Željko Bujas

Frequency Lists As Aids in Analysing the Etymological Composition of English

0.1. Establishing with any degree of certainty the proportions of elements of various etymologies that have gone into the making of the modern English vocabulary cannot ever be completely free from a great deal of drudgery. Extensive and adequately dispersed samples must be selected, the etymology of each individual word of these texts must be looked up in the dictionary of our choice, and endless numbing tally-sheets kept and wearily checked and re-checked. Even when we can use that great labour-saving device, the computer, hundreds of hours may be needed to put the data in the form acceptable to these fastidious, recently acquired friends of the linguist.

0.2. Small wonder then that, as recently as 1920, all we knew about the etymological proportions of the English vocabulary could be traced back to a single effort. The intrepid individual who did not — in that age of elegant academic attitudes in philology — shrink from actually and lowly counting and analysing several tens of thousands of words[1] of continuous-text samples was George Perkins Marsh, American philologist and diplomat.

0.21. Published in 1860, in *Lectures on the English Language* (pp 118—131), Marsh's results have stood the test of time surprisingly well. They were a definite improvement on his, ad-

[1] Perhaps as many as some 100,000 words. Marsh does not provide us with the exact number of words analysed, though the authors (a total of 30, plus the Bible) and their texts which he investigated have been carefully listed. Sections of texts analysed (chapters, lines) are often, but not invariably, indicated. The samples of Chaucer's texts analysed by Marsh, for instance, include a total of 13,800 words of continuous text.

mittedly few, predecessors.[2] Marsh carefully distinguished between a vocabulary-based approach (with repetitions of individual words not counted) and an actual-occurrence approach (repetitions counted). He was also keenly aware of the influence of style — though he never used the word — on the etymological proportions of the vocabulary in English. Finally, the size of his corpus would even now be considered quite respectable. Nevertheless, it is amazing to find Marsh's results serving as the basis for § 41, "Proportion of the Elements", of the Introduction to Webster's *New International Dictionary of the English Language* up to as recently as 1961. Or quoted by a number of authors over a period of one hundred years, some as recent as H. Alexander in *The Story of Our Language* (New York, 1962, p. 109), though he, like most of them, never mentions Marsh as the source of these data.[3]

1.0. However it may be, if no method had been devised to avoid at least part of the drudgery involved in an etymological proportion analysis of a large corpus, we should, very probably, even now be without any more extensive analysis than the one performed by Marsh over one century ago.

1.1. Luckily, the 1920's saw the full introduction of texts structured to serve as very efficient anti-drudgery devices: frequency lists resulting from very extensive word counts. Frequency lists take two principal forms: rank lists and alphabetical lists. A rank list is composed of all the different words of a sample in descending order of frequency; an alphabetical list is made up of all these words regardless of their frequency. Each different word *(word type)* in either list is accompanied by a figure, indicating the number of occurrences *(word tokens)* of the word in question.[4]

[2] English historian Sharon Turner (1768—1847), in *The History of Anglo-Saxons* (1799—1805) and Anglican Archbishop Richard C. Trench (1807—1886), in *English Past and Present* (1855) — if we discount the truly first analyst of the etymological proportions in English, the English philologist and theologian George Hickes (1642—1715). All their analyses were, however, based on very limited samples (Turner's on *ca* 1,500 words of continuous text, Trench's on 299 words, Hickes's on 67 words /Our Father/), so that statements about the etymological composition of the English vocabulary based on them have little value. More about these early analyses in this writer's article "Etimološke proporcije engleskog vokabulara. Analize i analizatori", *Filološki pregled,* 1—2, 1968 (Belgrade), pp. 71—98.

[3] For a survey of these authors and their works see n. 25 of the above article.

[4] Sometimes also by other figures, usually to show the number of samples containing the word (an important measure of dispersion).

1.11. To illustrate this, here are short sections from the beginning of two such lists, contained in Dewey's word count entitled *The Relativ*[5] *Frequency of English Speech Sounds* (Harvard University Press, 1923), one of the principal works in the genre to appear in the 1920's:

Rank List		Alphabetical List	
7,310	the	a	2,120
3,998	of	able	30
3,280	and	about	153
2,924	to	above	25
2,120	a	abroad	11
2,116	in	according	17
1,345	that	account	20
1,216	it	across	14
1,213	is	act	15
1,155	I	action	28

1.2. If we know that the number of running words in the corpus selected by Geoffrey Dewey for the above count totals 100,000, some quantitative conclusions are obvious. For instance, total occurrences of ten first words on the Rank List are 26,877, or 26.9 per cent of the entire text of the corpus. Next ten words from the Rank List bring the total of occurrences up to 33,421, or 33.4 per cent of all the words in the corpus. Put more simply, this means that the 20 most frequent words make up one-third of an average English text. Further calculations will provide us with the fact that 50 top-frequency words in English cover as much as 46.1 per cent of any text, and 100 such words account for 54.3 per cent of text.

