WORD DENSITIES

REX GOOCH
Welwyn, Herts, England

INTRODUCTION

Following Charles Bostick in the May 1974 Word Ways, we define the weight of a letter as its position in the alphabet (A=1, etc.), and the weight of a word as the sum of the weights of its letters. The density of a word is defined as its weight divided by the number of letters in it, so that the density of a word is none other than the average weight of letters it contains. Bostick's article challenged readers to find the lightest, heaviest, densest and least dense words (and other categories). Prior to this, Darryl Francis in November 1972 had listed the most dense and least dense words of lengths up to 15 letters, calling the article "Lightweights and Heavyweights". In the August 1995 Word Ways, Leonard Gordon extended the work to 24 letters, improving on two of Darryl's examples. In the November 1996 issue, Susan Thorpe provided 13 less dense examples and 12 denser ones.

To dispose of the first challenge quickly, readers will be familiar with the fact that A (weight 1) is the lightest word, followed by AA or B (weight 2), then AAA or AB or BA or C (weight 3), and we will not waste space arguing which of these are truly words. The heaviest word is "P45", listed below and weighing in at 560, followed by CONJUNCTIVO-DACRYOCYSTORHINOSTOMIZING (Stedman) at 490; many chemical and medical terms follow.

Turning to the second challenge, that of dense words, this article extends the list beyond length 24, improving upon quite a number of the previous champions, and offering a wider range of types of words, including a few 2-word expressions. However, the main purpose is to probe a little deeper, and ask what one might expect of densities.

Before leaving the subject of weights, it would seem much more reasonable to assign a weight based on how heavy a letter actually is. I seem to recall that on a proportionally-spaced typewriter, a W takes up 5 units of horizontal space, and an I, 2; all other letters have a value within this range. In this article I stay with established practice.

(Editor's note: the concept of letter-thickness was exploited in the August 1987 Word Ways by Donald Knuth, who discovered that the thickness of letters in running text often offered sufficient clues for the decipherment of those texts.)

WORDS OF GREATEST AND LEAST DENSITY FOR A GIVEN LENGTH

The following lists give a number of examples of the most and least dense words for each length. In part this is to permit the reader to select words to form their own lists, but it is also intended to indicate how exceptional (or not) the champion words are. This theme will be picked up again in the discussion which follows the list.

Words of length 32 and greater have been listed at the end without distinction as to least and greatest density, as there are so few words of these lengths that separate listings would lead to duplication.

Hyphens have been inserted only when I believe they must be present, although other words may also need to be hyphenated.

I have a problem with Z, ZZ, etc. In Collins COBUILD Bank of English (a modern corpus of English writings), ZZZZ appears 24 times, ZZZZZ 13 times, ZZZZZZ 9 times, and so on up to 13 Zs (from a review in the New Musical Express). Nevertheless, I have in general eschewed their use.

Clearly, any rearrangement of letters in a word (e.g., ABACA and CAABA) makes no difference to its weight or density.

Unless otherwise indicated, all words listed appear in Webster's Second, the Oxford English Dictionary, Stedman's, American Heritage or Pulliam and Carruth's Complete Word Game Dictionary. Labels include: ATHS=American Thesaurus of Slang; BIW=Wordsworth Book of Intriguing Words, by Paul Hellweg; Chamb=Chambers 20th Century Dictionary; Dorland=Dorland's Medical Dictionary; DFPF=Dictionary of Flowering Plants and Ferns; IM=Index Medicus (National Library of Medicine); LV=Language on Vacation, by Dmitri Borgmann; MI=Merck Index (7th Edition); NZ=Nomenclator Zoologicus; RI=The Ring Index (2nd Edition); Web 3=Webster's Third; Wyy-ppp=Word Ways year and page.

Least Dense Words

Length Density, word...

