RC 12193 (#54820) 10/1/86
Computer Science 11 pages

The Role of Etymology and Word Length in English Word Formation

Frank Anshen and Mark Aronoff
SUNY/Stony Brook
Stony Brook, New York 11794

Roy J. Byrd and Judith L. Klavans
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

Abstract: The attachment of affixes to English words is conditioned by both number of syllables and etymological origin. Using a computerized morphological analyzer on a large data base of English words, we show that both etymology and length in syllables effect some affix attachment categorically. For other affixes, the constraints appear to be variable. Our study results in a better understanding of the mechanisms that speakers use to analyze and generate complex words. We expect this knowledge to be useful for improving computerized morphological analysis and generation systems.

The Role of Etymology and Word Length in English Word Formation

This paper deals with the extent to which the attachment of affixes to English words is conditioned by number of syllables and etymological origin. The results to be reported derive from a continuing analysis of English word structure. The authors developed a computer program which analyzes English words into base words and affixes (Byrd(1983), Byrd, et al.(1986)). The data on which this paper is based came about by a several step process using this program and others developed by the authors. First the program was applied to the 6,000 most frequent words found in Kucera and Francis(1967). From this initial list, only morphologically simple content words (i.e., nouns, verbs, and adjectives) were retained. Then a large English word list consisting of nearly a quarter of a million words was run through the original program and the resultant bases were compared with the processed Kucera and Francis list. If a word was analyzed as composed of a base from the Kucera and Francis list plus one of the indicated affixes, a "one" was entered in an appropriate matrix. Then inert bases (those with no derivations marked) were eliminated and three matrices formed. The result was an adjective matrix of 832 bases by 14 affixes, a verb matrix of 1619 bases by 21 affixes, and a noun matrix of 1909 bases by 20 affixes.

For the purposes of this paper, the matrixes were then hand edited according to the following criteria. First, bases were marked for etymology. Bases were marked Germanic if they came into Middle English from Old English or another

TABLE 1
Adjective affixes by ethnicity and syllabicity

SAMPLE	GERMANIC		LATINATE		TOTAL	MONOSYLLABIC		POLYSYLLABIC		TOTAL
%	34.57		65.43		100	29.86		70.14		100
N	242		458		700	209		491		700
AFFIXES	%	N	%	N	SIG	%	N	%	N	SIG
-cy	0.00	0	100.00	59	**	0.00	0	100.00	59	**
-en	92.86	39	7.14	3	**	95.24	40	4.76	2	**
-ity	0.72	1	99.28	137	**	9.42	13	90.58	125	**
-ize	4.50	5	95.50	106	**	0.90	1	99.10	110	**
-ish	63.38	93	31.62	43	**	75.00	102	25.00	34	**
-ness	40.45	199	59.55	293	**	34.76	171	65.24	321	**
in-	0.00	0	100.00	106	**	7.55	8	92.45	98	**
inter-	0.00	0	100.00	38	**	7.89	3	92.11	35	**
non-	7.95	14	92.95	162	**	8.52	15	91.48	161	**
over-	46.80	95	53.20	108	**	51.23	104	48.77	99	**
pre-	8.49	9	91.51	97	**	0.94	1	99.06	105	**
sub-	0.00	0	100.00	72	**	11.11	8	88.89	64	**
un-	34.67	156	65.33	294	NS	29.11	131	70.89	319	NS
under-	52.38	22	47.62	20	*	59.52	25	40.48	17	**

* = sig <= .05, ** = sig <= .01

Germanic language, even if the word had been borrowed into the Germanic language from Latin. An example is *vine* which came into Modern English from Middle English, from the Old French *vigne*, from Latin *vinea* meaning "vine, vineyard". Bases were marked Latinate if they came into Middle English from a Romance language even if they had been originally borrowed from a Germanic language. Also, bases were marked for length in syllables. It was decided to make a binary division between monosyllables and polysyllables. The very small number of words of Germanic origin which are longer than two syllables precluded a finer analysis. This decision limits our analysis to some extent. For instance, we have reported elsewhere (Byrd, et al.(1986)) that there is a major distinction in English between *-ize* and *-ify* based on length in syllables. The suffix *-ify* goes on bases of two syllables or less while *-ize* goes on bases of two syllables or more. Obviously this distinction is lost when all polysyllables are considered as one category. Finally, in order to further clean up the data, three types of bases were eliminated. First are those bases with ambiguous or multiple etymologies. For example, a word like *pile* is said to be derived both from the Latinate term *pilum* "javelin" and from Old High German *pfil* "dart." In cases like these, it would be incorrect to chose etymological origin as ei-

