Željko Bujas

An Analysis of the Etymological Makeup of the English Core Vocabulary

INTRODUCTION

0.1. Etymological composition of the vocabulary of any European language is an intriguing aspect of that language's growth and of the cultural mechanisms involved. This is particularly relevant for English with its readily perceptible Germanic: Romance duality of vocabulary, open to erroneous popular interpretations. Various aspects of the etymological makeup of English have been the scope of four of my articles (Bujas, 1968a, 1968b, 1971, 1792). This effort is an attempt to refine my earlier analyses of correlation between vocabulary frequency levels and their etymological proportions.

SOME ANALYSES TO DATE

1.1. Curiosity about the etymological makeup of English is of long standing, beginning with the naive try by the English philologist and theologist G. Hickes of the 17th century. Analyzing (probably in his *Institutiones Grammaticae Anglo-Saxonicae et Maeso-Gothicae,* 1689) the text of Our Father (= 67 words), he was able to conclude that nine-tenths of the English vocabulary is of domestic, that is Anglo-Saxon, origin.

1.2. A number of 19th-century analysts engaged in similar endeavors, though on progressively larger texts. The crowning effort was by the American philologist and diplomat G. P. Marsh, who (in his *Lectures on the English Language,* New York, 1859) analyzed extensive samples from the Bible and a score of both classic and contemporary English and American authors. His conclusion was that in most of these texts

the percentage of words of Anglo-Saxon origin was between 80 and 95, exceptionally (as in Gibbon) falling off to 70%.[1]

1.2.1. Marsh is important in that he makes a clear distinction between analyzing the etymological makeup of a text *dynamically* (i. e. counting all repetitions of a word in the text) and *statically* (i. e. counting words only once, as dictionary entries).[2] He was thus able to point out the telling contrasts between the low share of Anglo-Saxon words in a writer's static vocabulary and their much higher proportions in the actual distribution in the text (in Milton, 33% contrasted with 80—90%; in Shakespeare, close to 60% oposed to 88—91%; in the Bible, 60% as against 90—96%).

1.3. Marsh's data are used (as a rule without acknowledgement) by a number of authors on the growth of English, between the 1860's and the present day (cf. n. 25 in Bujas, 1968 a: 80—91). Unfortunately, most of them took the philologist's approach, stressing (or exclusively quoting) the *static* proportions, which showed the English vocabulary to be made up of between 40 and 70 per cent of words of non-Anglo-Saxon (predominantly Romance) origin. Also, the entire § 40 *(Proportion of the elements)* in the introduction to *Webster's New International Dictionary of the English Language* was based on Marsh's figures and was present, unchanged, in all editions of this authoritative work betwen 1864 and 1961.

1.3.1. Seeping down, as it were, to another level of application — the textbooks and handbooks of rhetoric, with their combination (or rather jumble) or rhetoric, stylistics and prescriptive grammar — these data led to such arbitrary, impressionistic statements about the etymological composition of the English vocabulary as "about one-half of all English words are of foreign (mostly French) origin". This persistent, oversimplified popular idea about the English vocabulary is. unfortunately, still firmly established among the central ideas cherished by the average teacher of English about her or his subject.

---

[1] Marsh was perceptive enough to note the effect of style on the etymological proportions in the vocabulary, illustrating it on appropriate segments of literary text. His well-chosen examples were by the same author, Washington Irving, who used only 12% of non-Anglo-Saxon words in the semi-humorous *Stout Gentleman*, from *Bracebridge Hall*, while employing as many as 38% of such words in the solemn *Westminster Abbey*, from *The Sketch Book*.

[2] The current technical description of the dynamic approach would be "counting items as *word-tokens*", of static "counting items as *word-types*". Thus, the sentence *We are what we are* is composed of 5 tokens, but only 3 types (*we, are, what*), with two of them including 2 tokens each (or having a frequency of 2).

1.4. It was only in this century, beginning in the 1920's — the decade of great word counts — that a number of researchers tackled the problem of etymological composition of the English vocabulary, by meeting most of the requirements indispensable for a vigorus and realistic analysis. These include: a dynamic count of the text analyzed (with repetitions of the words as they occur in the text), differentiation of stylistic levels and taking into account the parameters of frequency.

Without going into particulars described elsewhere (Bujas, 1968 a: 83—98), I will only point out that the following major methods have been used:

a) running-word counts, with differentiation of stylistic levels

b) word-count lists, without frequency values

c) word-count lists, with static (total-sum) frequency values.

d) word-count lists, with dynamic (graded) frequency values.

