# A METHOD TO REDUCE LARGE NUMBER OF CONCORDANCES.

María Pozzi, Javier Becerra, Jaime Rangel, Luis Fernando Lara.

Diccionario del Español de México.
El Colegio de México
Camino al Ajusco #20 México 20,D.F.
MEXICO

## Summary

In order to help to solve the problem of analysing large number of concordances of a given word 'W', the 'Diccionario del Español de México' (DEM), has implemented a programme that

i) Reduces this number, as to obtain the maximum possible information with the minimum number of concordances to be handled.

ii) Sortes and rearranges the output so that similar concordances are printed out together.

This was done by comparing up to four words to the left and to the right of word W, through the whole set of concordances, associating together those which were repeated in a particular context. Once knowing this, some significant concordances were selected to be printed out, and the rest was discarded.

## I Introduction

In the composition of a dictionary,those involved in the definition of each word have to study very consciously its set of concordances, so that no meaning or use is missed.

there are, of course, some difficulties since on one hand, the sample is never large enough as to insure the occurrence of all the different meanings and uses of every word to be defined. This problem is solved by consulting other dictionaries and expertees on the particular subject.

On the other hand, there are words having a very large number of occurrences, making their analysis a very difficult task, since it is not possible to have present in mind everything that is being analysed. At first thought this could be solved by taking at random a smaller number of concordances; however, when reducing in this way, one is about to loose the grammatical and semantic information contained in all those concordances to be taken away; hence a method had to be implemented as to attain the maximum possible information.

In order to solve this problem, the DEM presents a method whose aim is to obtain optimal information with the minimum number of concordances to be handled.

This method consists of, for each concordance to analyse and compare four words to the left and to the right of word W together with their grammatical category associated, and establishing which one of them is identical to which other in a particular context: A tree structure is generated.

Having known this, it is proceeded to reduce the number, by selecting some of them considered to be representatives.

## II Preliminary Requirements

Our sample (Corpus del Español Mexicano Contemporáneo: CEMC), consists of 1,973,151 occurrences, resulting in 65,200 different types,[1] whose frequency vary from 1 to 68,252.[2]

Some preliminary work has been done consisting in the automatic labeling of each and every word of the corpus with its grammatical category,[2] in which from the total number of occurrences, 1,083,945 were automatically solved, and the rest had to be solved by hand, then the computer was fed with the results, obtaining in this form, the complete sample labelled. We took advantage of this work, since otherwise it would have been impossible to try to reduce the number of concordances in terms of the same grammatical category.

Next, was to implement a programme that produces, for any given word, its set of concordances; each word stating its own grammatical category. This is stored in a file called CONCUERDA, and it is organized in the following way:

Every concordance has three lines, each one of them consisting of:

- 6 characters (nnnnnn) reserved for the number of occurrence.
- 12 characters (tttppplll) reserved for the register of that line, according to the original text, and stating text code, page and line.
- 72 characters reserved for the actual text
- 18 characters for the label of each word of the line, stating the grammatical category code. The first two characters indicate the number of words in the line.

Figure number 1 shows part of file CONCUERDA and its organization.

## III The Algorithm

### 3.1 Association of the i-Concordance to table ORDENA.

For each concordance, a table ORDENA is associated in the following way:

- The word in question is located in the middle line and associated to ORDENA(5)
- Four words are selected to the right and to the left of W, since they are supposed to be carrying the most significant grammatical and semantic information about the word W.[3] We took this idea from the Centre du Trésor de la Langue française"s work concerning to the treatment of binary groupes
- Each of the next four words to the right of W will take its place in $0_{i+1}$ if and only if

  $w_{5+i} \in 0_{5+i}$ and $\nexists$ punctuation mark $p_i$ such that

  $w_{5+i-1} \; p_i \; w_{5+i}$ and $p_i \in \{.,;:¿?¡\}$

  as they are considered to break up the continuity of a context.
- In similar way, the words to the left of W are associated to their place in ORDENA.

Figure No. 2 shows how to construct table ORDENA from a given concordance.