TABLE 2
Verb affixes by ethnicity and syllabicity

	GERMANIC		LATINATE			MONOSYLLABIC		POLYSYLLABIC		
SAMPLE					TOTAL					TOTAL
%	34.85		65.15		100	49.98		51.02		100
N	429		802		1231	603		628		1231
AFFIXES	%	N	%	N	SIG	%	N	%	N	SIG
-able	21.99	137	78.01	486	**	38.68	241	61.32	382	**
-ance	2.33	2	97.67	84	**	11.63	10	88.37	76	**
-ant	1.33	1	98.67	74	**	18.67	14	81.22	61	**
-ed	35.40	143	64.60	261	NS	52.97	214	47.03	190	*
-ee	4.26	4	95.74	90	**	22.34	21	77.66	73	**
-ion	1.43	5	98.57	344	**	16.05	56	83.95	293	**
-ive	0.00	0	100.00	229	**	10.92	25	89.08	204	**
-ure	0.00	0	100.00	22	**	50.00	11	50.00	11	NS
-er	36.60	291	63.40	504	NS	54.34	432	45.66	363	**
-ing	52.95	323	47.05	287	**	69.34	423	30.66	187	**
-ment	10.63	17	89.38	143	**	20.00	32	80.00	128	**
de-	17.78	16	82.22	74	**	68.89	62	31.11	28	**
en-	33.70	31	66.30	61	NS	80.43	74	19.57	18	**
inter-	31.19	34	68.81	75	NS	60.55	66	39.45	43	*
mis-	18.78	37	81.22	160	**	37.56	74	62.44	123	**
over-	44.34	145	55.66	182	**	63.61	208	36.39	119	**
pre-	13.84	53	86.16	330	**	27.94	107	72.06	276	**
re-	26.79	172	73.21	470	**	44.55	286	55.45	356	**
sub-	25.89	29	74.11	83	*	57.14	64	42.86	48	NS
un-	40.85	134	59.15	194	**	63.41	208	36.59	120	**
under-	48.45	94	51.55	100	**	79.38	154	20.62	40	**

* = sig <= .05, ** = sig <= .01

ther Germanic or Latinate. Then we eliminated words formed by zero derivation with a limited range of occurrence, such as *cabinet* used as an adjective. We also eliminated rare bases derived incorrectly from the Kucera and Francis list such as *eth* from *ethics*. Some hand editing was done to eliminate false analyses such as *legal* from *leg+al*, and finally, we eliminated forms which do not occur in Webster's Seventh. The final result was to reduce the adjective matrix to 700 bases (by 14 affixes), the verb matrix to 1231 bases (by 21 affixes) , and the noun matrix to 1436 bases (by 20 affixes).

TABLE 3
Noun affixes by ethnicity and syllabicity

SAMPLE	GERMANIC		LATINATE		TOTAL	MONOSYLLABIC		POLYSYLLABIC		TOTAL
%	33.36		66.64		100	44.71		55.29		100
N	479		957		1436	642		794		1436
AFFIXES	%	N	%	N	SIG	%	N	%	N	SIG
-al	0.45	1	99.55	223	**	21.43	48	78.57	176	**
-an	22.03	13	77.97	46	NS	30.51	18	69.49	41	*
-ary	2.50	1	97.50	39	**	5.00	2	95.00	38	**
-cy	0.00	0	100.00	16	**	0.00	0	100.00	16	**
-ed	47.41	229	52.59	254	**	60.14	290	39.96	193	**
-ery	54.29	57	45.71	48	**	78.10	82	21.90	23	**
-ic	16.42	22	83.58	112	**	29.85	40	70.15	94	**
-ify	28.40	23	71.60	58	NS	55.56	45	44.44	36	*
-ize	0.69	1	99.31	144	**	0.00	0	100.00	145	**
-ous	24.05	19	75.95	60	NS	37.97	30	62.03	49	NS
-ful	52.89	128	47.11	114	**	67.77	164	32.32	78	**
-hood	40.00	14	60.00	21	NS	40.00	14	60.00	21	NS
-ish	58.11	86	41.89	62	**	73.65	109	26.35	39	**
-ism	0.85	1	99.15	116	**	5.13	6	94.87	111	**
-less	49.36	268	50.64	275	**	59.12	321	40.88	222	**
-like	58.46	197	41.54	140	**	67.95	229	32.05	108	**
-ship	14.63	12	85.37	70	**	17.07	14	82.93	68	**
non-	9.01	21	90.99	212	**	15.02	35	84.98	198	**
over-	52.11	185	47.89	170	**	68.17	242	31.83	113	**

