(ii) Overseas communities who could tap this reservoir only by expensive satellite transmission will be tragically disadvantaged. And will the system cope with accented letters and alphabets other than Roman? The English language has reached its present eminence by its remarkable suppleness and by the remarkable scientific performance of the anglophone communities. Let us take care lest leadership by excellence degenerate into dictatorship.

(iii) Most serious of all is the impediment the system offers to random reading. As I have stated earlier, science without browsing is not science.

I am no hidebound conservative. It may well be that the brilliant advocates of the Alternative will find ways to overcome these problems and still keep the system economic, and in that case I shall be just as happy as my neighbor to recycle my stacks of bound volumes. But let us not proclaim the millenium until we see the correct date on the calendar.

ACKNOWLEDGMENT

In preparing this polemic, I have had much help from concerned Australian colleagues, especially B. J. Walby and D. E. Boyd. (Boyd[8] has published a model similar to my own; it deals mainly with the signal-to-noise ratio of scientific communication.) J. Glover has supplied much helpful information on the financial structure of scientific institutions. I owe a special debt to Dr. M. Brogan, of the American Chemical Society, who first drew my attention to the King report and whose informed criticism demolished an earlier version of this paper.

REFERENCES AND NOTES

(1) "Economics of Scientific Publications", Council of Biology Editors, Washington, 1973.
(2) D. W. King and N. K. Roderer, "Systems Analysis of Scientific and Technical Communication in the United States. The Electronic Alternative to Communication through Paper-Based Journals", National Science Foundation, Washington, DC, 1978, with four annexes.
(3) D. H. M. Bowen, *Scholarly Publishing*, **11**, 43 (1979).
(4) M. Brogan, *Scholarly Publishing*, **11**, 47 (1979).
(5) Some authors with a penchant for rambling asides actually prefer tailnotes. It would be possible for a journal to distinguish between footnotes and tailnotes by numbering the first and labeling the latter with asterisks and daggers. The *Australian Journal of Chemistry* distinguishes "citation" and "information" footnotes in such a manner, but prints both at the foot of the page, so that the reader can assess their importance to him at a glance. The present tailnote has been inserted as a warning example, to show what annoyance such notes can cause.
(6) M. A. Tinker, "Legibility of Print", Iowa State University Press, Ames, 1963.
(7) R. Schoenfeld, *Chem. Aust.*, **44**, 206 (1977).
(8) D. E. Boyd, *Aust. Phys.*, **15**, 39 (1978).

Analysis of Keywords in Chemistry

SHIZUO FUJIWARA, MASANORI YOKOYAMA,* and SHUICHI UEDA†

Department of Chemistry, Faculty of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113, Japan

Received August 27, 1979; Revised Manuscript Received October 13, 1980

From the ~10 000 000 keywords in 3 volumes of *CA Condensates* for 1977 and 1978, 16 000 keywords were assembled from the 200 most frequently occurring ones in each of the 80 sections and analyzed for cross-correlation, characteristic features, and relationship to 10 000 terms in a separate Japanese Chemical Society (JCS) list.

Each abstract in *CA Condensates* is represented in condensed form by bibliographic data and keywords. The former consists of the title, author(s) name, and journal name with volume number, issue, page, and year of publication. A few keywords, representing the content of each article, convey to the reader the essence of the subject of the document. Accordingly, keywords profile the documents and thus each field of chemistry is described by its assembly of keywords.

Chemical Abstracts has 80 sections, each referring to a specific field of chemistry. According to Chemical Abstracts Service, the *CA* section arrangement has never been intended as a model for an overall classification scheme in chemistry but simply reflects the *CA* subject coverage based on the relative amount of information published in various general and specific fields. Hence, the section arrangement has been changed occasionally to reflect progress, changes, and trends occurring in the published world of chemistry. For instance, as applied chemistry and chemical engineering have developed over the years, more technology- and product-oriented *CA* sections, such as "Plastics Manufacture and Processing" and "Surface-Active Agents and Detergents", were created in contrast to the more traditional subdivisions of chemistry, such as "Inorganic Analytical Chemistry" and "Electrochemistry". Although we have "Subject Coverage and Arrangement of Abstracts by Sections in Chemical Abstracts" as a useful explanation for the classification, it still seems worthwhile to carry out an analysis of the *CA* classification.