1.3. The next conclusion, certain to be made by an analyst of etymological proportions, is that these proportions encountered among the top 20 (or 50, or 100) words in English are valid for 33.4 (or 46.1, or 54.3) per cent of any average text. Expressed otherwise, these proportions have a 33.4 (46.1, 54.3) — per cent reliability as a statement of the etymological composition of English as a whole.

1.4. Carrying this conclusion further logically, the same analyst will very soon realise that a larger, but still manageable, list may account for nearly all the occurrences of different words in any normal text. Dewey's word count actually shows that the 1,000 most frequent words in English account for 78.3 per

[5] This spelling (*Relativ* instead of *Relative*) is due to Dewey's interest in the Reformed Spelling (and shorthand). The entire textual content of this work has been published in the Reformed Spelling.

cent (500 for 71 per cent) of the entire 100,056-word corpus selected by Dewey at a variety of stylistic levels. A ten-thousand word list covers 91.8 per cent of any English text, as exemplified by the Kučera-Francis list from 1967 (see n. 23 and 34).

1.41. The above-90% reliability offered by a ten-thousand word list makes the list, consequently, a statistically and linguistically respectable tool, and a highly acceptable replacement for the gruelling and time-consuming technique of looking up the etymology of each word as it occurs in a continuous text.

1.5. No sophisticated estimates will be needed to make obvious the saving of time made possible by the use of frequency lists in an analysis of the etymological proportions of the English vocabulary. Analysing Dewey's corpus of 100,000 words would, for instance, require some 50,000 lookups — after an alphabetical list of 50 to 100 most frequent words has been established, and the need eliminated for looking them up each time they occured (though each occurrence would still have to be marked in a tally-sheet). Using a 10,000-word alphabetic frequency list of the same corpus would require only 10,000 alphabetic frequency list of the same corpus would require only 10,000 etymological lookups (but the same number of tally-sheet markings). Also, these lookups would be less time-consuming thanks to the alphabetical order of words on the list. In view of all this, we may safely assume that such an investigation of etymological proportions would require only a fifth or a sixth of the time needed for a continuous-text analysis of a 100,000-words corpus. The saving of time is truly spectacular with large corpora. Thus, for instance, Horn's corpus (see later) of over 5,000,000 words offers a contrast of about 2,000,000 lookups as against 36,400 lookups. Admittedly, no alphabetical list of all 36,400 different words in the corpus has been provided, and we can only use a 10,160-member list of top-frequency words supplied by Horn. This reduces even further the time needed for lookups (but at the same time makes the results only about 90 per cent reliable).

2.0. Awareness of these advantages may have varied in degree with prospective analysts of the etymological proportions in the English vocabulary. However, appearance of the first reliable large word counts[6] in the 1920's made very clear the principal time-saving implications of frequency lists. This triggered off a spurt in the field largely dormant for some sixty years. A four-year span (1922—1926) produced no fewer than

[6] A list of (seven) principal word counts published between 1904 and 1920 is presented by Dewey on pp. 3 and 4 (Preliminary Discussion) of his *Relativ Frequency of English Speech Sounds.* He also supplies brief critical comments on each of them.

six investigations of the etymological composition of the English vocabulary. Here they are, in chronological order:

1. Earle E. Franklin, "The Derivation of the Second 5,000 Words of the Thorndike's Teacher's Word Book", *School and Society,* vol. XV (1922), pp. 622—623; and (under a slightly modified title) in *The Classical Weekly,* vol. XVI (1923), p. 114.

2. Berthold L. Ullman, "Our Latin-English Language", *Classical Journal,* vol. XVIII (1922), pp. 82—90.

3. Wren J. Grinstead, "On the Sources of the English Vocabulary", *Teachers College Record,* vol. XXVI (1924), pp. 32——46.

4. Edward Y. Lindsay, "An Etymological Study of the 10,000 Words in Thorndike's Teacher's Word Book", *Indian University Studies,* vol. XII, Study No. 65 (March 1925), pp. 1—115.

5. Alexander Inglis, "Classical and Native Elements in the English Language", *Classical Journal,* vol. XX (1925), pp. 515——525.

6. Helen M. Eddy, "The French Element in English", *Modern Language Journal,* vol. X (1926), pp. 271—280.

2.1. As to the word counts themselves, the years between 1921 and 1926 saw the appearance of three large and reliable compilations: the already mentioned Thorndike, Dewey and Horn lists. Their various properties are presented in a chronologically arranged tabular survey (Table 1).

2.2. However, five of the six investigations from that period listed were based on Thorndike's list. Only Inglis's study was made on the material from Dewey's word count. H. M. Eddy, the only analyst who could (chronologically) have used Horn's list, also utilised Thorndike's count.

2.21. One might well wonder what useful could five different analysts have done with the same frequency list, within a five-year span at that. However, their probing into proportions of the etymological composition of the English lexis varied considerably in approach and emphasis. As a result, their findings are not uniform, containing some revealing or curious divergences.

2.22. What all these analysts had in common, though, was their primary interest in the share of classical (Greek and Latin) elements in the English vocabulary, particularly as a useful starting point in the methodology of language teaching. By profession, after all, they were professors of education (Inglis, Grinstead), of Latin (Ullman), French (Eddy), or teachers (Lindsay) and graduate students in education (Franklin).