1.5. The first of these approaches reveals best the correlation of a text's etymological makeup with its stylistic level. A project in 1964—1967, headed by me established the following interesting distributions (in %) on a 37,000-word corpus:

TABLE 1

Stylistic Level	AS	ON	Other Germ.	*Total Germ.*	Fr.	Lat.	Other Rom.	*Total Rom.*	Gk.	Other
Dialogue	85.1	1.5	1.6	*88.2*	5.8	5.0	0.3	*11.1*	0.3	0.4
Fiction	73.5	2.6	3.5	*79.6*	12.3	6.4	0.1	*18.6*	0.8	1.0
Newspapers	66.4	0.7	0.4	*67.5*	22.0	9.5	0.3	*31.8*	0.3	0.4
Scientific & Technical	58.4	1.6	1.7	*61.7*	12.1	23.3	0.7	*36.1*	2.1	0.1
Average	70.9	1.6	1.8	*74.3*	13.1	11.1	0.4	*24.4*	0.9	0.5

The table speaks for itself, unambiguosly revealing the reversely proportional ratios of AS and Fr/Lat etymologies, varying with levels of style.

1.6. Much drudgery is involved in the methodology just described, with hundreds of time-consuming and repetitive dictionary lookups required. Analyzing in this manner any corpus larger than some 50,000 running words is practically

unfeasible. A way around this was made possible in the 1920's with the arrival of first large word counts. Using word-count rank lists, one could go through, say, only, the top 10,000 words of English (based on a representative corpus of up to a couple of million running words), and be able to make statements that were statistically highly valid for *any average* English text or, by extension, for the English language as a whole.

Unfortunately, since most analysts of the etymological composition of English in those days were classicists or teaching methods specialists, primarily interested in the challenge of Latin and Greek words in English, they largely utilized the word-count rank lists as static vocabulary lists, ignoring the frequency values attached. Even so, the cumulative (rounded-off) results of five such analyses, carried out between 1922 and 1925 (cf. Bujas, 1968 a: 84—87), reveal a clear and intriguing correlation between frequency levels and the proportions of words of AS, French/Latin and Greek etymologies:

TABLE 2

Frequency Level	% of	AS	Fr/Lat	Gk
first 1,000		62	33	0.1
first 10,000		30—50	45—53	5—7
second 10,000 (10,001—20,000)		30	50	14
first 10,000 (17,000—30,000)		25	45	18

1.7. A step in the right direction is, no doubt, when a word-count rank list is used *with* the frequency values for each list item being etymologically classified. In this fashion, the rank list is used as a short cut to a dynamic analysis of texts which are viewed as composed of running entries rather than static vocabulary items. Only two major analyses, also in the 1920's, were based on this approach — with one of them producing less than reliable results, due to the analyst's misinterpretation of frequency values in the word count used (Bujas, 1968 a: 85—86). The other analysis was based on a relatively restricted word count (of a 100,000-word corpus with ten stylistic levels), but its results were close to those obtained by me (cf. Table 1) on actual text (at Fiction style level):

AS	74.4%	Lat/Fr.	19.7%	Other	2.1%
ON	2.2%	Gk.	1.6%		

1.8. Though the analysis just described did make use of frequency values to achieve a short cut to what this article (cf. 1.2.) has termed the dynamic vocabulary analysis of a corpus, the manner in which rank-list frequency values were used was in fact static. The frequency values were utilized only at their final (total sum) level, instead of being presented as sums of occurrence at graded levels of frequency for the corpus as a whole. To be sure, meeting this requirement with an already large corpus is a tall order for any manual procedure. This brings us to the first analysis of etymological composition of the English vocabulary to have used a computer for the purpose. It is A. Hood Roberts' impressive effort, presented in his book *A Statistical Analysis of American English* (1965).

1.8.1. In spite of its general title, the book in fact contains 21 quantitative analysis of a variety of aspects of the phonological structure of English, preceded by a somewhat lonely analysis of the "etymological composition of English according to proximate sources by thousands of frequency" (Roberts, 1965: 34). Based on Horn's word count[3] — of over 5,000,000 running words of private and business corespondence — Roberts' results necessarily reflect the limitations of this corpus. Nevertheless, thanks to this use of the computer, Roberts was able to reveal — more clearly, and incomparably more precisely, than any of his predecessors — the degree in which frequency levels affect the etymological composition of English.

I will present his findings in a much simplified and contracted[4] tabular survey:

[4] Roberts distinguishes as many as 104 various etymologies, including 66 hybrid and open-alternative types whose values were added (after wearying manual recomputation) to those of single etymologies. "Probable" etymologies were likewise merged with established etymologies. The category "other" was introduced to cover 13 low-frequency etymologies left over after these procedures. They were (in descending order): German, Arbitrary, Portuguese, Gypsy, Amerindian, Hindi, Arabic, Swedish, Walloon, Persian, Japanese, Chinese, Hawaiian and Flemish.

