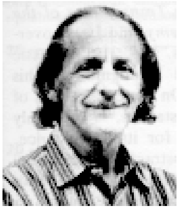
Essays of an Information Scientist, Vol:9, p.208, 1986
Current Contents, #27, p.3-8, July 7, 1986



Current Comments®

EUGENE GARFIELD
INSTITUTE FOR SCIENTIFIC INFORMATION®
3501 MARKET ST., PHILADELPHIA, PA 19104

ISI's Master List of Title Words Provides a Special Perspective on Science and Scholarly Activity. Part 1. The Lexicography of the Unique Word Dictionary

Number 27 July 7, 1986

In *George Orwell*'s *1984*[1] the government used *telescreens* to monitor people's thoughts and actions. The quality of everyday life in *Orwell*'s vision of the future was dismal; *individuality* was suppressed. If *Orwell* were alive today he might be confused to learn that in 1984 *ecstasy*, *freedom*, and *liberalism* increased while *disillusionment* decreased. He'd be less surprised that the use of *telescreens*, now called video display terminals (*VDTs*), also increased, although they are used for different purposes than those he envisioned. *Telescreens* help make *databases*, accessed through *modems*, available to the public. By *downloading*, *computer* users can then transfer data to their own personal *computer* memories.

Actually it was the number of times that these words were used that increased or decreased. Terms such as *modem* and *downloading* and dozens of others are unique to our age. The measure of their uniqueness is indicated by the number of times they appeared in the titles of articles indexed in the 1981 and 1984 *Science Citation Index*® (*SCI*®), *Social Sciences Citation Index*® (*SSCI*®), and *Arts & Humanities Citation Index*™ (*A&HCI*™). By examining the titles of articles indexed in these volumes, we can observe the rise and fall of various terms. Indeed, words, like citations, can be harbingers of change. But they should not be used as anything but elementary indicators of etymological evolution.

One reason for this caution is the difficulty of separating the many meanings of homonyms in language. Even a term like *ecstasy* has more than one definition. For example, *ecstasy* is a state of overwhelming emotion, especially rapturous delight.[2] (p. 395) But it is also the slang name for a mild psychedelic (3,4-methylenedioxymethamphetamine, or MDMA) used in conjunction with psychotherapy to encourage the expression of emotions.[3]

The Unique Word Dictionary

In this two-part essay we discuss ISI® 's Unique Word Dictionary (UWD), a computer file of words that occur in the titles of articles indexed in the *SCI*, *SSCI*, and *A&HCI* databases. In Part 1 we explain the mechanics of creating the UWD and list over 200 terms that appeared in it in 1981 and 1984. In Part 2 we will describe in detail the frequencies of several of these words. We will also examine terms used in a selected group of 1984 ISI research fronts.

The main function of the UWD, or master dictionary, is simply to provide an edited, or standardized, version of the source data we index in the *SCI*, *SSCI*, and *A&HCI* each year; it permits us to check the data for accuracy. It is, however, only one of several verification processes that we use. For example, every title is separately keyed by two different indexers. Each version is then matched against the other to check for keying errors. The UWD provides us with another filter for incoming source data. In addition to checking for keying errors and misspellings again, words new to the ISI database are also identified. Various permutations of these data are then used to create more comprehensive lists of words and word pairs such as those included in the *Permuterm® Subject Index* (*PSI*), described later.

Before any words are included in the UWD, however, our editors must verify their spelling and standardize any atypical forms to agree with accepted American usage. Our editors also check that each new term is in fact a real word and not just a misspelling.

Some words, such as *antitumor*, an adjective describing a substance or agent that counteracts *tumor* formation, are simply new combinations of "old" words. Others are brand-new terms or neologisms specifically coined to describe new specialties or phenomena. Scholarly and scientific articles abound with such terms. For example, *informatics*, which appeared 60 times in 1981 titles and 92 times in 1984 titles, was coined by A.I. Mikhailov, director of the All-Union Institute of Scientific and Technical Information (VINITI), USSR, and describes "the scientific discipline which studies the structure and general properties of scientific information and the laws of all the processes of scientific communication."[4] However, the meaning of *informatics* still varies among users in different countries because the field is relatively young.[4] And the term *medical informatics* has now gained wide popularity.