Table 1

Count	Size of corpus in running words	No. of different words[7]	Style levels in corpus	Words ranked by	Rank list limitations
E. L. Thorndike, *A Teacher's Word Book of 10,000 Words* (New York, 1921)	4,565,000	10,000	wide range of texts	"credit-number"	only 10,000 top-frequency words listed
G. Dewey, *The Relativ Frequency of English Speech Sounds* (Harvard Univ. Press, 1923)	100,056	10,161[8]	ten	number of occurrences	only 1,027 top-frequency
E. Horn, *A Basic Writing Vocabulary* (Iowa City, 1926)	5,137,000 (with estimates: 15,463,000)	36,373[9]	private and business correspondence	number of occurrences	only 10,160 top-frequency words listed

[7] By a "different word" (or "word type") we mean any combination of graphemes bounded by blanks. A considerable number of homographs, like *lead* (vb.) and *lead* (n.), is contained in these *heterographs*, as we might term them. A large majority of them, however, are of the same etymological origin — esp. the "grammatical" homographs, e. g. *lead* (vb. pres.) and *lead* (vb. pret.) — and can be disregarded in an analysis of etymological proportions.

[8] After the exclusion of 6,404 proper names, 2,102 numerals and 734 abbreviations, foreign words, etc. from the corpus.

[9] After the elimination of "names of persons and places". The total number of all different words may be guessed if we know the results arrived at by other analysts. Thus, for instance. Grinstead's study (see later) of 1,700,000 running words produced 57,700 different words, of which number 19,781 were proper names, 7,237 "still foreign" and 839 abbreviations. Kučera and Francis's recent (1967) computational analysis of a 1,014,232-word corpus (the Brown Corpus) produced a total of 50,400 undifferentiated "distinct graphic words (types)".

3.0. We shall do best to consider their analyses individually and in chronological order.

3.1. In the earliest of them, *The Derivation of the Second 5,000 Words of the Thorndike's Teacher's Word Book* by E. E. Franklin, the author's professed purpose was "to determine... the part which Latin may play in providing the vocabulary which should be attained by a pupil during the High School period".

3.11. Having estimated the (passive) vocabulary "of pupils entering a junior high school at 5,000 words or more", Franklin limited his analytical effort to the vocabulary range of 5—10,000 words of the recently appeared Thorndike list. Examining the first-origin[10] etymologies of a total of 4,829 words from Thorndike's seven (3—9) credit-number[11] groups, he presented his findings in the following table (of percentages):

Table 2

	9[12]	8	7	6	5	4	3	Total
Latin	47.3	50.1	47.7	48.8	46.9	47.7	51.9	48.6
Anglo-Saxon	34.4	24.6	28.5	27.9	27.7	25.9	25.9	27.3
French	6.4	9.5	8.7	8.3	10.5	9.5	8.1	8.9
German	0.5	1.4	1.0	0.9	0.8	0.9	0.4	0.8
All Others	11.6	14.4	14.1	14.1	14.1	16.0	13.7	14.4

3.12. The disproportionately large group "All Others" (the chief component of which is Greek, we are told) indicates an inadequate intuition of the analyst, and a generally innocent approach. Thus, Franklin says: "A rather surprising outcome of the study is the revelation of the very low percentage (less than one per cent) of words secured from the German. While no separate account was kept, it is the writer's impression that the Scandinavian tongues were somewhat more prolific in derivatives than the German". That "the very low percentage... of words... from the German" should be a "revelation" to any one in 1922, points to the tenacity of one of the 19th-century etymological fictions.

3.13. Nevertheless, the first, and therefore commendable, effort (and it did involve quite an amount of work) towards basing the research into etymological proportions on frequency lists,

[10] In Franklin's words: "...in the case of the words common to several languages: credit was given to the language in which the root originated".

[11] See **3.45.—3.456.**

[12] The present author has replaced Franklin's original group indices (I to VII) by the corresponding, and more relevant in this article, credit-number values.

should not have suffered the editorial treatment to which Franklin's brief article was subjected. It is an unusual treatment, to say the least, when an article is preceded by an ironical assesment of its value by the Editor-in-Chief, the assesment itself longer than the text being criticised. A good example, certainly, of the high-handed attitudes not untypical of the scholarly Establishment of that period.

3.2. Only five months later, one of the leading American latinists, B. L. Ullman, set himself the task to make others "realize the full extent and significance of this [Latin] element", and to prove that Latin is not a dead language.

3.21. Mentioning no particulars about his method of counting (or the name of his collaborator Lillian B. Lawler[13]), he only supplies the final results of etymological proportions in Thorndike's and Horn's lists: "After eliminating 668 proper names, we find that at least 46.8 per cent, and possibly 47.5 per cent, of the words are Latin in origin, 6 per cent Greek, 41 per cent Teutonic and 5.2 per cent miscellaneous. The comparatively low percentage of Latin words is due to the considerable number of words selected from children's literature and from the Bible. A similar list by Professor Horn, as yet unpublished, contains 8,951 words found in ordinary correspondence. Of these, 57,6 per cent are Latin and 4.8 per cent are Greek".