[3] E. Horn, *A Basic Writing Vocabulary*, Iowa City, 1926. For a detailed description of Horn's corpus and procedures see Roberts, 1965: 11—12.

TABLE 3

RELATIVE Etymological Proportions of 10,000 Items of Horn's List
(By thousands of frequency and as totals of word TOKENS)

	first 1,000	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	all 10,000
AS	83.4	35.1	31.1	29.5	27.7	29.7	25.3	29.0	28.8	28.5	78.3
French	11.8	48.0	48.4	47.1	48.4	44.0	47.8	43.7	43.5	44.7	15.3
Latin	1.9	11.7	15.5	18.6	17.3	20.4	18.4	18.8	18.4	18.9	3.2
ON	2.6	3.1	1.7	1.7	2.3	1.9	3.0	3.1	2.8	2.2	2.6
Top four etymologies	*99.7*	*97.9*	*96.7*	*96.9*	*95.7*	*96.0*	*94.5*	*94.6*	*93.5*	*94.3*	*99.4*
Unknown	0.09	0.55	0.53	0.93	0.81	1.26	1.19	1.31	1.83	1.17	0.15
Dutch	0.03	0.29	0.77	0.26	0.97	0.77	0.97	1.40	1.66	1.54	0.08
East Fries.	0.08	—	—	—	0.09	—	0.09	—	0.05	—	0.08
Low German	0.01	0.37	0.65	0.26	0.42	0.59	1.12	0.67	0.46	0.77	0.06
Greek	0.02	0.36	0.09	0.51	0.76	0.21	0.47	0.85	0.68	0.82	0.05
Italian	0.005	0.27	0.56	0.57	0.60	0.43	0.63	0.57	0.38	0.34	0.05
Celtic	0.025	—	0.18	0.06	0.28	0.06	0.10	0.18	0.19	0.08	0.03
Spanish	—	0.11	0.30	0.12	0.10	0.45	0.55	0.15	0.50	0.85	0.02
Other	—	0.17	0.22	0.48	0.27	0.38	0.37	0.32	0.73	0.28	0.024

1.9. Principal conculsions, leaping to the eye, are as follows:

a) In its real lexical distribution[5], English is a typical Germanic language. A grand total of 78.3% of Anglo-Saxon elements (growing to 81.2% when the shares of all other Germanic elements are added) is opposed by only 15.3% of French (or 18.6% of all Romance) words.

b) The total share of all other (non-Germanic, non--Romance and non-European) elements, with unknown etymologies and the arbitrarily formed words thrown in, account for a mere 0.24% of the text.

c) The four top etymologies (Anglo-Saxon, French, Latin and Old Norse) account for as much as 99.47% of the text — all remaining etmologies (21) for only 0.53%.

[5] That is to say, in a continuous text with word repetitions counted, as simulated by Horn's list in Roberts' analyses (cf. 1.6. and Bujas, 1978b: 129—132).

d) The share of Latin nearly doubles betwen the second and the fourth thousand (from 11.7% to 18.6%), continuing remarkably stable through the tenth thousand.

e) Old Norse element is similarly stable over all ten thousands of frequency.

1.9.1. The most intriguing observation to be made, however, is the decisive influence of frequency upon the etymological distributions in the English vocabulary. This is clearly seen in the fact that the distribution for the first thousand of frequency is still astonishingly close to that for all the ten thousand together (Anglo-Saxon 83.4—78.3%, French 11.8——15.3%, Latin 1.9—3.2%, Old Norse 2.6—2.6% /!/). And this is so in spite of the fact that each thousand of frequency between second and tenth has an equally stable *inverse* ratio of Anglo-Saxon to French (AS 25—35%, French 44—48%), with the share of Latin practically doubling 12—20%). All this is due to the effect of the mass of the first thousand words of English which, as we know (cf. 2. 4. 2. 2.), account for over two-thirds of any average English text. Thus, the cumulative etymological composition of all the nine thousand of lower frequency — however different each of them from the first thousand — cannot essentialy affect the average etymological proportion for the ten thousand as a whole.

A CURRENT CONTRIBUTION

2.1. For all practical purposes, Roberts' analysis of the etymological makeup of the English lexicon may be considered definitive. Still, the stylistic (and some other)[6] limitations of Horn's list make desirable a similar future effort based on a more representative corpus. It is unfortunate that in his search for an acceptable corpus he had to miss by a few years the much more representative, *and* computer-processed, Brown Corpus (cf. Kučera & Francis, 1971).[7] Though smaller in volume

[6] Such as the highly disproportionate share of certain items typical for the vocabulary of correspondence, notably the personal pronoun I that tops Horn's list with 13.9% of all occurrences, while coming only 20th (with a mere 0.51% of the mass) in the much more reliable and representative Brown Corpus. Also the top item in Brown Corpus the accounts for only 6.9% of the entire corpus mass.