The offhand, ad hoc words we create to describe our own tools or ideas can sometimes cause great frustration for etymologists. The so-called "Unique Word Dictionary" is, itself, one such absurd creation. All those at ISI who participated in the creation of this computer file know what it is, but to the outside world the name may sound absurd because the UWD is not a dictionary and the terms it contains are not unique. Even if it were a dictionary in the usual sense, containing definitions of words, it would still be rather redundant to call it a word dictionary. There are, of course, many different kinds of dictionaries in addition to the traditional sources we use every day. For example, the *Thorndike-Barnhart Comprehensive Desk Dictionary* contains over 80,000 words chosen by counting over 30,000,000 words of text in every field of general interest. These 80,000 terms constitute 99 percent of the words used in most written material with the exception of the very technical terms used in textbooks.[5] Of course, the *Thorndike-Barnhart* also includes word definitions, origins, pronunciations, synonyms, and spellings.

Space restrictions make it impossible for us to retain in our dictionary every word that appears in the annual indexes. So, at the end of each year we purge those words that have occurred fewer than three times. This reduces the file by nearly two-thirds. We then use this smaller, purged file as the basis for the following year's dictionary. It isn't likely that we lose any words of major importance by purging the UWD, since, in general, we assume that if a word is significant it will appear more than two times in a given year. And we don't eliminate from our indexes words that only occur once or twice. We simply index them as they appear in the original title.

Various studies have demonstrated that the words used most often in the written language constitute only a small percentage of the vocabulary. For example, a total of 260,430 words appeared in

James Joyce's *Ulysses*, but half of these were drawn from a group of just 135 unique words including *and*, *of*, and *the*. In addition, of the 260,430 total words, there are only 29,899 unique terms.[6] In 1949 George Kingsley Zipf, Harvard University, Cambridge, ranked these 29,899 words in descending order according to the number of times they occurred and found that multiplying the numerical value of each word's rank by its corresponding frequency gave him a product that was constant throughout the entire list of words.[7,8] Another example of the way we unconsciously limit our word choices is demonstrated in William Shakespeare's works. Computer counts have determined that he used 29,066 unique words in his plays but only 40 of these make up 40 percent of the occurrences.[6]

Although we include commonly used terms such as conjunctions, articles, and prepositions in the UWD, we don't record their frequencies because they simply appear too often in titles every year. And we exclude systematic chemical names from the UWD because they would inflate it each year by several hundred thousand entries.

The UWD is actually composed of several files: a short file listing words with 12 or fewer letters, a long file containing 13- to 30-letter words, a cross-reference file that includes the variant-to-preferred spelling of words, and a file consisting of two-element terms created by our editors. In 1981 these files combined contained 448,140 terms, and in 1984 they had 476,788 entries. Table 1 lists 272 words common to both years' files and includes the number of times the terms appeared in 1981 and 1984 titles. We selected these two years to demonstrate change over a four-year period. (When we began this study, 1985 data were not yet available.) Two earlier, similar studies published in *Current Contents*® (*CC*®)[9,10] contrasted title-word frequencies from 1973 and 1976 and from 1976, 1979, and 1980. Any terms from those studies that appear in Table 1 are designated by an asterisk.