3.22. Ullman's article contains half-a-dozen more analyses of the etymological proportions in the English vocabulary, but not based on a frequency list, so that are not of direct interest here.[14]

3.3. In one of them, however, Ullman makes use of a compromise approach in counting words for etymological proportions. Like Ramsay,[15] thirty years earlier, he analyses continuous text but counts each different word only once, thus reducing the proportion of Anglo-Saxon "structure words". The same approach was adopted by W. J. Grinstead, the next analyst who used frequency lists in his investigation of the etymological proportions among English words. In his article *On the Sources of the English Vocabulary* he reports on the Latin-English Word Count of the American Classical League (1922—1923).

[13] About whom we only get to know from the article by Carr, Owen and Schaeffer (see later).

[14] They include an analysis of the etymological makeup of American children's vocabularies; an investigation into the etymological proportions of the new words from the Addenda to the MW 1 (of 1900); and an illustration of the vocabulary interdependence between classical Latin and modern English.

[15] Ž. Bujas, *o. c.*, note 2, pp. 79—80.

3.31. "The present count has a primarily etymological aim", says Grinstead (p. 33), continuing: "It attempts . . . to determine the proportion of the Latin element in the English reading of the adolescent and adult at different ranges of distribution . . .". The ultimate motivation, however, vere the needs of secondary-school Latin instruction, as made specific in Grinstead's words: " . . . the Latin vocabulary in the early years should be based as far as practicable upon the Latin originals of the English vocabulary likely to be encountered by high school pupils".

3.32. The count was based on a very extensive corpus, totalling about 1,700,000 running words, and made up of the following samples:

a) 1,000,000 words from the *Encyclopaedia Britannica*
b) 250,000 words each from the popular magazines *Ladies' Home Journal* and *Saturday Evening Post*
c) 100,000 words from the serious *Literary Digest*
d) four samples (total: 100,000 words) from high-school textbooks

3.33. One more important feature that makes the word count under discussion essentially different from Ullman's and Franklin's counts is the factor of frequency present in it. The two earlier counts compute their proportions without the frequency values supplied by Thorndike's list, the list in fact serving as any other dictionary (i. e. simply a list of lexical entries) would have served. Thus, these earlier counts result in etymological proportion values basically obtainable from the etymological analysis of any regular dictionary of the same size (10,000 entries). Admittedly, the element of frequency is automatically present, owing to the structure of the corpus analysed (frequency list), but then it is equally present in any good ordinary dictionary, in which the selection of entries based on experience should closely correspond to a frequency list of the same span.[16]

3.34. However, the Thorndike frequency values — themselves relative[17] — are used by Grinstead with some modification. They are combined with dispersion (or range) values, i. e. with the number of samples containing the word in question. The combination of these two values is interpreted as the word's rank, placing it into one of ten classes, descending in frequency from 10 to 1. The complex combination procedures are described in detail, but we may safely skip the description. It should, perhaps, be emphasised only that the frequency and range of

[16] It seems self-evident to this author that dictionary-entry selection ought to be based, to a very large degree, on frequency lists.

[17] See **3.45.—3.456.**

Table 3

Rank	10	9	8	7	6	5	4	3	2	1	Total
A. WORDS NOT TRACED											
Abbreviations	—	5	3	10	11	20	24	68	158	540	839
Proper Names	1	106	72	98	127	227	403	822	2,485	15,440	19,781
Still Foreign	—	—	—	2	1	5	30	115	820	6,260	7,233
Total:	1	111	75	110	139	252	457	1,005	3,463	22,240	27,853
B. CLASSICAL WORDS											
English from Latin	83	1,395	695	681	632	761	1,042	1,372	2,024	5,760	14,835
L—E Hybrids	—	5	6	8	18	18	29	64	176		
L—Gr Hybrids	—	1	2	—	1	1	3	10	48		
English from Greek	3	83	100	76	102	127	193	374	652	2,260	4,008
Gr—E Hybrids	—	—	1	1	6	3	3	4	20		
Total:	86	1,484	804	766	759	910	1,270	1,824	2,920	8,020	18,843
C. NON-CLASSICAL WORDS											
Native English	406	654	341	368	317	374	515	608	953	2,980	7,516
Miscellaneous Foreign	7	124	106	116	143	177	229	426	600	1,560	3,488
Total:	413	778	447	484	460	551	744	1,034	1,553	4,540	11,004
GRAND TOTAL	500	2,373	1,326	1,360	1,358	1,713	2,471	3,863	7,936	34.800	57,700

Thorndike's 500 top words has not been analysed at all, and that they constitute class 10.

3.35. However it may be, Grinstead's analysis established a total of 57,726 different words in the 1,700,000-word corpus. Table 3 presents these words classified by frequency-cum-range rank and by principal etymological sources, with stress on classical etymologies. This author has modified Grinstead's original table by adding his figures for the class 1 column and the Total column.