[7] A more recent computer-processed corpus has appeared. It is the *AHI* (for *American Heritage Intermediate*) *Corpus*, presented in *Word Frequency Book* (cf. Carrol et al., 1971). But the AHI Corpus — with its 5,088,721 tokens and 86,741 types — was compiled to investigate

(1,014,232 running words), it covers a wide span of fifteen stylistic levels, literally from humor to theology. It also ensures a very favorable degree of dispersal through its five hundred 2,000-word samples, as well as temporal and linguistic uniformity secured by incorporating only texts published in the United States in 1961. Finally, while Horn's list suplies only 10,000 most frequent items,[8] Brown Corpus offers its total list of 50,406 items. Therefore, analyzing the etymological composition of this corpus would take us far beyond the first ten thousand of frequency and would, no doubt, reward our effort with new intriguing insights into the effect of higher frequencies on the etymological proportions of English (with Latin, and later Greek, probably taking the top place, and Anglo-Saxon ending fourth, after French).

2.2. Needless to say, an analysis of this magnitude would have to be computer-supported, because no amount of manual processing could quantify and tabulate the data with sufficient reliability, precision and versatility. It is to be hoped that some such project can be started in the not too distant future. At this point, I would like to present the results of a much more modest personal effort — *a manual analysis of the etymological composition of the first thousand words of Brown Corpus (by hundreds of frequency).* Though restricted in scope, to what may be termed the Core Vocabulary of English, its refinement of frequency levels and possibility of comparing its findings with the parallel range of Horn's list make it worthwhile.

2.3. My procedure followed those of Roberts in all the main points, such as the exclusion of proper names, homograph recomputation[9] and acceptance of his choice of *Webster's New*

---

and meet the vocabulary needs of American secondary-school students. As a results, it is more limited stylistically and therefore less generally representative of English as a whole than the smaller Brown Corpus.

[8] Both Horn's and Brown Corpus lists are not made up of straight dicitionary entries, but of *word types*, or graphically different words (*heterographs* would be a better term). Thus, for instance, *arrive, arrives, arriving,* and *arrived* are listed as four separate items, though clearly belonging to one dictionary word/entry.

[9] Since homography had not been resolved in the Brown Corpus Rank List, I used A. Hood Roberts' recomputations of Horn's list items. I was able to do this thanks to his kind loan of his copy of the latter list annotated for etymological origin and homograph proportions. Roberts had made use of I. Lorge's and E. L. Thorndike's *Semantic Count* (1938) for his computations which I refined in a few cases (also adding a couple of homographs missed) by reference to Lorge's (1949) and

World Dictionary as the principal authority on etymologies The only important point where I felt a different approach to be more warranted was the use of the ultimate rather than the immediate etyma, with the only exception (for obvious reasons) of the very numerous French-through-Latin words which were recorded as French. This exception was, I believe, an acceptable deviation from the principle and was more than offset by what was gained through recording such words as *problem* (OF<Lat<Gk) or *economic* (L<Gk) as being of Greek rather than French or Latin etymology. Finally, in addition to proper names, a number of special symbols (for instance **F to indicate any formula occurring in the text) and individual letters encountered in the Brown Corpus Rank List were also excluded. Numbers, both as digits and fully written forms, were not excluded.

2.4. The findings of my analysis are presented in the following tables:

Table 4, of *absolute* etymological proportions for the first 1,000 items of Brown Corpus (by hundreds of frequency and as totals of *tokens)*

Table 5, a contracted verson of Table 4

Table 6, of *relative* etymological proportions, also as *tokens,* of the same items as in Table 4

Table 7, a contracted version of Table 6

Table 8, of *absolute* etymological proportions of the same items as in Table 4 but as totals of *types*

Table 9, a contracted version of Table 8

Where necessary, the tables will be accompanied by comment.

M. West's (1953) semantic counts. I will use the item *found* to illustrate my transfer of Roberts' homograph recomputations to Brown Corpus frequency values. The item's total occurrence of 5,485 in Horn's list was split by Roberts into 4,975 for Anglo-Saxon (irregular forms of *to find)* and 501 for French *(to found).* I converted this to relative proportions of 91.7% : 9.3%, and applied these to 536, the frequency of the same item in Brown Corpus, which was then split into 486 for the Anglo--Saxon and 50 for the French item.