Length	Density, word...		
1	1.000 A	2.000 B	3.000 C
2	1.000 AA	1.500 AB	1.500 BA
3	1.000 AAA	1.333 ABA	1.333 BAA
4	1.000 AAAA (W92-162)	1.250 AAAB (W81-115)	1.250 AABA (NZ)
5	1.400 ABABA (NZ)	1.600 ABACA	1.600 CAABA
6	1.667 BACABA	2.333 ABBACE	2.500 BACCAE
7	2.429 ABEBAEA (NZ)	2.714 FABACEA (inferred singular)	2.857 CACHACA
8	3.000 FABACEAE	3.125 BABAJAGA	3.125 CABBAGED
9	1.778 ABBADABBA (W78-212)	3.222 BABE-FACED	3.333 BECCACCIA
10	3.200 CABBAGE-BED (OED)	3.900 FACE-BEDDED (LV)	3.900 GALACACEAE

11 3.545 CABBAGEHEAD — 4.091 DACCA BANANA? — 4.182 AECIDIACEAE
12 3.333 CABBAGEFACED (W74-116, hyp?) — 4.667 BIBBLEBABBLE — 4.750 GABBIE-LABBIE (W80-250)
13 3.692 CABBAGEHEADED (W72-226) — 5.077 ANACARDIACEAE — 5.462 BACKACHEBRAKE?
14 5.571 GAIADENDRACEAE (DFPF) — 5.786 HAMAMELIDACEAE — 5.857 BACILLARIACEAE
15 6.067 ECDEIOCOLEACEAE (DFPF) — 6.133 CHEMICAL BALANCE — 6.267 DICHAPETALACEAE
16 6.500 CABBAGELEAF MINER? — 6.563 HALFHEAD BEDSTEAD? — 6.563 LACTOBACILLACEAE(W95-154)
    6.750 ACETOBACTERACEAE? — 6.750 FIDDLE-COME-FADDLE
17 6.941 ABUNDANCE DECLAREE? — 7.000 BRACHIOCEPHALICAE — 7.118 ARCHIDIDASCALIANE?
18 6.167 BLACKFACED HIGHLAND — 7.111 FIDDLEBACK CHASUBLE — 7.611 ACHROMOBACTERACEAE
    7.611 LEAKAGE COEFFICIENT
19 7.526 PALAEACANTHOCEPHALA (W95-154) — 7.789 HELMINTHOCLADIACEAE
    7.842 CHLAMYDOBACTERIACEA?
20 7.700 CHLAMYDOBACTERIACEAE — 8.050 FACIOCEPHALALGICALLY — 8.200 ABRACHIOCEPHALICALLY
21 8.619 CHEIROBRACHIALGICALLY — 8.619 PERICARDIACOPHRENICAE — 8.810 CARBOANGIOCARDIOGRAPH
22 8.136 TACHYCARDIABRADYCARDIA — 8.455 CINEANGIOCARDIOGRAPHED — 8.591 CINEANGIOCARDIOGRAPHIC
23 8.435 CARBOANGIOCARDIOGRAPHED — 8.522 TETRACAIDECADELTAHEDRON (W68-109)
    8.565 CARBOANGIOCARDIOGRAPHIC
24 8.875 ARGININOSUCCINICACIDEMIA — 8.958 CARBOANGIOCARDIOGRAPHING
    9.000 CARBOANGIOCARDIOGRAPHICS
25 9.440 CHOLANGIOPANCREATOGRAPHED — 9.520 SACCHAROGALACTORRHEICALLY
    9.560 CHOLANGIOPANCREATOGRAPHIC
26 9.192 CINEANGIOCARDIOGRAPHICALLY — 9.769 ETHYLENEDIAMINETETRAACETIC
    9.885 CHEMOPALLIDOTHALAMECTOMIES — 9.885 CHOLANGIOPANCREATOGRAPHING
27 9.148 CARBOANGIOCARDIOGRAPHICALLY — 9.852 MAGNETOENCEPHALOGRAPHICALLY
    9.926 CHEMOPALLIDOTHALAMECTOMIZED — 9.926 ETHYLENEDIAMINETETRAACETATE
28 10.107 HEMANGIOENDOTHELIOBLASTOMATA — 10.214 ARGININOSUCCINICACIDURICALLY
    10.250 CHOLEDOCHOCHOLEDOCHOSTOMIZED
29 9.655 FLOCCINAUCINIHILIPILIFICATION — 9.966 CHOLANGIOPANCREATOGRAPHICALLY
    10.172 PARADIMETHYLAMINOBENZALDEHYDE
30 10.633 STEREOELECTROENCEPHALOGRAPHING — 10.667 STEREOELECTROENCEPHALOGRAPHICS
    10.733 STEREOELECTROENCEPHALOGRAPHIES
31 10.194 ADENOIDALPHARYNGEALCONJUNCTIVAL — 11.097 POLIOENCEPHALOMENINGOMYELITIDES
    11.194 DICHLORDIOXYDIAMIDOARSENOBENZOL

Most Dense Words

Length Density, word...