* = sig <= .05, ** = sig <= .01

The percentages of occurrence for each affix for bases of Germanic and Latinate origins and for monosyllabic vs. polysyllabic bases in the sample are given in Tables 1-3. At first glance, these tables seem to show that nearly all affixes are sensitive to both etymology and syllabicity. Of the 55 affixes considered, 46 show a preference for either Germanic or Latinate bases (sig<=.05) and 50 are sensitive to syllable length. In other words, of the 110 possible relationships only 14 are not significant at the .05 level. However, as can be seen from Table 4, there is a strong tendency in English for the Germanic vocabulary to be shorter in syllable length than the Latinate vocabulary. Although this phenomenon has often been remarked on, as far as we know, this is the first empirical demonstration of its validity. This means that we will have to do further analysis to untangle these associated factors as conditioning environments for affixation. Nevertheless, even with this complicating fact,

TABLE 4
Ethnic Origin vs. Length in Syllables
for English words by Part-of-Speech
(in percentages)

Part-of-Speech		Monosyllabic	Polysyllabic
Adjective	Germanic	60.74%	39.26%
	Latinate	13.54	86.46
Verb	Germanic	89.74	10.26
	Latinate	27.18	72.82
Noun	Germanic	86.22	13.78
	Latinate	23.93	76.07

we can draw some conclusions from the first breakdown in Tables 1 through 3. A limiting case of an affix's preference for certain types of bases is categorical or near categorical preference for those kinds of bases. Thus *-cy* attaches only to Latinate adjectives while *-en* attaches almost exclusively to Germanic adjectives. If we arbitrarily decide that we will call an affix categorical when it admits five or fewer exceptions, then we can note that certain affixes "categorically" prefer Latinate bases whereas others "categorically" prefer Germanic bases. Among the affixes that attach to adjectives, *in-*, *inter-*, *sub-*, *-cy*, *-ity*, and *-ize*, categorically prefer Latinate bases while *-en* prefers Germanic ones. Among the affixes that attach to verbs, *-ance*, *-ant*, *-ee*, *-ion*, *-ive*, and *-ure* are categorically Latinate while none are categorically Germanic. Among affixes which attach to nouns, *-al*, *-ary*, *-cy*, *-ism*, and *-ize* are categorical in their preference for Latinate bases. Turning to length in syllables, we find that among those affix attaching to adjectives, *in-*, *over- -cy*, and *-ity* categorically prefer polysyllables, while *-en* is categorical for monosyllables. Among the affixes which attach to verbs none are categorical for either monosyllables or polysyllables. Among the affixes which attach to nouns, *-cy* and *-ize* appear only on polysyllables while *-en* appears only on monosyllables. With the exception of *-en* attaching to adjectives, all of the categorical affixes are categorical in their preference for Latinate bases or polysyllabic bases. Not surprisingly then, with the exception of *-en*, all of the affixes with categorical preferences are themselves Latin in origin.

Categorical preference by an affix for Germanic or Latinate words presents some problems. If we accept the claim that modern speakers of English do not have a perfect knowledge of the etymology of their vocabulary (certainly true of the first author who spent many surprising hours looking up etymologies) then we are left with the question of how they can maintain this perfect, or nearly perfect, sort. There

are, for instance, 229 adjectives in our sample formed by adding *-ive* to a verb; all 229 of these verbs are of Latinate origin.

