DOCUMENT RESUME

ED 065 145	LI 003 766

AUTHOR	Schipma, Peter B.
TITLE	Term Fragment Analysis for Inversion of Large Files.
INSTITUTION	Illinois Inst. of Tech., Chicago. Research Inst.
PUB DATE	4 Jun 71
NOTE	17p.; (0 References)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Algorithms; Data Analysis; *Data Bases; Indexes (Locaters); Information Processing; Information Retrieval; Ratios (Mathematics); *Search Strategies; *Word Frequency
IDENTIFIERS	*Key Letter in Context; KLIC Index

ABSTRACT

Words and word fragments from the computer-readable data bases "Chemical Abstracts Condensates" and "Biological Abstracts Previews" were analyzed in terms of length, number, and frequency of appearance to determine some parameters upon which inversion of these data bases could be predicated. Types (unique words or fragments) and tokens (all appearances of types) were counted and type:token ratios calculated. A KLIC (Key-letter-in-Context) Index was also generated from each of the data bases. The paper discusses the impact of the various counts, ratios and projections on the problem of inverting the data bases for retrospective search purposes. (Author)

TERM FRAGMENT ANALYSIS FOR INVERSION OF LARGE FILES

by

Peter B. Schipma
IIT Research Institute
10 West 35 Street
Chicago, Illinois

June 4, 1971

IIT RESEARCH INSTITUTE

TERM FRAGMENT ANALYSIS FOR INVERSION OF LARGE FILES

1. Introduction

We have recently conducted some analyses of words and word fragments to obtain some information regarding the nature and contents of two data bases, Chemical Abstracts Condensates, and BA Previews. These analyses were made for two main purposes. First, we needed information for our profile coordinators to use when constructing profiles or helping others with profile construction. Second, we hope to use the information to help define a file structure and search technique for retrospective search of inverted data bases. The latter topic is the subject of this paper.

2. Parameters of the Analyses

2.1 Coverage

These analyses were made for one volume of Condensates, issues being selected from both volumes 72 and 73 for a total of 26 issues, representing about 150,000 citations; and about 20% of a volume of BA Previews, covering both Biological Abstracts and BioResearch Index. The terms CA and BA are used in reference to CA Condensates and BA Previews respectively. In both cases we covered titles and index terms only, omitting, for purposes of this study, the CODEN, journal titles, bibliographic citation and corporate authors. In the case of BA, titles are augmented and the CROSS Codes and Biosystematic

Codes are included. The analyses were made from the IITRI-formatted version of the tapes, and thus redundant index terms and phrases of CA had already been removed.

2.2 Word Definition

For purposes of this study, we defined a word as any combination of characters bounded by blanks, slashes or asterisks. This arbitrary definition naturally gives rise to some problems. In BA, for instance, words are split by blanks for the inverted file being continually updated at BIOSIS. This is especially true of chemical names such as LIPOPROTEIN, which appears in BA as LIPO PROTEIN. In a similar fashion BA links some words together with a hyphen, so that NEW YORK appears as NEW-YORK. Different problems arise in CA. Because of frequent use of various punctuation symbols, several versions of the same word appear. For instance "POLYMERS" appears many times as a term, but also appears as "POLYMERS," and "POLYMERS." because of punctuation in titles. To have limited words to alphamerics would have caused several other problems, significantly that of breaking a chemical name that contained commas and/or parentheses into several "words". Some convention was necessary. We chose our set of delimiters and got the problem situations outlined above, which have to be taken into account when evaluating the study.

2.3 Statistics Studies

2.3.1 The first analysis we made was a frequency count for purposes of determining the type/token ratio for the data bases. Listings in both alphabetical and frequency order were made. Figure 1 is a sample of the alphabetical listing. Types are defined as unique words, within the constraints for word definition given above, and tokens are the total number of appearances of the types.

2.3.2 The second product of the study was a Key-Letter-in Context, or KLIC, Index. This is a lexicographical ordering of all types in a data base by each character within each type. It resembles a KWIC index in appearance, but is done letter by letter rather than term by term. A sample page is shown by Figure 2.