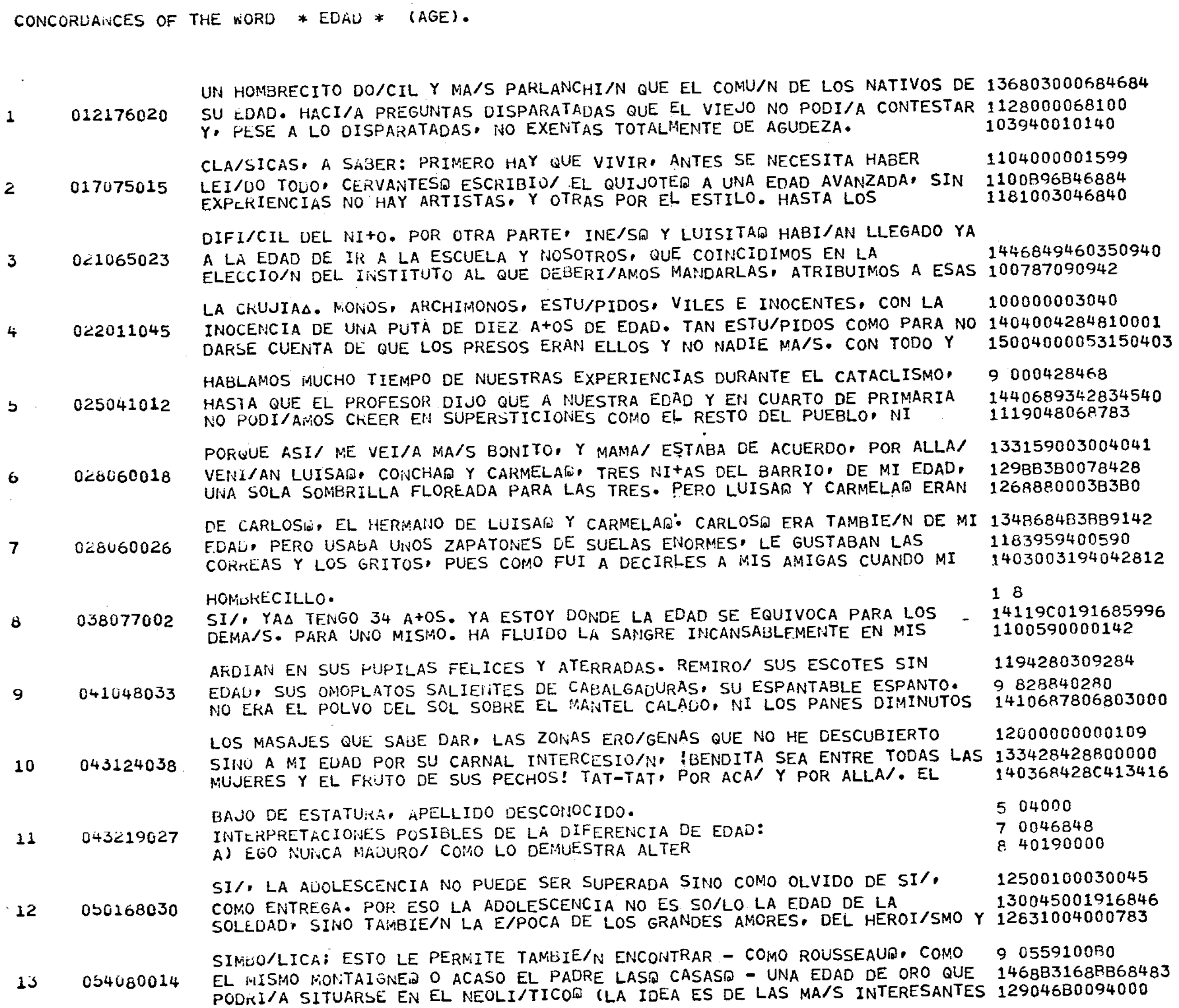
### 3.2 Generation of a Tree Structure starting from ORDENA.

Once obtained this set of up to nine words, it is proceeded to construct a tree structure for the words to the right of W and one for the words to the left of W.

It will only be described here the construction of the right branch of the tree. The left is generated immediately after, though in symmetric form:

- The tree has a root node which is the word W itself, and has five levels, being the root in level 5.
- A direct descendant of a node $w_i$ is given by the word $w_j$ such that $w_i w_j$ are adjascent, i.e. if $w_i \in$ ORDENA$_i$ and $w_j \in$ ORDENA$_{i+1}$ then $w_j$ is a direct descendant of $w_i$.
- The label of each node consists of:
  - Word w associated.
  - Its grammatical category.
  - Its frequency.