1 26.000 Z — 25.000 Y — 24.000 X
2 25.500 ZY — 24.000 WY — 24.000 XX? — 24.000 YW
3 26.000 ZZZ? — 25.000 XYZ (W79-218) — 25.000 ZYX (NZ, flea) — 25.333 YZY
4 26.000 ZZZZ (ATHS) — 23.750 VUZZ — 23.750 ZYXT
5 26.000 ZZZZZ (W70-252) — 25.400 ZZYZX, California — 24.200 WUZZY — 23.600 TUZZY
6 25.333 ZYZZYX (NZ) — 21.333 XYSTUS — 21.167 SYZYGY — 21.167 TRYSTY
   21.167 TRYYST — 21.167 YYYEZU (USSR)
7 21.571 ZYZZYVA — 21.286 ZYZOMYS? — 20.571 TWYNRYS
8 24.250 ZYZZYZUS (NZ) — 21.375 TIZZWOZZ — 21.250 ZYZZYVAS (plural)
9 20.778 UPSY-TURVY? — 19.444 SYSTYLOUS — 19.444 ZOROTYPUS (W72-227)
10 22.900 TUZZY-MUZZY — 22.500 FUZZYWUZZY — 21.300 TUZZIMUZZY?
   20.900 FUZZYGUZZY? — 20.500 SPYTUWYSLY
11 19.273 TWISTY-WISTY? — 18.818 TRUSTWORTHY — 18.545 TORTUROUSLY
12 19.250 TOOTSYWOOTSY — 18.917 VOLUPTUOUSTY (sic) — 18.500 TOPSYTURVILY (Chamb)
13 19.154 TOPSYTURVYIST? — 18.615 TOPSYTURVYISM? — 18.615 UNTRUSTWORTHY
14 18.214 UNTUMULTUOUSLY (inferred) — 17.643 CRYPTOZYGOUSLY — 17.571 TRUST TERRITORY
15 17.533 UNTRUSTWORTHILY — 17.000 POTRUSPYTOPHILY (W92-062) — 16.933 254 PROTOSTRONGYLUS?
16 17.250 UNPRESUMPTUOUSLY — 17.000 SYRINGOSYSTROPHY? — 16.875 QUANTITY SURVEYOR
17 17.000 QUANTITY SURVEYORS — 16.765 HYDROXYISOBUTYRYL — 16.647 UNTRUSTWORTHINESS
18 16.278 SOUTH-SOUTH-WESTERLY — 16.111 TYPHLOURETEROSTOMY — 16.000 POSTCOMMISSUROTOMY
19 16.263 HYPOPHYSEOPRIVOUSLY (W95-152) — 15.737 NITROSOSUBSTITUTION
   15.684 PSEUDOPOLYDYSTROPHY
20 15.650 PROSTATOCYSTOTOMIZES — 15.500 URETEROCYSTONEOSTOMY
   15.500 URETEROCYSTOSTOMIZES — 15.500 URETERONEOCYSTOSTOMY