3. Results of the Study

3.1 Word Frequency

3.1.1 BA Previews

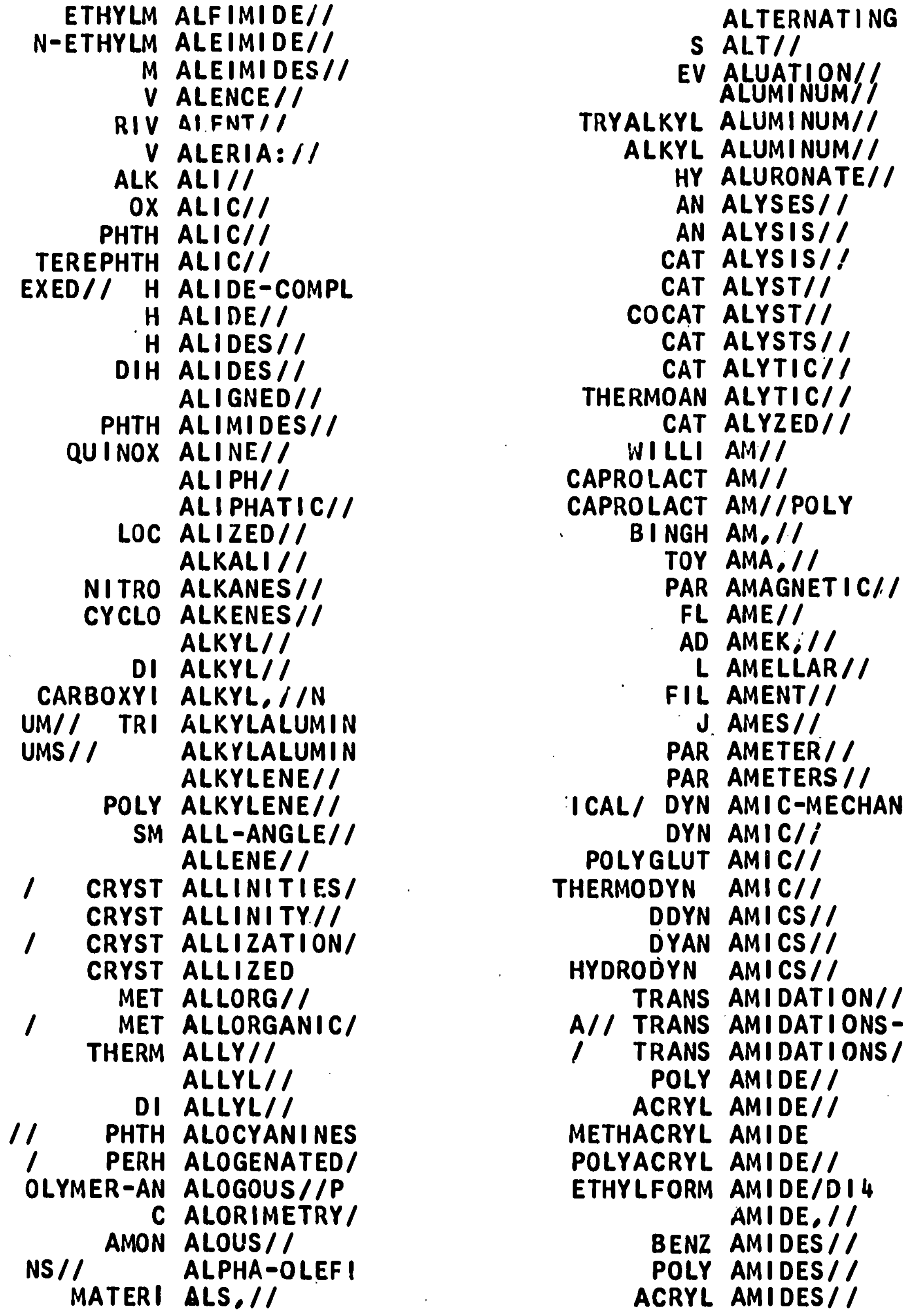
For a comparable number of citations (about 22,000) there are about 20% more word tokens and types in BA than in BIORI (See Table 1). The type/token ratios, however, were virtually the same. For the combined BA issues, the type/token ratio was 1:20.7. Thus each of

Figure 1: Alphabetical Listing of CA Terms and Frequencies

ABERCROMBIE,	1
ABERKHAEVA,	1
ABERRANT	1
ABERRATION	3
ABERRATIONS	7
ABESTUS	1
ABETALIPOPROTEINEMIA	1
ABF3	1
ABH	3
ABIES	1
ABIETATE-COMLNE	1
ABIETATES	1
ABIETATRIENES	1
ABIETELLA	1
ABIETENUATES	1
ABIETOPALEIC	1
ABIRO,	1
ABILITIES	4
ABILITY	40
ABIOGENESIS	1
ABIOGENESIS.	1
ABITZ,	1
ABKHAZ	1
ABKHAZIA)	1
ABLATING	3
ABLATION	6
ABLATIVE	4
ABLATIVES	3
ABLATOR	1
ABNORMAL	22
ABNORMALITIES	10
ABNORMALITY	1
ABO	2
ABOLITION	2
ABOMASAL	3
ABOMASUM	7
ABOMASUS	1
ABORINSKAYA,	1
ABORTIFACIENT	1
ABORTION	1
ABORTIONS	1
ABORTIVE	1
ABORTUS:	1
ABOSRPTION*	1
ABOU	1
ABOUL-FETOUM,	1
ABOUT	14
ABOVE	27
ABRADED	2

*Note: Misspelling present on Condensates data base.

