ARE REQUIRED.

THEREFORE , IT MAY BE NESSARY TO AMPLIFY HIGH FREQUENCY SIGNALS THAT HAVE A PULSE ENVELOPE SUCH AS THAT SHOWN IN FIGURE (6.1).

ALSO , PULSE TRANSMISSION SYSTEMS OFTEN INVOLVE CASCADES OF BOTH LOW PASS AND BAND PASS AMPLIFIERS.

IN SUCH CASES THE TRANSIENT RESPONSE CHARACTERISTICS OF BAND PASS AMPLIFIERS ARE OF INTEREST.

ALTHOUGH THERE ARE NOTABLE EXCEPTIONS , THE RESPONSE OF THE BAND PASS AMPLIFIER TO THE ENVELOPE OF THE HIGH FREQUENCY PULSE IS THE MATTER OF MOST CONCERN.

THIS WILL BE CALLED THE ENVELOPE RESPONSE OF THE AMPLIFIER.

IN OTHER WORDS , THE PERFORMANCE OF BAND PASS AMPLIFIER USED FOR PULSE TRANSMISSION IS MOST FREQUENTLY EVALUATED BY DETERMINING THE

FAITHFULNESS WITH WHICH THE ENVELOPE OF THE INPUT PULSE IS REPRODUCED IN THE OUTPUT.

A POSSIBLE RESPONSE TO THE HIGH FREQUENCY PULSE OF FIGURE XXX IS IS SHOWN IN FIGURE (6.3).

NOTE THAT A FINITE RISE TIME AND OVERSHOOT ARE INDICATED , BOTH AT THE BEGINNING AND THE END OF THE PULSE.

Fig. 5-6 Text 1 (Cont.)

THIS FIGURE ALSO SHOWS THAT THE EFFECTS PRODUCED BY OVERSHOOT ON THE LEADING AND TRAILING EDGES OF THE PULSE ARE DIFFERENT.

ASSUME THAT THE ENVELOPE RESPONSE CHARACTERISTICS OF BAND PASS AMPLIFIERS ARE TO BE DETERMINED.

BECAUSE THE ENVELOPE OF A HIGH FREQUENCY PULSE IS A VIDEO SIGNAL WHEN CONSIDERED SEPARATELY , IT SHOULD BE POSSIBLE TO STUDY THE ENVELOPE RESPONSE OF A BAND PASS AMPLIFIER IN TERMS OF THE TRANSIENT RESPONSE OF SOME VIDEO AMPLIFIER.

IN OTHER WORDS , IT IS SUGGESTED THAT A VIDEO AMPLIFIER COULD BE SPECIFIED TO HAVE THE SAME TRANSIENT RESPONSE CHARACTERISTICS AS THE ENVELOPE RESPONSE CHARACTERISTICS OF THE BAND PASS AMPLIFIER.

IF SUCH AN EQUIVALENCE CAN BE ESTABLISHED , THE DETERMINATION OF THE ENVELOPE RESPONSE WILL BE GREATLY SIMPLIFIED , BECAUSE THE TRANSIENT RESPONSE OF THE EQUIVALENT VIDEO AMPLIFIER CAN BE EVALUATED FAIRLY EASILY WITH THE AID OF THE RULES GIVEN IN THE PRECEDING SECTION.

A USEFUL CONCEPT CAN BE INFORMALLY DEDUCED BY COMPARING THE CHARACTERISTICS AND GAIN FUNCTIONS OF REPRESENTATIVE LOW PASS AND BAND PASS AMPLIFIERS.

FOR EXAMPLE , CONSIDER THE SIMPLEST CIRCUITS IN EACH CATEGORY , A RESISTANCE COUPLED AND A SINGLE TUNED AMPLIFIER.

THE ESSENTIAL FEATURES OF INTEREST IN THESE TWO AMPLIFIERS ARE SHOWN IN FIGURE (6.4).

IF YOU EXAMINE THE CURVES IN THIS FIGURE YOU CAN SEE THAT THEY ARE RELATED.

THAT IS , THE COMPLETE FREQUENCY RESPONSE CHARACTERISTIC OF A RESISTANCE COUPLED AMPLIFIER WOULD BE NEARLY IDENTICAL TO THAT OF A SINGLE TUNED AMPLIFIER IF THE NEGATIVE FREQUENCY RANGE COULD BE OBTAINED IN A PRACTICAL CASE.

Fig. 5-6 Text I (Cont.)

IN THE EARLY THIRTIES , LONG BEFORE SHANNON*S WORK ON INFORMATION THEORY , HOMER DUDLEY OF THE BELL LABORATORIES INVENTED SUCH A FORM OF SPEECH TRANSMISSION , WHICH HE CALLED THE VOCODER FROM VOICE CODER.

THE TRANSMITTING ANALYZER AND RECEIVING SYNTHESIZER UNITS OF A VOCODER ARE ILLUSTRATED IN FIGURE VII-4.

IN THE ANALYZER , AN ELECTRICAL REPLICA OF THE SPEECH IS FED TO 16 FILTERS.

EACH FILTER DETERMINES THE STRENGTH OF THE SPEECH SIGNAL IN A PARTICULAR BAND OF REQUENCIES.

EACH ONE THEN TRANSMITS A SIGNAL TO THE SYNTHESIZER WHICH GIVES THIS INFORMATION.

IN ADDITION , AN ANALYSIS IS MADE TO DETERMINE WHETHER THE SOUND IS VOICELESS OR VOICED , AND IF VOICED WHAT THE PITCH IS.

AT THE SYNTHESIZER , IF THE SOUND IS VOICELESS , A HISSING NOISE IS PRODUCED . IF THE SOUND IS VOICED A SEQUENCE OF ELECTRICAL PULSES IS PRODUCED AT THE PROPER RATE , CORRESPONDING TO THE PUFFS OF AIR PASSING THE VOCAL CORDS OF THE SPEAKER.

THE HISS OR PULSES ARE FED TO AN ARRAY OF FILTERS , EACH PASSES A BAND OF FREQUENCIES CORRESPONDING TO A PARTICULAR FILTER IN THE ANALYZER.

THE AMOUNT OF SOUND PASSING THROUGH A PARTICULAR FILTER IN THE SYNTHESIZER IS CONTROLLED BY THE OUTPUT OF THE CORRESPONDING ANALYZER FILTER SO AS TO BE THE SAME AS THAT WHICH THE ANALYZER FILTER INDICATES TO BE PRESENT IN THE VOICE IN THIS PARTICULAR FREQUENCY RANGE.

THIS PROCESS RESULTS IN THE REPRODUCTION OF INTELLIGIBLE SPEECH.

Fig. 5-7 Text II

IN EFFECT THE ANALYZER LISTENS TO AND ANALYZES SPEECH , INSTRUCTING THE SYNTHESIZER , WHICH IS AN ARTIFICIAL SPEAKING MACHINE , HOW TO SAY THE WORDS ALL OVER AGAIN WITH THE VERY PITCH AND THE VERY ACCENT OF THE SPEAKER.

MOST VOCODERS HAVE A STRONG AND UNPLEASANT ELECTRICAL ACCENT.

THE STUDY OF THIS HAS LED TO NEW AND IMPORTANT IDEAS CONCERNING WHAT DETERMINES AND INFLUENCES SPEECH QUALITY * WE CANNOT AFFORD TIME TO GO INTO THIS MATTER HERE.

EVEN IMPERFECT VOCODERS CAN BE VERY USEFUL.

FOR INSTANCE , IT IS SOMETIMES NECESSARY TO RESORT TO ENCIPHERED SPEECH TRANSMISSION.

IF ONE MERELY DIRECTLY REDUCES SPEECH TO BINARY DIGITS BY PULSE CODE MODULATION , 30 TO 60 THOUSAND BINARY DIGITS PER SECOND MUST BE SENT.

BY USING A VOCODER , SPEECH CAN BE SENT WITH AROUND 1500 BINARY DIGITS PER SECOND.

THE SORT OF VOCODER DESCRIBED SENDS INFORMATION CONCERNING FROM 10 TO 30 FREQUENCY BANDS.

SPEECH SOUNDS ACTUALLY HAVE ONLY A FEW VERY PROMINENT FREQUENCY RANGES CALLED FORMANTS.

THESE CORRESPOND TO THE RESONANCES OF THE VOCAL TRACT.

ONE CAN RECREATE INTELLIGIBLE SPEECH BY SENDING INFORMATION CONCERNING THE LOCATION AND INTENSITY OF TWO OR THREE FORMANTS.

SUCH A FORMANT TRACKING VOCODER CAN BE USED TO TRANSMIT SPEECH WITH EVEN FEWER BINARY DIGITS PER SECOND THAN THE CHANNEL VOCODER OF FIGURE VII-4 NEEDS.

Fig. 5-7 Text II (Cont.)

IN AN EVEN MORE ECONOMICAL , LESS INTELLIGIBLE VOCODER , CALLED THE THE FAILURE OF THE SOUND GENERATOR OF THE SYNTHESIZER TO ADEQUATELY FOLLOW PITCH , CHANGES FROM VOICED TO VOICELESS SOUND , AND OTHER QUALITIES OF THE EXCITATION OF THE SPEAKER*S VOCAL TRACT.

PHONEME VOCODER , THE ANALYZER , RECOGNIZING A NUMBER OF BASIC SPEECH SOUNDS CALLED PHONEMES , INSTRUCTS THE SYNTHESIZER TO SPEAK THESE.

FOR ORDINARY T' 'HONE USE , VOCODER QUALITY IS SCARCELY ADEQUATE.

THE UNNATURAL SOUND OF THE CHANNEL VOCODER APPEARS TO BE ASSOCIATED WITH BY SENDING A BAND A FEW HUNDRED CYCLES WIDE OF THE SPEECH TO BE RECREATED AND DISTORTING THIS AT THE SYNTHESIZER , A MORE SATISFACTORY SOURCE OF SOUND WITH WHICH TO FEED THE SYNTHESIZER FILTERS IS OBTAINED.

SUCH A VOICE-EXCITED VOCODER SOUNDS ALMOST AS GOOD AS REGULAR TELEPHONE SPEECH AND IT TAKES ONLY ONE-HALF AS MUCH CHANNEL CAPACITY TO TRANSMIT.

THE COST OF THE VOCODER EQUIPMENT WOULD PRECLUDE ITS USE ON ANY BUT LONG AND EXPENSIVE COMMUNICATION CIRCUITS , SUCH AS TRANSATLANTIC TELEPHONE CABLES.

WE WILL CONSIDER THE VOCODER FOR A MOMENT BEFORE LIAVING IT.

WE NOTE THAT TRANSMISSION OF VOICE USING EVEN THE MOST ECONOMICAL OF VOCODERS TAKES MANY MORE BINARY DIGITS PER WORD THAN TRANSMISSION OF ENGLISH TEXT.

PARTLY , THIS IS BECAUSE OF THE TECHNICAL DIFFICULTIES OF ANALYZING AND ENCODING SPEECH AS OPPOSED TO PRINT.

PARTLY , THIS IS BECAUSE IN THE CASE OF SPEECH WE ARE ACTUALLY TRANSMITTING INFORMATION ABOUT SPEECH QUALITY , PITCH , STRESS , ACCENT , AND SUCH INFORMATION AS THERE IS IN TEXT.

IN OTHER WORDS , THE ENTROPY OF SPEECH IS SOMEWHAT GREATER PER WORD THAN THE ENTROPY OF TEXT.

Fig. 5-7 Text II (Cont.)

THAT THE VOCODER DOES ENCODE SPEECH MORE EFFICIENTLY THAN OTHER METHODS DEPENDS ON THE FACT THAT THE CONFIGURATION OF THE VOCAL TRACT CHANGES LESS RAPIDLY THAN THE FLUCTUATIONS OF THE SOUND WAVES WHICH THE VOCAL TRACT PRODUCES.

THE EFFECTIVENESS ALSO DEPENDS ON LIMITATIONS OF THE HUMAN SENSE OF HEARING.

FROM AN ELECTRICAL POINT OF VIEW , THE MOST COMPLICATED SPEECH SOUNDS ARE THE HISSING FRICATIVES , SUCH AS F AND G OF FIGURE VII-3.

FURTHERMORE , THE WAVE FORMS OF TWO S*S UTTERED SUCCESSIVELY MAY HAVE QUITE A DIFFERENT SEQUENCE OF UPS AND DOWNS.

IT WOULD TAKE MANY BINARY DIGITS PER SECOND TO TRANSMIT EACH IN FULL DETAIL.

BUT , TO THE EAR , ONE WILL SOUND JUST LIKE ANOTHER IF IT HAS IN A BROAD
WAY THE SAME FREQUENCY CONTENT.

THUS , THE VOCODER DOESN*T HAVE TO REPRODUCE THE S SOUND THE SPEAKER UTTERED * IT HAS MERELY TO REPRODUCE AN S SOUND THAT HAS ROUGHLY THE SAME FREQUENCY CONTENT AND HENCE SOUNDS THE SAME.

WE SEE THAT , IN TRANSMITTING SPEECH , THE ROYAL ROAD TO EFFICIENT ENCODING APPEARS TO BE THE DETECTION OF CERTAIN SIMPLE IMPORTANT PATTERNS AND THEIR RECREATION AT THE RECEIVING END.

BECAUSE OF THE GREATER CHANNEL CAPACITY REQUIRED , EFFICIENT ENCODING IS EVEN MORE IMPORTANT IN TV TRANSMISSION THAN IN SPEECH TRANSMISSION.

PERHAPS WE CAN APPLY A SIMILAR PRINCIPLE IN TV.

THE TV PROBLEM IS MUCH MORE DIFFICULT THAN THE SPEECH TRANSMISSION PROBLEM.

Fig. 5-7 Text II (Cont.)

PARTLY , THIS IS BECAUSE THE SENSE OF SIGHT IS INHERENTLY MORE DETAILED AND DISCRIMINATING THAN THE SENSE OF HEARING.

PARTLY , THOUGH , IT IS BECAUSE MANY SORTS OF PICTURES FROM MANY SOURCES ARE TRANSMITTED BY TV , BUT SPEECH IS ALL PRODUCED BY THE SAME SORT OF VOCAL APPARATUS.

IN THE FACE OF THESE FACTS , PERHAPS SOME VOCODER-LIKE WAY OF TRANSMITTING IS POSSIBLE IF WE CONFINE OURSELVES TO ONE SORT OF PICTURE SOURCE , LIKE THE HUMAN FACE.

ONE CAN CONCEIVE OF SUCH A THING.

WE CAN IMAGINE THAT WE HAD AT THE RECEIVER A SORT OF RUBBERY MODEL OF A HUMAN FACE.

OR WE MIGHT HAVE A DESCRIPTION OF SUCH A MODEL STORED IN THE MEMORY OF A HUGE ELECTRONIC COMPUTER.

FIRST , THE TRANSMITTER WOULD HAVE TO LOOK AT THE FACE TO BE TRANSMITTED TO MAKE UP THE MODEL AT THE RECEIVER IN SHAPE AND IN TINT.

THE TRANSMITTER WOULD ALSO HAVE TO NOTE THE SOURCES OF LIGHT IN ORDER TO REPRODUCE THESE IN INTENSITY AND DIRECTION AT THE RECEIVER.

THEN , AS THE PERSON IN FRONT OF THE TRANSMITTER TALKED , THE TRANSMITTER WOULD HAVE TO FOLLOW THE MOVEMENTS OF HIS EYES , LIPS , JAWS , AND OTHER MUSCULAR MOVEMENTS , IN ORDER TO TRAMSNIT THESE SO THAT THE MODEL AT THE RECEIVER COULD DO LIKEWISE.

SUCH A SCHEME MIGHT BE VERY EFFECTIVE , AND IT COULD BECOME AN IMPORTANT INVENTION IF ANYONE COULD SPECIFY A USEFUL WAY OF CARRYING OUT THE OPERATIONS WHICH I HAVE DESCRIBED.

ALAS , HOW MUCH EASIEP IT IS TO SAY WHAT ONE WOULD LIKE TO DO THAN IT IS TO DO IT.

Fig. 5-7 Text II (Cont.)

IN OUR DAY OF UNLIMITED SCIENCE AND TECHNOLOGY , PEOPLE*S UNFULFILLED ASPIRATIONS HAVE BECOME SO IMPORTANT TO THEM THAT A SPECIAL WORD , POPULAR IN THE PRESS , HAS BEEN COINED TO DENOTE SUCH DREAMS.

THAT WORD IS BREAKTHROUGH.

MORE RARELY , IT MAY ALSO BE USED TO DESCRIBE SOMETHING , USUALLY TRIVIAL , WHICH HAS ACTUALLY BEEN ACCOMPLISHED.

IF WE TURN FROM SUCH DREAMS OF THE FUTURE , WE FIND THAT ALL ACTUAL PICTURE*TRANSMISSION SYSTEMS FOLLOW A COMMON PATTERN.

THE PICTURE OR IMAGE TO BE TRANSMITTED IS SCANNED TO DISCOVER THE BRIGHTNESS AT SUCCESSIVE POINTS.

THE SCANNING IS CARRIED OUT ALONG A SEQUENCE OF CLOSELY SPACED LINES.

IN COLOR TV , THREE IMAGES OF DIFFERENT COLORS ARE SCANNED SIMULTANEOUSLY.

THEN , AT THE RECEIVER , A POINT OF LIGHT WHOSE INTENSITY VARIES IN ACCORD WITH THE SIGNAL FROM THE TRANSMITTER PAINTS OUT THE PICTURE IN LIGHT AND SHADE , FOLLOWING THE SAME LINE PATTERN.

SO FAR ALL PRACTICAL ATTEMPTS AT EFFICIENT ENCODING HAVE STARTED OUT WITH THE SIGNAL GENERATED BY SUCH A SCANNING PROCESS.

THE OUTSTANDING EFFICIENT ENCODING SCHEME IS THAT USED IN COLOR TV.

THE BRIGHTNESS OF A COLOR TV PICTURE HAS VERY FINE DETAIL * THE PATTERN OF COLOR HAS VERY MUCH LESS DETAIL.

THUS , COLOR TV OF ALMOST THE SAME DETAIL AS MONOCHROME TV CAN BE SENT OVER THE SAME CHANNEL USED FOR MONOCHROME.

OF COURSE , COLOR TV USES AN ANALOG SIGNAL * THE PICTURE IS NOT REDUCED TO DISCRETE ON-OR-OFF PULSES.

Fig. 5-7 Text II (Cont.)

A PROPOSED METHOD FOR THE EFFICIENT ENCODING OF MONOCHROME TV IS TO SEND THE SLOW VARIATIONS OF THE SIGNAL IN GREAT DETAIL AND THE FAST VARIATIONS EITHER LESS ACCURATELY OR ONLY INTERMITTENTLY , AS THEY OCCUR.

THERE IS A GOOD DEAL OF CONTROVERSY AS TO HOW EFFECTIVE THIS IS.

IN TV , A COMPLETE PICTURE IS SENT EVERY 1/30 SECOND IN ORDER TO AVOID FLICKER.

ALTHOUGH IN MOTION PICTURES A NEW PICTURE IS USED EVERY 1/24 SECOND , IN ORDER TO AVOID FLICKER IT IS TURNED ON-AND-OFF BY A SHUTTER SEVERAL TIMES BEFORE THE NEXT PICTURE IS SUBSTITUTED.

IN THE CASE OF MANY SUBJECTS , SUCH AS A FACE , A NEW PICTURE EVERY 1/10 SECOND WOULD BE SUFFICIENT IF FLICKER COULD BE AVOIDED BY SHOWING IT SEVERAL TIMES.

THIS WOULD REQUIRE REPEATEDLY STORING A LENGTH OF SIGNAL CORRESPONDING TO A COMPLETE PICTURE AT THE RECEIVER.

AT PRESENT , THIS SEEMS TOO EXPENSIVE TO DO , BUT SUCH A SCHEME MIGHT CUT DOWN THE REQUIRED NUMBER OF BINARY DIGITS PER SECOND BY A FACTOR OF 3.

WE CAN SUPPOSE THAT THE VOLTAGE OF THE PICTURE SIGNAL VARIED WITH TIME AS SHOWN IN AA OF FIGURE VII-5.

A GREAT MANY SAMPLES , ALSO SHOWN , MIGHT BE USED TO REPRESENT IT.

INSTEAD , WE COULD PERHAPS USE A NUMBER OF STRAIGHT LINES TO APPROXIMATE THE PICTURE SIGNAL , AS IN B OF FIGURE VII-5.

THEN WE WOULD SEND ONLY THE HEIGHTS OF THE END POINTS OF THE LINES AND THE DISTANCES BETWEEN THE END POINTS OF THE LINES.

THIS IS QUITE AN OLD IDEA.

Fig. 5-7 Text II (Cont.)

IT HAS BEEN TRIED EXPERIMENTALLY RECENTLY , BUT THERE IS LITTLE AGREEMENT AS TO HOW EFFECTIVE IT IS.

WE MAY REMEMBER THAT IN TRASMITTING SPEECH BY PULSE CODE MODULATION IT IS EFFECTIVE TO ASSIGN CLOSELY SPACED AMPLITUDES OR LEVELS OF QUANTIZATION FOR SMALL SIGNALS AND MORE WIDELY SPACED LEVELS FOR LARGE SIGNALS.

THIS IS NOT EFFECTIVE IN THE CASE OF PICTURE TRANSMISSION , BECAUSE FINE DETAIL MAY OCCUR IN EITHER THE DARK
OR THE BRIGHT PART OF THE PICTURE , AT EITHER HIGH OR LOW SIGNAL LEVELS.

HOWEVER , IT IS NOT NECESSARY TO REPRODUCE LARGE CHANGES IN LIGHT INTENSITY AS ACCURATELY AS SMALL CHANGES.

THUS , IF WE SEND THE DIFFERENCES IN AMPLITUDE OF SUCCESSIVE SAMPLES , WE CAN USE CLOSELY SPACED LEVELS OF QUANTIZATION FOR SMALL DIFFERENCES AS IN HAIR.

SIMILARLY WE CAN USE COARSLY SPACED LEVELS FOR LARGE DIFFERENCES TO GET A
SAVING SIMILAR TO THAT ATTAINED IN SPEECH TRANSMISSION.

BY USING A REFINED FORM OF THIS SCHEME , IN WHICH ONE CAN CHOOSE TO SEND THE DIFFERENCE FROM AN ALREADY TRANSMITTED SAMPLE EITHER JUST ABOVE OR JUST TO THE LEFT OF THE SAMPLE TO BE SENT , ONE CAN DO ALMOST AS WELL WITH 3 BINARY DIGITS PER SAMPLE AS ONE CAN WITH 7 BINARY DIGITS PER SAMPLE IF THE AMPLITUDE OF EACH SAMPLE IS ENCODED TO BE SENT SEPARATELY.

REVIEWING WHAT HAS BEEN SAID , WE SEE THAT THERE ARE THREE IMPORTANT PRINCIPLES IN ENCODING SIGNALS EFFICIENTLY.

THE SIGNAL SHOULD NOT BE ENCODED ONE SAMPLE OR ONE CHARACTER AT A TIME.

Fig. 5-7 Text II (Cont.)

A CONSIDERABLE STRETCH OF SIGNAL SHOULD BE ENCODED AT A TIME , HYPERQUANTIZATION.

THE LIMITATIONS OF THE SOURCE SIGNAL SHOULD BE TAKEN INTO ACCOUNT.

ANY INABILITIES OF THE EYE OR EAR TO DETECT ERRORS IN A RECONSTRUCTION SHOULD ALSO BE TAKEN INTO ACCOUNT.

THE VOCODER ILLUSTRATES THESE PRINCIPLES EXCELLENTLY.

THE FINE TEMPORAL STRUCTURE OF THE SPEECH WAVE IS NOT EXAMINED IN DETAIL.

INSTEAD , A DESCRIPTION SPECIFYING THE AVERAGE INTENSITIES OVER CERTAIN RANGES OF FREQUENCIES IS TRANSMITTED , TOGETHER WITH A SIGNAL WHICH TELLS WHETHER THE SPEECH IS VOICED OR UNVOICED , AND IF IT IS VOICED , WHAT ITS PITCH IS.

THIS DESCRIPTION OF A SIGNAL IS EFFICIENT BECAUSE THE VOCAL ORGANS DON*T CHANGE POSITION RAPIDLY IN PRODUCING SPEECH.

AT THE RECEIVER , THE VOCODER GENERATES A SPEECH SIGNAL WHICH DOESN*T RESEMBLE THE ORIGINAL SPEECH SIGNAL IN FINE DETAIL BUT ON THE CONTRARY ,
SOUNDS LIKE THE ORIGHINAL SPEECH SIGNAL BECAUSE OF THE NATURAL LIMITATIONS OF OUR HEARING.

THE VOCODER IS A SORT OF PARAGON OF EFFICIENT TRANSMISSION DEVICES.

COLOR TV COMES NEXT PERHAPS , WHERE THE VARIATIONS OF COLOR OVER THE PICTURE ARE DEFINED MUCH LESS SHARPLY THAN VARIATIONS OF INTENSITY ARE.

THIS TAKES ADVANTAGE OF THE EYES INABILITY TO SEE FINE DETAIL IN COLOR PATTERNS.

Fig. 5-7 Text II (Cont.)

LAST YEAR , WE CHANGED THE NAME OF OUR COMPUTER•RELATED SCIENCE DEPARTMENT TO THE DEPARTMENT OF SYNNOETICS.

SINCE THEN , WE HAVE RECEIVED MANY INQUIRIES CONCERNING THIS DEPARTMENT FROM PEOPLE WHO THINK IT IS NEWLY FORMED RATHER THAN NEWLY NAMED.

TODAY , I WOULD LIKE TO TALK TO YOU ABOUT SYNNOETICS AT OUR UNIVERSITY.

IN DEALING WITH THE WORLD , MAN HAS BROUGHT TO BEAR ON HIS PROBLEMS AN ARSENAL OF BOTH HIS BASIC UNENHANCED MENTAL POWER AND THE ENHANCED PHYSICAL AND MENTAL PROWESS THAT IT PRODUCED.

UP TO ABOUT 1940 MAN , DRAWING ON THIS ARSENAL , JUST ABOUT HELD HIS OWN.

AFTER THAT , THE PROBLEMS , SOME OF WHICH WERE GENERATED BY THE VERY USE OF COMPUTERS AND OTHER ENHANCERS , WERE GETTING LARGER , MORE COMPLEX , MORE DIVERSE , AND MORE URGENT.

THE WORD SYNNOETICS , WHICH IS DERIVED FROM THE GREEK , MEANS TO POOL TOGETHER THE RESOURCES OF THE MIND.

WE MAY HAVE MAN•COMPUTER SYNNOESIS OR AUTOMATON•AUTOMATON SYNNOESIS , OR MAN•MAN SYNNOESIS.

IF PLANT OR ANIMAL ORGANISMS ARE INCLUDED IN A SYNNOETIC SYSTEM , WE MAY HAVE MAN•ORGANISM•AUTOMATON SYNNOESIS.

HOWEVER , SINCE THIS IS THE MOST POPULAR SYSTEM STUDIED NOW , USAGE MAY LATER CONFINE ITS MEANING TO A MAN•MACHINE SYNNOETIC SYSTEM.

THE PENALTIES FOR MISTAKES WERE MORE SEVERE.

Fig. 5-8 Text III

IN THE FIFTIES AND SIXTIES AND TO THIS DAY , THE PENALTY COULD BE THE EXTINCTION OF THE HUMAN RACE.

THUS , THE PROBLEM THAT BECAME THE LARGEST , MOST COMPLEX , MOST DIVERSE , AND MOST URGENT , WAS MAN'S NEED TO USE HIS INNATE AND HIS ENHANCED MENTAL POWER MORE EFFECTIVELY THAN HITHERTO AND TO SOMEHOW FURTHER ENHANCE HIS MENTAL POWER.

AMONG MAN'S EARLY FEATS IN THIS PURSUIT WERE THE DEVELOPMENT OF DIGITAL AND ANALOG COMPUTERS AND OPERATIONS RESEARCH.

AND I AM SURE YOU ALL RECALL THAT THESE WERE USED VERY SUCCESSFULLY IN THEIR DAY.