21 15.571 HYDROXYPHENYLPYRUVATE
15.238 HYDROXYKYNURENINURICS

15.429 TYPHLOURETEROSTOMIZES

22 15.000 TYPHLOURETEROSTOMIZING
14.909 DACRYOCYSTOSYRINGOTOMY

15.000 URETEROURETEROSTOMIZES

23 14.957 URETERONEOCYSTOSTOMIZES
14.609 URETEROURETEROSTOMIZING

14.913 URETEROPYELONEPHROSTOMY
then 3 more URETEROs!

24 15.333 TRANSURETEROURETEROSTOMY
14.667 COLPOURETEROCYSTOTOMIZES

14.667 COLPOCYSTOURETEROTOMIZES

25 14.760 URETEROTRIGONOENTEROSTOMY
14.320 COLPOURETEROCYSTOTOMIZING

14.320 COLPOCYSTOURETEROTOMIZING
14.320 URETEROPYELONEPHROSTOMIZE

26 15.385 COLPOCYSTOURETEROCYSTOTOMY
14.500 URETEROPYELONEPHROSTOMIZES

14.731 TRANSURETEROURETEROSTOMIZE

27 14.889 TRANSURETEROURETEROSTOMIZES
14.333 TRANSURETEROURETEROSTOMIZED

14.444 HYDROXYDESOXYCORTICOSTERONE (Web3)

28 14.571 TRANSURETEROURETEROSTOMIZING
13.964 URETEROILEONEOCYSTOSTOMIZING

14.393 URETEROTRIGONOENTEROSTOMIZES

29 14.103 URETEROTRIGONOENTEROSTOMIZING
13.586 CONJUNCTIVODACRYOCYSTOSTOMIES

13.828 CONJUNCTIVODACRYOCYSTOSTOMIZE
13.414 TRINITROPHENYLMETHYLNITRAMINE (Web3)

30 11.333 ETHANOLAMINEPHOSPHOTRANSFERASE
13.500 CONJUNCTIVODACRYOCYSTOSTOMIZED

14.000 CONJUNCTIVODACRYOCYSTOSTOMIZES
13.500 HYSTEROSALPINGOOOPHORECTOMIZES

31 13.742 CONJUNCTIVODACRYOCYSTOSTOMIZING
13.452 CHOLECYSTOJEJUNOCHOLECYSTOSTOMY (W84-032)

13.516 PSEUDOPSEUDOHYPOPARATHYROIDISMS

Least and Most Dense Longer Words

32 10.594 ENCEPHALOMYELORADICULOPATHICALLY
12.594 COUNTERIMMUNOELECTROPHORETICALLY
14.063 CONJUNCTIVODACRYOCYSTORHINOSTOMY

12.156 ARTHROPNEUMOROENTGENOGRAPHICALLY
12.781 PROSTATOSEMINALVESICULECTOMIZING

33 10.636 STEREOELECTROENCEPHALOGRAPHICALLY
11.485 ULTRAANTIDISESTABLISHMENTARIANISM (BIW)
13.152 LAPAROHYSTEROSALPINGOOOPHORECTOMY

11.121 TETRADECAMETHYLCYCLOHEPTASILOXANE (MI)

13.758 TETRAETHYLMONOTHIONOPYROPHOSPHATE

34 11.147 SUPERCALIFRAGILISTICEXPIALIDOCIOUS
11.559 DIAMINOPROPYLTETRAMETHYLENEDIAMINE (W95-8)
13.471 CONJUNCTIVODACRYOCYSTORHINOSTOMIES

11.382 PSEUDOANTIDISESTABLISHMENTARIANISM (BIW)

13.676 CONJUNCTIVODACRYOCYSTORHINOSTOMIZE

35 12.629 LAPAROHYSTEROSALPINGOOOPHORECTOMIES
13.400 CONJUNCTIVODACRYOCYSTORHINOSTOMIZED

12.829 LAPAROHYSTEROSALPINGOOOPHORECTOMIZE
13.829 CONJUNCTIVODACRYOCYSTORHINOSTOMIZES

36 12.583 LAPAROHYSTEROSALPINGOOOPHORECTOMIZED
13.611 CONJUNCTIVODACRYOCYSTORHINOSTOMIZING

13.000 LAPAROHYSTEROSALPINGOOOPHORECTOMIZES

37 11.892 HEPATICOCHOLANGIOCHOLECYSTENTEROSTOMY (BIW)
12.811 LAPAROHYSTEROSALPINGOOOPHORECTOMIZING

38 10.737 DIKETOHYDRINDYLIDENEDIKETOHYDRINDAMINE
11.105 GALACTOSYLGALACTOSYLGLUCOSYLCERAMIDASE (IM)

39 11.487 HEPATICOCHOLANGIOCHOLECYSTENTEROSTOMIES (Dorland)

41 10.805 TETRADECAHYDROTETRAZOLOAZACYCLOHEXADECINE (RI)
12.366 COMICONOMENCLATURIST DISPROPORTIONABLENESS (BIW)

45 12.444 PNEUMONOULTRAMICROSCOPICSILICOVOLCANOCONIOSIS ("factitious")

DISCUSSION

Graph 1 shows the least and most dense words of a given length (from the list above) plotted against length. As the length of the word increases, it becomes more and more difficult to find words consisting largely of the first or last few letters of the alphabet. In fact, the two lines converge to about 12.5, which compares with the average letter value of 13.5. This is even true at the extreme right, where there are few words to choose from. The implication seems to be that long words contain an almost random mix of letters, or at least letter weights.