Figure 2: Sample Key-Letter-in-Context Index Page



```
   ETHYLM ALFIMIDE//                          ALTERNATING
 N-ETHYLM ALEIMIDE//                        S ALT//
        M ALEIMIDES//                      EV ALUATION//
        V ALENCE//                            ALUMINUM//
      RIV ALENT//                    TRYALKYL ALUMINUM//
        V ALERIA://                     ALKYL ALUMINUM//
      ALK ALI//                            HY ALURONATE//
       OX ALIC//                           AN ALYSES//
     PHTH ALIC//                           AN ALYSIS//
 TEREPHTH ALIC//                          CAT ALYSIS//
EXED//  H ALIDE-COMPL                     CAT ALYST//
        H ALIDE//                       COCAT ALYST//
        H ALIDES//                        CAT ALYSTS//
      DIH ALIDES//                        CAT ALYTIC//
          ALIGNED//                  THERMOAN ALYTIC//
     PHTH ALIMIDES//                      CAT ALYZED//
   QUINOX ALINE//                       WILLI AM//
          ALIPH//                   CAPROLACT AM//
          ALIPHATIC//               CAPROLACT AM//POLY
      LOC ALIZED//                      BINGH AM,//
          ALKALI//                        TOY AMA,//
    NITRO ALKANES//                       PAR AMAGNETIC//
    CYCLO ALKENES//                        FL AME//
          ALKYL//                          AD AMEK,//
       DI ALKYL//                           L AMELLAR//
 CARBOXYI ALKYL,//N                       FIL AMENT//
UM//  TRI ALKYLALUMIN                       J AMES//
UMS//     ALKYLALUMIN                     PAR AMETER//
          ALKYLENE//                      PAR AMETERS//
     POLY ALKYLENE//                ICAL/ DYN AMIC-MECHAN
       SM ALL-ANGLE//                     DYN AMIC//
          ALLENE//                   POLYGLUT AMIC//
/   CRYST ALLINITIES/               THERMODYN AMIC//
    CRYST ALLINITY//                     DDYN AMICS//
/   CRYST ALLIZATION/                    DYAN AMICS//
    CRYST ALLIZED                    HYDRODYN AMICS//
      MET ALLORG//                      TRANS AMIDATION//
/     MET ALLORGANIC/               A// TRANS AMIDATIONS-
    THERM ALLY//                    /   TRANS AMIDATIONS/
          ALLYL//                        POLY AMIDE//
       DI ALLYL//                       ACRYL AMIDE//
//   PHTH ALOCYANINES               METHACRYL AMIDE
/    PERH ALOGENATED/               POLYACRYL AMIDE//
OLYMER-AN ALOGOUS//P                ETHYLFORM AMIDE/DI4
        C ALORIMETRY/                         AMIDE,//
     AMON ALOUS//                        BENZ AMIDES//
NS//      ALPHA-OLEFI                    POLY AMIDES//
   MATERI ALS,//                        ACRYL AMIDES//
```

Table 1: BA Type/Token Relationships

Issues	Citations	Tokens	Types	Type/Token Ratio
4 (BA 51: 18-21)	22,536	611,745	39,164	1:15.6
3 (BioRI 70: 9-11)	22,500	486,155	30,942	1:15.7
7 (Combined)	45,036	1,097,900	52,912	1:20.7

the 53,000 unique words that appeared in 45,000 citations appeared, on the average, about 21 times.

3.1.2 CA Condensates

For condensates, we did a series of these studies, using 2, 6, 13, another 13, and 26 issues. In this way we could get a curve of type/token ratio versus tokens. As would be expected, the type/token ratio increases with an increased number of citations. Each type appears, on the average, 5.48 times in 9000 citations taken from two issues, but 12 times for 134,000 citations taken from 26 issues. A summary is given in Table 2. The curve of type/token ratio versus tokens, plotted on a log scale, is a straight line (see Figure 3). Although it is probably not reasonable to project this line, if such is done the indications are that no new types would be added once the data base reached 45 million tokens (about 12 years worth of CA).