Table I. Statistics of the Volumes of *CACon*

Vol.	DC	KWC	no. KW
86	199 309	3 140 549	348 304
87	210 632	3 327 357	365 428
88	202 524	3 210 321	358 542

All keywords which appeared in volumes 86 (Jan–June, 1977), 87 (July-Dec, 1977), and 88 (Jan–June, 1978) of *Chemical* Abstracts were collected and examined. The frequency of occurrence of keywords was counted for each section of *CA Condensates*, and the 200 most frequently occurring keywords were collected for each section. The total number of keywords in each section is referred to as the keyword count number (KWC), and the total number of unique keywords per document as the number of keywords (No. KW). The number of articles was noted and referred to as the document count (DC). Cross-correlation of each keyword among the 80 sections was carried out. Keywords were grouped according to whether they can be used generally or only in chemical documentation.

* To whom correspondence should be addressed at Fujitsu Corporation, Ohta-ku, Tokyo.
† Department of Library Sciences, Keio University, Mita, Minato-ku, Tokyo.

0095-2338/81/1621-0066$01.25/0

Table II. Examples of Classification of the Keywords into General and Technical Terms

section number						
CA021		CA025		CA056		
general	technical	general	technical	general		technical
review	chemistry	acid	phenyl	alloy	mechanical	aluminum
organic	chem	catalyst	alkylation	metal	temp	nickel
book	ketone	reaction	benzene	review	transformation	copper
synthesis	carbonyl	synthesis	phenol	corrosion	growth	titanium
catalyst	hydrogenation	derivatives	N	coating	pressure	alloys
reaction	ed	prepn	aryl	welding	solidification	molybdenum
compd	ester	acids	ether	structure	crystal	tungsten
compounds	aldehyde	condensation	chloride	iron	formation	cobalt
reactions	chloride	compounds	arom	steel	molten	chromium
catalysis	beilstein	reactions	substituted	property	rolling	carbide
acid	olefin	addn	aromatic	fatigue	system	niobium
redn	carbon	salt	P	deformation	addn	zinc
cleavage	oxide	purifn	ester	properties	low	silicon
phase	Pt	cleavage	alkyl	metals	melt	zirconium
natural	As	compd	herbicide	heat	metallurgy	silver
complex	halide	crystal	xylene	casting	cutting	hydrogen
vol	sulfur	labeled	alpha	stress		magnesium
slide	polymer	review	aniline	composite		lead
synthetic	esters	presence	methyl	phase		eutectic
methods	fluorination	ring	O	surface		gold
metal	thiol	complex	nitro	powder		vanadium
laboratory	isomerization	structure	benzyl	high		oxidn
transfer	carboxylic	phase	beta	electron		superalloy
substitution	alkene		amino	treatment		elec
textbook	catalytic		hydroxy	strength		carbon
reagent	oxidative		hydrogenation	resistance		tantalum
handbook	alkyl		rearrangement	temperature		manganese
prepn	allylic		ketone	fracture		boron
transition	asym		carbon	friction		weld
new	carbonylation		toluene	creep		sintering
work	alumina		esters	diffusion		plastic
group	amine		sulfide	wear		uranium
structure	oxidation		chloro	materials		bronze
addn	ether		amine	grain		plasma
acids	ylide		preparation	internal		
selective	epoxidn		insecticide	mold		
enzyme	homologation		fungicide	film		
reagents	aldehydes		chlorination	brass		
syntheses	ketones		substitution	book		
techniques	cyclization		diphenyl	thermal		
agent	amino		anhydride	resistant		
exchange	cyclic		oxime	hardening		
cassette	anion		new			
system	vanadium		antiinflammatory	crack		
application	rhodium		hydrocarbon	aging		
	triphase		styrene	cast		
	catalyzed		dimethyl	melting		
	S		phenoxy	wire		
	hydrocarbon		deriv	dislocation		
	kinetics		cresol	tool		

DATA HANDLING

Volumes 86, 87, and 88 of the *CA Condensates* (*CACon*) were input on disk file (HITAC 8700/8800 computer at the University of Tokyo Computer Center) and converted into a data base. This data base can be accessed from terminals all over the country via public cable. The data base management system which handles the data was developed at the University of Tokyo, named TOOL-IR, Tokyo University On Line Information Retrieval. All keywords in the file were analyzed (Table I).

RESULTS AND ANALYSIS

Most Frequently Occurring Keywords. The 200 most frequently occurring keywords were determined for each section of *CACon* and listed in order of their frequency of occurrence. Examples of the first 100 keywords are shown in Table II.