3.36. Here is also, slightly compressed and modified, Grinstead's tabular presentation of the absolute and relative (in %) distribution of only the analysed words from his corpus (other analysts also normally exclude proper names, abbreviations and foreign words):

Table 4

Rank	Native English		Miscellaneous Foreign		Greek		Latin		Total Class
	Words	%	Words	%	Words	%	Words	%	Words
10	406	81,4	7	1,4	3	0,6	83	16,6	499
9	654	28,9	124	5,5	84	3,7	1,401	61,9	2,262
8	341	27,3	106	8,5	103	8,2	703	56,2	1,251
7	368	29,4	116	9,3	77	6,2	689	55,1	1,250
6	317	26,0	143	11,7	100	8,9	651	53,4	1,219
5	374	25,6	177	12,1	131	9,0	780	53,4	1,461
4	515	25,6	229	11,4	199	9,9	1,074	53,3	2,014
3	608	21,3	426	14,9	388	13,6	1,446	50,6	2,858
2	953	21,4	600	13,4	720	16,1	2,248	50,3	4,473
1	2,980	23,7	1,560	12,4	2,260	18,0	5,760	45,8	12,560
Total:	7,516	25,2	3,488	11,7	4,008	13,4	14,835	49,7	29,874

3.37. The above proportions, provided by the Latin-English Word Count of the American Classical League, differ significantly from those resulting from Ullman's and Franklin's analyses. This is due, as already stressed, to the factor of frequency present much more immediately in the efforts directed by Grinstead — though a full application of frequency values was not made.

3.38. Most interesting of Grinstead's proportions is the one for the top 500 words (class 10), with 81.4 per cent for Anglo-Saxon as opposed to 15.6 per cent for the non-native element in the English vocabulary. On all other levels, i. e. in all other frequency/range classes, we see a steady reverse ratio of 20—30 per cent of native and 70—80 per cent of foreign elements. Latin maintains itself at around 50—55 per cent of the vocabulary as an average, but shows a constant slight decrease. In contrast, "other non-classical etymologies" and Greek are on the increase, notably Greek (some 10 per cent of the vocabulary), which is understandable in view of professional terminology encountered in the more difficult or specialised texts.

3.4. The next year, 1925, saw the appearance of two more analyses which represented the peak of the 1920's effort in the field — by E. Y. Lindsay and A. Inglis.

3.41. In his extensive study (115 pp) *An Etymological Study of the 10,000 Words in Thorndike's Teacher's Words Book*, Lindsay is also motivated by a problem of methodology in language teaching. Like Franklin, he tries to establish the usefulness of learning Latin in secondary school for the expansion of the mother-tongue (i. e. English) vocabulary. To achieve this on the basis of reliable quantitative data, this analyst likewise turns to Thorndike's list. However, Franklin (and Ullman) used Thorndike's list as a mere glossary. Disregarding frequency figures for individual words, they in fact obtained etymological proportions valid for a dictionary of corresponding size (10,000 entries). Unlike them, Lindsay was the first user of individual-word frequencies from a frequency list.[18]

3.42. After an extensive and detailed analysis of Thorndike's list, Lindsay presents his result in this final (somewhat modified) table:

Table 5

Etymology	No. of different words	%	Index-number total	% of total word occurrence
Anglo-Saxon	3,209	35.15	95,101	50,52
Latin	4,198	45.98	67,992	36.12
Greek	657	7.19	8,962	4.76
Scandinavian	376	4.12	7,465	3.97
Low and Modern German	198	2.17	2,909	1.54
Other Germanic	90	0.98	1,336	0.71
Other Indo-European	239	2.62	3,177	1.69
Other languages	163	1.79	1,304	0.69
Total	9,130[19]	100.00	188,246	100.00

3.43. Using these values for etymological proportions in the English vocabulary, Lindsay irrefutably establishes in the end

[18] Grinstead used Thorndike's frequency figures, but only as average, group values.

[19] Instead of 10,160, because 870 words (largely proper names) have not been analysed.

the usefulness of learning Latin in secondary school for the extension of the pupils' English vocabulary. His conclusion (p. 8) is that the pupil who, after four years of Latin instruction, masters 2,000 words from G. Lodge's *Vocabulary of High School Latin* will automatically master the meanings of 3,100 (or 74.08$^{0}/_{0}$) words, with derivations, of Latin origin from Thorndike's list.

3.44. We are, however, more interested in Lindsay's analysis of frequency lists as a method of research into the etymological proportions of English. In Lindsay's study, the first to have made use of frequency figures for individual words, we are thus particularly interested in the figures in columns "Index-number total" and "$^{0}/_{0}$ of total word occurrence" from Lindsay's final table. These figures, based on the frequency of individual words with all their repetitive occurrences in a large number of extensive corpora of continuous English texts, indicate beyond doubt the prevalence of the Anglo-Saxon (50.52$^{0}/_{0}$) over the Romance (about 37$^{0}/_{0}$[20]) element in the English vocabulary.