SOCIETY HAS GONE A LONG WAY SINCE THEN IN PROVIDING US WITH AIDS FOR OUR MENTAL PROCESSES AND WE HAVE REAPED THE BENEFITS OF THE CONSEQUENT INCREASE IN OUR PHYSICAL AND INTELLECTUAL POWERS.

WE HAVE DONE THIS CHIEFLY BY ENLISTING THE AID OF SYSTEMS CONSISTING OF CONFIGURATIONS OF PEOPLE , MECHANISMS , AND AUTOMATA , MACHINES THAT EXHIBIT SOME MENTAL CHARACTERISTICS.

UNIVERSITY SCHOLARS HAVE CONTRIBUTED THEIR FAIR SHARE TO THESE PRACTICAL ADVANCES.

BUT THEIR MOST IMPORTANT CONTRIBUTION HAS BEEN THE DEVELOPMENT OF AN ORDERLY THEORY OF SYNNOETICS AND A COHERENT CURRICULUM WITH CORE COURSES FOR IT.

SYNNOETICS IS THE SCIENCE OF TREATING OF THE PROPERTIES OF COMPOSITE SYSTEMS CONSISTING OF CONFIGURATIONS OF PEOPLE , MECHANISMS , PLANT OR ANIMAL ORGANISMS , AND AUTOMATA''WHOSE MAIN ATTRIBUTE IS THAT ITS ABILITY TO INVENT , TO CREATE , AND TO REASON''ITS MENTAL POWER''IS USUALLY GREATER THAN THE MENTAL POWER OF ITS COMPONENTS.

Fig. 5-8 Text III (Cont.)

SYN IS A LATINIZED FORM OF A GREEK PREFIX MEANING TOGETHER * NOETICS IS DERIVED FROM THE GREEK , MEANING PERTAINING TO THE MIND OR INTELLECT * ICS IS A SUFFIX THAT IS AN ACCEPTED FORM WITH NAMES OF SCIENCES.

YOU ARE PROBABLY FAMILIAR WITH WHAT THE BIOLOGISTS CALL SYMBIOSIS , IN WHICH DISSIMILAR PLANT OR ANIMAL ORGANISMS LIVE IN ADVANTAGEOUS ASSOCIATION WITH EACH OTHER.

FOR EXAMPLE LICHEN IS A COMPOSITE PLANT FORMED OUT OF A FUNGUS AND AN ALGA GROWING TOGETHER TO PRODUCE AN ORGANISM ENTIRELY UNLIKE EITHER COMPONENT.

THE FUNGUS GAINS NUTRIENT FROM THE ALGA & THE ALGA GAINS AN INCREASED SUPPLY OF WATER.

SO IT IS IN SYNOETICS.

THE ETYMOLOGY OF SYNNOETICS IN NO WAY INDICATES THAT ONE IS ALWAYS TALKING ABOUT A MAN AND AN AUTOMATON AS THE COMPONENTS OF THE SYNNOETIC SYSTEM.

BY USING THIS DEFINITION OF SYNNOETICS AS A CRITERION , WE HAVE BEEN FAIRLY SUCCESSFUL IN DETERMINING WHAT SUBJECT MATTER IS IN ITS REALM AND WHAT SUBJECT MATTER BELONGS TO OTHER DISCIPLINES.

AUTONOMICS IS THE STUDY OF AUTOMATA , IN GENERAL , IN SYNNOETIC SYSTEMS.

SINCE ANALOG AND DIGITAL COMPUTERS ARE BUT ONE SPECIES OF AUTOMATA , ONE BRANCH OF SYNNOETICS IS THE THEORY AND PRACTICE OF THE DESIGN , PROGRAMMING , AND APPLICATION OF COMPUTERS.

THIS BRANCH IS CALLED THE COMPUTER SCIENCES.

Fig. 5-8 Text III (Cont.)

INTELLECTRONICS , THE IMPLEMENTATION OF SYNNOETIC SYSTEMS BY ELECTRONICS , IS TAUGHT IN THE ENGINEERING SCHOOL.

WE STUDY THE THEORY AND PRACTICE OF CONTROL AND COMMUNICATION IN SYNNOETIC SYSTEMS $ THIS BRAHCH IS CALLED CYBERNETICS.

IN THE UNIVERSITY COMMUNITY , RESULTS OF INVESTIGATIONS IN SYNNOETICS ARE USED TO INVENT PROCESSES AND GENERATE

IDEAS FOR SOLVING SCHOLARLY PROBLEMS AND ATTAINING SCHOLARLY GOALS IN WHATEVER DISCIPLINES THEY ARE APPLICABLE.

IN THIS SENSE , SYNNOETICS IS SUPRADISCIPLINARY RATHER THAN INTERDISCIPLINARY.

MODELS AND SIMULATORS OF VARIOUS KINDS ARE VERY POPULAR FOR ANALYZING AND SYNTHESIZING SYNNOETIC SYSTEMS , AND FOR SOLVING PROBLEMS IN VARIOUS UNIVERSITY DISCIPLINES.

THUS , THE STUDY OF THE THEORY AND PRACTICE OF MODELING , SIMULATING , ANALYZING AND SYNTHESIZING IS IN THE DOMAIN OF SYNNOETICS.

A POPULAR MODEL USED IN STUDYING THE FUNCTIONING OF AUTOMATA IS THE HUMAN BEING , JUST AS THE STUDY OF AUTOMATA HAS GIVEN INSIGHTS INTO HUMAN FUNCTIONING.

BIONICS IS THE BRANCH OF SYNNOETICS TREATING OF SUCH SUBJECTS.

SYSTEMS THEORY IS THE NAME OF THE STUDY OF GENERIC AND ANALYTIC AND SYNTHETIC METHODS.

TEACHING , LEARNING , AND THE COMMUNICATION OF IDEAS ARE CERTAINLY SUPRADISCIPLINARY SCHOLARLY ACTIVITIES.

Fig. 5-8 Text III (Cont.)

THUS , WE STUDY THE THEORY AND PRACTICE OF TEACHING AND TEACHING AIDS , OF LEARNING AND LEARNING AIDS , AND OF COMMUNICATION IN SYNNOETIC SYSTEMS.

NOTE THAT I SAID THAT WE ARE CONCERNED WITH TEACHING , LEARNING , AND COMMUNICATION IN SYNNOETIC SYSTEMS.

THUS , IN THE DEPARTMENT OF SYNNOETICS , THEY DO NOT STUDY THE PROBLEMS OF LEARNING AND TEACHING THAT HAVE LONG BEEN IN THE PROVINCE OF EDUCATIONAL PSYCHOLOGY.

NOR DO THEY STUDY NATURAL LANGUAGES , LIKE THE STRUCTURE , GRAMMAR , AND SYNTAX OF FRENCH OR GERMAN , WHICH ARE IN THE PROVINCE OF THE LINGUISTICS DEPARTMENT.

BUT , THEY DO STUDY THE FORMAL LANGUAGES USED IN THE COMMUNICATION BETWEEN COMPONENTS OF SYNNOETIC SYSTEMS.

SIMILAR COMMENTS MAY BE MADE ABOUT STUDIES IN OPERATIONS RESEARCH , GAME THEORY , INFORMATION STORAGE , ORGANIZATION AND RETRIEVAL AND AUTOMATIC PROGRAMMING.

ONE OF THE KEY ISSUES IN THE DESIGN , CONSTRUCTION , PROGRAMMING AND USE OF SYNNOETIC SYSTEMS , IS ERROR.

SPECIFICALLY , WE STUDY THE THEORY AND PRACTICE OF THE CONTROL , PREVENTION , MASKING , DETECTION , DIAGNOSIS , AND CORRECTION OF ERRORS IN THE DESIGN , CONSTRUCTION , PROGRAMMING , AND OPERATION OF SYNNOETIC SYSTEMS.

THE STUDY OF ERROR IS CALLED HAMARTIOLOGY , FROM THE GREEK HAMARTIA MEANING TO MISS THE MARK.

FINALLY , WE PAY CONSIDERABLE ATTENTION TO THE MENTAL CHARACTERISTICS OF SYNNOETIC SYSTEMS , AS WELL AS TO THEIR COGNITIVE , SELF-ORGANIZING , AND ADAPTIVE PROPERTIES.

Fig. 5-8 Text III (Cont.)

THE SUBJECTS OF SYNNOETICS PROVIDE TOOLS AND AIDS WHICH ARE BEING USED WITH INCREASING DEGREES OF SUCCESS BY PRACTITIONERS IN ENGINEERING , LAW , MUSIC , CHEMISTRY , PHYSICS , MEDICINE , PSYCHOLOGY , AND OTHER DISCIPLINES , IN WAYS WHICH WERE QUITE UNKOWN EVEN TEN YEARS AGO.

THESE TOOLS ARE ALSO USEFUL IN THE SOLUTION OF MANAGEMENT AND CONTROL PROBLEMS IN BUSINESS , INDUSTRY , LABOR AND GOVERNMENT.

I AM SURE YOU ALL RECALL HOW THE FAMOUS STRIKE OF 1970 WAS SETTLED WHEN ONE OF OUR FACULTY MEDIATORS USED AN AUTOMATON TO AID BOTH PARTIES IN AGREEING TO WHAT WAS AT ONCE AN OPTIMUM SETTLEMENT FOR BOTH SIDES.

THE REASON THAT WE WERE NOT SATISFIED WITH THE FORMER NAME COMPUTER-RELATED SCIENCES , WAS THAT THE APPEARANCE OF THE WORD COMPUTER WAS MISLEADING * ALTHOUGH WE WERE ACUTELY AWARE OF THE PUBLIC RELATIONS VALUE OF THIS WORD.

PEOPLE IGNORED THE QUALIFYING WORD RELATED AND ASSOCIATED THE NAME EXCLUSIVELY WITH THE COMPUTER SCIENCES , WITH THE DESIGN , PROGRAMMING , AND APPLICATIONS OF COMPUTERS , WHICH IS NOW ONLY A SMALL PART OF THE NUMBER AND VARIETY OF SUBJECTS WE INCLUDE IN SYNNOETICS.

THE OTHER NAMES VARIOUSLY USED , CYBERNETICS , INFORMATION SCIENCES , COMMUNICATION SCIENCES , HAD SIMILARLY RESTRICTED CONNOTATIONS.

THE SYNNOETICS DEPARTMENT WHICH IS ADMINISTRATIVELY SITUATED IN THE COLLEGE OF ARTS AND SCIENCES , OFFERS AN INTEGRATED AND COORDINATED SYLLABUS OF ABOUT FIFTEEN UNDERGRADUATE OURSES AND TWICE THAT NUMBER OF GRADUATE COURSES AND SEMINARS IN BOTH THEORETICAL AND APPLIED SUBJECTS.

THE RESEARCH PROGRAM IN THE DEPARTMENT OF SYNNOETICS IS EXTENSIVE.

Fig. 5-8 Text III (Cont.)

IT IS SOMETIMES DESCRIBED AS SPECTACULAR.

I SAY SPECTACULAR BECAUSE INTUITION IS OFTEN MISLEADING IN THIS FIELD , SO THAT MANY OF THE RESULTS ARE SURPRISING.

ONE EXAMPLE OF AN ENLIGHTENING RESULT BY ONE OF THE GRADUATE STUDENTS IN OUR PH D PROGRAM IS A PROOF OF THE THEOREM THAT THE MAXIMUM POTENTIAL EFFECTIVENESS OF TWO PEOPLE WORKING TOGETHER ON CERTAIN KINDS OF ABSTRACT PROBLEMS IS AT LEAST AS GREAT AS THE MAXIMUM POTENTIAL EFFECTIVENESS OF ONE OF THESE PEOPLE WORKING TOGETHER WITH AN AUTOMATON ON THESE PROBLEMS.

THERE ARE BETWEEN 50 AND 100 RESEARCH PROJECTS BEING PURSUED AT ANY ONE TIME BY APPROXIMATELY THE SAME NUMBER OF RESEARCH PEOPLE.

THE RESEARCH PROGRAM IS BALANCED , COVERING A VARIETY OF SUBJECTS.

STANDARDS OF EXCELLENCE ARE HIGH.

WE ARE THUS IN THE PLEASANT POSITION OF NOT HAVING TO CONSIDER SUPPORTING A RESEARCH PROGRAM MERELY BECAUSE IT IS A POTENTIAL SOURCE OF INCOME.

FOR SYNNOETICISTS NOT REQUIRING A LABORATORY , OR SPECIAL EQUIPMENT , BUT REQUIRING LIBRARY SERVICES AND OFFICE FACILITIES , WE HAVE A BUILDING OF WELL-EQUIPPED OFFICES AND A HIGHLY EFFICIENT LIBRARY AND DOCUMENT CENTER WITH OUR OWN RETRIEVAL SYSTEM FOR MATERIAL IN SYNNOETICS.

THIS LABORATORY IS SUPPLIED WITH TEST EQUIPMENT OF VARIOUS KINDS AND , OF COURSE , WITH A BATTERY OF MINIATURIZED COMPUTERS.

Fig. 5-8 Text III (Cont.)

ABOUT SEVEN MILLION DOLLARS HAVE BY NOW BEEN INVESTED IN THE FACILITIES FOR RESEARCH IN PURE AND APPLIED SYNNOETICS.

THE DEGREE PROGRAM IN SYNNOETICS IS ORTHODOX.

FOR THE APPLIED SYNNOETICISTS , WE HAVE A FOUR STORY BUILDING WITH 10,000 SQUARE FEET OF USABLE RESEARCH SPACE ON EACH FLOOR.

WE ENJOY THIS STATUS AND CAN INSIST ON SUCH STANDARDS CHIEFLY BECAUSE WE HAVE ENOUGH MORAL AND FINANCIAL SUPPORT FROM OUR OWN INSTITUTION , FROM THE INDUSTRIAL AND BUSINESS COMMUNITY , FROM THE GOVERNMENT , AND FROM OTHER INTERESTED INDIVIDUALS.

OUR GRADUATE AND UNDERGRADUATE STUDENTS , OUR FACULTY AND VISITING FACULTY MEMBERS AND VISITING RESEARCHERS FROM INDUSTRY AND FROM GOVERNMENT , ALL CARRY OUT RESEARCH PROGRAMS UNDER THE GUIDANCE AND CONTROL OF THE ACADEMIC RESEARCH STANDARDS COMMITTEE OF THE DEPARTMENT.

Fig. 5-8 Text III (Cont.)

THE MINDS OF THE ATOMIC PHYSICISTS AT LOS ALAMOS HAD BEEN GREATLY DISTURBED AND BEWILDERED BY THE NEWS OF THE BOMB DROPPED ON HIROSHIMA.

OP RP FRISCH REMEMBERS THAT ONE DAY HE SUDDENLY HEARD LOUD CRIES OF DELIGHT IN THE CORRIDOR OUTSIDE HIS STUDY.

WHEN HE OPENED THE DOOR HE SAW SOME OF HIS YOUNGER COLLEAGUES RUSHING ALONG WITH YELLS OF WHOOPEE , LIKE AN INDIAN WAR CRY.

THEY HAD JUST HEARD , OVER THE RADIO , PRESIDENT TRUMAN READING THE REPORT BY GENERAL GROVES OF THE SUCCESSFUL USE OF THE FIRST ATOM BOMB.

IT SEEMED TO ME THAT SHOUTS OF JOY WERE RATHER INAPPROPRIATE , FRISCH NOTED DRYLY.

IT WAS HE WHO , IN 1939 , HAD FIRST CALCULATED WHAT ENORMOUS ENERGY WOULD BE RELEASED BY SPLITTING THE ATOMIC NUCLEUS.

THAT ENERGY HAD NOW DESTROYED TENS OF THOUSANDS OF LIVES.

AUGUST 6 , 1945 , WAS A BLACK DAY FOR PEOPLE LIKE EINSTEIN , FRANCK , SZILARD AND RABINOWITCH , WHO HAD DONE THEIR BEST TO PREVENT USE OF THE BOMB.

BUT THE MEN AND WOMEN UP ON THE MESA WERE IN A QUANDARY.

AFTER ALL , THEY HAD WORKED DAY AND NIGHT TO ACHIEVE THEIR GOAL.

SHOULD THEY NOW BE PROUD OF WHAT THEY HAD DONE , AS IT WAS GENERALLY CONSIDERED THEY OUGHT TO BE , IN THIS FIRST MOMENT OF SURPRISE=

OR SHOULD THEY BE ASHAMED OF THEIR WORK WHEN THEY THOUGHT OF THE SUFFERING IT HAD CAUSED SO MANY DEFENSELESS PEOPLE=

OR AGAIN , IT WAS POSSIBLE , AND THIS POSITION WOULD BE THE STRANGEST OF ALL , REALLY ONLY COMPARABLE WITH THE CONTRADICTORY DATA OF ATOMIC PHYSICS , FOR ONE AND THE SAME PERSON TO FEEL PRIDE AND SHAME SIMULTANEOUSLY.

Fig. 5-9 Text IV

THE WHOLE BUSINESS BECAME STILL MORE CONFUSING WHEN ONE CONTRASTED THE CHARACTER OF THIS EVENT , SO DIFFICULT TO GRASP , WITH THAT OF THE MEN WHO HAD BROUGHT IT ABOUT BY THE EXERCISE OF THEIR INTELLIGENCE AND THEIR DELIBERATE CONCENTRATION ON THEIR EFFORT.

IN THE EYES OF THE WORLD THEY HAD NOW GROWN TO A STATURE WHICH NO LONGER CORRESPONDED WITH , IN FACT CONTRADICTED , THEIR TRUE PERSONALITIES.

THE GODLIKE MAGNITUDE OF THEIR PERFORMANCE HAD GIVEN THEM THE STANDING OF MYTHICAL FIGURES , MORE THAN LIFE SIZE , IN THE IMAGINATION OF THE PUBLIC.

THEY WERE CALLED TITANS AND COMPARED WITH PROMETHEUS , WHO HAD CHALLENGED ZEUS , THE CONTROLLER OF THE FATES.

THEY WERE ALSO CALLED DEVIL GODS.

BUT TO THEMSELVES AND THEIR NEIGHBORS THEY SEEMED THE SAME AS THEY WERE BEFORE , HUMAN BEINGS NOT DISTINGUISHED FOR ANY SPECIAL VIRTUE OR

WICKEDNESS , CONTRADICTORY BEINGS IN THE HABIT OF CALCULATING IN BUSINESS HOURS , UNDISTRACTED BY INCIDENTAL CONSIDERATIONS , THEIR BOMBS PROBABLE RANGE OF DESTRUCTION.

BUT IN THEIR LEISURE HOURS
THEY MIGHT BE : LIKE ALVIN GRAVES , THE MOST CAREFUL OF GARDENERS , RATIONING THEIR OWN DRINKING WATER TO SAVE ONE OF THEIR PLANTS FROM DRYING UP.

ROBERT BRODE , ONE OF THE AMERICAN PHYSICISTS WHO HAD STUDIED IN GOTTINGEN TWENTY YEARS BEFORE , TRIED TO DESCRIBE HIS OWN FEELINGS AND THOSE OF SOME OF HIS COMPANIONS AT LOS ALAMOS AT THAT TIME IN THE FOLLOWING TERMS.

Fig. 5-9 Text IV (Cont.)

WE WERE NATURALLY SHOCKED BY THE EFFECT OUR WEAPON HAD PRODUCED , AND IN PARTICULAR BECAUSE THE BOMB HAD NOT BEEN AIMED , AS WE HAD ASSUMED , SPECIFICALLY AT THE MILITARY ESTABLISHMENTS IN HIROSHIMA , BUT DROPPED IN THE CENTER OF THE TOWN.

IF I AM TO TELL THE WHOLE TRUTH , I MUST CONFESS THAT OUR RELIEF WAS GREAT.

FOR AT LAST OUR FAMILIES AND FRIENDS IN OTHER CITIES AND COUNTRIES KNEW WHY WE HAD DISAPPEARED FOR YEARS ON END.

THEY HAD NOW REALIZED THAT WE , TOO , HAD BEEN DOING OUR DUTY.

FINALLY WE OURSELVES ALSO LEARNED THAT OUR WORK HAD NOT BEEN IN VAIN.

SPEAKING FOR MYSELF , I CAN SAY THAT I HAD NO FEELINGS OF GUILT.

WILLIE HIGINBOTHAM , A THIRTY FOUR YEAR OLD ELECTRONICS SPECIALIST , THE SON OF A PROTESTANT CLERGYMAN AND SOON AFTERWARD PROMINENT AMONG THOSE ATOMIC SCIENTISTS WHO FELT POLITICALLY AND MORALLY RESPONSIBLE FOR THEIR WORK , WROTE FROM LOS ALAMOS , TO HIS MOTHER.

I AM NOT A BIT PROUD OF THE JOB WE HAVE DONE.

THE ONLY REASON FOR DOING IT WAS TO BEAT THE REST OF THE WORLD TO THE DRAW.

PERHAPS THIS IS SO DEVASTATING THAT MAN WILL BE FORCED TO BE PEACEFUL.

THE ALTERNATIVE TO PEACE IS NOW UNTHINKABLE.

BUT UNFORTUNATELY THERE WILL ALWAYS BE SOME WHO DON*T THINK.

I THINK I NOW KNOW THE MEANING OF MIXED EMOTIONS.

I AM AFRAID THAT GANDHI IS THE ONLY REAL DISCIPLE OF CHRIST AT PRESENT.

Fig. 5-9 Text IV (Cont.)

ANYWAY IT IS OVER FOR NOW AND GOD WILL GIVE US STRENGTH FOR THE FUTURE.

SOME OF THE ATOMIC PHYSICISTS AT WORK IN LOS ALAMOS KNEW THAT THE LAST OF THE ATOM BOMBS , ONLY THREE HAD BEEN COMPLETED SO FAR , WAS STORED ON THE ISLAND OF TINIAN , READY FOR USE.

IN CONTRAST TO THE BOMB DROPPED ON HIROSHIMA , CALLED THE THIN MAN ,

THIS WAS KNOWN AS THE FAT MAN.

THERE WAS EVERY REASON TO SUPPOSE THAT , WITH A SMALLER EXPENDITURE , IT WOULD BE EVEN MORE DESTRUCTIVE.

ONE OF THE CONSTRUCTORS OF THIS LAST BOMB , WHO FOR OBVIOUS REASONS DOES NOT WISH TO BE NAMED , ADMITS.

I DREADED THE USE OF THIS BETTER BOMB.

I HOPED THAT IT WOULD NOT BE USED AND TREMBLED AT THE THOUGHT OF THE DEVASTATION IT WOULD CAUSE.

AND YET , TO BE QUITE FRANK , I WAS DESPERATELY ANXIOUS TO FIND OUT WHETHER THIS TYPE OF BOMB WOULD ALSO DO WHAT WAS EXPECTED OF IT , IN SHORT , WHETHER ITS INTRICATE MECHANISM WOULD WORK.

THESE WERE DREADFUL THOUGHTS , I KNOW , AND STILL I COULD NOT HELP HAVING THEM.

TWENTY FIVE ATOMIC SCIENTISTS AND THEIR ASSISTANTS HAD MEANWHILE TRAVELED FROM LOS ALAMOS TO TINIAN , UNDER THE LEADERSHIP OF NORMAN RAMSAY , TO GET THE FAT MAN READY FOR USE.

SO LONG AS NO ONE ON THE ISLAND KNEW WHAT THE LONG , HAIRED GUYS WERE REALLY DOING IN THE BUILDINGS THEY OCCUPIED , SURROUNDED BY A SPECIAL GUARD , THE MILITARY PERSONNEL HAD CONSIDERED THEM MERELY OBJECTS FOR GOOD , NATURED RIDICULE.

Fig. 5-9 Text IV (Cont.)

BUT AS SOON AS THE NEWS OF THE DROPPING OF THE FIRST ATOM BOMB BECAME KNOWN , THEY WERE TREATED AS HEROES.

THERE WERE GOOD GROUNDS FOR THIS ATTITUDE.

FOR THE MEN OF THE MARINE CORPS STATIONED ON THE ISLAND HAD LEARNED THAT THEY WERE TO BEAR THE BRUNT , AS FRONT , LINE TROOPS , OF THE FORTHCOMING LANDING IN TOKYO BAY.

BUT THERE WAS NOW REASON TO HOPE THAT THIS OPERATION MIGHT NEVER TAKE PLACE.

A LARGE NUMBER OF JOURNALISTS BEGAN TO ARRIVE AT THE AIR BASE , AS WELL AS CERTAIN SENIOR OFFICERS , WHO DISTRIBUTED BADGES TO THE CREW OF THE ENOLA GAY , THE FIRST ATOM BOMB AIRCRAFT , WHICH WAS NAMED FOR THE MOTHER OF THE PILOT , PAUL TIBBETTS.

AMONG THE VERY IMPORTANT PERSONS WHO VISITED TINIAN AT THIS TIME WAS GENERAL TOOEY SPAATZ , COMMANDER IN CHIEF OF ALL THE AIR FORCES ENGAGED ON THAT FRONT.

HERBERT AGNEW , ONE OF THE ATOMIC EXPERTS ON THE ISLAND , RELATES THAT WE NATURALLY TOOK HIM , AMONG OTHER PLACES , TO THE HANGAR WHERE WE HAD GOT THE FIRST BOMB READY FOR RELEASE.

ONE OF MY COLLEAGUES SHOWED HIM THE LITTLE BOX IN WHICH THE CENTRAL MECHANISM OF THE BOMB HAD BEEN PACKED BEFORE WE FITTED IT.

THE GENERAL LOST HIS TEMPER.

HE TURNED TO HIS ADJUTANT AND SAID.

YOU CAN BELIEVE THIS BOY'S LINE OF SALES TALK IF YOU LIKE.

BUT HE DOESN'T PULL MY LEG.

THE GENERAL SIMPLY REFUSED TO BELIEVE THAT SUCH A LITTLE THING HAD CAUSED SUCH MIGHTY DESTRUCTION.

Fig. 5-9 Text IV (Cont.)

IT WAS ARRANGED THAT CERTAIN ATOMIC SCIENTISTS , INCLUDING ALVAREZ , AGNEW AND THE BRITISH BOMB EXPERT PENNEY , SHOULD ACCOMPANY THIS SECOND ATOMIC AIR RAID IN ANOTHER PLANE.

WHILE ALVAREZ AND HIS FRIENDS PHILIP MORRISON AND ROBERT SERBER WERE DRINKING A CAN OF BEER , SHORTLY BEFORE STARTING ON THE RAID , THEY HAD A SUDDEN BRAIN WAVE.

THEY DECIDED TO DROP A LETTER WITH THE BOMB ADDRESSED TO THEIR JAPANESE FRIEND PROFESSOR SAGANE , WITH WHOM THEY HAD WORKED IN CLOSE CONTACT AT THE RADIATION LABORATORY IN BERKELEY BEFORE THE WAR.

THREE COPIES OF THE LETTER WERE HANDWRITTEN IN GREAT HASTE , AND ONE COPY SECURELY FASTENED TO EACH OF THE THREE MEASURING INSTRUMENTS WHICH ALVAREZ WOULD RELEASE OVER THE TARGET.

WE ARE SENDING YOU THIS AS A PERSONAL MESSAGE , TO URGE THAT YOU USE YOUR INFLUENCE , AS A REPUTABLE NUCLEAR PHYSICIST , TO CONVINCE THE JAPANESE GENERAL STAFF OF THE TERRIBLE CONSEQUENCES WHICH WILL BE SUFFERED BY YOUR PEOPLE IF YOU CONTINUE IN THIS WAR.

YOU HAVE KNOWN FOR SEVERAL YEARS THAT AN ATOMIC BOMB COULD BE BUILT IF A NATION WERE WILLING TO PAY THE ENORMOUS COST OF PREPARING THE NECESSARY MATERIAL.

NOW THAT YOU HAVE SEEN THAT WE HAVE CONSTRUCTED THE PRODUCTION PLANTS , THERE CAN BE NO DOUBT IN YOUR MIND THAT ALL THE OUTPUT OF THESE FACTORIES , WORKING 24 HOURS A DAY , WILL BE EXPLODED ON YOUR HOMELAND.