TABLE 4

ABSOLUTE Etymological Proportions of the First 1,000 Items of Brown Corpus
(By hundreds of frequency and as totals of word TOKENS)

	first hundred	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	all 1,000
Anglo-Saxon	471,895	45,602	23,358	13,622	8,868	8,347	6,160	4,478	3,903	3,656	589,889
French	—	9,124	7,535	8,097	8,033	4,755	5,639	4,826	4,897	5,199	58,105
Latin	—	2,206	2,247	2,272	1,685	3,232	1,878	1,680	1,111	1,348	17,659
Old Norse	8,150	3,260	1,156	829	764	888	533	408	541	103	16,632
Greek	—	1,063	1,332	1,556	1,036	530	304	846	981	333	7,981
Top five etymologies	*480,045*	*61,255*	*35,628*	*26,376*	*20,386*	*17,752*	*14,514*	*12,238*	*11,433*	*10,639*	*690,266*
Common Germanic[10]	—	—	390	—	—	—	—	130	125	102	747
Frankish	—	464	—	—	—	—	—	—	—	—	464
East Friesian	—	—	—	242	—	—	143	—	—	—	385
Italian	—	—	—	—	204	—	—	—	—	107	311
Old High German	—	—	—	—	—	—	143	—	—	110	253
Middle Low German	—	—	—	—	—	—	—	143	—	110	253
Celtic	—	—	—	238	—	—	—	—	—	—	238
Unknown	—	—	360	—	—	—	73	269	—	—	702
Total analyzed	480,045	61,719	36,378	26,856	20,590	17,752	14,873	12,780	11,558	11,068	693,619
Excluded	1,010	721	362	—	400	—	447	802	718	—	4,460
GRAND TOTAL	481,055	62,440	36,740	26,856	20,990	17,752	15,320	13,582	12,276	11,068	698,079
GRAND CUMULATIVE TOTAL	481,055	543,495	580,235	607,091	628,081	645,833	661,153	674,737	687,011	698,079	698,079

[10] Words of common Germanic (Teutonic) origin that have reached Modern English through some other language.

TABLE 5

Contracted Version of Table 4

Level of frequency	Total Germanic	Total Romance	Other	Unknown	Total analyzed
1st hundred	480,045	—	—	—	480,045
2nd	49,326	11,330	1,063	—	61,719
3rd	24,904	9,782	1,332	360	36,378
4th	14,693	10,369	,1794	—	26,856
5th	9,632	9,922	1,036	—	20,590
6th	9,235	7,987	530	—	17,752
7th	6,979	7,517	304	73	14,873
8th	5,159	6,506	846	269	12,780
9th	4,569	6,008	981	—	11,558
10th	4,081	6,654	333	—	11,068
all ten hundred	608,623	76,075	8,219	702	693,619

Comment on Tables 4 and 5:

Though Table 4, and especially 5, already reveal characteristic distributions and etymological groupings, the same data presented as relative distributions (in %) in the next two tables facilitate observation further. Comment will, therefore, be made on Tables 6 and 7.

2.4.2.

TABLE 6

RELATIVE Etymological Proportions of First 1,000 Items of Brown Corpus

(By hundreds of frequency and as totals of word TOKENS)

	first hundred	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	all 1,000
Ang'o-Saxon	98.3	73.9	64.2	50.7	43.1	47.0	41.4	35.0	33.8	33.0	85.1
French	—	14.8	20.7	30.2	39.0	26.8	37.9	37.8	42.4	47.0	8.4
Latin	—	3.6	6.2	8.5	8.2	18.2	12.6	13.2	9.6	12.2	2.6
Old Norse	1.7	5.3	3.2	3.1	3.7	5.0	3.6	3.2	4.7	0.9	2.4
Greek	—	1.7	3.7	5.8	5.0	3.0	2.0	6.6	8.5	3.0	1.2
Top five etymologies	100.0	99.3	98.0	98.3	99.0	100.0	97.5	95.8	99.0	96.1	99.7
Common Ger.	—	—	1.10	—	—	—	—	1.02	1.08	0.92	0.11
Frankish	—	0.75	—	—	—	—	—	—	—	—	0.07
E. Friesian	—	—	—	0.90	—	—	0.96	—	—	—	0.06
Italian	—	—	—	—	0.99	—	—	—	—	0.97	0.05
O'd H. Ger.	—	—	—	—	—	0.96	—	—	—	—	0.04
Mid. L. Ger.	—	—	—	—	—	—	—	1.12	—	0.99	0.04
Celtic	—	—	—	0.89	—	—	—	—	—	—	0.03
Unknown	—	—	0.99	—	—	—	0.49	2.11	—	—	0.10

TABLE 7

Contracted Version of Table 6

Level of frequency	Total Germanic	Total Romance	Other	Unknown
1st hundred	100.0	—	—	—
2nd	80.0	18.4	1.7	—
3th	68.5	26.9	3.7	1.0
4th	54.7	38.7	6.7	—
5th	46.8	48.2	5.0	—
6th	52.0	45.0	3.0	—
7th	46.9	50.5	2.0	0.5
8th	40.3	51.0	6.6	2.1
9th	39.6	52.0	8.5	—
10th	35.8	60.2	3.0	—
all ten hundred	87.8	11.0	1.2	0.1