Compound Words

The hyphenated compounds that are listed in the short and long UWD files are those that were hyphenated by the authors of the articles. Seventeen hyphenated compounds appear in Table 1. Several would not be hyphenated if used as nouns, according to *Webster's Ninth New Collegiate Dictionary*.[2] But some, such as *fiber optics*, should be hyphenated if used as adjectives placed before nouns if a nonhyphenated construction might be ambiguous and prove confusing to the reader. If used in an unambiguous fashion, however, hyphens are not needed.[11]

Other hyphenated compounds, created by our editors by linking together related title words, are listed in the word-phrase file of the UWD. These sometimes arbitrary decisions about which form of a compound to include are made when the words are added to the dictionary. This is not always an easy process. ISI, like others who deal with problems of lexicology, must take into account the frequency of a word's use when deciding whether to include it. According to the editors of *Webster's Ninth New Collegiate Dictionary*, such decisions should be made only after examining several different examples of the word's use in citations that span a specified period of time and that appear in a wide range of publications. But "there is no magic number [of occurrences] that guarantees entry and no particular span of years that must be reached. To a great extent the judgment made here must rest on...insight and experience...."[2] (p. 29)

Of course, we cannot choose to add or eliminate hyphens in words such as *re-form* and *reform*. But if *antitumor* oc-

Table 1: Selected terms from the Unique Word Dictionary (UWD) and the number of times they appeared in the titles of 1981 and 1984 source items indexed by ISI®. Where indicated, all word forms and spelling variants are included in the counts for each term. An asterisk (*) indicates that the word appeared in a previous UWD study.

	1981	1984		1981	1984		1981	1984
AACR/2/II	63	5	*calmodulin	407	469	fractal/s	21	169
ABM/S	6	17	China	1036	1177	freedom	536	614
abortion	465	320	chiral	472	730	GaAs	732	1089
acetaminophen	131	138	cholesterol	1727	1271	GABA	508	476
acidification	146	191	cinema	344	397	gang/s	18	25
acquired	491	1065	*clone/s/ing/ed/al	1645	2258	gasohol	33	4
acyclovir	78	163	COBOL	43	14	*gene	2531	3808
Ada	86	133	*cocaine	93	132	Giacometti, Alberto	3	4
aerobic/s	283	316	coffee	200	219			
*aerosol/s/ized	995	837	*cognition*	272	266	*graffiti*	6	16
AI	36	60	comet	44	111	grammar	254	358
AIDS	453	1116	competition	937	971	GUT	354	337
*algorithm/s/ic	2366	2464	computer/s	5165	7086	*hadron/s	179	186
allergen/s	233	183	conservative	309	355	Haiti/an	51	53
allergy/ies	570	515	crowd/ed/ing	113	116	Hall-effect	7	9
*alpha-fetoprotein	342	232	cytomegalovirus	399	974	Halley/'s	24	50
Alzheimer/'s	175	330	database	436	696	halogen/s	163	178
*amniocentesis	70	87	daycare	19	30	harmony	77	78
amorphous	1313	1719	defense	959	1318	*herpes	850	909
anorexia	267	316	deficit	228	291	Higgs	88	110
anthelmintic/s	137	82	depression	1401	1498	homeless	13	25
antibody	2392	2718	dexamethasone	297	431	hominid	49	49
antigen/s	4939	4882	digital-analog	4	6	*homosexual/s/ity	245	308
antitumor	621	673	dioxin	25	55	hostage	55	11
anxiety	564	628	disillusionment	15	3	HTLV	4	60
apartheid	27	40	dissonance	15	27	hydroponic	12	11
apheresis	3	59	diversification	82	112	immunoassay	620	678
arrhythmias	596	813	divestiture	3	7	immunodeficiency	209	713
asbestos	340	307	*DNA/ deoxyribonucleic	5715	5585	individuality/ism/ ist/istic	113	79
autoimmune	483	530						
ballistic	61	89	downloading	1	19	inflation/ary	537	421
BASIC	2312	1734	drought	133	158	informatics	60	92
Beckett, Samuel	23	84	duplex	91	116	information	4991	5143
benzodiazepine/s	421	613	ecstasy	10	27	inositol	43	81
beta-blocker	41	78	electrophoresis	778	570	interfacing	71	95
bioengineering	61	23	*endorphin/s	175	65	*interferon	1101	998
*biofeedback	242	168	enzyme	3373	2994	interleukin/ -1/-2/-3	151	692
biotechnology	142	344	ergonomic/s	99	202			
boom	72	120	Ewing/'s	51	56	ion-beam/s	42	46
boson	109	164	*famine	54	43	irradiation	1803	1890
bulimia	15	95	fatigue	1020	1338	Kaluza-Klein	2	85
burnout	110	90	*fiber-optic/s	368	282	Kaposi/'s	58	188
bypass	937	960	fluoridation	71	33	keratin/s	106	195
caffeine	253	340	FORTRAN	142	77	lasers	884	1106