3.1.3 Comparison

From these figures it is very obvious that the type/token ratios for BA and CA are very different. For a similar number of citations they are 1:21 and 1:9 respectively. Each unique word in BA is used more than twice as often as each unique word in CA. A study of the

Table 2: CA Type/Token Relationships

Issues	Citations	Tokens	Types	Type/Token Ratio
2(Vol. 72)	9,067	91,760	16,753	1:5.48
6(Vol. 72)	31,402	479,856	60,876	1:7.88
13(Vol. 72)	67,456	877,734	92,216	1:9.52
13(Vol. 73)	66,796	963,698	100,498	1:9.59
26(Vol. 72 and 73)	134,252	1,841,432	153,268	1:12.01

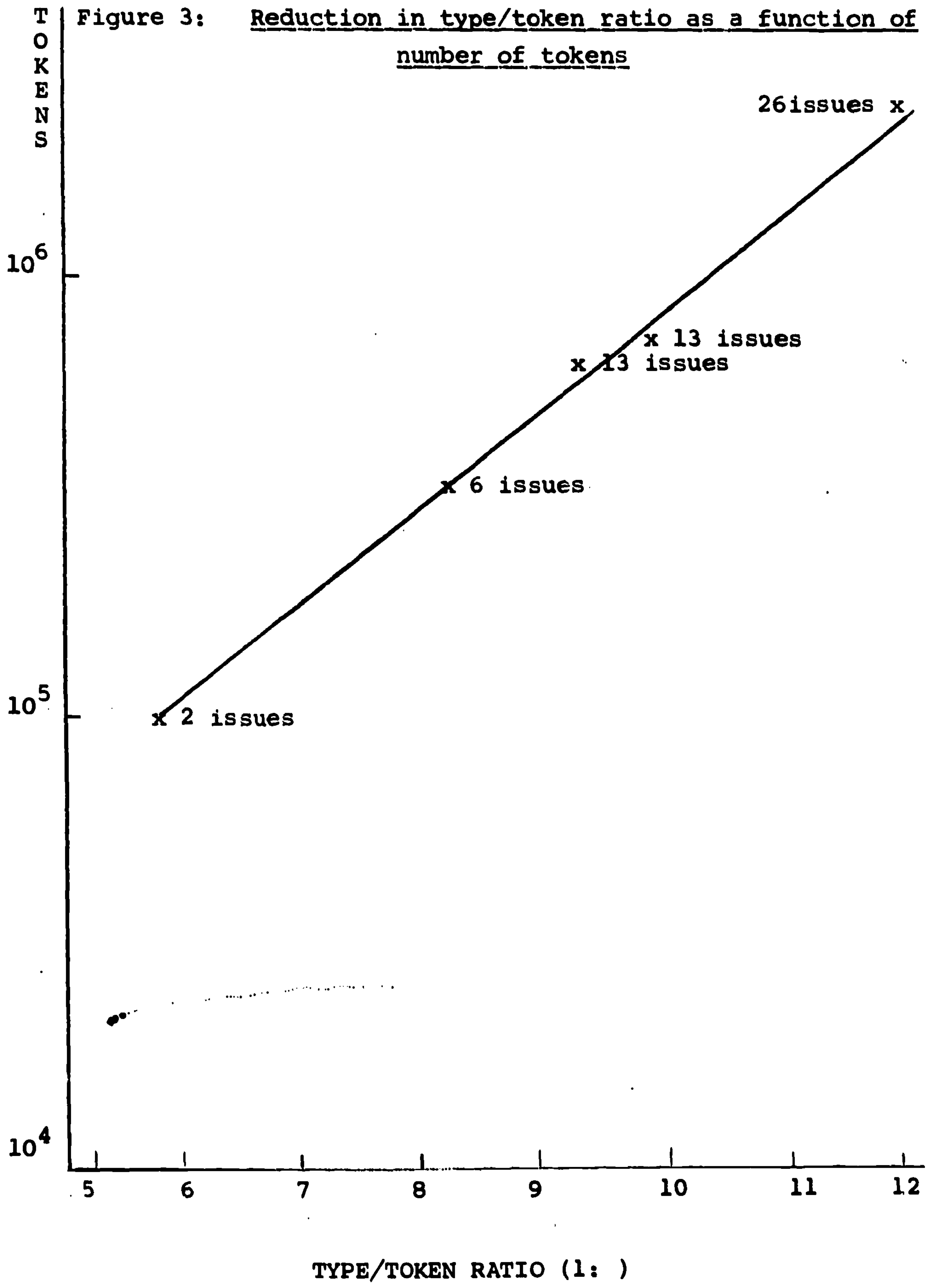


Figure 3: Reduction in type/token ratio as a function of number of tokens

listings of the types with their frequencies reveals three reasons for this:

- BA uses fewer chemical names than CA, as would be expected.
- Standardized spelling and abbreviations are used to a greater extent in BA. NMR always appears in that one form in BA, but in that form plus as N M R and N.M.R. in CA.
- More punctuation is used in CA.