Statistical Features of the Keywords per Document. The ratio of keyword counts or the number of keywords to the document count are shown in Tables III and IV. There is an average of about 15 or 16 KWC or 2 unique keywords per document. Some sections show abnormal values.

Cross-Correlation of Keywords among Sections. Cross-correlation of the sections can be evaluated in terms of the number of keywords found in any combination of two sections. If the sections of high degree of cross-correlation are clustered, those clustered sections may be taken as the sections which form a group. Figure 1 presents the results of the analysis of the cross-correlation. In Figure 1a, each star refers to the fact that the combination of two starred sections share more than 50 keywords. It is apparent in Figure 1a that there exist three or four groups of sections, one from 1 to 19, the second from 21 to 34, and the third from 35 to 80. The first one refers to the sections of pharmacology and biochemistry. Figure 1b presents the cross-correlation over 75 and Figure 1c over 100 keywords. If (a) and (b) are compared, one can see that the first group of Figure 1a is formed of two subgroups, one from

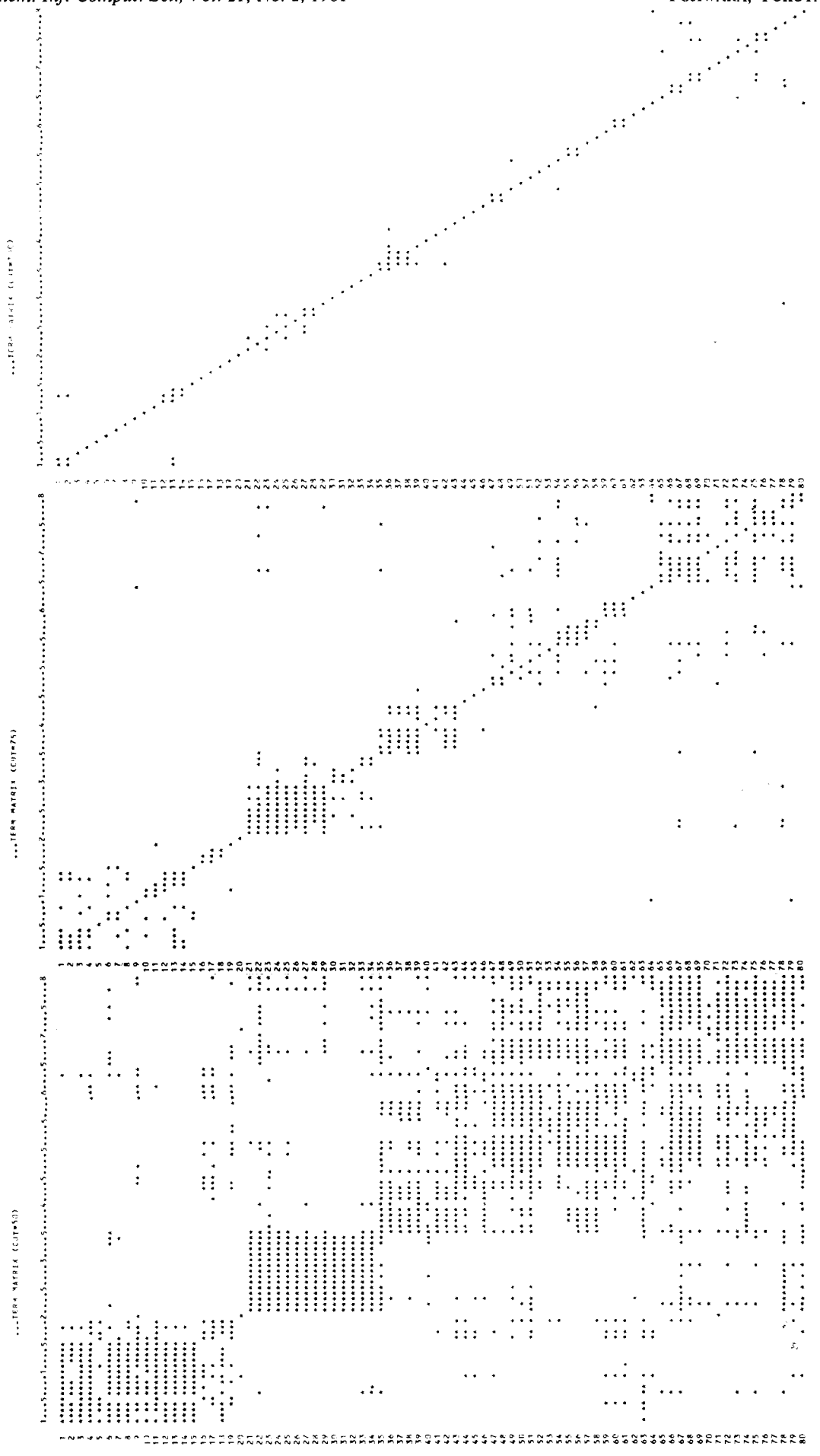

Figure 1. Cross-correlation of the sections as evaluated by the number of common key words, n. (a) $n \geq 50$, (b) $n \geq 75$, and (c) $n \geq 100$. The coordinates and the abscissa both refer to the numbers of the sections.

Table III. Statistical Distribution of DC, KWC, and No. KW in 80 Sections[a]