WITHIN THE SPACE OF THREE WEEKS WE HAVE PROOF FIRED ONE BOMB IN THE AMERICAN DESERT , EXPLODED ONE IN HIROSHIMA AND FIRED THE THIRD THIS MORNING.

WE IMPLORE YOU TO CONFIRM THESE FACTS TO YOUR LEADERS AND TO DO YOUR UTMOST TO STOP THE DESTRUCTION AND WASTE OF LIFE WHICH CAN ONLY RESULT IN THE TOTAL ANNIHILATION OF ALL YOUR CITIES , IF CONTINUED.

Fig. 5-9 Text IV (Cont.)

AS SCIENTISTS , WE DEPLORE THE USE TO WHICH A BEAUTIFUL DISCOVERY HAS BEEN PUT , BUT WE CAN ASSURE YOU THAT UNLESS JAPAN SURRENDERS AT ONCE THIS RAIN OF ATOMIC BOMBS WILL INCREASE MANYFOLD IN FURY.

ONE OF THESE MESSAGES WAS FOUND AFTER THE BOMBARDMENT OF NAGASAKI AND HANDED OVER TO THE JAPANESE NAVAL INTELLIGENCE DIVISION.

IT WAS NOT UNTIL MUCH LATER THAT IT REACHED THE MAN TO WHOM IT WAS WRITTEN.

IT IS NOT KNOWN TO WHAT EXTENT THIS LETTER CONTRIBUTED TO BRING ABOUT JAPAN*S CAPITULATION.

IN REALITY THE UNITED STATES HAD NOT A SINGLE ATOMIC BOMB IN RESERVE , READY FOR USE , AT THE TIME THE MESSAGE WAS DROPPED.

NOR COULD ANY FRESH BOMBS BE PRODUCED FOR SEVERAL WEEKS , POSSIBLY FOR SEVERAL MONTHS , AHEAD.

THE AMERICAN GENERAL STAFF HAD ONE OBJECT IN PARTICULAR IN RAIDING NAGASAKI.

IT WAS DESIRED TO GIVE THE ENEMY THE IMPRESSION THAT THE UNITED STATES ALREADY POSSESSED A WHOLE ARSENAL OF ATOM BOMBS AND THUS INDUCE THE JAPANESE TO LAY DOWN THEIR ARMS IMMEDIATELY.

THE BLUFF WAS SUBSTANTIATED , IN ALL INNOCENCE , BY THE MESSAGE WHICH THE THREE PHYSICISTS HAD COMPOSED FOR HUMANITARIAN ENDS.

CONSEQUENTLY , EVEN THE FRIENDSHIP AMONG SCIENTISTS OF DIFFERENT NATIONS HAD BEEN MISUSED AS A WEAPON.

LATE IN THE EVENING OF AUGUST 11 , 1945 , THE AMERICAN RADIO ANNOUNCED.

THE UNITED PRESS HAS JUST REPORTED FROM BERNE IN SWITZERLAND THAT THE JAPANESE GOVERNMENT HAS OFFERED UNCONDITIONAL SURRENDER....

THE NEWS CAUSED ECSTATIC REJOICING AT LOS ALAMOS.

Fig. 5-9 Text IV (Cont.)

ALL CONTRADICTORY FEELINGS AND DOUBTS WERE FOR THE MOMENT FORGOTTEN.

FURTHER BLOODSHED HAD BEEN PREVENTED BY THE TWO BOYS BORN ON THE HILL.

THE WAR WAS AT AN END.

A RUSH WAS MADE TO EXTRACT FROM THEIR HIDING PLACES THE SUPPLIES OF WHISKEY , GIN , VODKA AND OTHER ALCOHOLIC BEVERAGES , LONG SINCE , IN EXPECTATION OF THIS HOUR , SMUGGLED INTO THE CITY OF LABORATORIES , HITHERTO SUBJECT TO A STRICT PROHIBITION.

PEOPLE TOUCHED GLASSES HAPPILY AND DRANK TO PEACE.

AT THE CLIMAX OF ONE OF THE MANY IMPROVISED VICTORY PARTIES PROFESSOR KA , ONE OF THE LEADING SPECIALISTS IN THE EXPLOSIVES DEPARTMENT , ROSE TO HIS FEET , REELING SLIGHTLY , AND DASHED OUT INTO THE NIGHT BEFORE ANYONE COULD STOP HIM.

EVER SINCE AUGUST 6 HE HAD BEEN WORKING , UNKNOWN TO ALL BUT THE SECURITY AUTHORITIES , ON A SURPRISE OF HIS OWN , TO BE REVEALED THE DAY THE WAR ENDED.

A MOMENT LATER FLASHES AND ROARS CAME FROM ALL DIRECTIONS.

PEOPLE WHO RUSHED OUT OF THEIR HOUSES BEHELD A MAGNIFICENT SPECTACLE.

THE WHOLE OF THE TOWN OF LOS ALAMOS , PERCHED ON ITS PRECIPICE , WAS ILLUMINATED BY A BLINDING , SHIMMERING GLARE.

THE TOWERING RED ROCKS GLOWED IN THE REFLECTION OF THE FLAMES.

ARROWY FOUNTAINS OF SPARKS SHOT UP OUT OF THE CANYONS.

THERE SEMED NO END TO THE BANGS , LOUD REPORTS AND THUNDEROUS ECHOES.

PROFESSOR KA HAD CONNECTED BY WIRE TWO OR THREE DOZEN SMALL MUNITIONS DUMPS AT CONCEALED SPOTS $ BY PRESSING A BUTTON THEY WOULD EXPLODE.

Fig. 5-9 Text IV (Cont.)

AFTER THE VICTORY FIREWORKS HAD BURNED THEMSELVES OUT AND ONLY OCCASIONAL BELATED EXPLOSIONS COULD BE HEARD AS AN AFTERMATH TO THE MAIN DISPLAY , PEOPLE RETURNED TO THEIR HOUSES AND BEGAN TO LISTEN IN AGAIN IN THE HOPE OF HEARING MORE DETAILS OF THE SURRENDER.

THEY LEARNED , HOWEVER , THAT THE NEWS OF JAPAN*S CAPITULATION HAD UNFORTUNATELY BEEN PREMATURE.

FOUR DAYS LATER CAME THE ANNOUNCEMENT THAT JAPAN REALLY DID SURRENDER.

THIS TIME THERE WAS NO REJOICING AT FIRST , BUT AFTER A WHILE , DESPITE THE LATE HOUR OF THE ANNOUNCEMENT , A VICTORY PARADE WAS ORGANIZED AT LOS ALAMOS.

IT WAS LED BY A JEEP WITH MORE THAN A DOZEN OF THE YOUNGER SCIENTISTS CLINGING TO IT.

THE SLIM FIGURE OF WILLIE HIGINBOTHAM WAS SEATED ON THE SHOULDERS OF THE DRIVER.

HE PLAYED LIVELY TUNES ON HIS ACCORDIAN , BANGING A KETTLEDRUM MADE OF THE LIDS OF TWO DUSTBINS , TO MAKE SURE THAT THOSE WHO HAPPENED TO BE ASLEEP SHOULD HAVE NO DOUBT THAT PEACE HAD BROKEN OUT.

LIGHTS WENT ON AGAIN IN MOST OF THE HOUSES.

DORM PARTIES BEGAN IN THE BACHELOR SLEEPING QUARTERS.

DANCING WENT ON UNTIL DAWN.

THE STAFF WERE EXCUSED FROM WORK ON THE FOLLOWING DAY.

SO IT CONTINUED FOR TWO DAYS AND TWO NIGHTS.

BUT WHEN THE REJOICINGS CAME TO AN END , IT WAS FOUND THAT FOR THE PRESENT EVERYTHING WAS TO GO ON AS BEFORE.

Fig. 5-9 Text IV (Cont.)

THE WORLD MIGHT BE UNDER THE IMPRESSION THAT PEACE HAD COME AGAIN.

BUT SO FAR AS THE PEOPLE ON RESEARCH AT LOS ALAMOS , OAK RIDGE , HANFORD AND CHICAGO WERE CONCERNED , THE SAME STRICT RULES OF SECRECY PREVAILED AS HAD BEEN IN FORCE DURING THE WAR.

THE YOUNGER WORKERS ON THE MANHATTAN PROJECT , IN PARTICULAR , FOUND THESE CONDITIONS UNBEARABLE.

THEY BEGAN TO GRUMBLE.

A TYPICAL COMPLAINT CAME FROM HERBERT ANDERSON , A YOUNG AMERICAN PHYSICIST.

HE HAD TAKEN PART IN FERMI*S FIRST URANIUM EXPERIMENTS AT COLUMBIA UNIVERSITY , DURING WHICH HE HAD CONTRACTED LIFELONG BERYLLIUM POISONING.

SHORTLY AFTER THE WAR ANDERSON WROTE TO A FRIEND.

WE OUGHT TO RESIST EVERY ENCROACHMENT UPON OUR RIGHTS AS HUMAN BEINGS AND CITIZENS.

THE WAR HAS BEEN WON.

WE WISH TO BE FREE AGAIN.

THESE SCIENTISTS WERE NOT ONLY CONCERNED ABOUT THEIR PERSONAL FREEDOM.

THEY DESIRED IN PARTICULAR TO BE FREE TO ENLIGHTEN THEIR FELLOW MEN ABOUT THE TERRORS OF THE NEW WEAPON.

WHEN THEY READ IN THE NEWSPAPERS , AT THIS TIME , THAT MEMBERS OF CONGRESS WERE IN FAVOR OF THE UNITED STATES KEEPING THE SECRET OF THE ATOM BOMB TO THEMSELVES , THE PHYSICISTS WOULD HAVE LIKED TO RETORT THAT THERE WAS NO ATOMIC SECRET WHICH COULD NOT BE DETECTED WITHIN A VERY SHORT TIME BY ANY NATION SCIENTIFICALLY OF THE FIRST RANK.

Fig. 5-9 Text IV (Cont.)

THEY WOULD HAVE LIKED TO PRESS FOR THE IMMEDIATE CONVOCATION , ON AMERICAN INITIATIVE , OF AN INTERNATIONAL CONFERENCE ON THE CONTROL OF ATOMIC DEVELOPMENT , AS HAD BEEN DESIRED BY BOHR , SZILARD AND THE AUTHOR OF THE FRANCK REPORT.

A SPECIAL SUBJECT BROUGHT UP BY THE SCIENTISTS AT LOS ALAMOS WAS THE GAME OF HIDE AND SEEK PLAYED BY THE ARMY WITH THE PROBLEM OF RADIOACTIVITY.

EVEN BEFORE THE ATOMIC WEAPON HAD FIRST BEEN USED SOME PHYSICISTS HAD ENTREATED GENERAL GROVES TO ALLOW PAMPHLETS TO BE DROPPED AT THE SAME TIME AS THE BOMB , POINTING OUT THE UNFAMILIAR DANGERS OF RADIOACTIVITY ARISING FROM THE EXPLOSION OF THIS NEW WEAPON.

THIS REQUEST HAD BEEN REFUSED BY THE MILITARY AUTHORITIES , FOR THEY FEARED THAT SUCH WARNINGS MIGHT BE INTERPRETED AS A CONFESSION THAT THEY HAD BEEN EMPLOYING A TYPE OF WEAPON LIKE POISON GAS.

THEY PROCEEDED , PROBABLY FROM SIMILAR MOTIVES , TO TRY TO DIVERT ATTENTION FROM THE RADIOACTIVE EFFECTS OF ATOMIC BOMBARDMENT.

IT WAS EXPLAINED THAT THERE WAS NOW NO DANGEROUS RADIOACTIVITY TO BE FOUND IN THE RUINS OF HIROSHIMA , AND THE NUMBER THE VICTIMS WHO HAD BEEN EXPOSED , AT THE MOMENT OF THE EXPLOSION , TO A FATAL DOSE OF RADIATION OR ONE LIKELY TO CAUSE CHRONIC ILLNESS , WAS KEPT SECRET.

GROVES STATED OPENLY AT A CONGRESSIONAL HEARING THAT HE HAD HEARD DEATH FROM RADIATION WAS VERY PLEASANT.

SUCH OBSERVATIONS MADE THE LOS ALAMOS SCIENTISTS* BLOOD BOIL.

FOR AT THAT VERY MOMENT THEIR TWENTY SIX YEAR OLD COLLEAGUE HARRY DAGNIAN WAS STRUGGLING AGAINST THE MENACE OF A CRUEL DEATH FROM THE EFFECTS OF RADIATION.

ON AUGUST 21 , 1945 , DURING AN EXPERIMENT WITH A SMALL QUANTITY OF FISSILE MATERIAL , DAGNIAN HAD SET OFF A CHAIN REACTION FOR THE FRACTION OF A SECOND.

Fig. 5-9 Text IV (Cont.)

HIS RIGHT HAND HAD RECEIVED A HUGE DOSE OF RADIATION.

AFTER ADMISSION TO HOSPITAL WITHIN HALF AN HOUR OF THE ACCIDENT , THE PATIENT HAD AT FIRST NOTICED ONLY A CERTAIN LOSS OF SENSATION IN THE FINGERS , OCCASIONALLY SUPERSEDED BY SLIGHT TINGLING.

BUT SOON HIS HANDS GREW MORE AND MORE SWOLLEN AND HIS GENERAL CONDITION DETERIORATED RAPIDLY.

DELIRIUM SET IN.

THE YOUNG PHYSICIST COMPLAINED OF SEVERE INTERNAL PAINS , FOR IT WAS NOW THAT THE EFFECT OF THE GAMMA RAYS , WHICH HAD PENETRATED FAR BENEATH THE SKIN TO THE INTERIOR OF THE BODY , BEGAN TO BE PERCEPTIBLE.

THE PATIENT'S HAIR DROPPED OUT.

THE WHITE COUPUSCLES OF HIS BLOOD INCREASED RAPIDLY.

TWENTY FOUR DAYS LATER HE DIED.

FOR THE FIRST TIME DEATH BY RADIATION , WHICH THE MEN OF LOS ALAMOS HAD INFLICTED UPON THOUSANDS OF JAPANESE BY CONSTRUCTING THEIR WEAPON , HAD OVERTAKEN ONE OF THEMSELVES.

FOR THE FIRST TIME THE DANGEROUS EFFECTS OF THE NEW POWER HAD BEEN BROUGHT CLOSE , NOT IN THE FORM OF A DISTANT STATISTIC , BUT AS THE SUFFERING , PAIN AND FATAL SICKNESS OF ONE OF THEIR OWN GROUP.

THE ACCIDENT TO HENRY DAGNIAN INTENSIFIED THE MOVEMENT WHICH HAD BEGUN IN ALL THE ATOMIC LABORATORIES AMONG THOSE SCIENTISTS WHO INTENDED TO TELL THE WORLD THE WHOLE TRUTH ABOUT THE NEW WEAPON AND ENTREAT THEIR FELLOW MEN TO RENOUNCE ALL USE OF ATOMIC ENERGY IN WARFARE.

NINE DAYS AFTER DAGNIAN HAD BEEN TAKEN TO THE HOSPITAL SHED ON THE HILL , THE ASSOCIATION OF ATOMIC SCIENTISTS , HEADED BY HIGINBOTHAM , WAS FORMED IN LOS ALAMOS.

Fig. 5-9 Text IV (Cont.)

ABOUT A HUNDRED OF THE MEN IN RESEARCH IMMEDIATELY JOINED IT.

SIMILAR GROUPS HAD ALREADY ARISEN IN CHICAGO , AT OAK RIDGE AND IN NEW YORK.

THE GROUPS GOT IN TOUCH WITH ONE ANOTHER AND CAME TO A COMMON DECISION TO ENLIGHTEN THE PUBLIC AND THUS BRING STRONG PRESSURE TO BEAR ON THE STATESMEN OF THE COUNTRY . IN SPITE OF THE FACT THAT SUCH AN APPEAL WOULD CONSTITUTE AN INFRINGEMENT OF THE ARMY REGULATIONS TO WHICH THE MEMBERS OF THE ASSOCIATION WERE STILL SUBJECT.

SUCH WAS THE START OF THE MOVEMENT WHICH LATER BECAME KNOWN , IN A SOMEWHAT EXAGGERATED PHRASE , AS THE REVOLT OF THE ATOMIC SCIENTISTS.

SELDOM CAN JUBILATION HAVE MADE A MAN SO SAD AND ADULATION MADE A MAN SO SKEPTICAL AS THEY DID ROBERT OPPENHEIMER AS HE WATCHED THE FRENZIED DELIGHT WITH WHICH HIS COUNTRYMEN GREETED THE END OF THE SECOND WORLD WAR.

HE , KNOWN ONLY TO A SMALL CIRCLE OF HIS SCIENTIFIC COLLEAGUES AND A HANDFUL OF POLITICIANS , HAD SUDDENLY COME TO BE AN OBJECT OF MASS ADMIRATION.

THOUGH IT WAS A DESIGNATION WHICH HE ALWAYS REPUDIATED AS OVERSIMPLIFIED , THE LEARNED PHYSICIST WAS CALLED THE FATHER OF THE ATOM BOMB AND HE WAS SALUTED ON ALL SIDES AS A VICTORIOUS CAMMANDER IN CHIEF.

HE WAS REGARDED NOT ONLY AS THE MAN WHOSE MIRACULOUS WEAPON HAD SPARED THE COUNTRY THE DREADED PROSPECT OF HEAVY CASUALTIES IN AN INVASION OF JAPAN AND ANOTHER WINTER OF WAR , BUT ALSO AS A NEW KIND OF PEACEMAKER , WHOSE AMAZING DISCOVERY WOULD MAKE ALL ARMIES AND WARS SUPERFLUOUS FROM THIS TIME ONWARD.

OPPENHEIMER , HOWEVER , KNEW TOO MUCH TO BE ABLE TO ACQUIESCE IN THIS OVERWHELMING TIDE OF OPTIMISM ABOUT THE FUTURE.

Fig. 5-9 Text IV (Cont.)

HE MUST AT THAT TIME HAVE OBSERVED ALL THOSE WHO WERE NOT IN THE PICTURE AND SHOWED SUCH ENTHUSIASM FOR THE COMING PARADISE OF PEACE WITH THE SAME SADNESS WITH WHICH ADULTS SOMETIMES WATCH THE INNOCENT PLAY OF CHILDREN.

WHEN OPPENHEIMER SPECULATED ABOUT THE FUTURE , HIS MIND WAS OVERSHADOWED BY TWO COMPLEX SETS OF FACTS.

IN THE FIRST PLACE , IT WAS CLEAR TO HIM THAT THE TWO ATOM BOMBS WHICH HAD BEEN DROPPED ON HIROSHIMA AND NAGASAKI DID NOT REPRESENT THE HEIGHT OR EVEN AN EXTREME LIMIT , BUT ONLY THE BEGINNING , OF A NEW KIND OF WEAPONS DEVELOPMENT WHOSE LIMITS COULD STILL NOT BE SEEN.

EVEN BEFORE THE COMPLETION OF THE URANIUM BOMB HE HAD WRITTEN TWO LETTERS DATED SEPTEMBER 20 , 1944 , AND OCTOBER 4 , 1944 , TO A FRIEND , PROFESSOR TOLMAN , CHAIRMAN OF A RESEARCH COMMITTEE CONSTITUTED ALMOST A YEAR BEFORE THE END OF THE WAR TO STUDY THE FUTURE OF ATOMIC ENERGY , POINTING OUT THAT BECAUSE OF WARTIME CONDITIONS THEY HAD BEEN ABLE TO PRODUCE ONLY A RELATIVELY PRIMITIVE ATOMIC WEAPON.

THESE HAD BEEN HIS WORDS.

WHATEVER TECHNICAL SUPERIORITY THIS COUNTRY MAY AT PRESENT POSSESS IN DEALING WITH THE SCIENTIFIC AND TECHNICAL ASPECTS OF THE PROBLEM OF THE EXPLOITATION OF NUCLEAR REACTIONS TO PRODUCE EXPLOSIVE WEAPONS HAS RESULTED FROM A FEW YEARS OF WORK WHICH WAS , TO BE SURE , INTENSIVE , BUT INEVITABLY BADLY PLANNED.

SUCH SUPERIORITY CAN PROBABLY ONLY BE MAINTAINED THROUGH CONTINUED FURTHER DEVELOPMENT OF BOTH THE TECHNICAL AND THE UNDERLYING SCIENTIFIC ASPECTS OF THE PROBLEM.

FOR THIS PURPOSE BOTH THE AVAILABILITY OF RADIOACTIVE MATERIALS AND THE PARTICIPATION OF QUALIFIED ENGINEERS AND SCIENTISTS ARE EQUALLY INDISPENSABLE.

Fig. 5-9 Text IV (Cont.)

NO GOVERNMENT CAN ADEQUATELY MEET ITS RESPONSIBILITIES FOR DEFENSE IF IT RESTS CONTENT WITH THE WARTIME RESULTS OF THIS PROJECT.

IN THE SECOND PLACE OPPENHEIMER KNEW FROM PERSONAL EXPERIENCE , THE DEGRADING INTERVIEWS FORCED UPON HIM IN 1943 , THAT THE GERM OF ATOMIC RIVALRY BETWEEN THE TWO GREAT POWERS , THE UNITED STATES AND THE SOVIET UNION , THEN STILL ALLIES , ALREADY EXISTED.

UNLIKE HIS MILITARY CHIEF , GENERAL GROVES , WHO BELIEVED THAT IT WOULD BE TEN , TWENTY OR EVEN SIXTY YEARS BEFORE THE USSR COULD DEVELOP ITS OWN ATOM BOMB , OPPENHEIMER HAD A HIGH OPINION OF SOVIET RESEARCH.

HIS VIEWS HAD BEEN QUITE RECENTLY SUBSTANTIATED BY IRVING LANGMUIR , AN AMERICAN WHO HAD WON A NOBEL PRIZE FOR CHEMISTRY , ON HIS RETURN FROM MOSCOW , WHERE HE HAD BEEN THE GUEST OF THE ACADEMY OF SCIENCES.

LANGMUIR HAD NO DOUBT THAT THE RUSSIANS COULD , IF THEY WISHED , CONSTRUCT ATOM BOMBS WITHIN A RELATIVELY SHORT TIME AND MIGHT WELL HAVE DONE SO ALREADY.

HE EVEN CONSIDERED THAT THE SOVIET UNION , AS A TOTALITARIAN STATE , COULD EASILY INITIATE A BIGGER PROGRAM OF ATOMIC ARMAMENTS THAN WOULD BE POSSIBLE FOR THE UNITED STATES.

SUCH CONSIDERATIONS OF PRACTICAL POLITICS AT FIRST PREVENTED OPPENHEIMER , THE ATOMIC PHYSICIST WHOSE PUBLIC PRESTIGE PROBABLY STOOD HIGHER JUST AFTER THE WAR THAN ANY OTHER , FROM RAISING HIS VOICE TO JOIN IN THE STEADILY INCREASING CHORUS OF WARNINGS.

WHILE MEN LIKE EINSTEIN , SZILARD , FRANCK AND UREY TALKED OF THE NEED FOR AN UNDERSTANDING WITH RUSSIA , OPPENHEIMER WAS AT THE VERY SAME TIME TRYING TO ARRANGE FOR PATROLS OF AIRCRAFT FURNISHED WITH SENSITIVE MEASURING INSTRUMENTS TO DETECT ANY ATOMIC TEST EXPLOSION THAT MIGHT TAKE PLACE IN RUSSIA OR ANYWHERE ELSE IN THE WORLD.

DURING THE ACTUAL WEEK IN WHICH THE FIRST TWO ATOM BOMBS WERE DROPPED , OPPENHEIMER , COMPTON , FERMI AND LAWRENCE HAD ALREADY LAID DOWN THE LINES ON WHICH FUTURE ATOMIC ARMAMENT SHOULD PROCEED.

Fig. 5-9 Text IV (Cont.)

OPPENHEIMER HIMSELF STRENUOUSLY OPPOSED THE GROWING TENDENCY OF SCIENTISTS AND ALSO OF MANY GOVERNMENT OFFICIALS TO HAND BACK LOS ALAMOS TO THE DESERT FOXES.

IN PERSONAL CONVERSATIONS AND PUBLIC SPEECHES HE ENDEAVORED , USUALLY WITH SUCCESS , TO PERSUADE HIS COLLABORATORS TO REMAIN AT LOS ALAMOS FOR , AT ANY RATE , SOME TIME LONGER.

HE FELT HIMSELF MORE THAN EVER RESPONSIBLE FOR THIS EXTRAORDINARY SETTLEMENT ON THE EDGE OF THE WORLD.

HIS PERSUASIVE ABILITY AND DIPLOMATIC SKILL GAINED HIM NEW FRIENDS AMONG THE SOLDIERS STATIONED AT LOS ALAMOS.

THEY HAD EXPECTED A SPECIAL PUBLIC CITATION BY THE PRESIDENT AS A REWARD FOR THEIR SERVICES.

WHEN IT FAILED TO MATERIALIZE THEY GRUMBLED AND PROTESTED.

OPPENHEIMER LEARNED OF THEIR DISCONTENT S HE WROTE A PERSONAL LETTER OF THANKS , SIGNED WITH HIS OWN HAND , AND HAD A COPY DELIVERED TO EACH MAN.

THIS STEP MADE HIM MORE POPULAR THAN EVER WITH THE G.I.*S.

ON THE OTHER HAND OPPENHEIMER BEGAN TO LOSE MORE AND MORE FRIENDS AMONG HIS CLOSEST COLLEAGUES , WHO , WITH FEW EXCEPTIONS , HAD IDOLIZED HIM FOR YEARS.

THEY HAD HOPED THAT HE WOULD NOW ACT AS THEIR SPOKESMAN TO THE WORLD , SINCE THEY THEMSELVES WERE STILL SWORN TO SECRECY.

BUT WHENEVER THEY APPROACHED HIM , HE INVARIABLY REPLIED.

JUST HOW DELICATE QUESTIONS AS TO THE FUTURE CONTROL OF ATOMIC ENERGY ARE BEING DISCUSSED.

Fig. 5-9 Text IV (Cont.)

WE SCIENTISTS MUST BE CAREFUL NOT TO ROCK THE BOAT.

WE MUSTN•T INTERFERE.

THE DELAYING ANSWERS OPPENHEIMER GAVE TO THE WORRIED YOUNG SCIENTISTS OF LOS ALAMOS AND ALSO TO THOSE OF OAK RIDGE , WHEN HE PAID THAT ESTABLISHMENT A VISIT , RESEMBLED THE ADVICE PROFFERED BY A. H. COMPTON , HEAD OF THE METALLURGICAL LABORATORY IN CHICAGO , TO THE SCIENTISTS OF THE LABORATORY.

HE REPEATED AGAIN AND AGAIN.

YOU SHOULD NOT TAKE ANY ACTION.

IF YOU DO , YOU WILL ENDANGER IMPORTANT POLITICAL DEVELOPMENTS.

IT SEEMED CLEAR THAT HE COULD ONLY BE REFERRING TO SECRET NEGOTIATIONS WITH MOSCOW.

SO THE SCIENTISTS HELD THEIR TONGUES , AS COMPTON RECOMMENDED.

BUT TOWARDS THE END OF SEPTEMBER THE NEWS FILTERED THROUGH THAT NO CONVERSATIONS WHATEVER WITH THE RUSSIANS ON ATOMIC PROBLEMS HAD YET BEEN INITIATED FROM THE AMERICAN SIDE.

AT A CABINET MEETING ON SEPTEMBER 21 THE AMERICAN GOVERNMENT , WITH THE EXCEPTION OF THE FORMER VICE PRESIDENT AND PRESENT SECRETARY OF COMMERCE , WALLACE , HAD DECIDED FOR THE TIME BEING AGAINST ANY REVELATION OF ATOMIC SECRETS , REGARDED AS A SACRED TRUST.

WHAT , THEN , COULD COMPTON HAVE MEANT=

SZILARD DETERMINED TO FIND OUT.

IT WAS DUE TO HIS PERTINACITY THAT THE SCIENTISTS EVENTUALLY DISCOVERED THE TRUTH AT WHICH OPPENHEIMER AND COMPTON HAD ONLY HINTED.

CONVERSATIONS RELATING TO THE CONTROL OF ATOMIC ENERGY HAD IN FACT TAKEN PLACE IN WASHINGTON.