There are three possible explanations for the orderly behavior of English affixation. One possibility is that the vast bulk of the forms containing the affix in question entered English as a single unit, with only sporadic creation of new forms within English. We suspect that this is the case with such forms as *-cy*. A second possibility is that speakers have a notion of the ethnic preferences of affixes and use whatever heuristics are available to them to find suitable bases. Thus *non-* goes overwhelmingly, but not exclusively, with Latinate bases. A glance at such constructions as "non-transformational grammar" argues strongly that the form is productive in modern English. With this productivity it is not surprising that *non-* attaches to some Germanic bases, but it does not seem to attach to bases which are recognizably Germanic. One such set of bases are words with the non-productive Germanic prefixes *a-* (*awake*), *be-* (*believe*), and *for-* (*forgive*). None of these bases negate with *non-*. A third possibility is that speakers of English were more aware of the etymological sources of their vocabulary in Middle English times, when most of the borrowings were much more recent, than they are today. Thus for *-en*, which forms verbs out of adjectives, the only three forms of Latinate origin which take this affix (namely *neaten*, *quieten*, and *laten*) according to the OED were formed in the nineteenth century, while the bulk of the *-en* forms came into the language much earlier.

We now return to the problem of untangling the effects observed in Tables 1 through 3. Fortunately, our sample is large enough so that we can divide it up by etymological origins (i.e., "ethnicity"), and consider the effects of syllable length independent of etymology and visa versa. These results are presented in Tables 5 through 7. The first point to note is that we were correct: there was a confounding of effects in Tables 1-3. A number of affixes which appeared to show sensitivity to both etymological origin and syllable length turn out not to be associated with one of these factors when the other is controlled for. Thus, looking at the adjective affixes, in Table 5, we see that *-ity*, *in-*, *inter-*, and *sub-*, which all appeared to be sensitive to syllable length in Table 1, are not so when etymological origin is controlled for. Similarly, *over-* and *under-* are not sensitive to etymological origins when syllable length is controlled for. Among verb affixes, *-ee* and *mis-* turn out not to be sensitive to syllable length while *-ed* is not sensitive to ethnic origin. Among noun affixes, *-ic* is not sensitive to length in syllables, and *-ary* and *-cy* are not sensitive to etymological origin.

We have now shown that the effects of length in syllables and etymological origins of adjectives on the productivity of affixes are both strong and separable. An affix may be sensitive to both factors, e.g., *-ish* prefers to attach to adjectives which are monosyllabic and it prefers to attach to adjectives which are Germanic. An affix

TABLE 5
Affixes attaching to adjectives

	Significance of syllable length when bases are controlled for ethnicity:				Significance of ethnicity when bases are controlled for syllable length:			
	LATINATE		GERMANIC		MONOSYLLABIC		POLYSYLLABIC	
AFFIX	N	SIG	N	SIG	N	SIG	N	SIG
-cy	59	**P	0	NS	0	NS	59	**L
-en	3	**M	39	**M	40	**G	2	NS
-ity	137	NS	1	NS	13	**L	125	**L
-ize	106	**P	5	**P	1	NS	110	**L
-ish	43	**M	93	**M	102	*G	34	**G
-ness	293	**M	199	NS	171	NS	321	**G
in-	106	NS	0	NS	8	**L	98	**L
inter-	38	NS	0	NS	3	**L	35	**L
non-	162	**P	14	NS	15	*L	161	**L
over-	108	**M	95	**M	104	NS	99	NS
pre-	97	**P	9	**P	1	NS	105	**L
sub-	72	NS	0	NS	8	**L	64	**L
un-	294	NS	156	NS	131	NS	319	NS
under-	20	**M	22	*M	25	NS	17	NS

NS = not sig, * = sig<=.05, ** = sig<=.01
M = affix prefers monosyllabic bases at the indicated significance level
similarly for P = polysyllabic, L = Latinate, G = Germanic

may express a preference for the ethnicity of the adjectives it attaches to but not care about their length in syllables, e.g., *in-*, which prefers categorically to attach to adjective bases of Latinate origin but does not care about the length in syllables of the base. An affix may have a preference for length in syllables but be indifferent to etymological origin, e.g., *over-* which prefers to attach to monosyllabic adjective bases but doesn't care about the etymological origin. Finally, an affix may be completely promiscuous, attaching to bases with no thought of either ethnicity or length in syllables, e.g., *un-* which attaches to adjective bases of all descriptions.