section	DC	KWC	No. KW	section	DC	KWC	No. KW
1	7907	99911	10952	41	202	2935	849
2	3150	51641	5148	42	2221	36071	3729
3	1748	28324	5132	43	1456	23896	3259
4	3911	59672	7903	44	248	3824	1066
5	2580	34916	6571	45	115	1951	601
6	4476	64433	6981	46	507	7977	1637
7	3764	53887	6220	47	1059	12607	2427
8	847	12129	2515	48	2375	32593	3576
9	3471	53652	6326	49	1433	16514	2267
10	3156	44989	6172	50	637	7673	1767
11	3732	58170	8299	51	3393	53594	4981
12	1929	32722	5519	52	1354	17474	2162
13	6211	97541	7642	53	4588	72841	8319
14	4661	68163	6822	54	1649	24252	2543
15	1847	28107	3936	55	5438	85723	4647
16	1604	18092	3275	56	5709	82892	4869
17	2885	36864	5053	57	2238	26746	3061
18	2305	35550	4030	58	1490	17030	2330
19	2702	37233	4452	59	3980	66301	5589
20	1257	13421	3183	60	2973	46903	4437
21	272	4364	1163	61	2628	41456	4823
22	6781	124835	12178	62	664	9711	1666
23	1636	23866	4618	63	1876	23792	4875
24	888	12123	3211	64	464	6351	1578
25	1888	28304	6045	65	3467	61940	4241
26	393	5459	1910	66	1823	31685	3084
27	1808	26397	6754	67	1646	31189	3173
28	2803	38620	9483	68	2429	48894	3830
29	1413	25816	5255	69	1128	19336	2181
30	524	7170	2317	70	5843	87627	3770
31	357	5692	1790	71	4393	75989	5253
32	334	5483	1828	72	3092	57235	4505
33	839	13883	3344	73	9244	147732	8253
34	688	11896	2462	74	2000	26133	3259
35	3445	61958	6029	75	4519	80266	7567
36	4411	72866	6071	76	5769	104796	5668
37	1997	31116	3311	77	2483	43099	3208
38	1537	25814	3205	78	1853	36825	4637
39	2010	32147	3245	79	2442	57399	4268
40	554	8885	1717	80	570	13186	2252