For a given word length, the graph shows that the average of the least dense word and the most dense word of a given length is always about 12 or 13, which suggests that the distribution of word densities for a given word length is not terribly skewed. (If the most dense word were very exceptional, the average of it and the least dense would be well above the average density.) To test this idea, Graph 2 shows the distribution of densities for all 7-letter words (7 being chosen merely to avoid extremes): this distribution is symmetrical, and shows no long right tail, as may have been expected. Why should this be so? After all, a glance at any dictionary will show that there is a very uneven distribution of initial letters, with the Ss being very common. This distribution of initial letters is shown in Graph 3, which also shows the very different distribution of letters in words as a whole. As is the case with running text, the most common letter in (7-letter) dictionary words is the letter E (but note the next most common letters are not those seen in running text). Now densities arise from multiplying the weight of each letter by its frequency; as it is difficult to do this in one's head, the result is shown in Graph 4. The high frequency of E is offset by its low weight, and the low frequency of Y is boosted by its high weight. The consequence is that no single letter dominates the density, and none contributes as much as 12 per cent in average word density. Although we can scarcely call the graph random, it is more moderate than either weight or frequency alone, and helps to explain our results, which we now summarise.

SUMMARY

In general, the maximum and minimum densities for words of a given length are closely predictable. This is because multiplying the weight of a letter by its frequency has a moderating effect upon variability of letter weights. In addition, the densities of words formed from such letter weights follow a distribution much as if words consist of random letter weights, with no extreme outlying points. Obvious exceptions are the peak at length 8 in the upper line of Graph 1, which is due to the extraordinary word ZYZZYZUS, and at length 9 in the lower line, caused by ABBADABBA.
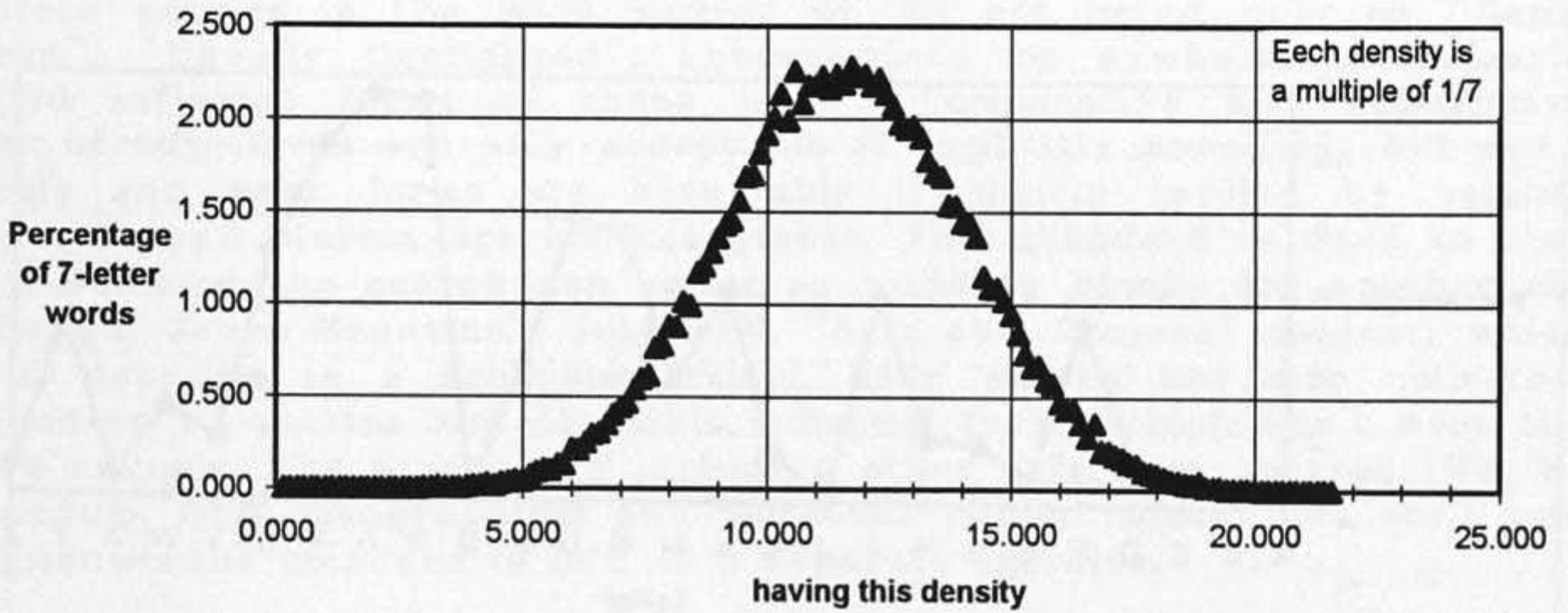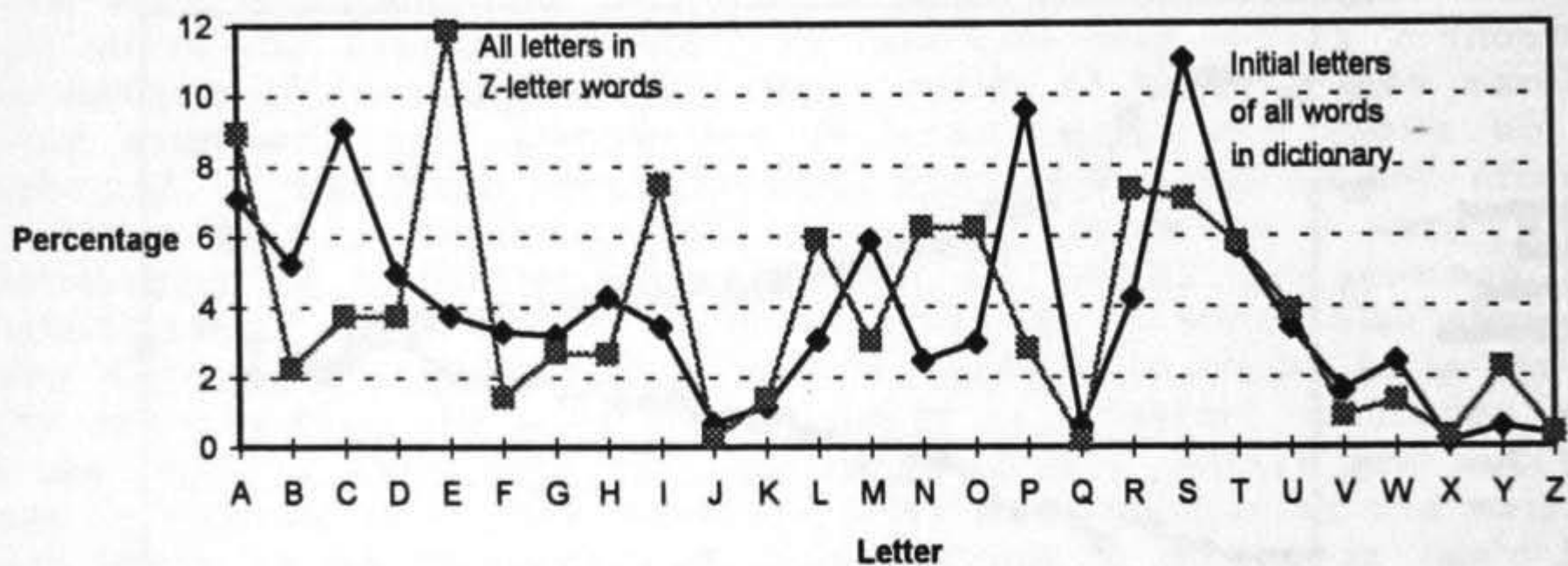
ACKNOWLEDGEMENTS