Fig. 5-9 Text IV (Cont.)

ONLY THEY HAD NOT DEALT WITH INTERNATIONAL CONTROL , AS HAD BEEN SUPPOSED , BUT WITH THE FORM OF CONTROL TO WHICH THE NEW POWER WAS IN FUTURE TO BE SUBJECT IN THE UNITED STATES.

ALMOST EVERY SCIENTIST AT THAT TIME WAS OF THE OPINION THAT THERE OUGHT TO BE SOME SORT OF PUBLIC SUPERVISION OF ATOMIC ENERGY.

NOW , FOR THE FIRST TIME IN HISTORY , SOMETHING HAD BEEN INVENTED WHICH IN IRRESPONSIBLE HANDS MIGHT IMPERIL THE LIVES OF ALL CITIZENS OF THE STATE AND PERHAPS OF THE ENTIRE POPULATION OF THE GLOBE.

BUT EVERYTHING DEPENDED UPON WHO , IN THE NAME OF THE NATION , WOULD EXERCISE SUCH CONTROL.

THE QUESTION WAS WHETHER DIRECTION OF THE NEW ATOMIC INDUSTRY BE PLACED , AS IN TIME OF WAR , IN THE HANDS OF THE MILITARY AUTHORITIES.

SZILARD GATHERED FROM COMPTON THAT SOME SUCH PLAN WAS IN VIEW.

THE LATTER ALSO REVEALED TO HIM , UNDER PRESSURE , THAT THE WAR DEPARTMENT , WHICH HAD FRAMED THE NEW LEGISLATIVE PROPOSALS FOR THE CONTROL OF ATOMIC ENERGY , CONSIDERED IT MOST IMPORTANT THAT THE BILL SHOULD PASS BOTH HOUSES OF CONGRESS WITHOUT DIFFICULTIES AND ALSO , IF POSSIBLE , WITHOUT DEBATE.

AT THIS NEWS SZILARD LOST HIS PATIENCE.

HE WENT STRAIGHT TO WASHINGTON IN ORDER TO FIND OUT WHAT EXACTLY THIS BILL , HITHERTO SO ANXIOUSLY KEPT SECRET FROM ALL THE WORLD , MIGHT CONTAIN.

BOB LAMB CIO REPRESENTATIVE IN WASHINGTON , PROCURED HIM A COPY OF THE BILL.

SZILARD WAS GREATLY AGITATED AT WHAT HE READ.

HIS NEGATIVE REACTION TO THE BILL'S CONTENTS WAS SUPPORTED BY THE LEGAL FACULTY OF HIS UNIVERSITY IN CHICAGO WHEN HE SUBMITTED THE DOCUMENT TO THAT BODY.

Fig. 5-9 Text IV (Cont.)

IF SUCH A LAW WERE TO BE PASSED BY THE REPRESENTATIVES OF THE AMERICAN PEOPLE , ALL FUTURE DEVELOPMENTS IN ATOMIC RESEARCH , INSTEAD OF BEING AT LAST DIRECTED TO THE PEACEFUL EXPLOITATION OF THIS RICH SOURCE OF ENERGY , WOULD BE FOR THE MOST PART MISUSED FOR THE PURPOSES OF ARMAMENT.

AND YET ATOMIC SCIENTISTS WERE SUPPOSED TO COMPLY WITH THE EXTREMELY STRICT SECRECY REGULATIONS APPLICABLE TO THEM , UNDER THE THREAT OF LONG PRISON SENTENCES FOR INFRINGEMENT.

IF THE BILL BECAME LAW THE RESULT WOULD SOON BE , AS CHESTER BARNARD , A DIRECTOR OF THE ROCKEFELLER FOUNDATION , HAD EXCLAIMED WITH MISGIVINGS WHEN HE FIRST HEARD OF THE ATOM BOMB , THE END OF DEMOCRACY.

THE SCHEME HAD BEEN INGENIOUSLY CONTRIVED.

THE WAR DEPARTMENT HAD DRAFTED THE NEW LEGISLATION UNDER THE DIRECTION OF KENNETH ROYALL , THE ASSISTANT SECRETARY OF WAR , AND WITH THE HELP OF GENERAL GROVES.

THE DEPARTMENT MANAGED TO INTRODUCE IT AS UNOBTRUSIVELY AS POSSIBLE INTO A CONGRESS OVERBURDENED WITH URGENT BILLS.

BUT UNDER THE CONSTITUTION PUBLIC HEARINGS WERE REQUIRED BEFORE ANY NEW LEGISLATIVE PROPOSAL WAS READ AND DEBATED IN CONGRESS.

AT SUCH HEARINGS QUALIFIED SUPPORTERS AND OPPONENTS OF THE BILL EXPRESSED THEIR OPINIONS.

CONGRESSMAN ANDREW MAY , A SMALL TIME ATTORNEY FROM KENTUCKY , WHOSE MANY YEARS IN THE HOUSE OF REPRESENTATIVES HAD SECURED HIM THE CHAIRMANSHIP OF THE MILITARY AFFAIRS COMMITTEE , SUCCEEDED IN HOLDING HEARINGS ON THE BILL , WHICH HE AND SENATOR JOHNSON OF COLORADO WERE TO INTRODUCE , WITHOUT ANY PUBLICITY.

ONLY FOUR PEOPLE HAD BEEN ASKED TO TESTIFY ON BEHALF OF THE BILL.

Fig. 5-9 Text IV (Cont.)

THEY WERE THE SECRETARY OF WAR , PATTERSON , AND GENERAL GROVES , WHO WERE BOTH NATURALLY IN FAVOR OF IT , TOGETHER WITH THE SCIENTISTS VANNEVAR BUSH AND JAMES CONANT , WHO HAD BOTH COLLABORATED AS CONSULTANTS IN THE FORMULATION OF THE BILL.

IT WAS ONLY AFTER SZILARD , AT THE LAST MINUTE , HAD ALARMED HIS COLLEAGUES , THAT MRMAY WAS COMPELLED , BY THE PRESSURE OF PUBLIC OPINION AROUSED BY STATEMENTS FROM THE SCIENTISTS , TO ARRANGE FURTHER HEARINGS TO BE ATTENDED BY WELL , KNOWN OPPONENTS OF HIS BILL.

ONE CAN IMAGINE HIS IRRITATION WITH SZILARD , WHO INSTANTLY PRESENTED HIMSELF AS THE FIRST WITNESS AGAINST THE PROPOSED LEGISLATION.

IT WAS JUST SIX YEARS SINCE SZILARD , ON HIS WAY TO EINSTEIN*S SUMMER HOME , HAD DOUBTED WHETHER HE OUGHT TO CONTINUE WITH HIS FATEFUL MISSION.

WHAT HE HAD FORESEEN THEN HAD COME TRUE.

THE MILITARY AUTHORITIES HAD NO INTENTION OF RELAXING THEIR CONTROL OF THE NEW SOURCE OF ENERGY.

AND HE HIMSELF , FOR DARING TO OPPOSE THEM , WAS NOW TREATED , IN SPITE OF HIS CONTRIBUTION TO THE DEVELOPMENT OF THE NEW POWER , WORSE THAN A DEFENDANT.

CONGRESSMAN MRMAY , WHO PRESIDED OVER THE HEARING , TRIED IN EVERY POSSIBLE WAY TO PROVOKE AND CONFUSE THE SCIENTIST.

HE PRETENDED HE HAD NOT CAUGHT SZILARD*S NAME PROPERLY , OR COULD NOT PRONOUNCE IT , AND PERSISTENTLY CALLED HIM SIGHLAND.

SZILARD TALKED FOR AN HOUR AND FORTY MINUTES , AND WAS CONTINUALLY INTERRUPTED AND DELIBERATELY MISUNDERSTOOD.

HE WAS RUDELY CALLED TO ORDER FOR NOT ANSWERING INTRICATE QUESTIONS WITH A PLAIN YES OR NO.

Fig. 5-9 Text IV (Cont.)

THE WITNESS WAS ALSO TOLD REPEATEDLY HE WAS TAKING UP TOO MUCH OF THE HEARING'S VALUABLE TIME.

SZILARD , BY NATURE A TEMPERAMENTAL MAN , CONTROLLED HIS INDIGNATION WITH REMARKABLE RESTRAINT.

HE SAW THROUGH THE TRAPS LAID FOR HIM.

HE ALLOWED NEITHER TAUNTS NOR ACCUSATIONS TO DISCONCERT HIM AND EVENTUALLY CONVINCED MOST OF THE MEMBERS OF THE COMMITTEE HE ADDRESSED THAT HIS RESISTANCE TO THE CONTINUANCE OF MILITARY CONTROL OF THE DEVELOPMENT OF ATOMIC ENERGY WAS WELL GROUNDED.

HE THUS WON THE FIRST SKIRMISH IN THE MONTH , LONG STRUGGLE OF THE ATOMIC SCIENTISTS TO ENSURE CIVIL CONTROL.

HIS ADVERSARY , CONGRESSMAN MRMAY , APPARENTLY SO DEVOTED TO THE INTERESTS OF THE MILITARY AUTHORITIES , WAS SOON AFTERWARDS FORCED TO RETIRE FROM PUBLIC LIFE AND SERVE A PRISON SENTENCE FOR SHOWING FAVORS TO AN INDUSTRIALIST WHO HAD GOT ARMY CONTRACTS BY CORRUPT PRACTICES.

WHEN COPIES OF THE MAY AND JOHNSON BILL REACHED THE ATOMIC LABORATORIES AND THE UNIVERSITIES , THE MEMBERS OF THE NEW SCIENTIFIC ASSOCIATIONS , MOSTLY THE YOUNGER GENERATION OF SCIENTISTS , DETERMINED TO SEND DELEGATES TO NEW YORK AND WASHINGTON.

THEY WERE ANXIOUS TO ENTER THE POLITICAL ARENA TO CAMPAIGN FOR MORE SATISFACTORY LEGISLATION FOR THE CONTROL OF ATOMIC POWER.

BY THE MIDDLE OF NOVEMBER THE LOCAL GROUPS HAD COMBINED INTO A SINGLE BODY , THE FEDERATION OF ATOMIC SCIENTISTS.

THE WORD ATOMIC WAS LATER REPLACED BY AMERICAN , FOR A GREAT MANY OF THE MEMBERS HAD NOTHING WHATEVER TO DO WITH NUCLEAR RESEARCH.

BUT AT THAT TIME , IN THE AUTUMN OF 1945 , THE OMINOUS ADJECTIVE WAS STILL INDISPENSABLE.

Fig. 5-9 Text IV (Cont.)

IT WAS STILL A NAME TO BE USED IN CONJURING.

ALL DOORS OPENED AT THE WORD ATOMIC , THE NEW SUPERLATIVE.

SENATOR TYDINGS , FOR INSTANCE , DECLARED THAT AN ATOMIC SCIENTIST IS ONE OF THE FEW PERSONS WHOSE INTELLECTUAL DEVELOPMENT IN MANY RESPECTS , AND ESPECIALLY IN THE SCIENTIFIC FIELD , BEARS THE SAME RELATION TO THAT OF THE REST OF US AS A RANGE OF MOUNTAINS BEARS TO A MOLEHILL.

THE ATOMIC SCIENTISTS HAD BECOME IMPORTANT PEOPLE.

THAT WAS THEIR FIRST DISCOVERY WHEN THEY RETURNED FROM THEIR LABORATORIES TO THE WORLD AT LARGE.

BEFORE THE WAR WE WERE SUPPOSED TO BE COMPLETELY IGNORANT OF THE WORLD AND INEXPERIENCED IN ITS WAYS.

BUT NOW WE ARE REGARDED AS THE ULTIMATE AUTHORITIES ON ALL POSSIBLE SUBJECTS , FROM NYLON STOCKINGS TO THE BEST FORM OF INTERNATIONAL ORGANIZATION . ONE OF THE REMARKED WITH MILDLY IRONIC SELF DETACHMENT , AFTER HE HAD BECOME SOMEWHAT ACCUSTOMED TO BEING DAZZLED BY FLASH BULBS AND CONFRONTED BY MICROPHONES AND NEWSREEL CAMERAS.

THE MORE SENSITIVE OF THESE SCIENTISTS SUFFERED INCREASED PANGS OF CONSCIENCE WHEN THEY REALIZED , AS THE BIOLOGIST DR. THEODOR HAUSCHKA PUT IT IN A BITTER OPEN LETTER TO OPPENHEIMER , THAT THEIR PRESTIGE CAME CHIEFLY BECAUSE THEY HAD BEEN BRILLIANT COLLABORATORS WITH DEATH.

BUT WHENEVER THEY STARTED TO CONFESS THEIR SINS , PUBLIC INTEREST IN THEM INCREASED.

THOSE WHO UNBURDEN THEIR HEARTS CAN NEARLY ALWAYS COUNT ON A SYMPATHETIC AUDIENCE WHICH NOT ONLY PARDONS BUT ADMIRES THEM.

MANY OF THE SCIENTISTS VERY SOON PERCEIVED THAT THIS ASSET OF ACCUMULATED ATTENTION AND RESPECT MIGHT PERHAPS BE CONVERTED INTO THE CURRENT COIN OF A GENUINE POLITICAL INFLUENCE.

Fig. 5-9 Text IV (Cont.)

THEY ACCORDINGLY BEGAN THE LAST CRUSADE , AS THEIR EFFORTS WERE CALLED BY MICHAEL AMRINE , AN IDEALISTIC YOUNG WRITER WHO PLACED HIMSELF AT THEIR DISPOSAL IN THOSE DAYS.

IT WAS A CRUSADE UNDERTAKEN BY MEN WHO WERE CHILDREN IN POLITICAL AFFAIRS AND YET , OR POSSIBLY FOR THAT VERY REASON , GRADUALLY MADE HEADWAY IN WASHINGTON AGAINST CUNNING POLITICIANS AND APPARENTLY INVINCIBLE VESTED INTERESTS.

AMRINE , THE LOYAL HISTORIAN OF THIS UNUSUAL MOVEMENT , DESCRIBES THE MOOD THAT INSPIRED IT.

THESE MEN HAD REDISCOVERED THEIR PERSONAL , HUMAN CONSCIENCES AND WERE DETERMINED TO OVERCOME ALL OPPOSITION IN ORDER TO GUIDE SOCIETY BACK TO THE ROAD OF PROGRESS AND DIVERT IT FROM THAT WHICH LED TO ANNIHILATION.

THE MANIFESTO IN WHICH THEY ANNOUNCED THIS AIM WAS A SMALL SHEET OF PAPER WRITTEN IN SINGLE SPACE ON EACH SIDE.

A RADIO REPORTER REMARKED LATER THAT IT SEEMED TO HAVE BEEN DUPLICATED WITH A WET HANDKERCHIEF.

HE COULD NOT HAVE KNOWN , OF COURSE , THAT THE SCIENTISTS ONLY POSSESSED AN OFFICE WHICH HAD BEEN LENT TO THEM ON THE FOURTH FLOOR OF A HOUSE WITHOUT AN ELEVATOR.

THEY HAD ONLY ONE ROOM , WHERE THERE WERE NOT ENOUGH TABLES AND CHAIRS , SO THAT WORLD RENOWNED NOBEL PRIZEWINNERS AND STUDENTS HAD TO SQUAT ON THE FLOOR WHILE THEY PASSED TO ONE ANOTHER THE STATEMENTS AND PETITIONS WHICH WERE SUBSEQUENTLY HEARD BY THE ENTIRE WORLD.

SUCH WAS THE BEGINNING OF AN AMAZING CAMPAIGN CARRIED OUT IN THE FACE OF INDIFFERENCE FROM THE WHITE HOUSE , THE STATE DEPARTMENT AND CONGRESS , AND AGAINST POWERFUL AND WELL ORGANIZED OPPOSITION.

Fig. 5-9 Text IV (Cont.)

EXPERIENCED PEOPLE IN WASHINGTON ALL SHOOK THEIR HEADS.

THEY WARNED THE LEAGUE OF FRIGHTENED MEN, WHICH THE SCIENTISTS WERE CALLED, NOT TO EXPECT THEIR UNDERTAKING TO SUCCEED.

DURING THE WINTER OF 1945 THE SCIENTISTS' VISION OF A NEW WORLD WITHOUT HUNGER OR COLD WAS BEING JOTTED DOWN BY MEN IN THICK OVERCOATS IN AN UNHEATED OFFICE IMMEDIATELY ABOVE LARRY'S COFFEE SHOP ON L STREET.

THESE MEN LEARNED THE LANGUAGE OF POLITICS WITH SURPRISING RAPIDITY.

FOR INSTANCE, THEY FIRST WROTE.

THE TRANSFORMATION OF MASS INTO ENERGY HAS FUNDAMENTALLY CHANGED OUR CONCEPTION OF THE NATURE OF THE WORLD.

BUT THAT WAS MUCH TOO ABSTRACT AND CAUTIOUSLY FORMULATED A SENTENCE TO MAKE ANY IMPRESSION.

SOON AFTERWARDS THEY WERE ADDRESSING POLITICIANS IN THE FOLLOWING JAZZED-UP TERMS.

SENATOR, IF A SINGLE ONE OF THE NEW BOMBS WERE TO BURST ON THE RAILWAY STATION AT WASHINGTON, THE MARBLE ON TOP OF THE CAPITOL HERE WOULD BE GROUND TO POWDER.

YOU YOURSELF AND MOST OF YOUR COLLEAGUES WOULD PROBABLY BE DEAD WITHIN THE FIRST FEW MINUTES.

THESE WORDS WERE EFFECTIVE.

WHAT THE YOUNG SCIENTISTS LACKED IN POLITICAL EXPERIENCE THEY MADE UP FOR BY AN ENTHUSIASM AND SINCERITY WHICH DEEPLY IMPRESSED THE POLITICIANS AND IN PARTICULAR THE REPRESENTATIVES OF THE PRESS IN WASHINGTON.

Fig. 5-9 Text IV (Cont.)

IT WAS KNOWN THAT THIS STRANGEST OF ALL LOBBIES WAS FINANCED ONLY BY VOLUNTARY CONTRIBUTIONS FROM THE SCIENTISTS AND THAT MANY OF THEM WHO HAD BEEN GIVEN NO LEAVE FOR YEARS WERE NOW DEDICATING THEIR FIRST FREE TIME TO THIS PUBLIC QUESTION.

THAT THEY WERE REALLY INDEFATIGABLE IS PROVED BY THE ENTRIES IN A GREAT , COVERED , OBLONG LOGBOOK IN WHICH EVERY SCIENTIST WORKING FOR THE FEDERATION WROTE DOWN HIS DOINGS AT THE END OF THE DAY.

ATOMIC SCIENTISTS WERE THE FIRST TO ENTER THE ANTEROOMS OF CONGRESSMEN EARLY IN THE MORNING.

LATER ON THEY VISITED EDITORIAL OFFICES TO DISTRIBUTE THE STATEMENTS WHICH THEY HAD THEMSELVES TYPED AND DUPLICATED.

AT NOON THEY GAVE LUNCHTIME LECTURES TO ALL SORTS OF SOCIETIES , ANSWERING SUCH QUESTIONS AS WHAT THE COLOR OF PLUTONIUM WAS.

IN THE AFTERNOONS THEY SOMETIMES EVEN VENTURED INTO THE LIONS* DEN ITSELF , THE ARMY HOSPITAL , OR ATTENDED THE TEAPARTIES GIVEN FOR THEM BY MRS. PINCHOT , A POLITICALLY INFLUENTIAL MEMBER OF WASHINGTON SOCIETY.

LATE IN THE AFTERNOON THEY WERE TO BE FOUND AT COCKTAIL PARTIES WHERE THEY MIGHT MEET IMPORTANT PERSONS.

SOME ALSO CONDUCTED EVENING CLASSES IN NUCLEAR PHYSICS FOR CONGRESSMEN AND GOVERNMENT OFFICIALS.

OTHERS DISCUSSED THEIR MISSIONARY TASK , FAR INTO THE NIGHT , WITH DOCTORS , SOCIOLOGISTS AND REPRESENTATIVES OF THE CHURCH , THE PRESS AND THE FILM WORLD.

THE FIRST RESULT OF ALL THESE ACTIVITIES WAS A SUBSTITUTE FOR THE MAY , JOHNSON BILL , FRAMED BY THE SCIENTISTS IN COLLABORATION WITH SENATOR MCMAHON AND NOW LAID BEFORE CONGRESS.

Fig. 5-9 Text IV (Cont.)

THE NEXT PROBLEM WAS AN ADDITIONAL RIDER ATTACHED TO THE BILL BY SENATOR VANDENBERG , WHO HAD TAKEN THIS INDIRECT METHOD OF SMUGGLING IN MILITARY CONTROL AGAIN.

THE SCIENTISTS CONTRIVED TO HAVE THIS SMOTHERED UNDER AN AVALANCHE OF LETTERS OF PROTEST FROM MANY INDIGNANT VOTERS.

AT LAST , IN JULY 1946 , WHEN THE MCMAHON BILL , WHICH HANDED OVER CONTROL OF ATOMIC RESEARCH DEVELOPMENT IN THE UNITED STATES TO A CIVIL COMMISSION , WAS MADE LAW , THE SCIENTISTS WERE ABLE TO TASTE THE FRUITS OF VICTORY.

BUT THAT VICTORY VERY SOON TURNED OUT TO HAVE BEEN A PYRRHIC ONE.

Fig. 5-9 Text IV (Cont.)

A LIST STRUCTURE CAN BE MADE , IN MANY INSTANCES , CONSIDERABLY MORE EFFICIENT THAN HAS BEEN DEMONSTRATED IN THE PAST.

ALL THAT IS NEEDED IS SOME COMMON SENSE AND INGENUITY ON THE PART OF THE USER.

IN AN AUTOMATIC PROGRAMMING EFFORT AT MS SPACE GUIDANCE CENTER , OWEGO , N Y , IT IS FOUND THAT THE TECHNIQUE OF ORGANIZING A COMPUTER MEMORY INTO LIST STRUCTURES , AN APPROACH INTRODUCED BY NEWELL , SIMON , AND SHAW , WAS PARTICULARLY PERTINENT.

THIS REPORT INTRODUCES THE CONCEPT OF THE MULTIWORD LIST ITEM , WHICH WAS DEVELOPED TO OFFSET THE INEFFICIENCIES OF SINGLEWORD ITEMS.

FOLLING A BRIEF DESCRIPTION OF STANDARD LIST STRUCTURES WITH SINGLEWORD ITEMS , CALLED SINGLETS , THE MULTIWORD ITEMS ARE INTRODUCED.

THEN VARIABLELENGTH ITEMS ARE CONSIDERED , ALONG WITH THE CORRESPONDING PROBLEMS INVOLVING THE UTILIZATION OF AVAILABLE SPACE.

SEVERAL EXAMPLES ARE GIVEN ILLUSTRATING THE USE OF MULTIWORD LISTS.

THERE ARE TWO OBJECTIVES OF THIS REPORT.

FIRST , AND MOST IMPORTANT , IT WILL SHOW HOW MULTIWORD LIST ITEMS CAN PROVIDE CONSIDERABLE SAVINGS , BOTH IN TERMS OF EXECUTION TIME AND IN THE MEMORY SPACE REQUIRED OVER NORMAL SINGLEWORD ITEMS.

Fig. 3-10 Text V

SECONDLY ,
IT WILL HELP THE
UNINITIATED READER TO APPRECIATE THE SIMPLICITY OF
LIST STRUCTURES IN THEMSELVES , AND THE EASE WITH WHICH THEY MAY BE
APPLIED.

A LIST IS A CONNECTED SEQUENCE OF ITEMS FN .

IN THE FORM IN
WHICH LISTS WERE ORIGINALLY INTRODUCED BY NEWELL , SIMON AND SHAW
REF3 , AND IN WHICH THEY HAVE GENERALLY BEEN USED , AN ITEM CONSISTS
OF ONE COMPUTER WORD.

THESE ITEMS ARE CONNECTED THROUGH A FIELD
WITHIN EACH WORD WHICH CONTAINS THE ADDRESS OF
THE SUCCEEDING
ITEM.

SINCE EACH ITEM POINTS TO ITS SUCCESSOR , OR CONTAINS THE ADDRESS OF
ITS SUCCESSOR , SUCCESSIVE LIST ITEMS NEED NOT BE AND IN GENERAL
ARE NOT CONSECUTIVE WORDS IN MEMORY.

IN FACT , ONE OF THE POWERFUL
FEATURES OF LISTS IS THEIR
ABILITY TO UTILIZE ARBITRARY , DISJOINT
SECTIONS OF MEMORY.

SUCH A LIST CAN BE DIAGRAMMED AS IN FIG AB.

THE ARROW INDICATES THE SUCCESSOR OF EACH ITEM.

LATER IT WILL BE
USEFUL TO INDICATE THE FIELD IN THE ELEMENT WHICH CONTAINS THE
POINTER.

Fig. 5-10 Text V (Cont.)

THIS IS THE ADDRESS OF THE SUCCESSOROR IN ADDITION , AND SINCE
THE LIST ITEMS ARE ARBITRARY STORAGE WORDS , EACH LIST IS PROVIDED
WITH A HEAD , WHICH IS A KNOWN LOCATION IN MEMORY.

IT IS A
WORD THE LOCATION OF WHICH IS KNOWN TO THE PROGRAMMER , AND
THROUGH
WHICH HE IS ABLE AT ALL TIMES TO LOCATE THE LIST SINCE IT POINTS
TO THE FIRST ITEM ON THE LIST.

SIMILARLY , A SPECIAL MARK IS REQUIRED
TO INDICATE THE END OF A LIST.

THIS IS USUALLY DONE BY USING A
POINTER OF ZERO.

THUS , A SIMPLE LIST
CONTAINING THREE ITEMS WOULD
APPEAR AS IN FIG B.

A LIST HEAD CONTAINING A ZERO POINTER IS AN
EMPTY LIST.

THE REMAINING PORTION OF EACH ELEMENT MAY CONTAIN DATA , MAY
POINT TO A FULL WORD OF DATA , OR MAY POINT TO ANOTHER
LIST CALLED
A SUBLIST.

HERE , THE FIRST ITEM POINTS TO
A DATA WORD , THE SECOND POINTS TO A ITEM SUBLIST , AND THE FOURTH
POINTS TO A ITEM SUBLIST , WHICH IN TURN POINTS TO A DATA WORD.

SUCH A CONFIGURATION IS CALLED A LIST STRUCTURE.

Fig. 5-10 Text V (Cont.)

IT IS NECESSARY
THAT EITHER THE PROGRAMMER KNOWS WHAT EACH ITEM REPRESENTS , OR ELSE
THAT THERE IS SUFFICIENT INFORMATION STORED IN THE LIST IN ORDER THAT
THE PROGRAM MAY FIND OUT WHAT EACH REPRESENTS.

WE SHOULD NOTE THAT , UNLESS THE COMPUTER WORD IS
LONG ENOUGH TO HOLD TWO
POINTERS , ONLY THE MOST RUDIMENTARY OPERATIONS CAN BE PERFORMED ON
THE SINGLET LISTS.

HOVER , A POINTER NEED NOT BE OF FULLADDRESS
LENGTH.

IT MAY BE SEVERAL BITS SHORTER , BUT WITH A CORRESPONDING
DECREASE IN THE AMOUNT
STORAGE AVAILABLE FOR USE ON LISTS.

AN
ALTERNATE APPROACH IS THE USE OF MULTIWORD ITEMS , WHICH ARE
DESCRIBED LATER.

IN NORMAL USE , ALL AVAILABLE STORAGE IS INITIALLY PUT ON A
SPECIAL LIST OF AVAILABLE SPACE.

AS THE PROGRAM PROGRESSES
BUILDING VARIOUS LISTS , IT OBTAINS EMPTY ITEMS FROM THE AVAILABLE
SPACE LIST.

THE SPACE LIST AND ANOTHER LIST ARE SHOWN DIAGRAMMATICALLY
IN FIGCA.

SUPPOSE IT IS DESIRED TO TAKE AN ITEM FROM THE SPACE LIST
AND INSERT IT BETWEEN THE FIRST AND
SECOND ITEMS OF THE OTHER LIST.

Fig. 5-10 Text V (Cont.)

