[Mechanical Translation and Computational Linguistics, vol.9, no.2, June 1966]

Structural Definition of Affixes from Multisyllable Words

by Lois L. Earl,* Lockheed Missiles and Space Company, Palo Alto, California

In a recent paper by H. L. Resnikoff and J. L. Dolby, "The Nature of Affixing in Written English," 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 Dictionary. *Fourteen strong prefixes and twelve strong suffixes and 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. *Thirteen additional weak suffixes, nineteen weak prefixes, seventeen strong prefixes, one strong suffix, and twelve possible suffix-compounding elements were found.*

In this paper, as in the preceding one,[1] the aim 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, though 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 their 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 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 that did not occur in the two-vowel-string data set (used in reference one) will be defined, along with the two-vowel-string affixes that 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): Given a word of the form $C_1V_1C_2V_2C_3V_3$ . . ., if either C_2 or C_3 is an inadmissible consonant string, there is a mandatory syllabic 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 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 as a "suffix possibility." Then the definition for suffix probability and for strong suffix is the same as for prefixes above, in

TABLE 1

A. 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		

B. Admissible Final Consonant Strings of CVC Words not Ending with E

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

* This work was accomplished under the Office of Naval Research and the Lockheed Independent Research Program. The author wishes to thank Dan L. Smith for writing many of the computer programs used in deriving the affixes.

[1] J. L. Dolby and H. L. Resnikoff, "The Nature of Affixing in Written English," *Mechanical Translation,* Vol. 8, Nos. 3, 4 (June and October, 1965), pp. 84-89.

which 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 of two or more consonant strings in defining an affix from a probability was established by Dolby and Resnikoff. This criterion was established heuristically and has been retained here not only for the sake of consistency but also because it was 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 on a mandatory one. Because such probable breaks are not interior to a consonant string, weak prefixes end with a vowel and weak suffixes begin with one. 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 criterion by which an affix possibility becomes an affix is 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, in order 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 it 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 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 we are here 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 that 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 that 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

and so on. 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 set had at least four members. When this test was applied to the five suffixes *a, age, ah, ent,* and *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 2, which reviews the prefixes and suffixes defined by Resnikoff and Dolby, uses the two-vowel-string words as the data set. Table 3 shows the new suffixes defined using four-, five-, six-, and seven-vowel-string words, with the preceding letter strings and occurrence counts that established them as suffixes. Surprisingly, there is only one that can be considered a strong suf-

TABLE 2
AFFIXES FROM TWO-VOWEL-STRING WORDS

PREFIXES		SUFFIXES			
Strong	Weak	Strong	Weak		
ac	a	ful	a	eon	ive
ad	be	land	age	er	o
al	cy	ler	ah	et	ock
con	de	less	al	ic	on
dis	e	let	an	ie	or
en	i	ling	ant	ier	ot
ex	re	lock	ar	ile	ow
in		ly	ard	in	ue
mis		man	at	ine	um
out		ment	ed	ing	ure
sub		ness	ee	ion	us
sun		ward	el	is	
trans			en	ish	
us			ent	ite	

TABLE 3
SUFFIXES FROM MULTISYLLABLE WORDS

Suffix	Preceding Letter String	No. of Occurrences of Suffix Following the 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

TABLE 4
ELEMENTS COMBINING WITH SUFFIXES

Suffix-Compounding Element	Associated Terminal Letter String	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

fix, 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* (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 be obtained. Instead, a number of sequences were observed that 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,* or *ticide,* as in *elliptical, asepticism, didacticism, ascepticize, romanticize,* and *infanticide.* Such interior sequences that meet the occurrence criteria set up for suffixes are listed in Table 4. It is expected that these sequences will have little syntactic meaning but may be helpful in word-hyphenation techniques.

Table 5 shows the prefixes defined using four-, five-, six-, and seven-vowel-string words, with the following letter strings and occurrence counts that 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 com-

monly recognized, but *ana, apo, deni,* and *irre* are not. *(Irre* is no doubt a combination of the recognized prefixes *i* and *re.)* None of the one-syllable weak prefixes *(au, ca, hy, ma, mi, lu, pro, sa, su, vi)* is familiar as a meaningful prefix except for *pro.* Therefore, the next step, in which the part of speech implications of the structurally defined affixes is investigated, will be especially interesting for this group. It is, in fact, in the next steps, 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.

TABLE 5
PREFIXES FROM MULTISYLLABLE WORDS
A. WEAK PREFIXES

Prefix	Following Letter String	No. of Occurrences of Prefix Preceding the 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
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

B. 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
ec	*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

Received December 8, 1965