Graph 1: Highest and lowest densities of words by word length
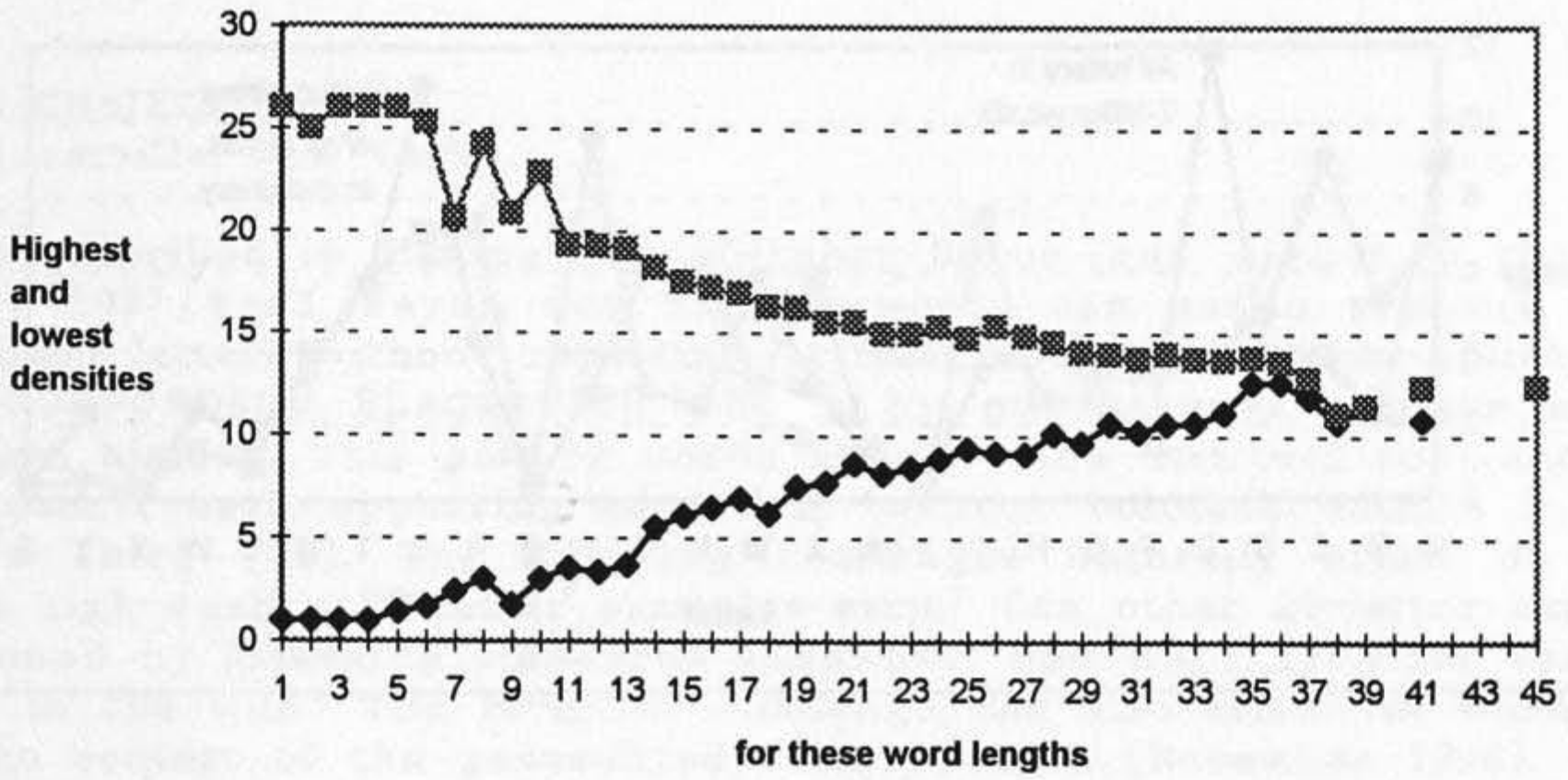
Graph 2: Densities of 7-letter words

Graph 3: Percentage of letters

Graph 4: Percentage contribution of each letter to the densities of 7-letter words