We turn now to cases where an affix is sensitive to length in syllables for bases of one etymological origin but not for bases of the other. Limiting our search to affixes which attach to at least twenty Latinate bases and at least twenty Germanic bases, we find sixteen cases in which length in syllables is significant for one group but not the other. In all sixteen of these cases, the affix is sensitive to syllable length for the Latinate vocabulary but not the Germanic vocabulary. We might offer the possible explanation that affixes just are not sensitive to length in syllables in the

TABLE 6
Affixes attaching to verbs

	Significance of syllable length when bases are controlled for ethnicity:				Significance of ethnicity when bases are controlled for syllable length:			
	LATINATE		GERMANIC		MONOSYLLABIC		POLYSYLLABIC	
AFFIX	N	SIG	N	SIG	N	SIG	N	SIG
-able	486	*P	137	NS	241	**L	382	**L
-ance	84	**P	2	NS	10	**L	76	*L
-ant	74	*P	1	NS	14	**L	61	*L
-ed	261	NS	143	NS	214	NS	190	NS
-ee	90	NS	4	NS	21	**L	73	*L
-ion	344	**P	5	**P	56	**L	293	**L
-ive	229	**P	0	NS	25	**L	204	**L
-ure	22	*M	0	NS	11	**L	11	NS
-er	504	**M	291	NS	432	**L	363	*G
-ing	287	**M	323	**M	423	**G	187	**G
-ment	143	**P	17	**P	32	**L	128	NS
de-	74	**M	16	NS	62	**L	28	NS
en-	61	**M	31	NS	74	**L	18	NS
inter-	75	**M	34	NS	66	**L	43	NS
mis-	160	NS	37	NS	74	**L	123	NS
over-	182	**M	145	NS	208	NS	119	**G
pre-	330	**P	53	NS	107	**L	276	**L
re-	470	NS	172	NS	286	**L	356	**L
sub-	83	**M	29	NS	64	**L	48	NS
un-	194	**M	134	NS	208	*L	120	**G
under-	100	**M	94	*M	154	NS	40	NS

NS = not sig, * = sig<=.05, ** = sig<=.01
M = affix prefers monosyllabic bases at the indicated significance level
similarly for P = polysyllabic, L = Latinate, G = Germanic

Germanic portion of the vocabulary. Or, we might appeal to the fact that there are just fewer Germanic words in our sample and we thus are less likely to find statistically significant relationships (the ratio of Latinate vocabulary to Germanic is about 2:1 for all three parts of speech). However, two facts argue against these explanations. One is the existence of eight affixes which are sensitive to length in syllables for Germanic bases (and which do attach to more than twenty Germanic bases). Second is the fact that there are several affixes among the 16 mentioned above which attach to a large number of Germanic bases (for instance *-less*, which attaches to 268 Germanic nouns). Generally, affixes are more likely to be sensitive to length in syl-

TABLE 7
Affixes attaching to nouns

	Significance of syllable length when bases are controlled for ethnicity:				Significance of ethnicity when bases are controlled for syllable length:			
	LATINATE		GERMANIC		MONOSYLLABIC		POLYSYLLABIC	
AFFIX	N	SIG	N	SIG	N	SIG	N	SIG
-al	223	NS	1	*P	48	**L	176	**L
-an	46	*P	13	NS	18	NS	41	*L
-ary	39	**P	1	NS	2	NS	38	NS
-cy	16	*P	0	NS	0	NS	16	NS
-ed	254	**M	229	NS	290	*G	193	**G
-ery	48	**M	57	NS	82	NS	23	*G
-ic	112	NS	22	NS	40	*L	94	NS
-ify	58	**M	23	NS	45	**L	36	NS
-ize	144	**P	1	*P	0	NS	145	**L
-ous	60	NS	19	NS	30	*L	49	NS
-ful	114	**M	128	NS	164	NS	78	**G
-hood	21	NS	14	**P	14	NS	21	**G
-ish	62	**M	86	NS	109	NS	39	**G
-ism	116	**P	1	*P	6	**L	111	**L
-less	275	**M	268	NS	321	**G	222	**G
-like	140	**M	197	NS	229	**G	108	**G
-ship	70	*P	12	**P	14	NS	68	NS
non-	212	**P	21	**P	35	**L	198	**L
over-	170	**M	185	**M	242	*G	113	*G