2.4.2.1. Comment on Tables 6 and 7:

a) Etymological proportions supplied by these tables parallel Robert's findings, as presented in Table 3:

	AS	Fr.	Lat.	ON	Gk	*Top 5 etym.*	Other	Unknown
lst '000 in Horn's list	83.4	11.8	1.9	2.6	0.05	*99.45*	0.46	0.09
lst '000 in Brown Corpus	85.1	8.4	2.6	2.4	1.2	*99.70*	1.20	0.10

b) The more pronounced differences in the proportions of items of French and Greek etymologies are due to Roberts' choice of the proximate rather than the ultimate etymological source. This is particularly true of the Greek element with its absolute prevalence of the Greek-through-Latin elements over those directly from Greek. The disparity between the proportions under the heading 'Other' is also easily traceable to this difference of approach. The ultimate-etymology criterion yields naturally more etymologies, some of which would have otherwise disappeared under one of the few 'funneling' etymologies (such as Latin, French or, for Amerindian items, Spanish).

The distributions become closer when grouped by linguistic families:

	Total Germanic	Total Romance			Other	Unknown
1st '000 in Horn's list	86.15	13.71	0.05	*99.70*	1.28	0.09
1st '000 in Brown Corpus	87.15	10.97	1.19	*99.70*	1.28	0.10

c) The persistent gap between the values for Total Germanic (+1.6%) and Total Romance (—2.74%) must be interpreted as reflecting the ultimate, overall difference in the makeup of the two corpora, with the stylistically narrower Horn's corpus incorporating a more formal vocabulary (normally found in letters, notably business correspondence).

d) The most important observations, however, are those that can be made about the interdependence of frequency levels and individual etymology proportions. Whereas Anglo--Saxon falls from 98.3% for the first hundred down to 33.0% for the tenth hundred, the combined French and Latin rise from 18.4% to 59.2%, while ON and Greek keep fairly stable. This, viewed by itself, apperars to do no more than herald the trend of the second-to-tenth thousand range in Horn's list (as shown by Roberts' analyses), where the corresponding drop was 83.4—25.8% and rise 13.7—63.5%.

e) But there is another, more significant, phenomen in Horn's list echoed, as it were, by Brown Corpus proportions. It is the manner in which the first decile overwhelmingly affects the other nine, so that the ultimate (average) etymology proportions are not significantly different from those in the first decile. Thus, while the Anglo-Saxon element in the first hundred of Brown Corpus, as already noted, accounts for 98.3%, falling off to a mere 33.0% in the tenth hundred, the *average* share of Anglo-Saxon for all ten hundred is still as high as 85.1%. The corresponding values for Horn's list are: 83.4% (first thousand), 28.5% (tenth thousand) and 78.3% (all ten thousand). This in spite of the fact that, in the remaining nine deciles, the proportions of French were either consistently higher than those for Anglo-Saxon (Horn's list, 43.5——48.4% as against 25.3—35.1%), or were roughly equivalent in mid deciles, with French gaining the upper hand in the last three deciles (Brown Corpus).

2.4.2.2. We know why this is so: the cumulative totals of the first 1,000 items account, owing to their high frequences, for over two-thirds of the entire mass of any average English text (68.6%, to use the exact findings of Brown Corpus). The remaining nine thousand together represent no more than 22% of the same mass (between the 68.8% and the 90.8% covered by the first ten thousand Brown Corpus items).

2.4.2.3. If we now closely observe the proportions of individual hundreds in Table 4, we will notice that the first hundred (with its 481,055 tokens) by itself accounts for as much as 68.91% of the whole mass (698,079 tokens). The second (8.95%) and third (5.26%) hundred bring this share up to 83.12%. To this may be added the information that the first 100 items account

for close to one-half (47.43%) of the entire Brown Corpus. This requires us to modify the conclusion reached earlier (cf. 1.9.1.) that *the first txhousand words (items) of English are decisive for the etymological proportions of any average English text,* by adding: *with the first hundred decisive within that first thousand.* Merging and rewording these findings, we obtain a simpler, more expressive, statement:

> *The mere 100 most frequent words of English significantly affect the etymological proportions of any average English text, while the first 1,000 words are decisive for these proportions.*

2.4.3. All tables of Brown Corpus proportions (Nos. 4 to 7), both absolute and relative, have so far reflected only the *dynamic* distribution of items, by word tokens (cf. 1.2. and n. 2). In the next three tables, each item is observed *statically,* as a word type, with items counted only once. There is obviously no need for parallel tables of relative distribution since the absolute totals can be reinterpreted as relative values with facility (thus, 509 for Anglo-Saxon is read as 50.9%). Finally, decimal values in the tables reflect cases of combined etymologies when the items involved were split.