3.45. This ratio, however, struck the present writer as still incorrect and unfair to the Anglo-Saxon element. Doubting intuitively any ratio with the share of the native element below 70$^{0}/_{0}$, we began to examine critically Lindsa's procedures in utilising Thorndike's *credit-number* values. Very soon, Lindsay's mistake was revealed. He had interpreted Thorndike's *credit-numbers* (or, as he calls them, *index-numbers*) as actual quantitative data, not as relative indices of frequency hierarchy which is how Thorndike uses them. It should, however, be pointed out that Thorndike himself does not say precisely anywhere in the book how he has obtained his *credit-numbers* which misled Lindsay, or what their precise quantitative value (absolute frequency) is.[21] Indeed, he expressly says[22]: "The reader is asked to accept arbitrarily these credits, since an explanation of the

[20] Latin element (36.11$^{0}/_{0}$), plus part of the value in the column "Other Indo-European languages" (1.69$^{0}/_{0}$), where Italian, Spanish and Portuguese are included.

[21] This remains unaltered through later, expanded, editions of Thorndike's word count: in 1931 (with 20,000 words) and 1944 (30,000 words). To be true, in the third edition (co-authored by I. Lorge) *credit-numbers* were replaced by estimates of occurrence over a total of 4,565,000 running words. J. Alan Pfeffer, in *Basic (Spoken) German Word List. Grundstufe* (Prentice-Hall, 1964) has the following footnote (n. 5, on p. 2): "Professor Thorndike, it appears, was the first to introduce the aspect of range, and it was he who, together with Professors V. A. C. Henmon and Peter Sandiford, devised the credit-number formula:

$$\frac{\text{Frequency}}{10} + \text{Range}".$$

[22] As quoted by Dewey, *o. c.*, p. 5.

method by which they were obtained is too involved to be giveı here".

3.451. At any rate, the *credit/index-number* total in Lindsay is 188,246. This should have been ample warning that Thorndike's figures could not have been absolute-frequency values for individual words. In that case, their total should have been 4,565,000 — the number of running words contained in all corpora on which Thorndike's first list is based.

3.452. If the relative values (in %) of *credit-numbers* consistently reflected the proportions of absolute frequencies (i. e. of total occurrences) of individual words, Lindsay's final table would no doubt very reliably indicate the real etymological composition of the English vocabulary. However, the degree in which Thorndike's *credit-numbers* are inconsistent and unrealistic, especially with top-frequency words — exactly where a misinterpretation can cause the severest quantitative distortions — can be observed best with ten most frequent words in English (according to Thorndike).

3.453. Their values, expressed in *credit-numbers,* are all within a narrow span, ranging between 204 and 211 (i. e. varying by a mere 3.3%). But if we compare them with the precise absolute frequencies from Horn's, Dewey's or Kučera-Francis's[23] word counts, the unrealiability of Thorndike's *credit-numbers* becomes obvious:

	Thorndike	Horn	Dewey	Kučera-Francis
in	211	265,531	2,116	21,341
and	210	519,583	3,280	28,852
that	209	194,645	1,345	10,595
a	208	359,119	2,120	23,237
the	208	560,601	7,310	69,971
to	208	496,776	2,924	26,149
with	208	106,784	727	7,289
be	206	147,612	846	6,377
of	205	332,710	3,998	36,411
as	204	123,469	782	7,250

[23] The top-frequency word from Horn's list is the personal pronoun *I*, which has obviously reached the top owing to the stylistically less than neutral level of Horn's corpora (correspondence). In the stylistically neutralised Dewey's word count (spanning ten stylistic levels) we meet this pronoun in the 10th place. In Kučera-Francis (15 stylistic levels; total corpus 1,014,232 words, or ten times the size of Dewey's) it ranks 20th. If we were to replace this actual top word in Horn's list with the runner-up *the* (560,601; top word in Dewey's and Kučera-Francis lists) we should find it to be 450 times greater in absolute frequency than the 1,000th member on the same list.

3.454. Whereas, as already pointed out, the difference in the absolute frequency between the bottom *(as)* and the top *(in)* members of the above Thorndike's sublist amounts to a negligible 3.3%, this difference in Horn's list is as much as 115%, with Dewey 176%, and with Kučera-Francis as high as 195%. If we entirely disregard Thorndike's hierarchy and arrange the same words according to their individual absolute frequencies, we obtain:

Horn	*(the — with)*	425%
Dewey	*(the — with)*	906%
Kučera-Francis	*(the — be)*	997%

3.455. Another drastic measure of the unreliability of Thorndike's list when dealing with top-frequency words is obtainable through a comparison of the span between the first and the 1,000th member of the list. Whereas with Thorndike the frequency of the first member (*credit-number* 211) is only 4.3 times greater than that of the 1,000th member (49), in other lists we have:

Horn	(1,367—715,130[23])	523	times	greater
Dewey	(11—7,310)	665	,,	,,
Kučera-Francis	(106—69,971)	665	,,	,,

3.456. To make a long matter short, we must agree with Dewey's words[24]: "The complex credit-number scheme by which the quantitativ data ar reported is glaringly defectiv and misleading, with respect to the most common words at least".

3.5. The next, and the most reliable, analysis, that of A. Inglis, appeared only a few months after Lindsay's study in the form of a posthumous article. Unlike Lindsay and Grinstead, Inglis based his investigation into the proportions of various etymologies in the vocabulary of English on Dewey's word count. As this count ranks its members according to absolute (i. e. actual) frequencies of individual words, Inglis's analysis of etymological proportions provides us with the first quantitatively precise statement in that area.