NS = not sig, * = sig<=.05, ** = sig<=.01
M = affix prefers monosyllabic bases at the indicated significance level
similarly for P = polysyllabic, L = Latinate, G = Germanic

lables among the Latinate vocabulary of English than in the Germanic section. Specifically, at least in our sample, any affix which is sensitive to syllable length among the Germanic portion of the vocabulary will also be so among the Latinate portion, if it attaches to at least twenty Latinate bases.

Now consider those affixes which meet the criterion of attaching to at least twenty monosyllables and twenty polysyllables and which show an etymological preference in only one of the two cases. There are fourteen such affixes and, unlike the previous case, they may show their preference in either of the two groups. For eight of the affixes (*de-*, *inter-*, *mis-*, *sub-*, *-ic*, *-ify*, *-ment*, and *-ous*), there is an etymological preference among the monosyllables only. For the other six (*a-*,

un-, *-ery*, *-ful*, *-ish*, and *-ness*) the preference shows up only among the polysyllables. However, again there is a regularity: in all of the cases where the preference is expressed only among the monosyllables, the preference is for Latinate bases. In all of the cases where the preference is expressed only among the polysyllables, the preference is for Germanic bases. We may be able to make some sense out of these facts if we rephrase "preference for some bases" into "bias against other bases" and note that in all eight of the cases of bias against Germanic bases, the affix in question is Latinate in origin and in five of the six cases of bias against Latinate bases, the affix in question is Germanic in origin. The exception is *-ery*. In our discussion of categorical preferences in etymological origin, we suggested that speakers, without a perfect knowledge of etymology, might still have some heuristics to help them sort words by etymology. We suggested that the occurrence of certain affixes on particular words might be one such heuristic. Here we might suggest that, given the facts in Table 4 along with the facts just noted, length in syllables may well be another such heuristic, monosyllabicity suggesting Germanic origin and polysyllabicity suggesting Latinate origin. That is, even though we have shown that length in syllables may act as an independent factor in determining the productivity of certain affixes, it may also be a factor in the classification of some bases by etymological origin. Of course, this heuristic is not fool-proof, as evidenced by the cases of *neaten*, *quieten*, and *laten*, mentioned above, where *en* attaches to short Latinate words rather than short Germanic words. In addition to word length, there may be other factors which correlate to some degree with etymological origin.

We have shown that some properties of words, including specifically etymological origin and length in syllables, can categorically condition the attachment of affixes. We have also shown that these same properties can have non-categorical influence on the attachment of affixes. The categorical cases can be expressed in the various rule formalisms familiar to linguists and used by computational linguistics to build word recognition and generation systems. However, the non-categorical cases (of the form "there is a tendency for such-and-such an affix to attach to such-and-such a base") will require different notational devices and computational mechanisms. It is not clear to us at this time if these facts may best be stated in the variable rule formalisms such as proposed by Labov(1969), or if they require new notational devices. We plan to exploit our improved understanding of speakers' intuitive knowledge of constraints on affixation in the creation of more accurate morphological analysis and generation systems.

Bibliography

Byrd, R. J. (1983) "Word formation in natural language processing systems," *Proceedings of IJCAI-VIII*, 704-706.

Byrd, R. J., J. L. Klavans, M. Aronoff, and F. Anshen (1986) "Computer Methods for Morphological Analysis," *Proceedings of the Association for Computational Linguistics*, pp. 120-127.

Kucera, H. and W. N. Francis (1967) *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island.

Labov, W. (1969) "Contraction, Deletion, and Inherent Variability of the English Copula," in *Language* 45:4:715-762.

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