curs in one title, and *anti-tumor* in another, how should this word appear in the UWD? Some experts might expect to find it listed under *tumor*, others under *anti*, while perhaps it should be listed under *antitumor*. Generally, it is impossible to standardize every word variant even within a single issue of *CC*. Only when we have assembled a number of possibly conflicting uses can we make an arbitrary decision. But this is complicated by the fact that librarians and others aspire to all-inclusiveness and standardization in what is essentially an endless task.

One place where we list all possible pairings of key title words is in the *PSI*, one of the indexes included in each year's *SCI*, *SSCI*, and *A&HCI*. This index includes every significant title term, hyphenated or otherwise, that appeared in the year covered by that particular volume. Each term is permuted with other terms in the same article title to

	1981	1984
*lemming/s	19	13
lemon	22	28
*lepton/s	136	121
leukotriene/s	90	233
liberalism	101	127
light-scattering	59	44
lipoprotein/s	1677	1344
love	488	515
lumpectomy	1	9
lupus	1099	1137
lymphoma	1025	1176
machismo	3	5
mainframe/s	24	91
malaria	477	388
malathion	59	79
medfly	13	3
Medline	3	7
melanoma	901	1019
meson	95	115
metonymy	5	7
metric	312	217
micelle	121	110
microwave	1147	1089
modem/s	31	106
monoclonal	1491	3541
*monopole/s	144	317
Monte-Carlo	367	494
morality	190	230
MPP	1	5
MPTP	0	32
MTV	2	4
multimedia	22	49
mutagenesis	351	309
MX	35	24
myc	0	43
myelitis	15	27
naphthol/s	11	9
naturalism	46	46
*neonate/s	666	580
networking	39	100
neuroleptic/s	316	306
neurolinguistic/s	11	16
*neuropeptide/s	136	280
neutrality	49	65
neutrino/s	311	315
NMR	2490	2948
nociception/ive	72	112
oligoglycosides	6	5
olive/s	78	111
oncogene/s	28	394
online	1009	1201
open-heart	41	48
orange	185	224
Orwell, George	28	95
Orwellian	1	8
osteoporosis	144	206
ozone	464	380
panda/s	9	24
pascal	142	125
pasta	8	19
patriarchy	26	35
PCB	51	91
peptide	1135	1309
*pharmacokinetic/s	1705	1772
phenomenology/ ical	356	332
phorbol	178	253
phosphorylation	925	964
piracy	13	31
plaque	535	382
plasmapheresis	186	240
pluralism	96	118
PMS	5	18
polyacrylamide	254	242
polymer/ization	4527	5029
polymorphism	509	548
postmodern/ism	17	34
prayer	25	58
prolactin	1483	1017
PROLOG	1	80
psychoneuro-immunology/ic	4	3
QCD	216	282
rain	373	522
Reaganomics	16	25
realism	213	264
*recombinant	221	439
retinoid/s	293	206
retrovirus/es	195	215
Reyes	72	68
*RNA/ribonucleic	3528	2324
robotic/s	63	388
Salle, David	3	7
sarcoma	836	766
SATCOM/s	7	71
schistosomiasis	173	168
schizophrenia/ic	973	722
SDI	4	12
*self-help	94	99
semiconductor	817	1094
semiotic/s	163	208
*sexism	50	27
shuttle	160	348
signifier	4	8
silicon	2143	3030
software	1088	1917
solar	3282	2932
solar-wind	2	5
solidarity	55	71
somatization	10	16
sonography	210	258
spin-glass/es	69	105
squatter/s	11	21
steroid/s	1903	1559
string/s	203	295
structuralism	44	86
success/ful	1184	1566
suicide	370	430
*superconductor/ s/ing	721	607
supergravity	68	220
supersymmetry	43	208
syndrome	6309	6463
*T-cell/s	2160	2791
technocrat/s/ic	8	18
telescreen	3	0
tenure	60	75
*terrorist/s/ism	126	165
thin-film	145	169
third-world	250	452
TOE	81	69
tofu	4	11
tomography	2046	1825
toxic-shock	53	14
translocation	512	502
transplant	461	545
trauma	1095	1073
trimethoprim	86	70
ulcer/s	1103	1073
*ultrasound	1271	1202
VDT/s	14	42
venectomy	0	3
video	406	634
*Vietnam/ese	229	298
VLSI	244	787
volcano/es/ic	332	426
*winter	618	799