[a] (1) Pharmacodynamics, (2) hormone pharmacology, (3) biochemical interactions, (4) toxicology, (5) agrochemicals, (6) general biochemistry, (7) enzymes, (8) radiation biochemistry, (9) biochemical methods, (10) microbial biochemistry, (11) plant biochemistry, (12) nonmammalian biochemistry, (13) mammalian biochemistry, (14) mammalian pathological biochemistry, (15) immunochemistry, (16) fermentations, (17) foods, (18) animal nutrition, (19) fertilizers, soils, and plant nutrition, (20) history, education, and documentation, (21) general organic chemistry, (22) physical organic chemistry, (23) aliphatic compounds, (24) alicyclic compounds, (25) noncondensed aromatic compounds, (26) condensed aromatic compounds, (27) heterocyclic compounds (one hetero atom), (28) heterocyclic compounds (more than one hetero atom), (29) organometallic and organometalloidal compounds, (30) terpenoids, (31) alkaloids, (32) steroids, (33) carbohydrates, (34) synthesis of amino acids, peptides, and proteins, (35) synthetic high polymers, (36) plastics manufacture and processing, (37) plastics fabrication and uses, (38) elastomers, including natural rubber, (39) textiles, (40) dyes, fluorescent whitening agents, and photoensitizers, (41) leather and related materials, (42) coatings, inks and "and related products", (43) cellulose, lignin, paper and "and other wood products", (44) industrial carbohydrates, (45) fats and waxes, (46) surface-active agents and detergents, (47) apparatus and plant equipment, (48) unit operations and processes, (49) industrial inorganic chemicals, (50) propellants and explosives, (51) fossil fuels, derivatives, and thermal energy technology, (52) electrochemical, radiational, and thermal energy technology, (53) mineralogical and geological chemistry, (54) extractive metallurgy, (55) ferrous metals and alloys, (56) nonferrous metals and alloys, (57) ceramics, (58) cement and concrete products, (59) air pollution and industrial hygiene, (60) sewage and wastes, (61) water, (62) essential oils and cosmetics, (63) pharmaceuticals, (64) pharmaceutical analysis, (65) general physical chemistry, (66) surface chemistry and colloids, (67) catalysis and reaction kinetics, (68) phase equilibriums, chemical equilibriums, and "and solutions", (69) thermodynamics, thermochemistry, and thermal properties, (70) nuclear phenomena, (71) nuclear technology, (72) electrochemistry, (73) spectra by absorption, emission, reflection, or magnetic resonance and "and other optical properties", (74) radiation chemistry, photo-chemistry, and photographic processes, (75) crystallization and crystal structure, (76) electric phenomena, (77) magnetic phenomena, (78) inorganic chemicals and reactions, (79) inorganic analytical chemistry, and (80) organic analytical.

sections 1–4, pharmacological chemistry, and the other 13 and 14, biochemistry. The second cluster in Figure 1a, formed by sections 21–29, shows a tight clustering even in Figure 1b. The sections under investigation refer to organic chemistry. The third cluster in Figure 1a is divided into subgroups in Figure 1b, i.e., one from 35 to 39 or 43, and the others 65–69, 72 and 73, 72–79. Figure 1c refers to the cross-correlation of more than 100 KW. Clusters are found in combinations of sections: (1,2), (1,13), (2,13), (12,13), (13,14), (12,13,14), (21,23), (23,25), (24,28), (25,27), (25,28), (27,28), (35,36), (36,37), (36,39), (37,38), (35,36,37,38), (36,42), (47,48), (49,54), (49,55), (55,56), (60,61), (64,80), (65,73), (67,78), (68,69); (68,75), (68,78), (73,75), and (75,76).

Comparison of *CA* Keywords with the JCS List. The Japanese Chemical Society (JCS) formed a committee on chemical terminology in 1961 to represent the physicochemical, inorganic, organic, and chemical engineering fields and technology, such as oils, paints, dyes, ceramics, etc. The committee published a list of chemical terms, 1st edition in 1964, 2nd in 1968, and 3rd in 1974. The committee chose the most appropriate terms, minimum in number and most appropriate for general use in all fields of chemistry. The task was not easy. For example, the published list is of terms in two categories, one of fundamental importance, which can also be used

Table IV. Ratios KWC/DC (B/A), KW/DC (C/A), and KWC/No. KW (B/C).

section	B/A	C/A	B/C	section	B/A	C/A	B/C
1	14	1.5	9.2	41	15	4.2	3.5
2	16	1.6	10.0	42	16	1.7	9.7
3	16	2.9	5.5	43	16	2.2	7.3
4	15	2.0	7.6	44	15	4.3	3.6
5	14	2.5	5.3	45	17	5.2	3.2
6	14	1.6	9.2	46	16	3.2	4.9
7	14	1.7	8.7	47	12	2.3	5.2
8	14	3.0	4.8	48	14	1.5	9.1
9	16	1.8	8.5	49	12	1.6	7.3
10	14	2.0	7.3	50	12	2.8	4.3
11	16	2.2	7.0	51	16	1.5	10.8
12	17	2.9	5.9	52	13	1.6	8.1
13	16	1.2	12.7	53	16	1.8	8.8
14	15	1.5	10.0	54	15	1.5	9.5
15	15	2.1	7.1	55	16	0.9	18.4
16	11	2.0	5.5	56	15	0.9	17.0
17	13	1.8	7.3	57	12	1.4	8.7
18	15	1.7	8.8	58	11	1.6	7.3
19	14	1.6	8.4	59	17	1.4	11.9
20	11	2.5	4.2	60	16	1.5	10.6
21	16	4.3	3.8	61	16	1.8	8.6
22	18	1.8	10.3	62	15	2.5	5.8
23	15	2.8	5.2	63	13	2.6	4.9
24	14	3.6	3.8	64	14	3.4	4.0
25	15	3.2	4.7	65	18	1.2	14.6
26	14	4.9	2.9	66	17	1.7	10.3
27	15	3.7	3.9	67	19	1.9	9.8
28	14	3.4	4.1	68	20	1.6	12.8
29	18	3.7	4.9	69	17	1.9	8.9
30	14	4.4	3.1	70	15	0.7	23.2
31	16	5.0	3.2	71	17	0.7	14.4
32	17	5.5	3.0	72	19	1.5	12.8
33	17	4.0	4.2	73	16	0.9	17.9
34	17	3.6	4.8	74	13	1.6	8.0
35	18	1.8	10.3	75	18	1.7	10.6
36	17	1.4	12.0	76	18	1.0	18.5
37	16	1.7	9.4	77	17	1.3	13.4
38	17	2.1	8.1	78	20	2.5	7.9
39	16	1.6	9.9	79	24	1.7	13.4
40	16	3.0	5.2	80	23	4.0	5.9