TABLE 8

ABSOLUTE Etymological Proportions of First 1,000 Items of Brown Corpus

(By hundreds of frequency and as totals of word TYPES)

	first hundred	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	all 1,000
Anglo-Saxon	96	71	61	50	42.5	47	40.5	34	30.5	36.5	509
French	—	14.5	22	31	38	26	36.5	35.5	40	46	289.5
Latin	—	5	7.5	8	8	18	12	12.5	9	10	90
Old Norse	3	5.5	3	3	3.5	5	3.5	3	4.5	0.5	34.5
Greek	—	2	3.5	6	5	3	2	6	8	3	38.5
Top five etymologies	*99*	*98*	*97*	*98*	*97*	*99*	*96.5*	*91*	*92*	*96*	*961.5*
Common Ger.	—	—	1	—	—	—	—	—	2	—	3
Frankish	—	1	—	—	—	—	—	—	—	—	1
East Friesian	—	—	—	1	—	—	1	—	—	—	2
Italian	—	—	—	—	1	—	—	—	—	1	2
Old H. Ger.	—	—	—	—	—	—	1	—	—	1	2
Middle L. Ger.	—	—	—	—	—	—	—	1	—	1	2
Celtic	—	—	—	1	—	—	—	—	—	—	1
Unknown	—	—	1	—	—	1	0.5	2	—	1	5.5
Total analyzed	99	99	99	100	98	100	97	94	94	100	980
Excluded	1	1	1	—	2	—	3	6	6	—	20
GRAND TOT.	100	100	100	100	100	100	100	100	100	100	1,000

TABLE 9

Contracted Version of Table 8

Level of frequency	Total Germanic	Total Romance	Other	Unknown	Excluded
1st hundred	99	—	—	—	1
2nd	77.5	19.5	2	—	1
3rd	65	29.5	3.5	1	1
4th	54	39	7	—	—
5th	46	47	5	—	2
6th	52	44	3	1	—
7th	46	48.5	2	0.5	3
8th	38	48	6	2	6
9th	37	49	8	—	6
10th	39	57	3	1	—
all ten hundred	553.5	381.5	39.5	5.5	20

Comment on Tables 8 and 9:

As was to be expected, Anglo-Saxon and French (or, viewed more broadly, Total Germanic and Total Romance) continue in their inverse proportions. However, since each hundred — and later each thousand — represent an exactly equal portion (10%) of the entire mass under analysis, there cannot be any decisive influence of the proportions in the first one or two hundreds upon the ultimate (average) proportions. Consequently, Anglo-Saxon ends up with the average share of 50.9% as against 96% in the first hundred, while French, beginning from zero in the first 100 (and still only at 14.5% in the second) grows to an average of 28.95% for the first 1,000. That this trend continues is demonstrated by the following table for the first 10,000 items, again based on the simplified and contracted (cf. n. 3) Roberts' figures.

2.4.4.

TABLE 10

ABSOLUTE Etymological Proportions of 10,000 Items of Horn's List (By thousands of frequency and as totals of word TYPES)

	first 1,000	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	all 10,000
AS	518.5	345	309	297	283	301	254	292.5	291.5	288	3,179.5
French	360	483	487	469	486	443	479	437	437.5	443.5	4,525
Latin	79	125	157	192.5	175	200	187	190.5	185	189	1,680
ON	39.5	31	19	17.5	25.5	23	30	30.5	27.5	22.5	266
Top four etymologies	997	984	972	976	969.5	967	950	950.5	941.5	943	9,650.5
Unknown	2.5	6	6	9	8	13	12	13	18.5	12.5	100.5
Dutch	2	2.5	8	2.5	9.5	7.5	9.5	14	16	15	86.5
East Fries.	2	—	—	—	1	—	1	—	0.5	—	4.5
LG	1	3.5	7.5	3	4.5	5.5	12	7	5	3.5	52.5
Greek	1	3	1	5	8	2	5	9	7	8	49
Italian	0.5	4	5.5	5.5	5.5	4	6.5	6	4	3.5	45
Celtic	1	—	1.5	0.5	3	0.5	1	2	2	1	12.5
Spanish	—	1	3.5	1	1	4.5	5.5	1.5	5	8.5	31.5
Other	—	2	2	4.5	3	4	3.5	3	7.5	3	32.5
Total	1,007	1,006	1,007	1,007	1,013	1,008	1,006	1,006	1,007	998	10,065[11]