produce all possible pairs. Words in a subtitle, however, are usually permuted separately with only the words that appear in that subtitle and not with those in the main title. All terms are then cross-referenced; that is, they are listed in the *PSI* under their own heading as well as under each of the terms with which they are paired.

Most entries in the *PSI* are the exact words used by the authors in their article titles. So the *PSI* user must remember to also check variants as well as synonyms and related terms when conducting a literature search. For example, *adenosine triphosphate*, *adenosine-triphosphate*, and *ATP* may be listed separately in the *PSI*. However, certain frequently used compounds, identified by UWD counts, are combined in the *PSI* to facilitate their retrieval. For example, *breast-cancer* is one such term. *Breast-cancer* occurs as such, but so do many articles involving *cancer* of the *breast*. Other examples are *breeder-reactor* and *brush-border*. This combining of words makes searching of such terms more convenient for the user and provides greater subject specificity, since a third title word can then be linked to the sometimes artificially hyphenated term, for example *brush-border* with *membrane*.

Conclusion

This concludes the first part of our study on ISI's master dictionary. In Part 2 we will more closely examine specific words that appeared in this file in 1981 and 1984.

* * * * *

My thanks to Abigail Grissom and Janet Robertson for their help in the preparation of this essay.

REFERENCES

1. **Orwell G.** *1984.* New York: New American Library, 1949. 267 p.
2. **Mish F C,** ed. *Webster's ninth new collegiate dictionary.* Springfield, MA: Merriam-Webster, 1985. 1563 p.
3. MDMA—a multidisciplinary investigation. *Reports from the medical, scientific, and regulatory communities.* Lafayette, CA: Earth Metabolic Design Laboratory, 1985.
4. **Mikhailov A I, Chernyi A I & Giliarevskii R S.** *Scientific communications and informatics.* Arlington, VA: Information Resources Press, 1984. p. 371-8.
5. Preface. (Barnhart C L, ed.) *Thorndike-Barnhart comprehensive desk dictionary.* Garden City, NY: Doubleday, 1962. p. xi-xiii.
6. **Kenner H.** Neatness doesn't count after all. *Discover* 7(4):86-93, 1986.
7. **Zipf G K.** *Human behavior and the principle of least effort.* New York: Hafner, 1972. 573 p.
8. **Garfield E.** Bradford's law and related statistical patterns. *Essays of an information scientist.* Philadelphia: ISI Press, 1981. Vol. 4. p. 476-83.
9. --------------. ISI's master dictionary aids scientific etymology and reflects changes in science. *Ibid.*, 1980. Vol. 3. p. 393-9.
10. --------------. Another look at ISI's master dictionary—aiding scientific etymology and reflecting changes in science. *Ibid.*, 1983. Vol. 5. p. 288-95.
11. *The Chicago manual of style.* Chicago: University of Chicago Press, 1982. p. 163-4.