3.51. Inglis was acutely aware of the importance of the frequency level in any statement about the etymological composition of the English vocabulary. In fact, about half the space in his article is devoted to a painstaking interpretation of Dewey's results. This interpretation included also all the words occurring from 1 to 10 times (and accounting for 21.4% of all occurrences, but covering 91.3% of all different words). These

[24] Dewey, *ib.*

words had not been included in Dewey's list in its published form, and Inglis analysed them in manuscript.

3.52. The following table is a simplified version of Inglis's findings about the etymological proportions of English:

Table 6

Etymology	Different Words		All Occurrences	
	Total	%	Total	%
Anglo-Saxon	3,069	30.2	74,434	74.4
Latin (direct)	2,054	20.2	5,476	5.4
Latin through French	3,318	32.7	14,291	14.3
Greek	507	5.0	1,599	1.6
Scandinavian	332	3.3	2,193	2.2
All others	881	8.6	2,063	2.1
Grand Total	10,161	100.0	100,056	100.0

3.53. Based on a relatively modest sample (ca. 100,000 running words), but spanning some ten stylistic levels, Dewey's corpora, as interpreted by Inglis's analysis, enable us much better than any earlier effort to make a truly reliable, condensed statement of the etymological make-up of English:

3.531. *About one-third of the average English vocabulary is Anglo-Saxon; about one half of it is of Romance origin. But if all occurrences of each word are counted, Anglo-Saxon accounts for three-quarters of an average English text, Romance words covering only an approximate one-fifth.*

3.54. The influence of frequency levels on the etymological proportions in the English vocabulary is noticeable from the above table, but Inglis makes it still more evident with this survey of relative distribution (slightly modified by the present writer):

Table 7

Etymology	Words occurring:		
	1—5 times	51—100 times	over 100 times
Anglo-Saxon	27.2	79.0	97.8
Latin (direct)	21.9	3.2	—
Latin through French	34.2	12.4	0.7
Greek	5.1	1.8	—
Scandinavian	3.4	3.6	1.5
All others	8.2	—	—
Total	100.0	100.0	100.0

3.55. A simple statement based on this second table might run as follows:

3.551. *The Anglo-Saxon element, barely one-third in rare words, rises to four-fifths in medium-frequency words and accounts for practically all occurrences in very common words. Romance elements, on the other hand, while covering over one half of all rare-word occurrences, are reduced to one-sixth in medium-frequency words, practically disappearing with very common words.*

3.6. The last of these analysts of frequency lists in the '20s, Helen May Eddy, made use of Thorndike's list. Motivated by the needs of foreign-language teaching, she analysed the shares of various etymologies in the English vocabulary. Her aim was to determine the measure in which it would be possible to master Latin vocabulary in the American high school through learning French.

3.61. However, her method — as observable in her article "The French Element in English", *Modern Language Journal*, vol. X (1926), pp. 271—278 — was no step forward. Neglecting individual word frequencies (even in the form of Thorndike's credit-numbers), this analyst simply computed the totals of different words, thus treating all of them as equal vocabulary units and duplicating the efforts of Franklin, Ullman and Lindsay. The only difference was in her emphasis on French (as immediate source for English words) against Latin (further source). As a result, Eddy's findings were:

Table 8

Etymology		1st 5,000 wds		2nd 5,000 wds		All 10,000 wds	
Direct from	French	2,036	(40.9%)	1,833	(41.4%)	3,869	(41.4%)
„	Latin	508	(10.3%)	829	(18.7%)	1,337	(14.3%)
„	Greek	8	(0.2%)	23	(0.5%)	31	(0.3%)
„	Anglo-Saxon and other Germ. languages	2,113	(44.5%)	1,342	(30.3%)	3,455	(37.0%)
Other languages		4,925	(100.0%)	4,427	(100.0%)	9,349	(100.0%)

3.62. Following is a comparison of Eddy's findings about the principal etymological proportions of the English vocabulary with those of the other analysts of the '20s who used the same material (Thorndike's entire list) for their analyses:

Table 9

Etymology	B. L. Ullman	E. Y. Lindsay	H. M. Eddy
All Germanic languages	41.0%	42.4%	37.0%
Latin	46.8—47.5%	46.0%	14.3%
French	—	—	41.4%
Greek	6.0%	7.2%	0.3%

3.63. The differences, considerable at first sight, between Eddy's and the other two analysts' findings have, naturally, been caused by their different approach (direct, not ultimate, etymological source chosen by Eddy).

3.631. Thus, if an expectable 6—7% for Greek[25] is subtracted from the total proportion of Romance elements (which include words of ultimate Greek etymology), we are left with some 48—49% for Romance elements, and this is quite close to Ullman's and Lindsay's findings. The slight difference observable in the above table can easily be explained by the analysts' choice of various etymological authorities (dictionaries) — not specified by any of them — and by their varying criteria in the elimination of proper names from Thorndike's list.