THE POINTERS WOULD BE REPRESENTED AS IN FIG B.

HOWEVER , THE NEW
DIAGRAM WOULD APPEAR AS IN FIGCC.

THUS , THE MOST FREQUENT LIST OPERATIONS, INSERTING
AND DELETING ITEMS FROM A LIST , ARE
ACCOMPLISHED THROUGH SIMPLE
MANIPULATION OF THE POINTERS.

NOTE THAT WHEN AN ITEM IS REMOVED
FROM A LIST , IT IS RETURNED TO THE LIST OF AVAILABLE SPACE , FOR
POSSIBLE LATER USE.

NOW , IT CAN BE OBSERVED THAT IN USING SINGLET LISTS , SOMEWHAT
LESS THAN ONE HALF OF EVERY
SINGLET MUST BE GIVEN OVER TO USE AS
A POINTER , THUS DECREASING THE NUMBER OF BITS AVAILABLE TO HOLD
DATA OR OTHER INFORMATION.

TO ILLUSTRATE THIS , ASSUME THAT A
POINTER REQUIRES EXACTLY ONE HALF OF A WORD OF A SINGLE ITEM.

TABLE BBB
SHOWS THE NUMBER OF
SINGLET LIST ITEMS REQUIRED TO HOLD A GIVEN
AMOUNT OF DATATYPE INFORMATION.

THERE ARE TWO WAYS IN WHICH ONE
AND ONEHALF WORDS OF DATA CAN BE STORED.

SIMILAR SITUATIONS HOLD FOR LARGER AMOUNTS OF DATA.

THUS , ABOUT HALF
OF STORAGE IS USED UP IN POINTERS.

Fig. 5-10 Text V (Cont.)

IN MORE
COMPLEX PROBLEMS , THIS APPARENT WASTAGE OF STORAGE IS CONSIDERABLY REDUCED BY THE AMOUNT OF INFORMATION IMPLIED IN THE STRUCTURE , AND THROUGH THE USE OF BORROWED SUBLISTS.

HOWEVER , THESE
TOPICS
ARE NOT PERTINENT TO THIS PAPER.

THE OTHER MAIN PROBLEM WITH SINGLET LISTS INVOLVES THE ACCESS TIME FOR DATA.

IT IS NOT POSSIBLE TO DIRECTLY COMPUTE THE LOCATION OF ANY PARTICULAR ITEM ON A LIST.

IT FOLLOWS ,
THEN , THAT IF EACH ITEM CAN ONLY BE LOCATED THROUGH A POINTER , THE NUMBER OF INSTRUCTIONS WHICH MUST BE EXECUTED TO OBTAIN A GIVEN AMOUNT OF DATA DEPENDS DIRECTLY UPON THE NUMBER OF LIST ITEMS REQUIRED TO STORE IT.

IT WILL BE SHOWN IN THE NEXT SECTION THAT MULTIWORD ITEMS FOR LISTS WILL GREATLY ALLEVIATE BOTH OF THESE PROBLEMS.

A MULTIWORD ITEM IS A SINGLE LIST ITEM WHICH IS ORGANIZED IN STORAGE AS A SEQUENCE OF TWO OR MORE
CONSECUTIVE WORDS OF MEMORY.

IT CAN BE OBSERVED FROM THE FOLLOWING DISCUSSION THAT THE MULTIWORD ITEM CONCEPT IS THE BASIS FOR THE APPLICABILITY OF THE NCOMPONENT ELEMENT OF ROSS REFERENCED ABOVE.

THE SIMPLEST FORM OF A MULTIWORD ITEM IS THE TWORD
ITEM ,
OR DOUBLET FN , WHICH CONSISTS OF TWO CONSECUTIVE WORDS IN STORAGE.

Fig. 5-10 Text V (Cont.)

DIAGRAMMATICALLY, TWO WORDS SEPARATED BY AN ASTERISK ARE
IN CONSECUTIVE MEMORY LOCATIONS , AS IN FIGGA.

THUS, A DOUBLET LIST HAS THE APPEARANCE
SHOWN IN FIG B.

THE EFFECT OF THIS SIMPLE STEP IS ILLUSTRATED FOLLOWING.

HERE
A SINGLET LIST REQUIRES FOUR ITEMS AND THEREFORE FOUR WORDSO TO
STORE ONE AND ONEHALF WORDS OF INFORMATION $ A DOUBLET LIST REQUIRES
ONLY ONE TWOWORD ITEM.

ALSO , WHERE A SINGLET LIST REQUIRES EIGHT
TO TEN INSTRUCTIONS TO FETCH THE INFORMATION , THE DOUBLET REQUIRES
ONLY THREE.

THUS , THE MULTIWORD ITEM REDUCES THE SIZE OF LIST MEMORY , REDUCES
THE PROGRAM SIZES AND DECREASES THE PROGRAM
EXECUTION TIME.

PERHAPS IT IS APPROPRIATE NOW TO MAKE A COMMENT ON PROGRAMMING
TECHNIQUES.

WITH SINGLET LISTS , IT IS SOMETIMES DEBATABLE WHETHER
THE PROCESS OF SEQUENCING DOWN A LIST THROUGH SELECTION AND THE
USE OF POINTERS SHOULD BE ACCOMPLISHED BY BRUTE FORCE MODIFICATION
OF ADDRESSES SUCH AS STORING A POINTER AS THE ADDRESS PORTION OF A
FETCH INSTRUCTIONO OR BY THE USE OF INDEX REGISTERS WHEREBY THE
POINTER IS LOADED INTO AN INDEX REGISTER FN.

HOWEVER , WITH
MULTIWORD ITEMS THE INDEXING TECHNIQUE IS NECESSARY TO OBTAIN THE
INDICATED SPEED INCREASES.

Fig. 5-10 Text V (Cont.)

IN THE CASE OF THE TRIPLET , THE INDEX
REGISTER NEED BE LOADED ONLY ONCE , AND THEN FETCH INSTRUCTIONS
WITH TAGGED ADDRESSES OF ZERO
AND TWO WILL MAKE AVAILABLE THE
FIRST , SECOND AND THIRD WORDS , RESPECTIVELY , OF THE THREEWORD
ITEM.

THIS IS CALLED REVERSE INDEXING BY ROSS AND IS DISCUSSED BY
HIM IN DETAIL REF1.

FIGURE AREFA SHOWS ON THE LEFT HALFO A SINGLET LIST
STRUCTURE
CONTAINING SINGLET ITEMS.

HOWEVER , IT APPEARS THAT THE SAME
INFORMATION COULD BE STORED ON A SIMPLE LIST CONTAINING ONLY
WORDS WHEN SET UP WITH FOURWORD ITEMS.

THIS IS ILLUSTRATED IN
FIGA.

FOR AN EXAMPLE OF THE USE OF A DOUBLET
LIST , WE WILL CONSIDER THE
PREPARATION OF A SYMBOL TABLE.

FOR A REASONABLY STANDARD TWOPASS
ASSEMBLY PROGRAM FOR A MACHINE WITH RANDOM ACCESS MEMORY , THE FIRST
PASS INVOLVES BUILDING A TABLE OF ALL DIFFERENT SYMBOL NAMES ,
RESTRICTING THEIR LENGTH TO
WORDO AND ASSIGNING TO EACH AN
ABSOLUTE MACHINE ADDRESS ONEHALF WORDO.

THE SECOND PASS THEN
REQUIRES THAT EVERY SYMBOL BE LOOKED UP IN THE TABLE AS IT OCCURS
AND THE CORRESPONDING ABSOLUTE ADDRESS BE INSERTED INTO THE
INSTRUCTION BEING ASSEMBLED.

Fig. 5-10 Text V (Cont.)

THE DOUBLET LIST WORKS VERY NICELY
IN THIS CASE , AS ILLUSTRATED IN FIG1.

AS THE SYMBOL TABLE IS

BEING BUILT , SYMBOLS MAY BE SORTED INTO THE LIST ALPHABETICALLY ,
IF DESIRED, OR MERELY PUT INTO THE LIST IN THEIR ORDER
OF OCCURRENCE.

ON THE SECOND PASS OF THE ASSEMBLY PROGRAM , A SIMPLE SEARCHING
ROUTINE WILL EFFECT THE TRANSLATION.

TO FURTHER ILLUSTRATE THE AD HOC USE OF LISTS , THE OBVIOUS LACK
OF HIGH EXECUTION SPEED TO SUCH A SYMBOL TABLE CAN BE OFFSET BY
A
TABLE OF LISTS.

HERE A TABLE IS SET UP WITH LIST HEADS , ONE
FOR EACH LETTER OF THE ALPHABET , AND ORDERED ACCORDING TO THE
NUMERICAL CODE CORRESPONDING TO THE LETTERS.

THUS , A SYMBOL
TABLE WOULD APPEAR AS SHOWN IN FIGB.

THE ADVANTAGE
OF THIS IS THAT
THE FIRST CHARACTER OF THE SYMBOL CAN BE EXTRACTED , AND THE LOCATION
OF THE HEAD OF THE APPROPRIATE LIST IN THE TABLE CAN BE COMPUTED
IMMEDIATELY.

THIS IS ESSENTIALLY A TECHNIQUE OF BLOCK
SORTING , WHICH HAS BEEN USED ON
OCCASIONS WITH TABLES RATHER
THAN LISTS.

BUT IN SUCH CASES , IT REQUIRES THE PROVISION OF
SPECIAL OVERFLOW PROCEDURES IF A TABLE BECOMES FULL.

Fig. 5-10 Text V (Cont.)

IF ALL
LETTERS HAVE AN EQUAL PROBABILITY OF OCCURRING AS THE FIRST
CHARACTER OF A SYMBOL , THE
SEARCH TIME IS REDUCED BY A FACTOR OF 2.

THIS TECHNIQUE CAN BE EXTENDED TO HANDLE ALMOST ANY PROBLEM OF
SORTING WITHIN THE HIGHSPEED MEMORY.

FOR EXAMPLE , IF IT IS
DESIRED TO READ IN A DECK OF CARDS AND SORT THEM ON SOME FIELD PRIOR
TO SOME OTHER OPERATION , AS EDITING A TAPE FILE IN THE 700
COMPUTER , THIRTEEN AND ONEHALF WORDS OF MEMORY ARE REQUIRED TO
STORE THE CONTENTS OF ONE CARD.

BY SIMPLY USING A LIST WITH
WORD ITEMS THE ADDITIONAL ONE HALF WORD HOLD THE
LIST POINTERO ,
THE SORTING MAY BE ACCOMPLISHED WITH A MINIMUM OF EFFORT AND , SINCE
THE ONE HALF WORD WOULD NORMALLY BE UNUSED , NO ADDITIONAL STORAGE IS
NECESSARY.

IN THE LAST TWO EXAMPLES , THE LISTS ARE OF A CONTINUOUSLY
GROWING NATURE.

NO INDIVIDUAL ITEMS ARE REMOVED FROM THE LIST
UNTIL THE FUNCTION OF THE WHOLE LIST IS COMPLETE.

IN SUCH CASES ,
IT IS NOT NECESSARY TO MAINTAIN THE EFFORT OF
CONSTRUCTING AND MAINTAINING A LIST OF AVAILABLE SPACE.

IT SUFFICES
TO
REMEMBER THE STARTING POINT OF THE AS YET UNUSED PORTION OF
MEMORY , AND TO TAKE A NEW ITEM AS IT IS REQUIRED.

Fig. 5-10 Text V (Cont.)

PROBABLY THE MOST SIGNIFICANT CONTRIBUTION OF MULTIWORD ITEMS
TO THE PROCESSING OF LIST STRUCTURES LIES IN THE AVAILABILITY OF
MULTIPLE POINTERS.

FOR EXAMPLE , A SINGLET LIST IS A ONEWAY DEVICE.

IT IS POSSIBLE TO START AT THE HEAD AND MOVE DOWN THE LIST , BUT
IT IS NOT POSSIBLE TO MOVE BACK UP THE LIST.

THIS MOVEMENT IS EASILY
ACCOMPLISHED WITH MULTIWORD
ITEMS , BY SIMPLY MAKING THE ITEM LONG
ENOUGH TO CONTAIN A POINTER FOR THIS PURPOSE.

THUS , A TWOWAY LIST
OF THREEWORD ITEMS MIGHT APPEAR AS IN FIG4.

IT WOULD REQUIRE
SOMEWHAT MORE WORK TO EFFECT THE INSERTION AND DELETION OF ITEMS.

HOWEVER , THIS IS READILY ACCOMPLISHED , AND THERE ARE SITUATIONS
WHERE A TWOWAY LIST IS EXACTLY WHAT IS DESIRED.

AN EXAMPLE OCCURRED IN A FIXEDPOINT FORMULA TRANSLATION ROUTINE.

FOLLOWING THE NORMAL PROCEDURE , AN ARITHMETIC
STATEMENT IS DECOMPOSED
INTO A LIST OF ADDRESS TYPE OPERATIONS.

A TWOWAY LIST WITH WORD
ITEMS WAS USED , AS SHOWN IN FIG10.

THE OPERATION WORD ALSO
CONTAINED SOME CODED INFORMATION ON THE SCALING OF THE TWO OPERANDS.

THE TWOWAY POINTERS
WERE IMPORTANT FOR TWO REASONS.

Fig. 5-10 Text V (Cont.)

SINCE
PROGRAM EFFICIENCY WAS OF PRIMARY IMPORTANCE THEY WERE REALTIME CONTROL PROGRAMS $ CONSIDERABLE SCANNING AND RESCANNING OF THIS LIST WAS DONE , TRYING TO IMPROVE THE OBJECT PROGRAMS.

SINCE THE
PROGRAMS WERE FOR FIXED-POINT ARITHMETIC , ALL OPERANDS WERE TO BE AUTOMATICALLY SCALED IE , SHIFTED THE PROPER NUMBER OF PLACES BEFORE AND AFTER THE OPERATION , WHICH REQUIRED CONSIDERABLE MANIPULATION OF THE SEQUENCE OF LIST ITEMS.

THE TWOWAY
LIST WAS
IDEALLY SUITED TO THE PROBLEM.

IN REFA , THE ORGANIZATION FOR AN INFORMATION PROCESSING SYSTEM TO BE USED FOR INFORMATION RETRIEVAL IS PROPOSED.

THE BASIC IDEA
IS THAT A LARBE FILE OF ITEMS IS GIVEN.

ASSOCIATED WITH EACH ITEM ARE SEVERAL CHARACTERISTICS SUCH AS HEIGHT , WEIGHT, AGE, AND SEX.

ASSOCIATED WITH EACH CHARACTERISTIC IS A SET
OF VALUES.

FOR EXAMPLE, IN A GIVEN FILE, THE CHARACTERISTIC AGE MAY TAKE ON CONSECUTIVE VALUES.

FOR EACH SUCH
VALUE , THERE IS A
LIST , WHICH CONNECTS EVERY ITEM WHICH HAS THAT VALUE.

THUS , EVERY
ITEM CONTAINS A POINTER FOR EACH CHARACTERISTIC , CONNECTING IT TO THE LIST WHICH REPRESENTS THE PROPER VALUE OF THAT CHARACTERISTIC.

Fig. 5-10 Text V (Cont.)

SUCH ITEMS ARE READILY OBTAINABLE THROUGH
MULTI*ORD TECHNIQUE.

FIGE ILLUSTRATES A POSSIBLE ITEM OF THE TYPE.

IF THE FILE OF
SUCH ITEMS REPRESENTS A COLLEGE STUDENT BODY , AN ENTERPRISING
BASKETBALL COACH MAY WISH TO OBTAIN THE NAMES OF ALL MALE STUDENTS
WHO ARE OVER 5FEET 11INCH IN
HEIGHT AND OVER 0 POUNDS IN WEIGHT.

IN THIS CONFIGURATION , IT IS NECESSARY TO SEARCH THE LIST REPRESENTING
THE PROPER VALUES OF EACH CHARACTERISTIC IN SEQUENCE SELECTING THOSE
ITEMS WHICH APPEAR ON ALL.

IN THIS WAY , IT IS UNNECESSARY TO SEARCH
ALL
ITEMS IN THE FILE EXCEPT THOSE WHICH HAVE THE PROPER VALUE IN AT
LEAST ONE CHARACTERISTIC ARE SEARCHED.

IT IS NOT PROPOSED AT THIS TIME
TO CONSIDER WHETHER OR NOT THIS IS THE BEST TECHNIGUE FOR SOLVING
THIS TYPE OF PROBLEM.

HOWEVER , IT SHOULD BE
POINTED OUT THAT
WHEREAS IN REFN A MACHINE TO OPERATE IN THIS FASHION IS PROPOSED ,
MULTI*ORD ITEMS WITH MULTIPLE POINTERS ALLOW THE SAME FLEXIBILITY
ON A GENERALPURPOSE MACHINE WITH A MINIMUM OF SPACE.

ANOTHER POSSIBLE APPLICATION OF MULTIPLE
POINTERS MIGHT ARISE

IN *CKDING WITH VERY LARGE BUT VERY SPARSE MATRICES.

Fig. 5-10 Text V (Cont.)

Fig. 5-10 Text V (Cont.)

FOR EXAMPLE ,
SOME LINEAR PROGRAMMING PROBLEMS MAY INVOLVE A 000 BY 000 MATRIX
WITH ONLY WWW TO XXX NONZERO ELEMENTS.

ONE WAY TO STORE ONLY
THE NONZERO ELEMENTS MIGHT BE TO HAVE A LIST FOR EACH ROW.

THEN
THE NONZERO MATRIX ELEMENTS WOULD EACH BE REPRESENTED BY AN ITEM
OF THE FORM SHOWN IN FIGG.

TO LOCATE THE ELEMENT AT THE
COORDINATES W AND J , IT IS NECESSARY TO SEARCH THE
LIST REPRESENTING
ROW I FOR AN ITEM WITH COLUMN NUMBER J.

HOWEVER , TO PERFORM THE
MORE GENERAL COMPUTATIONS OF LINEAR PROGRAMMING IN A REASONABLY
EFFICIENT WAY , IT MIGHT BE NECESSARY TO PROVIDE A LIST FOR EACH
COLUMN AS WELL , AND CONVERT THEM INTO
TWO-WAY LISTS.

THIS CAN ALSO
BE DONE QUITE READILY , AS INDICATED IN FIGHB.

THUS , HAVING
LOCATED ELEMENT JJJ , IT IS RELATIVELY EASY TO LOCATE ELEMENTS
XXX AND RRR.

A SPECIFIC APPLICATION WHERE THE MULTIPLE POINTERS
PROVED MOST
VALUABLE IN ACTUAL USE WAS IN ANOTHER PORTION OF THE FIXEDPOINT
COMPILER MENTIONED PREVIOUSLY.

THE SOURCE LANGUAGE OF THE COMPILER DEFINED
ALL CONTROL AND STATEMENT SEQUENCING IN
THE FORM OF A FLOW DIAGRAM.

Fig. 5-10 Text V (Cont.)

A STYLIZED VERSION OF THE FLOW DIAGRAM , USING A FIXED BUT EXPANDABLE FORMAT , WAS PUNCHED INTO CARDS , AND THEN TRANSFERRED TO TAPE.

IT
WAS DESIRED TO LOAD THE FLOW DIAGRAM INTO CORE MEMORY IN SUCH A WAY AS TO PRESERVE
COMPLETELY THE TOPOLOGY OF FLOW.

THIS WAS ACCOMPLISHED THROUGH THE USE OF A LIST STRUCTURE WITH FOUR WORD ITEMS AND WAY POINTERS.

THE ITEM FORMAT IS SHOWN IN FIGX ,
WHERE EACH NONTRIVIAL BLOCK IN THE FLOW DIAGRAM IS REPRESENTED BY ONE ITEM.

ONE ITEM POINTS TO ANOTHER ONLY IF THERE IS A LINE ON
THE FLOW DIAGRAM CONNECTING THE TWO CORRESPONDING BLOCKS.

WITH THE
SOURCE PROGRAM DEFINED IN THIS WAY , IT IS QUITE SIMPLE TO ACCOMPLISH A GLOBAL TRACI , FOR INTERSTATEMENT OPERATION
OF THE OBJECT PROGRAM.

TO DETERMINE THE LONGEST PATH THROUGH THE PROGRAM ,
AND TO COMPUTE AN OPTIMUM ORDER FOR
THE PROGRAMMING OF THE PATHS IS DIFFICULT.

IT REPRESENTS A VERY COMPLEX BIT
OF DATA
MANIPULATION , PROBABLY NOT FEASIBLE WITHOUT MULTIPLE WORD LIST ITEMS WITH MULTIPLE POINTERS.

IT HAS BEEN IMPLICITLY ASSUMED IN THE PRECEDING DISCUSSION
THAT IF, FOR EXAMPLE, THREEWORD ITEMS ARE BEING USED, THEN ALL LISTS
HAVE THREEWORD ITEMS AND THE SPACE LIST PROBLEM IS QUITE SIMPLE.

Fig. 5-10 Text V (Cont.)

NAMELY , IT IS ITSELF A LIST WITH THREEWORD ITEMS.

HOWEVER , THIS
NEED NOT BE THE CASE.

THERE MIGHT BE SEVERAL LISTS SIMULTANEOUSLY
IN STORAGE , EACH WITH DIFFERENT SIZE
ITEMS.

ALTERNATIVELY , THERE
MAY ONLY BE ONE LIST STRUCTURE WHICH ITSELF HAS ITEMS OF VARIOUS
SIZES.

IN EITHER CASE , THE PROBLEM OF HOW TO HANDLE THE SPACE LIST
BECOME SIGNIFICANT.

ONE APPROACH IS TO PROVIDE A SEPARATE SPACE LIST FOR EACH
SIZE
ITEM.

HOWEVER , TO DO THIS DIRECTLY REQUIRES THAT THE PROGRAMMER
DECIDE HOW MUCH SPACE WILL BE ALLOTTED TO EACH LIST , AND THIS IS
DIRECTLY CONTRARY TO ONE OF THE BASIC IDEAS BEHIND THE LIST STRUCTURE
CONCEPT NAMELY $ THE PROGRAMMER DOES NOT
DECIDE THIS , AND IF THERE
IS ANY SPACE AVAILABLE , IT CAN BE USED ANYWHERE.

A MODIFICATION OF THIS APPROACH WILL IMPROVE THINGS SOMEWHAT.

SUPPOSE THERE ARE THREE TYPES OF ITEMS * SINGLETS , DOUBLETS AND
TRIPLETS.

INITIALLY , ALL AVAILABLE SPACE IS PLACED ON THE
TRIPLET
SPACE LIST WITH BOTH COUBLET AND SINGLET SPACE LISTS BEING EMPTY.

Fig. 5-10 Text V (Cont.)

THEREAFTER , IF A DOUBLET IS REQUIRED , AND THE DOUBLET SPACE LIST IS EMPTY , A TRIPLET CAN BE OBTAINED AND DIVIDED INTO A DOUBLET AND A SINGLET , OR A DOUBLET CAN BE USED TO GET SINGLETS.

WHEN SINGLETS OR DOUBLETS ARE RETURNED , THEY GO ONTO THE PROPER SPACE LIST.

THIS , TO A CERTAIN EXTENT , ALLIVIATES THE PROBLEM INDICATED ABOVE.

NAMELY, AS LONG AS THERE IS A TRIPLET AVAILABLE, A DOUBLET OR SINGLET MAY BE OBTAINED.

HOWEVER , IT IS NOT UNREASONABLE TO EXPECT THAT SOME PROBLEMS MAY REQUIRE TRIPLETS WHEN THE ONLY SPACE LEFT CONSISTS OF SINGLETS AND DOUBLETS.

THE FINDING OF NEW TRIPLETS FROM A SET OF DOUBLETS AND SINGLETS CAN BE ACCOMPLISHED , BUT IT IS QUITE INCONVENIENT.

WE CONSIDER NOW A GENERALIZED SPACE LIST , ONE FROM WHICH AN ITEM OF ARBITRARY LENGTH CAN BE TAKEN , IF THAT MANY CONSECUTIVE WORDS EXIST ANYWHERE IN AVAILABLE SPACE.

SUCH A LIST CAN BE REALIZED BY HAVING IT CONSIST OF VARIABLELENGTH ITEMS.

INITIALLY , IT CONTAINS ONLY ONE ITEM WHICH IS MADE UP OF ALL WORDS OF AVAILABLE SPACE.

WHEN A SPECIFIC ITEM IS REQUIRED , THE PROPER NUMBER OF WORDS ARE REMOVED FROM THE SPACE LIST ITEM , THUS REDUCING ITS SIZE.

Fig. 5-10 Text V (Cont.)

LATER WHEN THE
SPACE LIST CONTAINS MANY ITEMS , THE LIST MUST BE SEARCHED FOR THE
FIRST ITEM WHICH IS LARGE ENOUGH * THEN , WHEN AN ITEM IS TO BE
RETURNED , SINCE IT IS DESIRED TO DETERMINE WHETHER OR NOT THIS ITEM
FITS CONSECUTIVELY ONTO SOME CURRENT SPACE ,
IT IS SORTED
ONTO THE SPACE LIST.

TO ILLUSTRATE HOW SUCH A LIST MIGHT BE HANDLED , SUPPOSE FIG A
IS A PORTION OF THE SPACE LIST.

THE NUMBER REPRESENTS THE LENGTH OF
THE ITEM WHICH IS REQUIRED FOR THE SORTING PROCESS.

Z AND X REPRESENT
THE MEMORY LOCATIONS OF THE FIRST WORD OF THE TWO ITEMS.

WE POSTULATE THAT A
TWOWORD ITEM , SHOWN IN FIG B , IS TO BE PUT BACK ONTO THE SPACE LIST.

HOWEVER , IF POSSIBLE , IT IS TO BE PUT BACK IN ORDER TO MAKE A LONGER
ITEM.

THEREFORE , IT
IS NECESSARY TO FIRST SEARCH DOWN THE ITEMS OF
THE SPACE LIST TO THE POINT WHERE Z EQUALS X.

THEN , THE DOUBLET CAN GO ONTO
THE LIST IN FOUR POSSIBLE WAYS.

IN THIS WAY , THE SPACE LIST ITEMS ARE ALWAYS OF MAXIMUM
LENGTH.

Fig. 5-10 Text V (Cont.)

ADMITTEDLY , THIS OPERATION TAKES A SMALL BUT SIGNIFICANT AMOUNT OF TIME , SINCE THE SPACE LIST MUST BE SEARCHED BOTH COMING AND GOING.

HOWEVER , IN THOSE CASES WHERE SIMPLER TECHNIQUES ARE NOT SUFFICIENT , IT IS THE PRICE TO BE PAID.

Fig. 5-10 Text V (Cont.)

III

STUDIES IN PHONETIC ENGLISH

6. STATISTICS OF OPERATIONALLY DEFINED HOMONYMS OF ELEMENTARY WORDS*

B. V. Bhimani, L. L. Earl, and R. P. Mitchell

Words that are pronounced the same but have different spellings and meanings, as for example pail and pale, generally called homonyms, have long been of interest to punsters. Systematic investigation of the number and nature of these words shows that they are also of more general and serious interest. Of the approximately 5,700 "elementary" words in the dictionaries studied,[1] about 3,000 can be ambiguous in their spoken form. Moreover, many of these words are common words; in the 503 words in Godfrey Dewey's word list[2] with a text frequency of more than 20 in a sample of 100,000, only 222 words are not part of a homonym set. Thus, homonyms are a significant class of words not to be overlooked in the study of the English language.**

For purposes of this study, a homonym set was defined as a set of different orthographic forms having an identical phonetic transcription as provided by a specified authoritative source. Any member of a homonym set is called a homonym. An exhaustive compilation of all such sets was made, by computer program, from the 5,757 elementary words listed in the five dictionaries considered, each of which provides an authoritative phonetic transcription.[4-8]

---

*Supported by the Lockheed Independent Research Program.

**According to the 2nd edition of Fowler's A Dictionary of Modern English Usage,[3] Robert Bridges published an essay on homophones in 1919, as Tract II of the Society for Pure English, in which he compiled lists of words that are pronounced alike but have "different origin and signification." His lists contained 835 entries comprising 1,775 words (not limited to one-syllable words, and not including words that were originally the same but have acquired different meanings), which led him to the propositions that homophones are a nuisance and that English is exceptionally burdened with them. He proposed also, however, that homophones are self-destructive and tend to become obsolete, a proposition which may be questioned in the light of our recent compilations.