generally, and the other of significance in each specific field of chemistry.

The 3rd edition has 10066 terms which contain 6306 unique keywords. When the most frequently appearing keywords in *CA*, Vol. 86 (16000 or 200 × 80 sections), are compared to the 6306 words of JCS, the overlap is 1910. Thus, the JCS list holds one-half of the list of the most frequently appearing keywords of CA; however, two-thirds of the JCS list is not included in the *CA* list of the most frequently appearing keywords. Roughly speaking, the *CA* and JCS overlap refers to the keywords of "general" terms. In other words, two-third of the JCS list represents the technical terms which can be assigned to the sister member societies of JCS.

DISCUSSION

It must be recognized that the keywords assigned to a document are by no means limited to the subject matter of the *CA* section in which the corresponding abstract is placed. On the contrary, the most useful function of the keyword assignment is to cover the *total* subject content of the document. Frequently, the subject matter of a document is broad enough to justify its placement in more than one *CA* section, and yet only one section, corresponding to the main thrust of the document, can be chosen. Obviously, at times the choice is quite arbitrary. It is the keywords that allow the retrieval of documents on a given topic, irrespective of their abstracts' placement. The results of the present analysis of the keywords at least suggest that the keywords play another function. For example, the whole assembly of keywords can be used as a measure of the potential for clustering the sections. The sections of organic chemistry, from 21 to 34, show a high cross-correlation among themselves; they are clustered and form a family. Those from 1 to 19 represent pharmacology and biochemistry. The average number of keywords per document is exceedingly high for the sections of organic chemistry in comparison to other fields (section 22 is an exception). Another feature which distinguishes the organic chemistry sections from others is the fact that the technical terms dominate the list of keywords of high-frequency occurrence. These features have been revealed by the present computer handling of 10000000 keywords in chemistry.

The real basis of the cross-correlation of the sections will be a subject of further investigation. Another finding of the present analysis is that the keywords of chemistry can be classified into two categories, general and technical. After careful examination of the general keywords, it may be possible to establish a list of keywords of basic importance which will be useful for chemical education, text editing, and the production of chemical thesauri. Predominance of either general terms or technical terms has been recognized with each of the 80 sections and reflect the nature of the subjects covered in each section. Finally, it is noted that the 16000 keywords, which have been chosen as the 200 most frequently occurring keywords in each of the 80 sections, could be taken as a profile of the current activity of chemistry. Further investigation on the details of the structure of the chemical keywords assembly is being carried out in this laboratory.

ACKNOWLEDGMENT

Thanks are due to Chemical Abstracts Service of the American Chemical Society for making the present investigation possible. S.F. expresses his thanks to Dr. Kenjiro Kimura, Emeritus Professor of the University of Tokyo, who was a member of the JCS committee on chemical terms and who recognized the importance of keywords in chemistry. The help of Professor Malcolm Bersohn, University of Toronto, Dr. Hiroshi Ozawa of the University of Tokyo Computer Center, Mr. Masamitsu Negishi of the Research Center for the Information and Library Sciences, and Mr. Toshikatsu Sugiura, University of Tokyo and the Japan Association for the International Chemical Information, and the financial support for the research by the Ministry of Education are also acknowledged.