Comment on Table 10:

a) As readily observable, the figures for the whole span of 10,000 items are markedly different from those for the first 1,000. After falling off to 28.8% by the 10th decile, the proportion of Anglo-Saxon is pulled up slightly (to 31.8%) for the total value of Anglo-Saxon, but this is quite a comedown compared with the 51.9% in the first decile. On the other hand, French, beginning lower (36.0% in the first

[11] After Horn's list items had been keypunched, transferred onto a magnetic tape and a printout of them produced, their total turned out to be 10,065 (Roberts, 1965:31). However, this minor discrepancy can be neglected when the figures in this table are read as relative proportions (cf. 2.4.3.). Thus, the figure for Anglo-Saxon in the first column is 518.5 which converts to 51.85%, as against the correct value of 51.89% (since 1,007 is the actual total of this decile). Similarly, the total figure for Anglo-Saxon (3,179.5) is interpreted as 31.80% compared to the correct 31.59%.

1,000), maintains a steady rate between the 2nd and 10th deciles (48.7—43.7%) and finishes on top of Anglo-Saxon (45.3%). This ratio of 45.3%: 31.8% changes further, and quite heavily, in favor of French after individual etymologies are grouped into families, becoming 62.9% : 36.4% (Romance to Germanic).

b) The steady proportions of Old Norse are surprisingly parallel to their relative-distribution values (cf. Table 3), with the totals almost identical (2.66% here as against 2.62% in Table 3).

c) The totals of four top etymologies, for each decile and overall, are likewise closely parallel to those in Table 3.

3. CONCLUSION

3.1. A painstaking analysis of the etymological composition of the English core vocabulary (first 1,000 words) has been carried out, making a modest and hopefully useful contribution to an aspect of English, through a number of statements made possible by the findings of this investigation.

3.2. The following, condensed, statement reflects the most important findings of the effort described in this paper:

The mere top-frequency words (items) of English — with their almost exclusive Anglo-Saxon origin — significantly affect the etymological proportions of any average English text, while the first 1,000 words are decisive for these proportions. As a result, Anglo-Saxon accounts for 85%, with French plus Latin covering less than 11% at the first--thousand level, and the ratio is still 78% : 18% at the ten--thousand level. Naturally, when these words are observerd as static dictionary entries (without the weight of their distributional frequencies) this influence is weaker, and the overwhelmingly Anglo-Saxon character of the first 100 items (96%) is changed to AS 52%: Fr/Lat. 44% for the first 1,000, (the ratios reversing for the second thousand (35% : 61%) and staying so with little change for the entire ten-thousand span (32% : 62%).

3.3. This statement and other, more detailed, findings of the present analysis refine those already made in several earlier investigations (Bujas, 1968 a: 97—98, 1968 b: 144—145 and 1972 : 593).

3.4. In conclusion, may I repeat the hope that further, more comprehensive and computer-supported, efforts will be possible in the future, shedding more light on this particular area of English vocabulary studies.

REFERENCES

Bujas, Ž. 1968a

"Etimološke proporcije engleskog vokabulara. Analize i analizatori" [Etymological Proportions of the English Vocabulary. Analyses and Analysts], *Filološki pregled*, Belgrade, 1—2, 1968, pp. 71—98 (English abstract).

Bujas, Ž. 1968b

"Frequency Lists As Aids in Analysing the Etymological Composition of English", *Studia Romanica et Anglica Zagrabiensia*, Zagreb, 25—26, 1968, pp. 129—148.

Bujas Ž. 1971

"Chronological and Area Survey of Foreign Element in the English Vocabulary", *Studia Romanica et Anglica Zagrabiensia*, Zagreb, 29—32, 1970—71, pp. 131—188.

Bujas, Ž. 1972

"A 'Time' Magazine Vocabulary Study", *Studia Romanica et Anglica Zagrabiensia*, Zagreb, 33—36, 1972—73, pp. 579—594.

Carroll, J. B., Davies, P. and Richtman, B. 1971

Word Frequency Book, Houghton Mifflin & American Heritage, New York, 1971, 856 pp.

Kučera, H. and Francis, W. N. 1967

Computational Analysis of Present-Day American English, Brown University Press, Providence, R. I., U.S.A., 1967, 424 pp.

Lorge, I. and Thorndike, E.L. 1938

A Semantic Count of English Words, Teachers College, Columbia University, New York, 1938, 4 v. (mimeo.).

Lorge, I. 1949

The Semantic Count of the 570 Commonest English Words, Teachers College, Columbia University, New York, 1949, 187 pp. (mimeo.).

Roberts, A. H. 1965

A Statistical Linguistic Analysis of American English, Mouton, The Hague, 1965, 437 pp.

West, M. 1953

A General Service List of English Words, Longmans, London, 1953, 588 pp.