References 4 through 8, respectively, will be referred to by the following abbreviations:

- MW3
- KK
- ACD
- JON
- SOX

The homonym sets were derived separately for each dictionary, so that differences in the phonetic symbology of the dictionaries did not cause any problems. For each compilation, all 5,757 elementary words listed were considered, even though each word did not appear in all five dictionaries. Before the homonym sets were compiled, each pronunciation of each word was identified by dictionary source and also by class of dialect when applicable.* For words missing from one or more of the dictionaries, the missing phonetic transcriptions were generated by algorithm and marked with an indicator so they could be readily identified as special cases.**

---

*SOX and JON represent speech patterns in Great Britain; sometimes variant British pronunciations are given in JON. The other three dictionaries represent speech patterns in the United States. ACD represents the midwestern speech pattern, with occasional variant pronunciations given. KK presents separately the pronunciation of words in eastern, southern, and midwestern "dialects." MW3 presents speech in regions considered by KK and also in regions of New York City (e.g., Brooklyn and Bronx).

**Instead of transcribing the phonetics from the dictionaries, a highly accurate algorithm (better than 93 percent accurate) was devised for automatically generating the phonetic form for each dictionary from the graphic form. The generated forms were then checked against the dictionaries, and errors were corrected. Corrected words were marked with a D indicator. The phonetic representations of words missing from a given dictionary could not be directly checked, however, and were marked with (1) an N indicator if the algorithm had functioned correctly in deriving the SOX phonetics of that word or (2) an M indicator if the algorithm had given incorrect results on this dictionary, in which case the probable error had been corrected. Thus, the M indicator is almost equivalent to an N + D marker. The algorithms for generating phonetic transcriptions are described in two not-yet published manuscripts, "Acoustic Phonetic Transcription of Written English," by B. V. Bhimani and J. L. Dolby, and "The Operational Relation Between the Phonetic Forms of Elementary Words," by B. V. Bhimani and R. P. Mitchell.

The statistics of the homonym compilation in each of the five dictionaries are given in Table 6-1 and graphically in Fig. 6-1. (Note the 10 to 1 change in scale in Fig. 6-1 between sets of 3 and sets of 4.) Figure 6-2 is a sample page from one of the homonym printouts. The first three columns give the graphic form split into consonant and vowel strings; the next three columns give the code for the phonetic representation; the seventh column indicates the set of algorithmic rules by which the phonetic representation was derived,[2] and the final column indicates the source of the phonetic data used. A blank line separates the homonym sets.

Table 6-1

NUMBER OF HOMONYM SETS IN FIVE DICTIONARIES

Dictionary	MW3	KK	ACD	JON	SOX
No. 2 Word Sets	1889	1402	717	727	661
No. 3 Word Sets	380	268	133	142	117
No. 4 Word Sets	99	55	33	31	27
No. 5 Word Sets	18	11	4	8	3
No. 6 Word Sets	9	5	2	0	0
No. 7 Word Sets	1	1	0	0	0
No. 8 Word Sets	1	0	1	1	0
No. 9 Word Sets	0	1	0	0	0
No. 10 Word Sets	1	0	0	0	0

Surprisingly, both the number of sets and number of total words involved in homonym sets differ considerably from dictionary to dictionary, and a word which may be in a homonym set according to the phonetic representation in one dictionary may not have a homonym according to another dictionary. Accordingly, a homonym comparison table of the 5,757 words considered was prepared by a computer program, showing in which dictionaries each word occurs in a homonym set, and how many and which phonetic representations were involved. Table 6-2 summarizes the phonetic

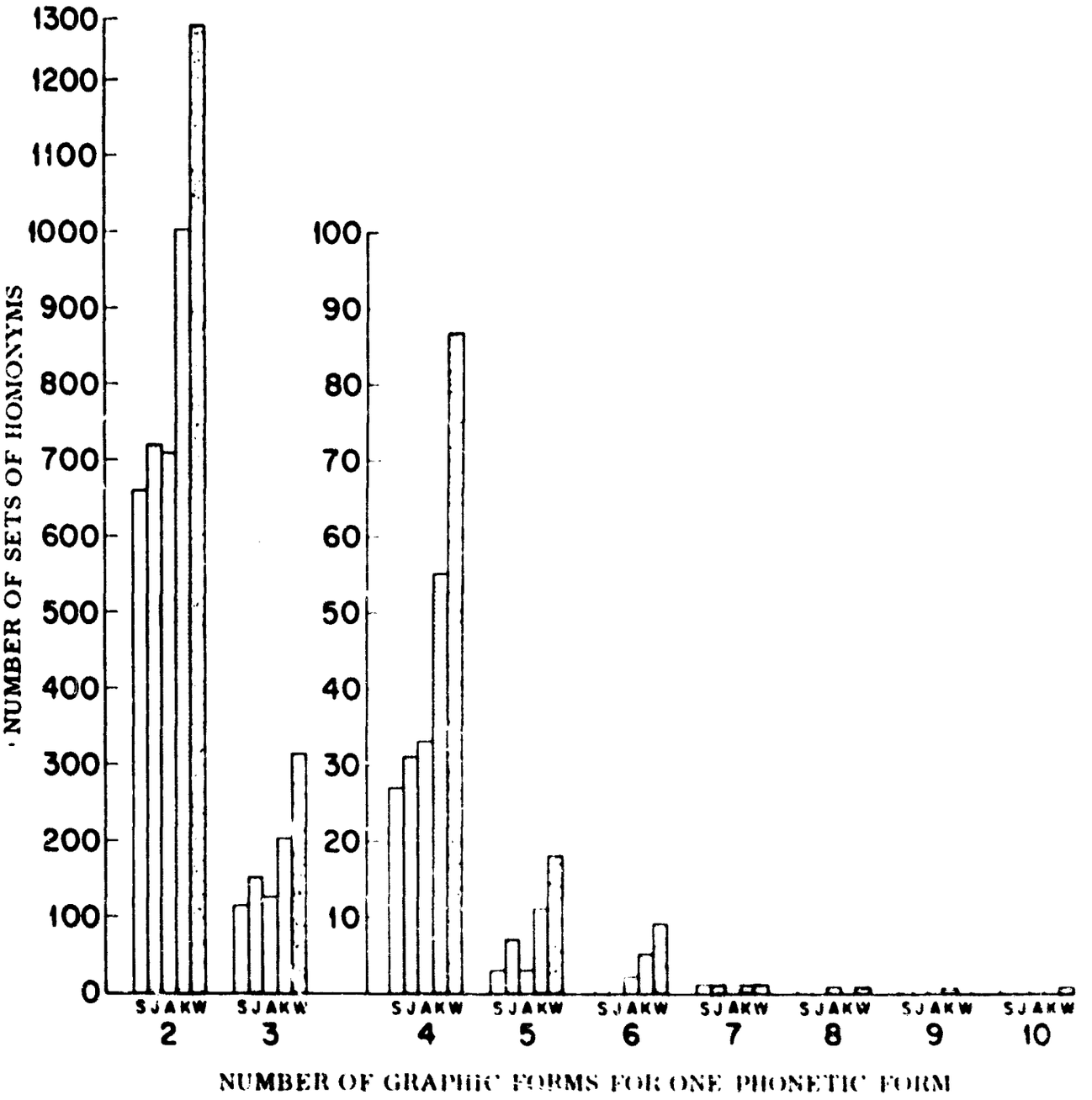

Fig. 6-1 Graphic Presentation of Number of Homonym Sets in Five Dictionaries

CH	E	RE	CH	A3	R	*NAD ACD1
CH	A	RE	CH	A3	R	CMSA ACD1
CH	AI	R	CH	A3	R	CMSA ACD1
D	E	RE	D	A3	R	CMSA ACD1
D	A	RE	D	A3	R	CMSA ACD1
E	A	RE	E	A3	R	CMSA ACD1
E	AI	R	E	A3	R	CMSA ACD1
EH	A	RE	E	A3	R	*NAD ACD1
FL	AI	R	FL	A3	R	CMSA ACD1
FL	A	RE	FL	A3	R	CMSA ACD1
GL	A	RE	GL	A3	R	CMSA ACD1
GL	EI	R	GL	A3	R	*NAD ACD1
GL	AI	R	GL	A3	R	CMSA ACD1
H	AI	R	H	A3	R	CMSA ACD1
H	AI	RE	H	A3	R	*NAD ACD1
H	A	RE	H	A3	R	CMSA ACD1
I	EI	R	I	A3	R	*NAD ACD1
I	E	RE	I	A3	R	*NAD ACD1
J	E	RE	J	A3	R	*NAD ACD1
K	A	RE	K	A3	R	CMSA ACD1
K	AI	R	K	A3	R	*NAD ACD1
L	E	RE	L	A3	R	*NAD ACD1
L	AI	R	L	A3	R	CMSA ACD1
L	A	RE	L	A3	R	*NAD ACD1
M	EA	R	M	A3	R	*NAD ACD1
M	A	RE	M	A3	R	CMSA ACD1
M	A	RC	M	A3	R	+ACD ACD2
N	E	RE	N	A3	R	*NAD ACD1
N	EE	R	N	A3	R	SVSA ACD1
N	A	RE	N	A3	R	*NAD ACD1
P	AI	R	P	A3	R	CMSA ACD1
P	A	RE	P	A3	R	CMSA ACD1
P	EA	R	P	A3	R	CMSA ACD1
SC	A	RCE	SK	A3	RS	*NAD ACD1
SC	A	RCE	SK	A3	RS	SVSA ACD1
SP	A	RE	SP	A3	R	*NAD ACD1
SP	A	RE	SP	A3	R	CMSA ACD1
ST	A	RE	ST	A3	R	CMSA ACD1
ST	AI	R	ST	A3	R	CMSA ACD1

Fig. 6-2 Sample Page of Homonym Printout

Table 6-2

PHONETIC REPRESENTATION CODES

Code	Interpretation	Dictionary
JON 1	1st pronumciation	JON
JON 2	2nd pronunciation	JON
ACD 1	1st pronunciation	ACD
ACD 2	2nd pronunciation	ACD
101SK	Midwestern pronunciation	KK
102SK	First Variant pronunciation	KK
103SK	East and South pronunciation	KK
104SK	East pronunciation	KK
105SK	Second Variant pronunciation	KK
106SK	Third Variant pronunciation	KK
107SK	Fourth Variant pronunciation	KK
101SW	Midwestern pronunciation	MW3
102SW	First Variant pronunciation	MW3
103SW	Boston R Dropper pronunciation	MW3
104SW	Brooklyn R Dropper pronunciation	MW3
105SW	L Dropper pronunciation	MW3
106SW	Second Variant pronunciation	MW3
107SW	Third Variant pronunciation	MW3
108SW	Fourth Variant pronunciation	MW3
109SW	Fifth Variant pronunciation	MW3
20XSW	Consonant variant pronunciation on the 10x pronunciation of	MW3
20XKK	Consonant variant pronunciation on the 10x pronunciation of	KK

codes used. Note that dictionaries do not all give the same number of phonetic variations, nor are their phonetic classes always the same. SOX usually gives only one pronunciation, and therefore there are no SOX entries in Table 6-2. Figure 6-3, a sample page from the homonym comparison table, indicates, for example, that the word fon is involved in a homonym set only according to the MW3 pronunciation. Yet the word fort is involved in six MW3 homonym sets, four KK sets, one JON set, one ACD set, and no SOX set. (In general, SOX has the fewest homonyms, indicating perhaps that the SOX phonetic transcription system is finer.)

The total number of words in the homonym comparison table is 2,966, showing that 2,966 of 5,757 words are in a homonym set according to at least one dictionary. Thus, over 50 percent of the elementary words are ambiguous in their spoken form. The homonym comparison table points up two significant findings, the apparent disparity among dictionaries and the large percentage of elementary words distinguished in the graphic but not the spoken form (as recorded by the dictionaries).

Before exploring the possible reasons for the disparity in homonym sets according to the dictionary from which derived, some possibilities can be eliminated. First, since all these dictionaries were published at approximately the same time, and since it is generally recognized that their contents are periodically updated, historic vowel changes are not expected to cause discrepancies. Also, vowels which are consistently pronounced one way according to one dicationary and another way (but always the same other way) according to a second dictionary, will affect the homonym compilation very little. For example, break and brake are homonyms whether the vowel is given a British pronunciation as indicated by "b r e i k" in JON or an American pronunciation

F	I	NN	101SW	101SK	JON1	ACD1	SOX
F	I	R	101SW 102SW 103SW	103SK 101SK	JON1	ACD1	SOX
F	I	RN			JON1		
F	I	RTH	101SW	103SK 101SK	JON1	ACD1	SOX
F	I	SC	101SW	101SK	JON1	ACD1	SOX
F	I	SK	101NW	101SK	JON1	ACD1	SOX
F	I	T	201SW 101SW	101SK	JON1	ACD1	SOX
F	I	ZZ	101SW	101SK	JON1	ACD1	SOX
F	IE		101SW	101SK	JON1	ACD1	SOX
F	IE	L	101NW 102NW	101NK	JON1	ACD1	SOX
F	O	H	101SW	101NK	JON1	ACD1	SOX
F	O	HN	101MW	101NK		ACD1	
F	O	LD			JON1		
F	O	N	101SW				
F	O	NE	101NW	101NK	JON1	ACD1	SOX
F	O	NT	101SW			ACD1	
F	O	R	102DW 101DW 103DW	101DK 103DK	JON1 JON2 JON1	ACD2	
F	O	RD	103SW 104SW				
F	O	RE	101SW 103SW 104SW 102SW	101SK 103SK 102SK 104SK	JON2	ACD1	SOX
F	O	RT	101SW 203SW 103SW 204SW 104SW 102SW	101SK 103SK 102SK 104SK	JON1	ACD1	
F	O	RTE	101SW 203SW 103SW 104SW 102SW 204SW	101DK 103DK 102DK 104DK	JON1	ACD1	
F	O	RTH	101SW 103SW 102SW	101SK 103SK 102SK	JON1	ACD1	SOX

Fig. 6-3 Sample Page of Homonym Comparison Table

as indicated by "b r e k" in KK. The list below gives the phonetic symbols for this sound from each of the five dictionaries and the corresponding code used for machine purposes. (JON and KK use the International Phonetic Alphabet.)

SOX	b r e^{i} k	B R E1419 K
JON	b r e i k	B R E I K
ACD	b r ā k	B R A 4 K
KK	b r e k	B R E K
MW3	b r ā k	B R A 4 K

Thus, consistent changes from dialect to dialect will not cause significant discrepancies in homonyms.

What then will cause discreparcies from dictionary to dictionary? When several dialects are considered together in the compilation of homonyms, as in KK and MW3, extra homonym sets or larger sets can be produced across the dialects. For instance, two words which are not homonyms in either the southern or eastern dialects may become homonyms when the southern pronunciation of one is compared with the eastern pronunciation of the other. By removing the dialect pronunciations from the homonym sets, two objectives are met:

- The ambiguity-producing effects of dialects are shown.
- Homonym disparities between ACD and KK or MW3 which result from the inclusion of dialects are removed.

In removing dialects, some difficulty is encountered in identifying true dialectal pronunciations. The 103SK, 104SK, 20XSK (where X is any number), 103SW, 104SW, 105SW, 30XSW, and 20XSW pronunciations (Table 6-2) were considered to be true dialects by the dictionaries in which presented and were, therefore, removed by computer program from the homonym sets. The homonym comparison program was run

again on the homonyms after the removal of the dialectal pronunciations to produce another homonym comparison table of the same form as shown in Fig. 6-3. The results show the expected reduction in the number of sets containing a given word and in the number of words that appear in homonym sets, but these reductions are not as large as was expected.

To show the relationships among the five dictionaries from the point of view of the involvement of the words in homonym sets, some statistics of homonym membership were compiled and are given in Table 6-3. Since the statistics were compiled from the homonym comparison tables, which were compiled before and after the removal of the dialects, the effect of the dialect removal is shown. Note that with the dialects removed the number of elementary words which are in homonym sets is reduced only about 5 percent, from 52 to about 47 percent. Note also that the relationships among the various sets named in Table 6-3 not change significantly. In particular, the ratio between the words forming a homonym in all dictionaries and the words forming a homonym in any dictionary changes only from 0.5074 to 0.5467 when dialects are removed. Thus, the dialects are not the main reason for the large number of homonyms, nor are they the major cause of discrepancies among the dictionaries.

It is also revealing to consider the actual occurrence of ambiguity introduced by the dialects, and because they are not numerous we have prepared tables which give them all. In Table 6-4, Part A shows all new sets introduced by the dialect pronunciations of KK; Part B shows all words or sets added to nondialectal homonym sets by a dialect pronunciation of KK. The starred items were not removed by the program but seemed to the authors to be dialect forms and were removed later. Table 6-5 shows all the dialectal pronunciations removed from MW3, but here we have divided them into nine significant categories.

Table 6-3

STATISTICAL SUMMARY OF WORDS INVOLVED IN HOMONYM SETS, SHOWING EFFECT OF DIALECT REMOVAL

Set Description	No. of Words in Set	
	With Dialects	Without Dialects
Total Set: Words forming a homonym in at least one dictionary	2966	2714
Words forming a homonym in one dictionary	746	535
Words forming a homonym in two dictionaries	236	214
Words forming a homonym in three dictionaries	189	184
Words forming a homonym in four dictionaries	290	297
Words forming a homonym in all dictionaries	1505	1484
Words forming a homonym in SOX	1754	1754
Words forming a homonym in ACD	1937	1937
Words forming a homonym in JON	2039	2039
Words forming a homonym in MW3	2600	2297
Words forming a homonym in KK	2140	2096

Table 6-4

WORDS INVOLVED IN HOMONYM SETS IN KK BECAUSE OF DIALECTAL PRONUNCIATIONS

Part A

Graphic	Phonetic	Dictionary Code
MUZZ	MA6Z	101NK
MUS		201NK
DAZE	DEZ	101SK
DASE		201NK
GREITH	GREP1	201NK
GRAETH		101DK
NAIS	NEZ	201NK
MAZE		101NK
CLEAR	KLE1E2(R	103SK
CLARE		104SK
REAR	RE1E2(R	105SK
RARE		104SK
MY	ME2	103 or 104SK
	or	
MAC	MI	105 or 106SK
BROOSE	IUIZ	202NK
BRUISE		102DK
CHESE	T$I1Z	201NK
CHEESE		101NK
CROZE	KROZ	101NK
CROSE		201NK
SHORE	$0E2(R	104 or 103SK
SURE		105 or 102SK
LAUSE	HOIZ	201NK
HAWSE		101NK
BROOSE	BRU1Z	201NK
BRUISE		101DK
COUTH	KU1P1	101SK
COOTH		201MK
JEER	DZ1E1E2(R	105SK
*GEER	DZ1E1E2(R	105SK
JEER	DZ1E1E2(R	105SK

Graphic	Phonetic	Dictionary Code
*FEAR	FE1E2(R	105SK
FEER	FE1E2(R	105SK
*FLEAR	FLE1E2(R	105SK
FLEER	FLE1E2(R	105SK
HEAR	HE1E2(R	107SK
*HEER	HE1E2(R	107SK
HERE	HE1E2(R	108SK
*LEAR	LE1E2(R	105SK
LEER	LE1E2(R	105SK
TEAR	TE1E2(R	106SK
*TEER	TE1E2(R	105SK
TIER	TE1E2(R	105SK
*WEIR	WE1E2(R	105SK
WERE	WE1E2(R	105SK
*TROTH	TRA3P1	105SK
TROUGH		106DK
*BUM	BA6M	101SK
BOMB		102SK

Part B

NEEZE	NIIZ	101NK
*WERE	WA1E2(R	107SK
*OUR	A2U(R	105SK
*EAR	E1E2(R	105SK
*BIER	BE1E2(R	105SK
BEER	BE1E2(R	105SK
*BLEAR	BLE1E2(R	105SK
*DEER	DE1E2(R	105SK
DEAR	DE1E2(R	105SK
*KIER	KE1E2(R	105SK
*MEER	ME1E2(R	105SK
*PEER	PE1E2(R	105SK

Table 6-4 (Cont.)

Graphic	Phonetic	Dictionary Code
*SPEAR	SPE1E2(R	105SK
*SPEERE	SPE1E2(R	105SK
*CHEER	T$E1E2(R	105SK
*AND	E2N	106SK
*WEAR	WIE2(R	106SK
*POOR	POE2(R	105SK
*PRYSE	PRAIZ	201SK
*BLOUSE	BLAUZ	201SK
*CLOUGH	KLA2F	103DK
*DON	DA3N	103SK
*WOT	WA3T	103SK
*SHARE	$E1E2(R	103SK
*CERE	SE1E2(R	104SK
*ERR	E3(R	103SK
*YAIR	JE1E2(R	104NK

Table 6-5

WORDS INVOLVE' IN HOMONYM SETS IN MW3 BECAUSE OF DIA :CTAL PRONUNCIATIONS

Set A

PUT	DROWTE	SATE	SNOT	WET	CLEAT	LIT	QUOTE
PUD	DRAD	SADE	SNOD	WED	CLEAD	LID	QUOD
NEWT	CLOUT	SLATE	TROT	CHUT	LEASE	MITT	TOTE
NUDE	CLOUD	SLADE	TROD	CHAD	LEESE	MOD	TOAD
FAT	CROUT	TRAIT	BET	GLUT	PLEAT	WRIT	BROUGHT
FAD	CROWD	TRADE	BED	BLOOD	PLEAD	ROD	BROAD
GAT	LOUT	BRAT	BRUTTE	HUT	SPETE	SKIT	BRAUGHT
GAD	LOUD	BROD	BUD	HUD	SPEED	SKID	BRAUD
HAT	BLATE	DOT	FET	CRUT	TWEET	FRIGHT	SQUAT
JAD	BLADE	DOD	FED	CRUD	TWEED	FRIED	SQUAD
CAT	DATE	CLOT	GET	MUTT	WEET	KRAIT	SHAT
CAD	DADE	CLOD	GED	MUD	WEED	CRIED	SWAD
GNAT	DASE	POT	KET	SHUT	WHIT	PIGHT	WATT
NAD	DAZE	POD	KED	SHOULD	WHID	PIED	WAD
PAT	SOOT	PLOT	PET	SCUT	BRIT	SNITE	FEUTE
PAD	SUD	PLOD	PED	SCUD	BRID	SNIDE	FOOD
PLAT	CADE	SOT	STET	STUT	GRIT	TIGHT	HOOT
PLAID	CATE	SOD	STEAD	STUD	GROD	TIDE	HOOD
RAT	PATE	SKOT	THREAT	BLEAT	KIT	TRITE	MOOT
RAD	PAID	SKOD	THREAD	BLEED	KID	TRIED	MOOD
WAT	RATE	SQUAT	TRET	CHESE	QUIT	CROSE	FOOT
WAS	RAID	SQUAD	TREAD	CHEESE	QUID	CROZE	FOOD

Table 6-5 (Cont.)

Set B

SHRAG	CARVE	SHRINE	MON
SWAG	CALVE	SWINE	MUM
CHERT	CLART	SHRIVE	MONT
CHAT	CLAUT	SWIVE	MENT
HAULSE	MARL	SOURCE	PURSE
HOUSE	MALL	PSOAS	PUS
HAULT	PARSE	FAULT	SHONG
HOUT	PASS	FOUGHT	SHUN
GOLF	SCARP	GAULT	THIS
GOFF	SCAUP	GHAUT	THUS
ARSE	SMARM	SURE	AL
ASS	SMALM	SHRR	ILE
BARGH	SPAR	SPEARE	DEE
BAFF	SPA	SPHERE	DIT
BARM	TAR	SAVLE	LA
BALM	TA	SERVE	LAW
BARSE	HEARSE	UGH	DRAUGHT
BASS	HUSS	HER	DROUTH
BARTH	SIR	FUM	THEE
BATH	SO	FROM	THY
CHAR	SEER	DUD	TIE
CHA	SEA	DID	TAILLE
DART	SHRIFT	GUN	KINE
DOT	SWIFT	GON	KIN
GAR	SHRILL	HUFF	FETCH
GAW	SHILL	HAVE	REACH
HAUGH	SHRINK	HUZZ	ILL
HARK	SWINK	HAS	ILE
JAR			
JAH			

Table 6-5 (Cont.)

Set B

AT	WADE	CUT	DID	CODE
BAT	KOD	NOULD	FID	LOTE
BRAT	GOD	PUD	GED	NODE
DRAD	QUAT	RUT	KIT	TOTO
LAD	NOD	SHOOT	CID	SHOAT
MAD	ROD	ESE	BIDE	BOOT
SCAD	SWAD	FEED	BRIDE	BROUD
BLOUSE	TOD	GLEET	GUIDE	LEUD
FADE	WAD	GREED	HIDE	HARD
GADE	BAWD	NEAT	SIDE	CARD
GRADE	RET	REIT	SICE	SAID
HADE	SAID	CEASE	SLIDE	BIDE
LATE	IDE	SWEDE	WIDE	BRIDE
MATE	BUD	IT	OAT	GUIDE
SPADE	FUD	BID	BODE	HIDE

Set D

DALT	HARRE	AR	*CAID	*MORE
AR	HARM	(	*HOLD	*HOW
BARK	CARF	AYAH	*HAULM	*YOUR
GUAD	MAR	SOY	*HORSE	

Table 6-5 (Cont.)

Set E

AIT
EIGHT
EYGHT

AUGHT
GUTTE

GOT
GHAIT

GRETT
BRET

DEBT
PETTE

LET
LETT

GUT
GUTTE

BEAT
BEET

HEAT
HETE

LEET
LEAT

METE
MEET
MEAT

CETE
SEAT

NIGHT
KNIGHT

WRITE
RIGHT

GOAT
GOTE

MODE
MOD

BOUGHT
BOTT

ROOD
RUDE

Set F

BAR
BARR

PAR
PARR

EARN
URN

Set G

CHAD

DOWD

BIRT

HERD

FORD

CORT

WARD

Set H

SWATH
SWATHE

FART
FAD

SPOUT
SPALD

SWEAT
SHED

GIRT
GIRD

CURT
CURD

WORT
WORD

GIRT
BIRD

CURT
CURD

CEAT
SURD

TIT
TEAT

SORT
SWORD

Set I

GHAUT
GALT

Set A. New homonym sets in which a pronunciation of type 20X is involved. These reflect confusion between T and D or S and Z sounds, which may not be strictly a dialectal phenomenon.

Set B. New homonym sets in which a pronunciation of the type 20X is not involved.

Set C. Words in which a pronunciation of the type 20X adds one to the number of homonyms in a nondialectal homonym set.

Set D. Same as C, except a non-20X dialectal pronunciation is responsbile for an extra member of a homonym set. (Starred items were added by hand as in Table 6-4.)

Set E. New homonym sets caused by a pronunciation of the type 20X, where each of these sets has the same pronunciation as a nondialectal homonym set. Thus, these words add more than one member to a nondialectal set.

Set F. Same as E, except a non-20X dialectal pronunciation is responsible for the extra members to homonym sets.

Set G. Words in which a dialectal pronunciation causes confusion with words already in sets B or D. Thus, a dialectal pronunciation of chert causes the homonym set chert, chat. A dialectal pronunciation of chad adds to the set, making it chert, chat, chad.

Set H. New homonym sets in which two dialectal variations combine to form a homonym group.

Set I. New homonym sets in which two dialectal variations combine to form a homonym group, where each of these groups has the same pronunciation as a nondialectal homonym set.

To summarize our results, it has been shown, using phonetic representations from five dictionaries, that approximately half of the elementary words of English are

ambiguous according to at least one dictionary, and that this figure is not significantly changed by removal of predefined dialectal pronunciations. The words whose dialectal pronunciations have affected the homonym sets have been listed. Discrepancies in homonym data among the five dictionaries have been made apparent. It has been indicated that neither historic vowel changes nor consistent vowel changes can be considered to be a major cause of these discrepancies. Also, it has been shown that the dictionary-defined dialectal vowel variations account for only a small proportion of these discrepancies.