In terms of inverting these data bases, much more storage is required for CA than for BA, or more computer time will be necessary to reduce non-unique types to unique ones (NMR, N M R and N.M.R. should all be put in one form, for example). In either case CA will be more expensive to invert than BA because of this type/token ratio discrepancy. And it should be expected that the chemical name additions each year will always keep the type/token ratio for CA relatively low.

3.2 KLIC Indexes

The KLIC indexes provide even greater insights into the nature of a data base as it relates to file inversion for retrospective search. Using the KLIC Index, one can discover how the use of a given search fragment will be affected by the contents of the data base. Some

Interesting examples are *FLUORO*, *ALKYL* and *ENE*, all of which are used as search terms by our users. In 13 issues of CA, *FLUOR* appears in a total of 607 terms that have a total frequency of occurrence of 1174. Of the 607 terms, 204 are right-truncated, 18 are left-truncated and 385 are truncated on both sides. They represent 548, 63 and 563 occurrences respectively (see Table 3).

Thirteen issues were used for this count, and from that small sample it is obvious that the numbers would be quite high for just one year of Condensates. The other two term fragments, *ALKYL* and *ENE* appeared in 729 and 3004 terms respectively, in the 13 issues.

From the overall totals of the KLIC Indexes (See Table 4) we also found that the average BA word is shorter than the average CA word (8.7 characters as opposed to 10.8) and that there are more words in BA title and index terms than in CA (24 as opposed to 14).

4. Application to Inversion

4.1 Size of data base

For 1 year of CA there would be nearly 3,700,000 tokens, representing some 220,000 types, if the graph in Figure 3 is extended from one volume to two volumes. We

Table 3: *FLUOR* Appearances in CA

	FLUOR*	*FLUOR	*FLUOR*	TOTAL
Terms	204	18	385	607
Total Frequency in 13 issues	548	63	563	1174

Table 4: KLIC Index Entries

Issues	Citations	KLIC Entries	Average Word Length
13 CA	67,456	993,264	10.8
26 CA	134,252	1,343,100	N/A
7BA	45,036	462,317	8.7

can then make the following set of assumptions:

- code for each type could be 10 characters
- code for each posting could be 6 characters
- there are about 300,000 citations for the year
- there are about 14 non-trivial words per citation.

The storage requirement for the types would then be 2,200,000 characters; and that for the postings would be 25,200,000 characters for a total of 27.4 million characters for a one-year file.

This does not include authors, corporate authors, etc., nor does it take into account the overhead of the file. Obviously, just a five-year inverted file of unique words and postings to them would be of very large proportions, in the order of 100 million characters.

These numbers, large as they are, are within the realm of current storage devices. This very brief exercise does not, of course, give any indication of the file design, compression and/or coding techniques that would be required, nor of a search strategy for efficient operations upon files of this size, just storage requirements.

4.2 On File Organization

A more interesting result of an analysis of a KLIC index is the light it throws on the problems of file organization. If we assume that we are going to make retrospective searches on such a large file of uncontrolled vocabulary content and retain the capability of searching on word fragments, with both right- and left-hand truncation, some method will have to be found of pointing to all words in the file containing any given fragment. If someone used the term fragment *FLUORO* we would need some way to point to all the terms containing that fragment (607 in 13 issues, more in a larger file). One way would be to store the term types in a KLIC format, but this would increase either the storage requirement by about an order of magnitude, or require a great deal of pointer processing. This does not, therefore, appear to be a very feasible approach.

Having posed the problem and giving some of its dimension, I have to state here that we do not at present have a solution. We do feel that this type of analysis has been valuable, however, in giving us some idea of the extent of the problem. We are beginning to explore some possible solutions to the problems posed by this study.

4.3 Exploration Points

4.3.1 The first possibility is to reduce term types via a hash code to some fairly unique representation.

One implementation of this idea would be to assign numbered bits to the various characters and character combinations within a term type. A bit could be switched for each letter in the term, another for each occurrence of additional vowels, another for certain diphthongs, etc. This technique would do away with the truncation problem.

4.3.2. A second possibility, which is an offshoot of the first, would be to tie the hash coding algorithum to the relative frequency of appearance of various letters and/or letter combinations in the data base. We are currently using a search algorithum in our SDI programs that makes use of some of the information on frequency that we have gleaned from these studies, and it is a very effective technique. We hope we can make similar use of the knowledge in our research on searching large inverted files.