3.7. With Eddy's analyses ends the important series from the 1920s of quantitative researches into the etymological composition of the English lexis. The success of these researches is best witnessed by the fact that for as long as sixteen years no new extensive analyses were attempted in the field, and when they did take place their object was no original approach but a mere complementation of the existing data.

3.71. The analysis in question is a three-man effort, entitled "The Sources of English Words". This brief article appeared in *The Classical Outlook,* vol. XIX (1942), pp. 45—46. It is based on Thorndike's expanded list published in 1932[26], thus covering the 20,000 most frequent words in English. It seems a pity that this effort could not have waited two more years for the appearance of Thorndike's third list,[27] which would have offered it a 30,000-word basis.

[52] Cf. the share of Greek elements, with different-word (no-repetition) approach, in tables 2, 4, 5 and 7; also Ullman's findings in p. 9.

[26] Edward L. Thorndike, *A Teacher's Word Book of the 20,000 Words Found Most Frequently and Widely in General Reading for Children and Young People,* New York (Teachers College, Columbia University), 1932.

[27] Edward L. Thorndike and Irving Lorge, *The Teacher's Word Book of 30,000 Words,* New York (Teachers College, Columbia University), 1944.

3.72. Unfortunately, the authors — W. L. Carr (first 10,000 words), Elvion Owen (second 10,000 words) and Rudolf F. Schaeffer (general supervision) — did not improve as much upon their predecessors as the sixteen-year lapse warranted. Thus, no frequency values were utilised, and the words were counted statically, as simple dictionary entries with equal individual weight. On the other hand, the issue of homography was not ignored, and a semantic count[28] was used to discern among homographs of various etymologies.

3.73. Carr et al. present their findings in five tables contracted by the present author into the following two:

Table 10

Number of Words

	Lat.	Gk.	Germ.[29]	Celt.	Misc.	Imit.	Dub.	Total
First 10,000	4,155	669	3,744	74	171	128	415	9,356[30]
Second 10,000	4,709	1,174	2,073	52	168	58	439	8,673[30]
Total	8,864	1,843	5,817	126	339	186	854	18,029
Difference	*+544*	*+505*	*—1,671*	*—22*	*—3*	*—70*	*+14*	*—683*

Table 11

Relative Distribution (in %)

	Lat.	Gk.	Germ.	Celt.	Misc.	Imit.	Dub.	Total
First 10,000	44.41	7.15	40.02	0.79	1.83	1.37	4.43	100.00
Second 10,000	54.30	13.53	23.90	0.60	1.94	0.67	5.06	100.00
Total	49.16	10.22	32.27[31]	0.70	1.88	1.03	4.74	100.00
Difference	*+9.89*	*+6.38*	*—16.12*	*—0.19*	*+0.11*[32]	*—0.70*	*+0.63*	

[28] Irving Lorge and Edward L. Thorndike, *A Semantic Count of English Words*, New York (Teachers College, Columbia University), 1938.

[29] Germanic (including Anglo-Saxon, Scandinavian, Dutch, German, etc.).

[30] Instead of 10,000, because proper names, etc. have been omitted. (Cf. notes 8, 9 and 19)

[31] The original figure here (37.76) was a mistake in the calculation of percentages discovered and replaced by the present author.

[32] A seeming paradox, since the difference in this column of the preceding table is negative (—3). Here, however, we have to do with two different aspects. The absolute distribution (number of words) figures indeed record a decrease, but the relative distribution values — indicating the share in the more severely reduced total of the second 10,000 words — correctly show actual increase in the proportion of these etymologies.

3.8. This was followed by another long lapse of time before the next analysis of etymological proportions in English based on frequency lists. Twelve years were to pass before F. G. Cassidy used the composite frequency list by Fawcett and Maki[33] in his revision of S. Robertson's book *Development of Modern English* (New York, 1954, p. 155).

3.81. Cassidy's effort was however, very modest. He only investigated the etymological composition of the 1,000 top-frequency words in English, as presented by Fawcett and Maki. His results were:

	Anglo-Saxon	French	Latin (Direct)	Scand.	All Others
Percentages	61.7	30.3	2.9	1.7	2.9

4.0. Cassidy brings us to the present period and the first attempt to use computers in analysing frequency lists for their etymological make-up. This was part of the impressive effort by A. Hood Roberts, entitled *A Statistical Linguistic Analysis of American English* (The Hague, 1965, 437 pp.).

4.1. However, describing Roberts's results — a nesessarily lengthy analysis — would take us beyond the scope of this article. The more so as another important, and more recent, effort by Francis and Kučera,[34] with their extensive and reliable computer-compiled frequency list, also invites discussion and utilisation in further analyses of the etymological composition of the English vocabulary (notably in the above-10,000 frequency ranges). The present author hopes he may turn his respectful attention to these latest efforts in the near future.

[33] L. Fawcett and Itsu Maki, *A Study of English Word Values*, Tokio, 1932.

[34] Henry Kučera and W. Nelson Francis, *Computational Analysis of Present-Day American English* (Providence, 1967; 424 pp.). Cf. note 9.