REFERENCES

1. J. Dolby and H. Resnikoff, "On the Structure of Written English Words," Language, 40, 2, April–June 1964
2. Godfrey Dewey, Relative Frequency of English Speech Sounds, Cambridge, Mass., Harvard University Press, 1923
3. H. W. Fowler, A Dictionary of Modern English Usage, 2nd ed., revised and edited by Sir Ernest Gowers, New York and Oxford, Oxford University Press, 1965
4. Webster's Third New International Dictionary of the English Language, Springfield, Mass., G. C. Merriam Company, Publishers, 1961
5. J. S. Kenyon and T. A. Knott, A Pronouncing Dictionary of American English, Springfield, Mass., G. C. Merriam Company, Publishers, 1958
6. The American College Dictionary, New York, Random House, Inc., 1962
7. Daniel Jones, Everyman's English Pronouncing Dictionary, 12th ed., New York, Dutton, 1963
8. The Shorter Oxford English Dictionary on Historical Principles, 3rd ed., revised with addenda, Oxford at the Clarendon Press, 1959

7. ACOUSTIC PHONETIC TRANSCRIPTION OF WRITTEN ENGLISH*

B. V. Bhimani and J. L. Dolby

INTRODUCTION

The current spelling of an English word is the symbolization of a traditionally preserved form of its pronunciation, even though it may seem to be an imperfect representation. We shall investigate the accuracy of this representation by a detailed examination of all of the one-syllable words given in the Shorter Oxford Dictionary.[1] In particular, we shall show that it is possible to construct a computable algorithm that provides the correct phonetic representation (according to the source dictionary) for 93 percent of these words, given only the written form of the word.

The essential feature of this algorithm is that it makes use of what we shall here call the "marking system of written English." In some writing systems explicit markers are used to indicate vowel duration and stress. Sanskrit, for instance, uses a phonetic alphabet and numerals for indication of vowel duration along with markers for stress. Thus the English word ALMS would be represented as आ३म्स् . In French, diacritic markers are used for similar effects (e.g., HÔPITAL).

In English, however, the orthography is limited to the 26 letters of the alphabet and the marking system is more subtle. One well-known feature of the English marking system is the use of the final E (represented here as Ɇ) as a marker operating on the preceding vowel string. Proper use of the markers is necessary in written English for phonetic transcription of its words.

The initial restriction of this study to the one-syllable words of English was made to enable us to study the marker system for precise transcription of vowel articulation

*This work was supported by the Lockheed Independent Research Program.

and duration and to resolve certain consonantal ambiguities without analyzing the added complications introduced by the stress-marking system necessary for polysyllabic words. The stress-marker system of written English is intimately connected with the systems for carrying important grammatical signals.[2,3] As has been noted elsewhere,[4] the one-syllable words (except for the small but important set of structure words) are generally grammatically homogeneous.

THE SCHEMATIC STRUCTURE OF THE ALGORITHM

The algorithm considered here has been programmed on a digital computer. This device uses a limited number of symbols. As a result, it was necessary to replace the phonetic coding system of the Shorter Oxford Dictionary (SOX) with a set of alphanumeric codes acceptable to the machine. These codes are given in Fig. 7-1. It will be noted that the transformation from the dictionary codes to the machine codes is one-for-one so that no essential information is lost by this step in the procedure. Moreover, the alphanumeric codes were chosen to ensure that all possible codes given by the dictionary would be representable. Only 38 of the 150 codes actually occurred in the one-syllable words.

The algorithm itself is shown in schematic form in Fig. 7-2. The first step consists of a simple classification of the written symbols into the system of graphemic vowels, consonants, and markers given in Fig. 7-2. In this system, a final E is classed as a marker, all other occurrences of E together with all occurrences of A, I, O, U, and Y are classed as (graphemic) vowels. All remaining characters are classed as (graphemic) consonants.

Step two consists of an analysis in context to resolve consonantal ambiguities such as those which occur with the graphemic C and G. In the third step, the letter strings

Symbol	Symbol	No Marking	Code Markings								
			–	⌣	ˈ	..	Inv. Period	Super Letter	ˉ	˘	˗
a	A		4	5	6	7	8	9	64	65	74
æ	A1		4	5	6	7	8	9	64	65	74
a	A2		4	5	6	7	8	9	64	65	74
ɒ	A3		4	5	6	7	8	9	64	65	74
e	E		4	5	6	7	8	9	64	65	74
e	E1		4	5	6	7	8	9	64	65	74
ə	E2		4	5	6	7	8	9	64	65	74
ə	E3										
i	I		4	5	6	7	8	9	64	65	74
i	I1		4	5	6	7	8	9	64	65	74
O	O		4	5	6	7	8	9	64	65	74
O	O1		4	5	6	7	8	9	64	65	74
§	O16,										
u	U		4	5	6	7	8	9	64	65	74
u	U1		4	5	6	7	8	9	64	65	74
Break	/										
:	+		.	.	.	.	.	.	.	.	.
Inv. Per	.										

a. Coding of Vowel Sounds and Pertinent Markings for Pronunciations of English Words

Fig. 7-1 Alphanumeric Coding for the Phonetics of the Shorter Oxford Dictionary

g	G	ṅ	n8
ḥ	H	l^y	LY9
r	R	n^y	NY9
ɹ	R1	x	X
s	S	b	B
w	W	d	D
hw	HW	f	F
y	Y	k	K
p	P1	l	L
ɗ	D1	m	M
ʃ	$	n	N
tʃ	T$	p	P
ʒ	Z1	t	T
dʒ	DZ1	v	V
ŋ	N1	z	Z
ŋg	N1G		

b. Coding for Consonant Sounds

Fig. 7-1 Alphanumeric Coding for the Phonetics of the Shorter Oxford Dictionary (Cont.)

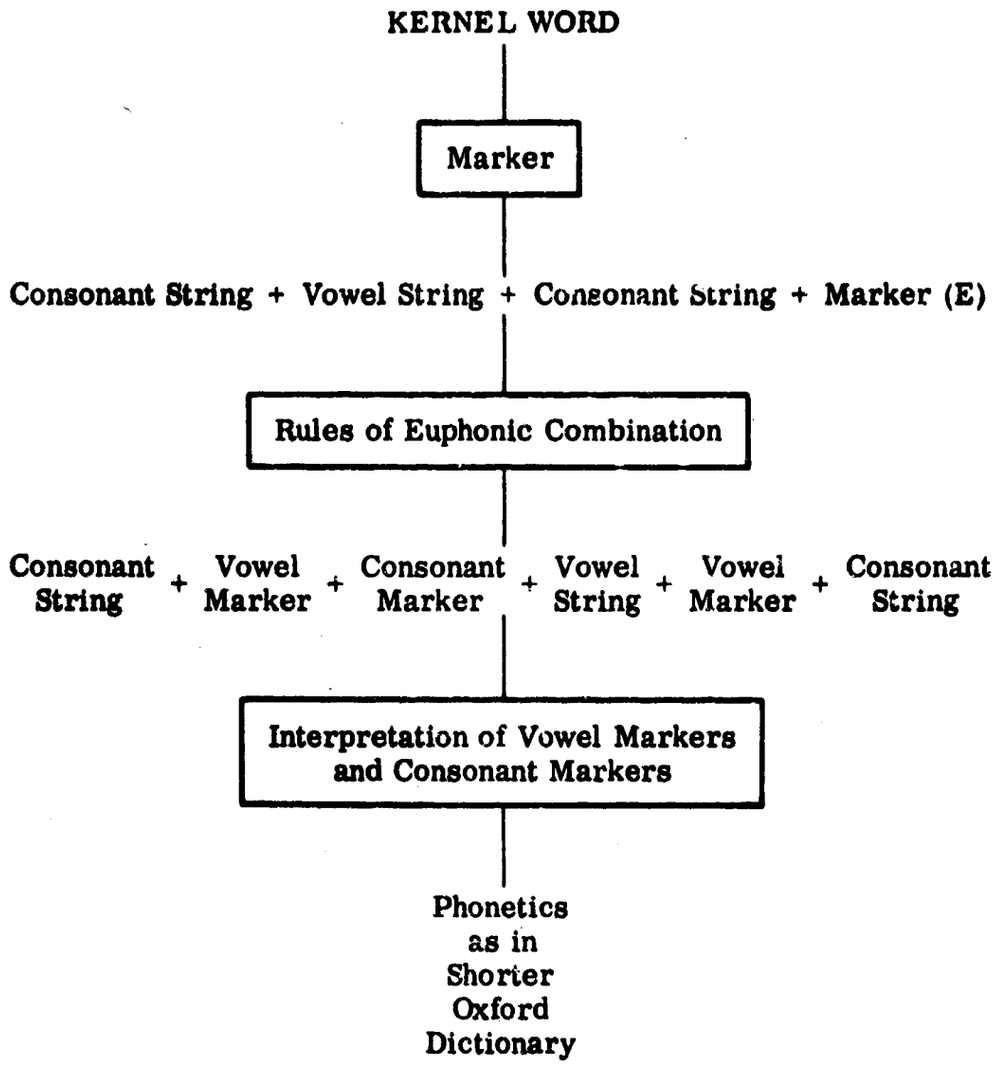

Fig. 7-2 Algorithm in Schematic Form

are processed through the rules of euphonic combination, reported previously,[5] to change the letter strings into consonant strings, vowel strings, and markers. Step four consists of the necessary rules to resolve the vowel ambiguities, which are the main portion of the problem. In step five the vowel markers are used to transform the graphic symbols into the phonetic symbols as given by SOX.

An illustration of the operation of the program is given in Fig. 7-3 for the word NICE. Figure 7-4 illustrates the processing of the word SMUDGE. A typical page of computer output for the phonetics of SOX is given in Fig. 7-5. The first column is the orthographic form of the word, the second column is the compiled phonetic representation, and the third column specifies the rule used for the resolution of the vowel ambiguities. The resulting phonetic codes were checked individually against the source dictionary and correction cards (identifiable by the English words following the asterisk) and were added to the output deck where errors appeared (see, for instance, BLAE in Fig. 7-5).

A total of 407 errors were detected in the 5,757 one-syllable words given in the source dictionary. Some of these errors were a result of errors in the syllable counting routine used to obtain the one-syllable words from a magnetic tape listing of SOX.[6] The word BLASE is the one example of this sort shown in Fig. 7-5. Many of the remaining errors (such as BLAE) occurred in obscure words and words of limited current interest. To obtain a quick check on the expected accuracy of the program on words of greater usage, a random sample of 50 words was chosen from the subset of the one-syllable words having a standard meaning in both the source dictionary and Webster's Third International Dictionary.[7] Only one error was found in this sample (the vowel of CRASS was incorrectly equated to the vowel of BRASS).

In the remainder of the paper we discuss the derivation of the rules necessary to resolve the various ambiguities of written English.

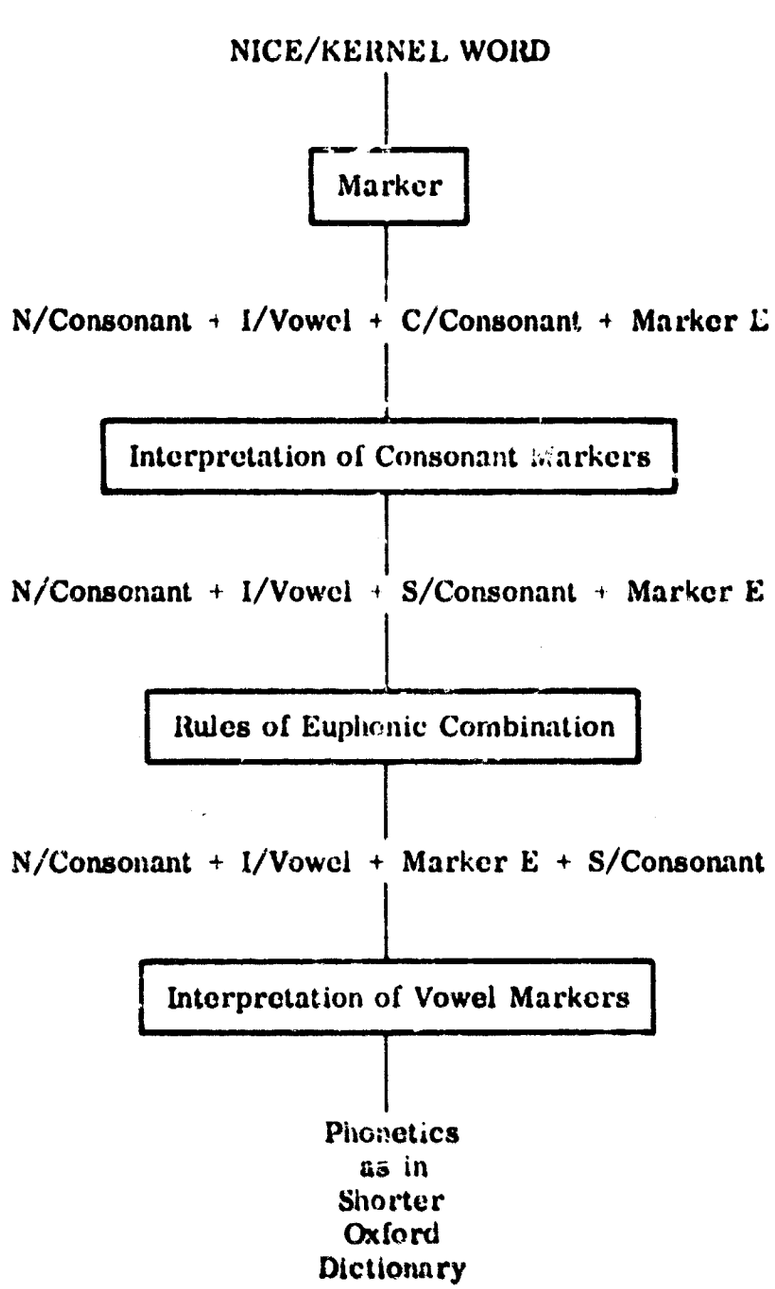

Fig. 7-3 Operation of the Program for the Word NICE

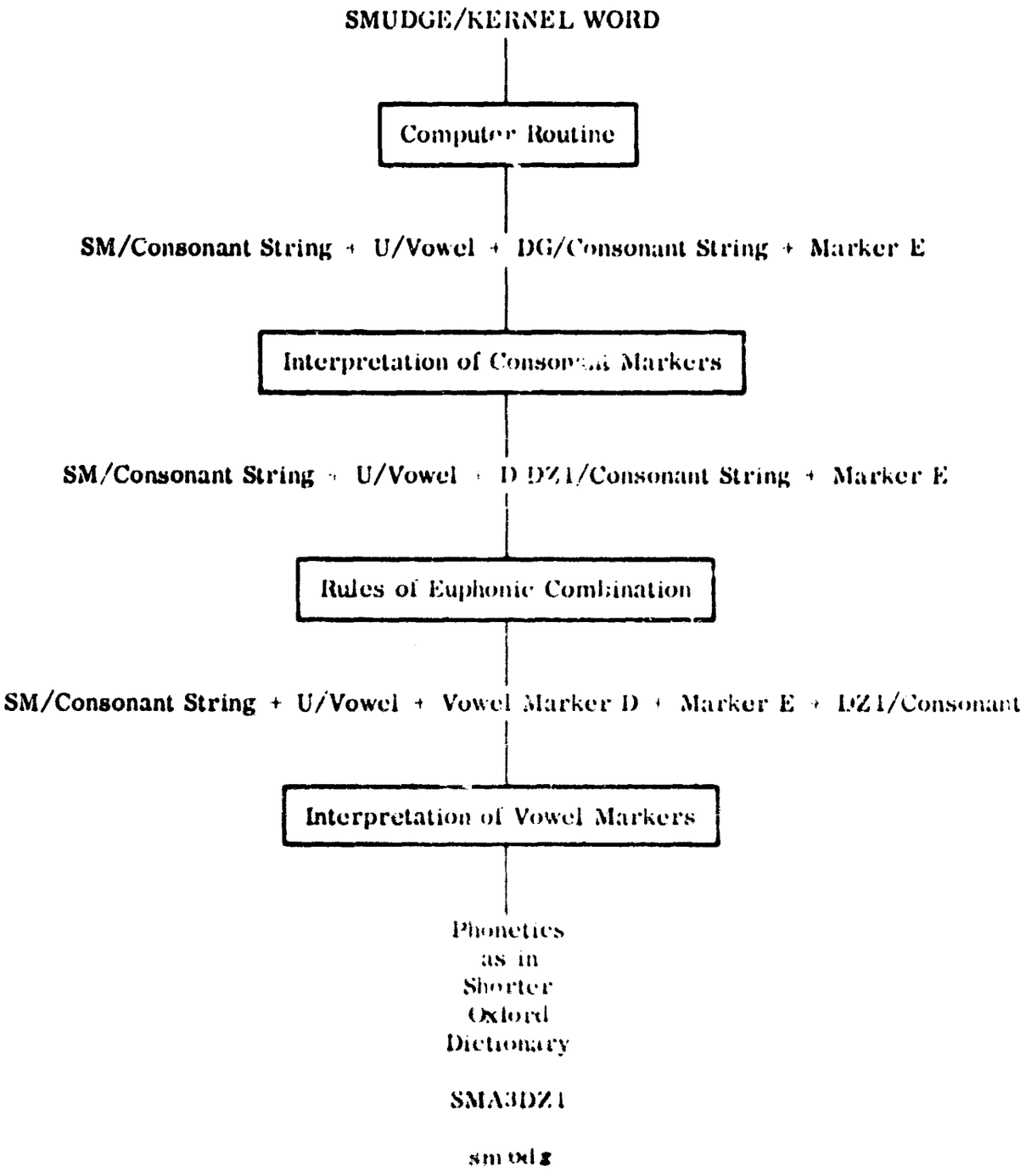

Fig. 7-4 Processing of the Word SMUDGE

ORTHOGRAPHIC FORM	PHONETIC RENDERING	OPERATING RULES
BLACK	BLA1K	•S / .53 , A • •. •
BLAD	BLA1D	•S / .53 , A • •. •
BLADE	BLE1419D	•S / .38 , A • •. •
BLAE	BLE1419	•S / .44 , A • •. •
BLAE	BLE14	•VOWEL ERRORS
BLAIN	BLE1419N	•S / .2 , AI • •. •
BLAKE	BLE1419K	•S / .38 , A • •. •
BLAME	BLE1419M	•S / .38 , A • •. •
BLANCH	BLA1N%	•S / .53 , A • •. •
BLANCH	BLA2N%	•SINGULARITIES
BLAND	BLA1ND	•S / .53 , A • •. •
BLANK	BLA1N1K	•S / .53 , A • •. •
BLARE	BLE4F29R1	•S / .5 , A • •. •
BLAS	BLA1SORZ	•S / .53 , A • •. •
BLASE	BLE1419SOMZ	•S / .38 , A • •. •
BLASE	BLA2.ZE1	•POLYSYLLABIC
BLAST	BLA2ST	•S / .33 , A • •. •
BLATE	BLE1419T	•S / .38 , A • •. •
BLAY	BLE1419	•S / .4 , AY • •. •
BLAYK	BLE1419K	•S / .4 , AY • •. •
BLAZE	BLE1419Z	•S / .38 , A • •. •
BLEACH	BLI14T%	•S / .46 , EA • •.
BLEAK	BLI14K	•S / .46 , EA • •.
BLEAR	BLI14F29R1	•S / .38 , EA • •.
BLEAT	BLI141	•S / .46 , EA • •.
BLEB	BLER	•S / .15 , E • •. •
BLECK	BLEK	•S / .15 , E • •. •
BLEE	BLI14	•S / .2 , EE • •. •
BLEED	BLI14D	•S / .2 , EE • •. •
BLENCH	BLEN%	•S / .15 , E • •. •
BLEND	BLEND	•S / .15 , E • •. •
BLENDE	BLEND	•S / .15 , E • •. •
BLENK	BLEN1K	•S / .217 , E • •.
BLENT	BLENT	•S / .15 , E • •. •
BLERE	BLF4F29R1	•S / .8 , E • •. •
BLESS	BLES	•S / .13 , E • •. •
BLEST	BLEST	•S / .15 , E • •. •
BLET	BLET	•S / .15 , E • •. •
BLICK	BLIK	•S / .14 , I • •. •
BLIGHT	BLE2IT	•S / .4 , I • •. •
BLIN	BLIN	•S / .14 , I • •. •
BLIND	BLE2IND	•S / .9 , I • •. •
BLINK	BLIN1K	•S / .14 , I • •. •
BLINT	BLE24R1T	•S / .2 , I • •. •
BLISS	BLIS	•S / .14 , I • •. •
BLITE	BLE2IT	•S / .3 , IE • •.
BLITHE	BLE2ID1	•S / .3 , IE • •.
BLOAT	BLO14U0T	•S / .7 , OA • •. •
BLOB	BLO1AB	•S / .92 , O • •. •
BLOCK	BLO16K	•S / .92 , O • •. •
BLOK	BLO1AK	•S / .92 , O • •. •
BLOKE	BLO14U0K	•S / .81 , O • •. •
BLOND	BLO1AND	•S / .92 , O • •. •
BLOOD	BLU1D	•S / .1 , OO • •. •
BLOOM	BLU14M	•S / .4 , OO • •. •
BLOOTH	BLU14D10RP1	•S / .8 , OO • •. •
BLORE	BLO4F29R1	•S / .44 , O • •. •
BLOT	BLO1AT	•S / .92 , O • •. •
BLOTCH	BLO1AT%	•S / .92 , O • •. •
BLOTE	BLO14U0T	•S / .81 , O • •. •

Fig. 7-3 Typical Computer Output for Phonetics of the Shorter Oxford Dictionary

THE CONSONANT STRING MAPPING

As noted in Reference 4, it is always possible to represent the graphemic form of a one-syllable word in the form CVC where C represents a string of consonants and V represents a string of vowels. The only conventions necessary to accomplish this are the conventions whereby the final E is treated as a marker and words beginning or ending with a vowel are augmented with the "blank consonant" Ø. The blank consonant is also used in the phonetic form of the word. The relation between the graphemic and phonetic forms of the word can then be studied as a composite of the three mappings that carry the initial consonant string, the vowel string, and the final consonant string from the graphemic form to the phonetic form. For instance, if we consider the word STRAIGHT we obtain the following triple mapping:

	Graphemic	Phonetic
Initial Consonant String	STR	STR
Vowel String	AI	E14I9
Final Consonant String	GHT	$(GH)/_{MARKER}T$

The complexity of this mapping then becomes the central issue. The writing system of English might be considered purely phonetic if one could find that every written form with the initial consonant string STR would also have STR as its initial phonetic consonant string. An even more stringent requirement for an "ideal" writing system would require that each symbol (rather than each string) map into a unique phonetic symbol. It would also be convenient if this map were invertible. That is, if graphemic STR and only STR mapped into phonetic STR. However, numerous examples exist to show that this is not the case (e.g., graphemic F and graphemic PH both into phonetic F). However desirable this might be from the viewpoint of the linguist, it is clear that

it is not in keeping with the written English form. However, the use of the subtle marking system accomplishes these objectives, as discussed next.

When we assign a specific phonetic value for a specific graphic symbol, only two consonants lead to ambiguous situations; namely the graphic consonants C and G. The former maps into either phonetic K or phonetic S, and the latter maps into either phonetic G, or DZ1. There is, however, a subtle consonant marking system that readily resolves this ambiguity with a very high degree of accuracy, and it is tabulated next.

	Graphic	Phonetic
If C is followed by A, O, U	C	K
Otherwise,	C	S
If G is followed by E	G	DZ1
Otherwise	G	G

Such mappings of graphic to phonetic values for the initial strings produce the correct phonetic form in all but 58 of the one-syllable words of SOX. The 58 errors include those words with "uncommon" consonant strings, the words where the C or G rule fails, and those words where such specific mapping fails on the consonant strings. The most notable case of the latter set is the graphic string TH which maps into P1 or D1 and the only simple algorithm suggested works for about 90 percent of the cases, and it is to map initial TH into P1.

The mapping for the terminal consonant strings is similar to that for the initial strings; however, it becomes necessary to treat separately the large number of strings that are indicated as being difficult to pronounce by the rules of euphonic combination.[5] For this reason, the terminal consonant strings are first mapped into corresponding phonetic consonant strings in a manner similar to that described for the initial strings.

The resulting phonetic strings are processed by rules of euphonic combination[5] and separated into pronouncable consonant strings and consonants that act as vowel markers to be used in the mapping of the vowel strings. Such a processing provides accurate mapping of all but 155 words in the set of one-syllable words studied, and it identifies the vowel marking consonants to be discussed under vowel string mapping.

THE VOWEL STRING MAPPING

Table 7-1 shows the possible phonetic vowel strings for each of the 19 graphic vowel strings, after uncommon strings have been removed. The only graphic strings that provide a specific phonetic map are:

Graphic	Phonetic
AI	E1419
EY	E1419
OI	OI
OY	OI

For most other cases, it becomes necessary to use the vowel marking consonants identified in the processing of the terminal consonant strings; one of the important exceptions being the initial consonant W which influences pronunciation of the following vowel as evident in the pronunciations of the words AS and WAS.

Since a detailed listing of all the necessary maps to resolve ambiguity in the other cases would be of limited interest, we will here content ourselves with a few examples to show how the vowel marker system operates in the simpler cases. The strings EE and OA, for instance, illustrate the importance of a following R as a marker. In both cases the potential ambiguity is resolved by the presence or absence of a following R.

Table 7-1

GRAPHIC TO PHONETIC MAPPINGS OF VOWEL STRINGS

Graphic String	Phonetic Strings
A	Λ, A1, A2, A4, E14I9, E4E29, O164, O6, O64
AI	E14I9
AU	A2, O64
AY	AI, E14I9
E	E, E14I9, E24, E4E29, IU14, I14, U14
EA	E, E14I9, E24, E4E29, I14, I4E29
EE	I14, I4E29
EI	E14I, E14I9, E2I, E4E29, I14
EY	E14I9
I	E2I, E2IE29, E24, I, I14
IE	E2I, E2IE, E2IE2, I14, I4E29
O	A2, A2U, A3, A34, O, O1, O14U9, O16, O164, O4, O4E29, O6, U14
OA	O14U9, O4E29
OA	OI
OO	U, U14, U4E29
OU	A2U, A2UE29, A3, O14U9, O16, O4E29, O64, U, U1, U14, U4E29
OY	OI
U	A3, A34, IU14, IU4E29, I9U14, U, U14
Y	E2I, E2IE29, I

	Graphic	Phonetic
If EE is followed by R	EE	I4E29
Otherwise	EE	I14
If OA is followed by R	OA	O4E29
Otherwise	OA	O14U9

In such cases R induces the so-called visarga vowel, as has been noted in Reference 8. Thus R can act as a "silent" consonant used only for the purposes of marking vowels just as does the final E, the DG form, the GH of the GHT form, and so forth.

To get a more complete picture of the vowel marking system, consider the graphic vowel E.

If E is followed by RE	E	E4E29
Otherwise if E is followed by R	E	E24
If E is followed by WØ and preceded by L or R	E	U14
Otherwise if E is followed by W	E	IU14
Otherwise if E is followed by a single consonant which is, in turn, followed by E	E	I14
Otherwise	E	E

This rule provides the correct results for all but seven of the words in this set. Note that the final E applies only when there is but a single consonant preceding it, and that the W marker is modified both by a following blank consonant and a preceding L or R. In essence, the markers must be used in conjunction with one another by way of a set of precedence relations, and this may in part be responsible for the general feeling that English orthography does not present a nice means of representing phonetic values.

Overall, the mapping of specific graphic strings into corresponding specific phonetic strings with the necessary use of vowel markers provides the correct result in all but 151 of the one-syllable words of SOX. Thus the total mapping of the consonant strings and vowel strings provides the correct phonetic value in all but 363 of the 5,757 one-syllable words when compared to the phonetic values given by SOX.

SUMMARY

When one considers the mappings from the orthographic forms of the elementary words into their corresponding phonetic forms, as in the Shorter Oxford Dictionary, it becomes apparent that the consonant mappings seem straightforward, but that proper processing of these strings by the rules of euphonic combination identifies the pronouncable consonants and the consonants that function as vowel markers. By such identification of the vowel markers, it is possible to obtain a highly accurate mapping of the graphic vowel strings into their corresponding phonetic values with complexity that does not exceed the mapping of any specific graphic string into a corresponding specific phonetic string. When such a mapping is constructed, some 93 percent of the elementary words are interpreted correctly. The residue consists primarily of obscure forms, important structure words that have a unique spelling (relative to their pronunciation) such as the word ARE, and a set of ambiguous forms that can only be resolved by examining the surrounding context of the word in a given usage, such as the words BOW and HOUSE.

With this analysis we see that (at least for the elementary words) the English orthography is a highly developed phonetic system that provides information about the precise pronunciation of the consonants and vowels and the duration of the vowels and certain consonants. Other work indicates that the necessary stress information is also

available in the graphic forms. No other published phonetic system in use at present can claim to accomplish this without the use of a very large population of phonetic symbols in addition to the necessary incorporation of diacritic markers for indication of duration and stress.

REFERENCES

1. The Shorter Oxford Dictionary on Historical Principles, 3rd ed., revised with addenda, Oxford at the Clarendon Press, 1959
2. J. L. Dolby and H. L. Resnikoff, "On the Structure of Written English Words," Language, 40 (1964)
3. H. L. Resnikoff and J. L. Dolby, "On the Nature of Affixing in Written English," presented at the second annual meeting of the association for Machine Translation and Computational Linguistics, Bloomington (1964)
4. J. L. Dolby and H. L. Resnikoff, "On the Graphic Structure of Word Breaking," presented at the first annual meeting of the Association for Machine Translation and Computational Linguistics, Denver (1963)
5. B. V. Bhimani, "A Multidimensional Model for Automatic Speech Recognition," Final Report of Air Force Contract AF (19)-628-2766 Defense Documentation Center Document Number
6. J. L. Dolby and H. L. Resnikoff, "A Tape Dictionary for Linguistic Experiments," Proceedings of the Fall Joint Computer Conference, Las Vegas (1963)
7. Webster's Third New International Dictionary of the English Language, Springfield, Mass., G. C. Merriam Company, Publishers, 1961
8. B. V. Bhimani, "Visarga Vowels," Journal of the Acoustical Society of America, 35, 1889(A) (1963)

8. COMPUTER STUDY OF TRANSCRIBED ENGLISH PHONETICS: A PROGRESS REPORT

R. P. Mitchell

This report is intended to be a brief summary of an important phase of the computer-oriented research within the field of English phonetics which is now in progress at the LMSC Information Sciences Laboratory. The Laboratory research effort in English phonetics is not restricted to the work reported here; it also includes studies of the techniques of generalized speech spectrum analyzers and the instrumented palate. Nevertheless, the work reported here is basic and indispensable to the total planned program in English phonetics at the Laboratory.

This report is not a detailed "work paper." Rather, it is a summary of background and progress; its purpose is to indicate the nature of the research effort in sufficient detail that its significance can be evaluated. Detailed research results have been reported elsewhere, and are contained in the papers listed in the references.

Of prime importance to an adequate understanding of this research is the role of machine processing of data. Unlike many applications of data-processing techniques in linguistics, the role of the computer here is not primarily the role of "accountant." To be sure, we are handling a fairly large group of data, and these data are growing larger in both volume and complexity, so that accounting services of some sophistication are always needed; but the part which the computer is asked to play is not limited to this role.

The computer is used primarily to compute. Its inputs are not, of course, conventional numeric data as they would be, for example, for an orbit trajectory computation. In this application, the computer inputs are Hollerith characters, and remain alphamerics throughout the processing cycle. For the basic data-generation programs,

the inputs are elements of a subset of English words in conventional spelling form and the outputs are phonetic transcriptions of the inputs. As the research project now stands, what takes place within each processing cycle is the computation of the English word's pronunciation according to a given recognized authority; this computation is efficient, accurate, and requires no external input other than the conventional form of the English word. The processing cycle is not a trivial one of matching inputs with elements in a large memory-core; it is a nontrivial cycle involving a computation of the phonetic transcriptions of the input-word using all the relevant information contained in the graphemic structure of the word.

In addition to the basic data-generation programs, as much of the analysis of the phonetic data as is possible is carried out by various formatting and counting programs. These programs fall largely within the category of "accounting services." They are essential, though unspectacular, services which the computer is asked to perform. At this writing, the subset of English words from which basic phonetic data has been generated is the set of elementary words as defined by Dolby and Resnikoff,[1] which is a set of some 5700 words. The pronunciations of each of these words has been determined for each of five authoritative sources, so that our actual set of data consists of some 50,000 entries. This is a sizeable file to maintain and edit as it now stands, but our immediate plans are to extend our results to the set of words which contain one and only one phonetic vowel according to the transcriptions in the Shorter Oxford Dictionary. This would increase the input-word set to about 8,000 entries, with a corresponding minimum of 90,000 phonetic entries. To process files of alphameric data of this size, without extravagant use of computer time, requires extremely efficient and reliable data-processing techniques.

In this respect, the research on programming techniques carried out elsewhere in the Information Sciences Laboratory has been most helpful in this project. Two of

the basic data-generation programs were originally coded in CHARM, developed by Lois Earl of this Laboratory for the purpose of processing linguistic data. CHARM is easy to learn, easy to use, and efficient for files of medium size. Many of our programs were coded in MPL1, a multipurpose language developed by Roger Stark, also of this Laboratory. MPL1, which uses the XPOP compiler, generates extremely efficient codings for processing files of all sizes. For our formatting and analysis programs, MPL1 was found to be a most helpful tool.

The basic source of data input for this project is the word-list generated by Dolby and Resnikoff;[2] the set of elementary words, defined by the same authors in Reference 1, was chosen as a basic set with which to investigate relationships between orthographic and phonetic representations of English words. The choice was governed partly by the fact that this set is easily defined in terms of orthographic structure, and partly also by the fact that it is a sufficiently large set of lexed items with which to begin the study. B. V. Bhimani, Information Sciences Laboratory consultant and the principal investigator for phonetics research at the Laboratory, initiated this investigation and has carried it through to its present form.

The set of elementary words corresponds approximately to the set of all lexed items of the form (consonant string)(vowel string)(consonant string). In this definition, the set of graphic vowels contains the elements a, e, i, o, u, and y, final e, denoted by the symbol e', is an element of the set of nonvowels, or consonants; and either or both of the consonant strings may be empty, but the vowel string is non-empty. For convenience, in this paper we use parentheses to mark off orthographic elements. Thus, (e), (ball), and (x)(y)(e') are all orthographic elements. Note that there is a difference between the elements (x)(y) and (x)(y)(z); in the former, we mean to indicate consideration of the string "xy" as an integral element, while in the latter

we intend to consider the contingent concatenation of the graphic elements (x) , (y) , and (z) . The set of graphic vowels is denoted by V and the set of graphic nonvowels is designated by C .

In this notation, the set (X) , where X is either C or V , is the set of all elements (x) , (x(w)) , ((x)w) , (xw) , (x(w(z))) , (x((w)z)) , . . . , where x , w , and z are elements of either C or of V , but not both. The work of Dolby and Resnikoff on the graphemic structure of English words is basically the study of written English in terms of sequences of the form (C)(V)(C)(V)(C)...(C) . In other words, a lexed item, except for convenience of reference, is never considered as a distinct unit of the form (CVCVC...C) .

Approximately, then, the study of elementary words is the study of strings of the form (C)(V)(C) , with the study restricted first to those lexed items in the Shorter Oxford Dictionary which can be represented as an element of the set (C)(V)(C) , and second to certain statistical restraints which eliminate obscure words and unusual consonantal combinations. There was, however, a third restraint operating in this investigation – viz., the general purpose of the investigators in attempting to discover relationships between the orthographic structure of words and their possible grammatical properties. To obtain as nearly grammatically homogeneous partitionings of the set (C)(V)(C) as possible, the authors were led to criteria which eliminate a small but important set of lexed items strictly elements of (C)(V)(C) . On the whole, however, the set of elementary words as defined by Dolby and Resnikoff is representative of English (C)(V)(C) words.

The question of whether there exist computable relations between the graphic representation of English words and their sets of pronunciations is a topic of considerable intrinsic and economic interest. It is not necessary here to elaborate on the possible important implications of an affirmative answer to this question. To begin to make

the question a sensible scientific problem and further a meaningful computational problem, there must exist phonetic data in some reasonably tractable form. A major part of linguistic science has always been concerned with various aspects of the problem of representing phonetic data. We thus have the various phonetic alphabets, phoneme studies, the notion of "distinctive features," and so on, as attempts to give a certain structure to an apparently otherwise unstructured mass of data. As nearly as we can determine, and without questioning the scientific merit of these efforts, the kinds of structure which they induce upon the data are not computationally practicable, even when they happen to be computable in principle. It seems reasonable to assume that if the question is meaningful at all, it is meaningful at some lower level of structure among the several which have appeared in the literature. Such a level is exemplified by the pronunciations of words as recorded and transcribed for various dialects of English by different authorities. Clearly, there is no "proper" operational definition of a pronunciation of a word, and just as clearly there exist certain limits within which communication by means of oral articulation of the word is signified. There is an automatic interplay within the communication process of the various aspects of speech production, perception, recognition, and representation, to name but a few of the larger aspects of the process. There do exist transcriptions of speech which attempt to represent the pattern of speech as perceived by the transcriber and communicate the perceived pattern to others (within the limits of the phonetic alphabet used). These transcriptions have inherent limitations, and are clearly not intended to be reproductions of speech in the same sense that tape recordings and wave forms are physical representations of the speech patterns they reproduce.

The value of transcribed phonetics to the question we are considering lies in the fact that the transcriptions constitute "intermediate data" within this question. Not all of the physically significant features of speech production are represented in the

transcription, but enough of the phonetically significant features of the speech pattern are present in quality transcriptions to make the pattern meaningful as data. Our first task, therefore, was to discover whether there exist relations between the graphic structure of lexed items and the phonetic transcriptions of those items as transcribed by recognized authorities. If the answer should be negative in any practical sense, then the problem of relating phonetic and graphic structures must rely for its solution upon analysis of some lower level of phonetic data. On the other hand, if the answer should be affirmative, and if the relations turn out to be practicably computable as well, then the problems of analysis and synthesis are simplified and a valuable intermediate structure is made available.

An authoritative source of phonetic transcriptions of lexed items lay in the various phonetic dictionaries and the conventional dictionaries. Bhimani, using work completed before joining the Lockheed consulting staff, was able to construct an algorithm relating the orthographic structure of elementary words to their corresponding phonetic transcriptions in the Shorter Oxford Dictionary with 93 percent accuracy. This result was reported in Reference 3. The term "93 percent accuracy" means that, on the average, the algorithm will fail to yield the Shorter Oxford phonetic transcription for 7 words out of 100. This result was sufficiently encouraging that he turned his attention to the problem of constructing similar algorithms for other authoritative transcriptions. Using the results for the Shorter Oxford algorithm, he obtained algorithms for the phonetic transcriptions of this set of words according to five other authorities, representing individually and collectively several sets of dialectal variants of spoken English.

The success of this work depended upon a proper understanding of the rules of euphonic combination for English and the manner in which conventional spelling utilizes graphic symbols to indicate vowel duration and stress. A well-known example of this

"marking system" is the ∉ marker. In the word "scrape," for example, the algorithm first splits the word into its graphic CVC form, then looks for markers which (if present) yield a modified pronunciation of the word. Thus "scrape" is split into (scr)(a)(p∉) , ∉ is recognized as a marker in this context with no others present, and the resultant transcription is processed. It is of interest that, to obtain these accurate results via the marking system, the general orthographic domain is not the set (C)(V)(C) as it was for purely orthographic studies. The general orthographic domain for obtaining a proper phonetic image is the union of the sets (C)(V(C)) and ((C)V)(C) , a much larger set. This certainly does not imply, however, that the set of phonetic images is equivalent to the set of allophones. This is because, first, not every individual combination of graphic vowel and graphic consonant needs to be interpreted precisely and, second, not all combinations of phonetic vowel and phonetic consonant need to be interpreted precisely when a mapping is effected from one phonetic source to another.

Upon completion of these algorithms and their implementation to obtain calculated phonetic transcriptions, the results were checked entry by entry against each source, errors were noted and corrected, and the entire corrected output placed on a single file of magnetic tape. The next step was to reformat the data in such a manner that all pronunciations of each word could be displayed at once. This was accomplished in a single MPL1 program which required only 5 minutes of 7094 computer time. A section of the output of this program is shown in Fig. 8-1.

We could now evaluate the calculated transcriptions "across the board" by comparing each of the five sets of pronunciations obtained for each word against the other four. It had already been obvious that each of the five sources possessed its own interpretation of the marking system for the language. How do these interpretations

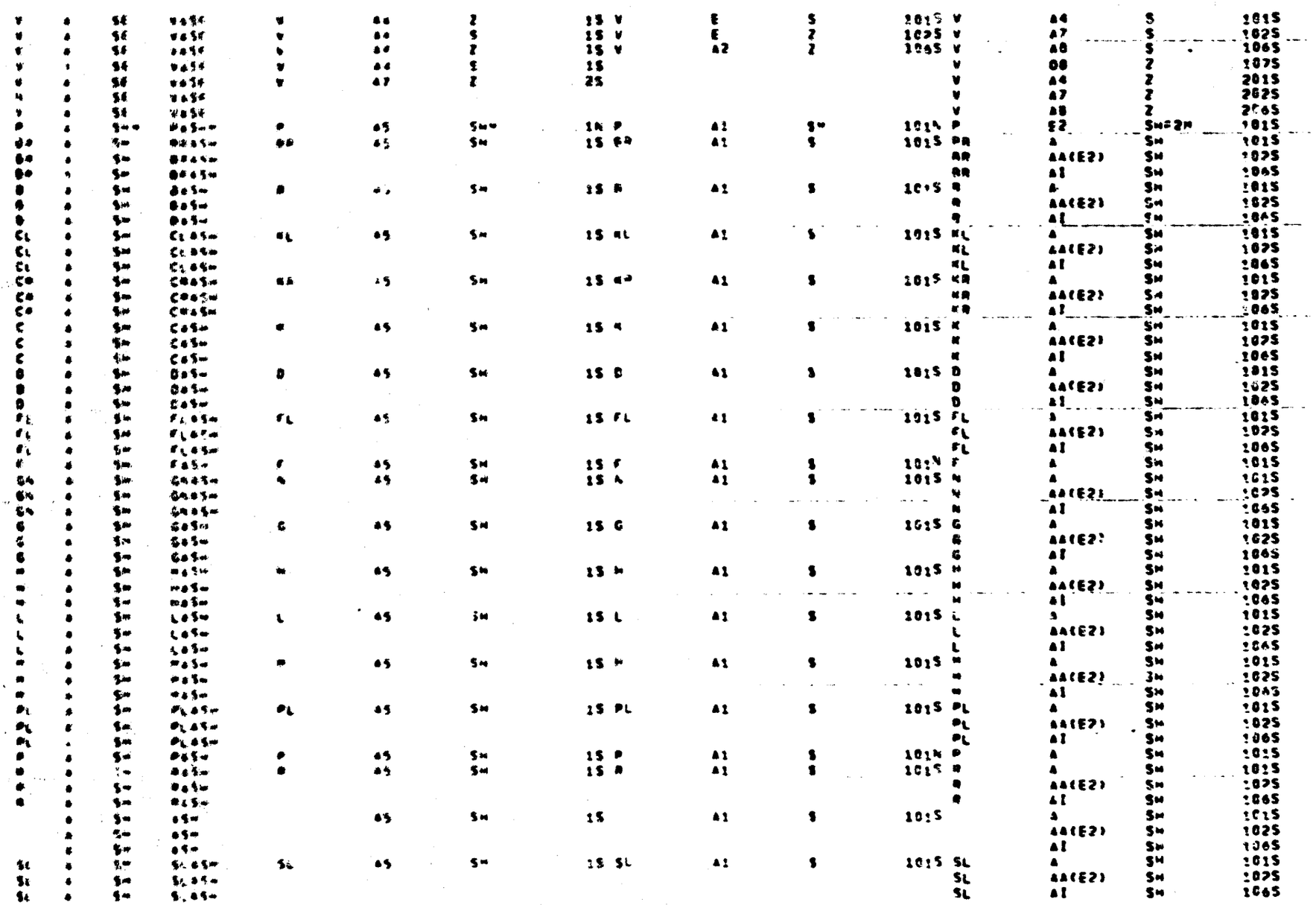

Fig. 8-1 Section of Output of "Format" Program Showing American Dialects

differ and what are the effects of these differences? Are the differences so great that it is impossible to construct an algorithm which calculates all of the given transcriptions at once for each word? These and other related questions concerning the value and consistency of the phonetic data are currently in the process of being answered. In order to provide material for statistical analysis, and also to provide users of the dictionary a key to the status of each phonetic entry with respect to the algorithm, a code was inserted which describes the extent to which any given entry agrees with the algorithm for that word (Fig. 8-2). This phase of the work has been completed and preliminary results were reported in References 4 and 5. A brief summary of these results follows.

Suppose that the observed differences among the authorities studied are caused by "sound change" for the dialects recorded. This hypothesis could be tested with our data by determining what transcription patterns in four of the sources correspond to a given pattern in any single source. The results clearly indicate the following facts: (a) Any algorithm which could be constructed to relate the phonetic transcriptions of one authority to those of any of the other four, would require more rules and a more complex logic than the algorithm which produces the transcription from the orthographic form of the word. (b) The data show that rules of the type which simply substitute phonetic vowels for graphic vowels and phonetic consonants for graphic consonants are gross oversimplifications and lead to erroneous data. The data further show that rules of the type which substitute phonetic vowels in one source for phonetic vowels in another, and phonetic consonants in one source for phonetic consonants in another, are imprecise rules to effect a mapping from one source to another. (c) The data indicate a predictable dependency of vowel values upon surrounding consonant values. Note that these results were obtained for corrected data, and not just the data-generator outputs. This result tends to confirm the existence of functional relationships between

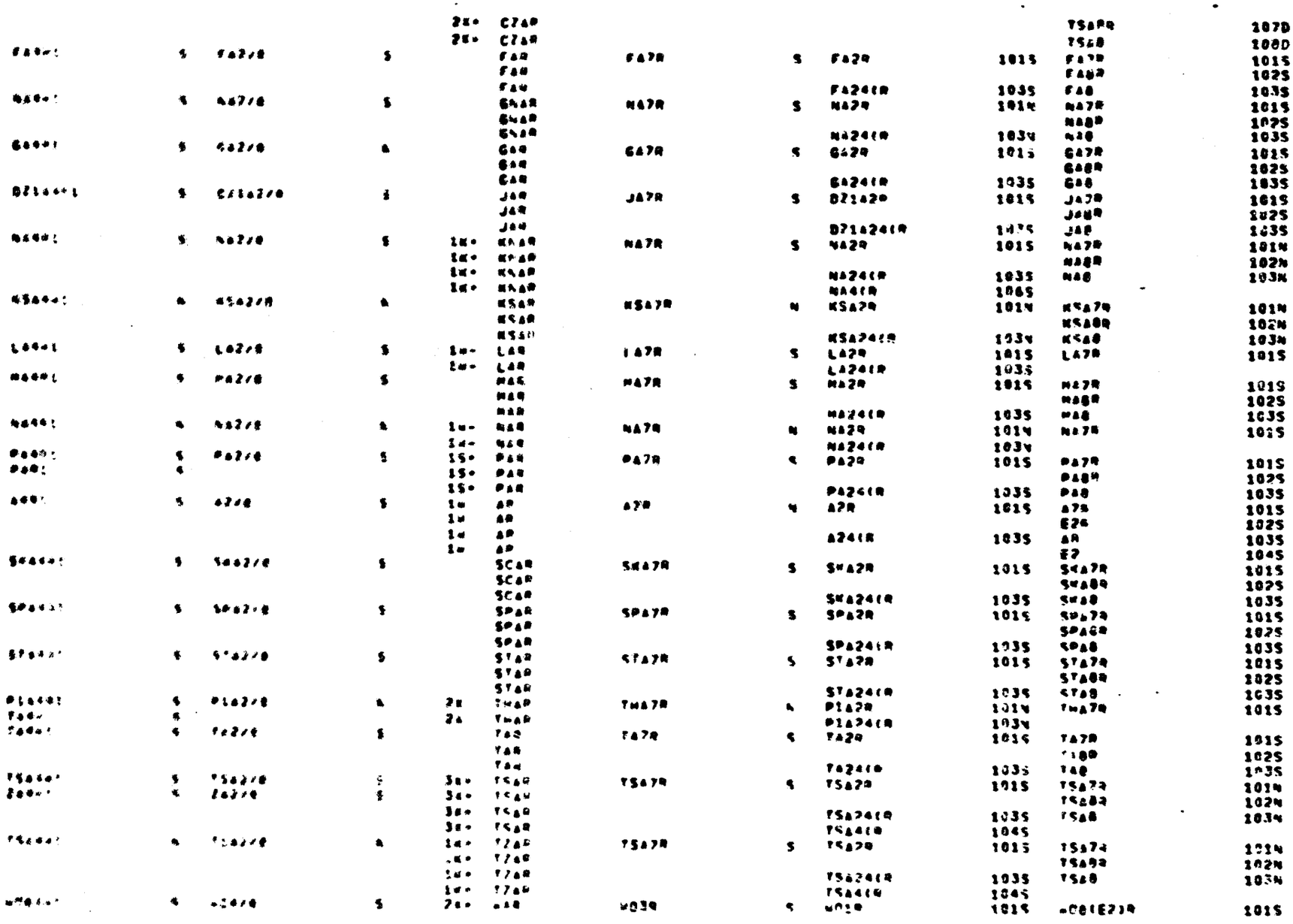

Fig. 8-2 Section of Output of "Display" Program

vowel and consonant values which the marking system for English orthography predicts.

Certainly, a legitimate question which can be, and ought to be, raised is how accurate and consistent are the transcriptions themselves in each of the five sources we have considered. Obviously, we cannot check the accuracy of each transcription with respect to what was perceived when the transcriptions were made and the dictionaries compiled, but we can expect to distill a gross picture of some of the perception and compilation problems which the authorities individually encountered for this set of words. Having available the corrected dictionary, we sorted each source separately on the phonetic fields in terminal rhyme order and examined the result: There exist, in each of the five sources studied, unexpectedly large sets of orthographically distinct words which have the same phonetic transcription.

Two orthographically distinct words with the same transcription are said to be "homonyms," and the set of all such words for a given transcribed pronunciation is called the "homonym set" for that transcription. Our preliminary results show that even removal of "extreme" variants in pronunciation as well as recognized dialect variants in each of the sources, does not significantly affect the high incidence (averaging over 40 percent) of homonyms in this set of data. Quite apart from their number, no great agreement was found among the five authorities regarding the sets of homonyms. Evidently, many factors are at work to produce this result; problems of speech production and perception, possible constraints in transcription techniques, and other factors enter into this phenomenon. Nevertheless, though the extent of homonym sets was surprising to us, the existence of homonyms is predictable in terms of the orthographic marking system.

At this point of our study, we possessed three distinct sets of relations: (a) The set of relations between the orthographic forms and the phonetic transcriptions in each

of five sources; (b) the set of relations between the various phonetic forms for each of the recorded dialects; and (c) the set of relations for the generation of transcriptions in the other four authorities from the Shorter Oxford phonetics. Further, each of these sets is a set of computable relations. It turns out that the segments of phonetic forms in each of these sets of relations are identical. It will be recalled from earlier discussion that these relations are not vowel-for-vowel and consonant-for-consonant relations, but of the form typified by the orthographic domain: ((C)V) and (V(C)) . The fact that the segments are identical for the three sets of relations means that they are independent of the properties of any particular transcription. These segments provide the necessary and sufficient conditions, without having to resolve homonyms, for the definition of the minimum segment of speech perception. In view of this result, the problem of resolving homonyms takes on a new importance. Obviously other information than that which has been used is required to resolve, or minimize, sets of homonyms. Our current efforts are in the direction of determining whether grammatical properties of the words are helpful. It is already apparent, however, that simply listing and comparing possible parts of speech for homonyms is not going to be very helpful. Larger context than the word in isolation may be required for effective resolution of homonyms.

It is important to understand that the existence of calculable speech segments which provide minimum segments of speech perception has been demonstrated entirely without reference to theories of perception and linguistic analysis at a higher level than the rules for euphonic combination, and analysis of phonetic data published in dictionaries. There appears to be some confusion in the minds of linguists who have seen these results and insist that we are studying graphemic-phonemic relationships and that our results [illegible] unique (having been obtained earlier). This is simply not true. First, our level of analysis is much lower than that level of abstraction on which

phonemic analysis proceeds. We have no need for phonemic analysis, and it is meaningless to discuss phonemes at the level of our data. Second, our results are obtained from completely self-contained algorithms. So far as we have been able to determine, no other operational procedure in this field obtains so much from so little. We have shown that the results obtained by our methods do not depend upon any particular transcription, and furthermore the transcriptions we have used are readily available in any library; it is not impossible to reproduce our data and our results. In any event, this is demonstrably not possible for procedures based upon phonemic theories. Third, we have seen no comparable results anywhere else.

REFERENCES

1. J. L. Dolby and H. L. Resnikoff, "On the Structure of Written English Words," Language, Vol. 40, 1964
2. J. L. Dolby and H. L. Resnikoff, "A Tape Dictionary for Linguistic Experiments," Proceedings, Fall Joint Computer Conference, Las Vegas, 1963
3. B. V. Bhimani and J. L. Dolby, "Acoustic Phonetic Transcription of Written English," presented at the 68th meeting of the Acoustic Society of America, Austin, 1964
4. B. V. Bhimani, L. Earl, and R. Mitchell, "Operationally Defined Homonyms of One Syllable in the English Language," presented at the 69th meeting of the Acoustical Society of America, Washington, D.C., 1965
5. B. V. Bhimani and R. Mitchell, "Operational Relation Between the Phonetic Forms of Elementary Words," presented at the 70th meeting of the Acoustic Society of America, St. Louis, 1965

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Lockheed Palo Alto Research Laboratory
Lockheed Missiles & Space Company
Palo Alto, California

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP
N/A

3. REPORT TITLE

Annual Report: Automatic Indexing and Abstracting Part I

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Annual Progress Report

5. AUTHOR(S) (Last name, first name, initial)
Rudin, B. D., Principal investigator
Bhimani, B. V. Earl, L. L. Resnikoff, H. L.
Dolby, J. L. Mitchell, R. P.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO OF REFS
March 1966	195 text pages	Various

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Nonr 4440(00)	
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	M-21-66-1

10. AVAILABILITY/LIMITATION NOTICES The complete report is available in the major Navy technical libraries and can be obtained from the Defense Documentation Center. A few copies are available for distribution by the author.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Department of the Navy Office of Naval Research

13. ABSTRACT The second annual report on automatic indexing and extracting consists of 8 papers summarizing progress in three areas of investigation: (1) Application of English word morphology to automatic indexing and extracting; (2) Use of combined syntactic and entropy selection criteria in automatic indexing; (3) Studies in phonetic English. The first four papers are concerned with the relationship between the part of speech of words and their graphic form. An operational definition of affixes is given, the usefulness of affixes in the automatic determination of parts of speech is discussed, and an algorithm is outlined for determining parts of speech with a dictionary look-up of less than 200 affixes and less than 800 words. The inflection of adjectives is also discussed, anticipating the need for future refinement of the part-of-speech algorithm, which at present identifies 11 part-of-speech categories. For some objectives these categories may be inadequate, necessitating further breakdown, for example adjectives might be further distinguished as relative, comparative, etc. The fifth paper is a progress report on the development of a method for automatic indexing without reference to any pre-prepared dictionary, thesaurus, etc. It shows the current results on five text excerpts. The final three papers are concerned with the relationship between English phonetics and English morphology. One of the papers is concerned with homonyms, which represent a problem area in transformation from phonetic to graphic English. Another discusses a function for mapping written English into spoken English, and the third describes a computerized study of transcribed English phonetics as given by different dictionaries.

DD FORM 1473

UNCLASSIFIED
Security Classification