  And pointers to:
  - Direct ascendant.
  - First direct descendant.
  - Next node whose direct ascendant is the same as the one of itself.
  - Another file called CONCORD, where it is stored the number of the concordance or concordances where that word in that



```
CONCORDANCES OF THE WORD  * EDAD *  (AGE).


                   UN HOMBRECITO DO/CIL Y MA/S PARLANCHI/N QUE EL COMU/N DE LOS NATIVOS DE 136803000684684
1     012176020    SU EDAD. HACI/A PREGUNTAS DISPARATADAS QUE EL VIEJO NO PODI/A CONTESTAR 1128000068100
                   Y, PESE A LO DISPARATADAS, NO EXENTAS TOTALMENTE DE AGUDEZA.            103940010140

                   CLA/SICAS, A SABER: PRIMERO HAY QUE VIVIR, ANTES SE NECESITA HABER      1104000001599
2     017075015    LEI/DO TODO, CERVANTES@ ESCRIBIO/ EL QUIJOTE@ A UNA EDAD AVANZADA, SIN  1100B96B46884
                   EXPERIENCIAS NO HAY ARTISTAS, Y OTRAS POR EL ESTILO. HASTA LOS          1181003046840

                   DIFI/CIL DEL NI+O. POR OTRA PARTE, INE/S@ Y LUISITA@ HABI/AN LLEGADO YA
3     021065023    A LA EDAD DE IR A LA ESCUELA Y NOSOTROS, QUE COINCIDIMOS EN LA          1446849460350940
                   ELECCIO/N DEL INSTITUTO AL QUE DEBERI/AMOS MANDARLAS, ATRIBUIMOS A ESAS 100787090942

                   LA CRUJIA∆. MONOS, ARCHIMONOS, ESTU/PIDOS, VILES E INOCENTES, CON LA    100000003040
4     022011045    INOCENCIA DE UNA PUTA DE DIEZ A+OS DE EDAD. TAN ESTU/PIDOS COMO PARA NO 1404004284810001
                   DARSE CUENTA DE QUE LOS PRESOS ERAN ELLOS Y NO NADIE MA/S. CON TODO Y   1500400005315O403

                   HABLAMOS MUCHO TIEMPO DE NUESTRAS EXPERIENCIAS DURANTE EL CATACLISMO,   9 000428468
5     025041012    HASTA QUE EL PROFESOR DIJO QUE A NUESTRA EDAD Y EN CUARTO DE PRIMARIA   1440689342834540
                   NO PODI/AMOS CREER EN SUPERSTICIONES COMO EL RESTO DEL PUEBLO, NI       1119048068783

                   PORQUE ASI/ ME VEI/A MA/S BONITO, Y MAMA/ ESTABA DE ACUERDO, POR ALLA/  133159003004041
6     028060018    VENI/AN LUISA@, CONCHA@ Y CARMELA@, TRES NI+AS DEL BARRIO, DE MI EDAD,  129BB3B0078428
                   UNA SOLA SOMBRILLA FLOREADA PARA LAS TRES. PERO LUISA@ Y CARMELA@ ERAN  1268880003B3B0

                   DE CARLOS@, EL HERMANO DE LUISA@ Y CARMELA@. CARLOS@ ERA TAMBIE/N DE MI 134B684B3BB9142
7     028060026    EDAD, PERO USABA UNOS ZAPATONES DE SUELAS ENORMES, LE GUSTABAN LAS      1183959400590
                   CORREAS Y LOS GRITOS, PUES COMO FUI A DECIRLES A MIS AMIGAS CUANDO MI   1403003194042812

                   HOMBRECILLO.                                                            1 8
8     038077002    SI/, YA∆ TENGO 34 A+OS. YA ESTOY DONDE LA EDAD SE EQUIVOCA PARA LOS  -  14119C0191685996
                   DEMA/S. PARA UNO MISMO. HA FLUIDO LA SANGRE INCANSABLEMENTE EN MIS      1100590000142

                   ARDIAN EN SUS PUPILAS FELICES Y ATERRADAS. REMIRO/ SUS ESCOTES SIN      1194280309284
9     041048033    EDAD, SUS OMOPLATOS SALIENTES DE CABALGADURAS, SU ESPANTABLE ESPANTO.   9 828840280
                   NO ERA EL POLVO DEL SOL SOBRE EL MANTEL CALADO, NI LOS PANES DIMINUTOS  1410687806803000

                   LOS MASAJES QUE SABE DAR, LAS ZONAS ERO/GENAS QUE NO HE DESCUBIERTO     12000000000109
10    043124038    SINO A MI EDAD POR SU CARNAL INTERCESIO/N, !BENDITA SEA ENTRE TODAS LAS 133428428800000
                   MUJERES Y EL FRUTO DE SUS PECHOS! TAT-TAT, POR ACA/ Y POR ALLA/. EL     140368428C413416

                   BAJO DE ESTATURA, APELLIDO DESCONOCIDO.                                 5 04000
11    043219027    INTERPRETACIONES POSIBLES DE LA DIFERENCIA DE EDAD:                     7 0046848
                   A) EGO NUNCA MADURO/ COMO LO DEMUESTRA ALTER                            8 40190000

                   SI/, LA ADOLESCENCIA NO PUEDE SER SUPERADA SINO COMO OLVIDO DE SI/,     12500100030045
12    050168030    COMO ENTREGA. POR ESO LA ADOLESCENCIA NO ES SO/LO LA EDAD DE LA         130045001916846
                   SOLEDAD, SINO TAMBIE/N LA E/POCA DE LOS GRANDES AMORES, DEL HEROI/SMO Y 12831004000783

                   SIMBO/LICA; ESTO LE PERMITE TAMBIE/N ENCONTRAR - COMO ROUSSEAU@, COMO   9 05591000B0
13    054080014    EL MISMO MONTAIGNE@ O ACASO EL PADRE LAS@ CASAS@ - UNA EDAD DE ORO QUE  1468B3168BB68483
                   PODRI/A SITUARSE EN EL NEOLI/TICO@ (LA IDEA ES DE LAS MA/S INTERESANTES 129046B0094000
```

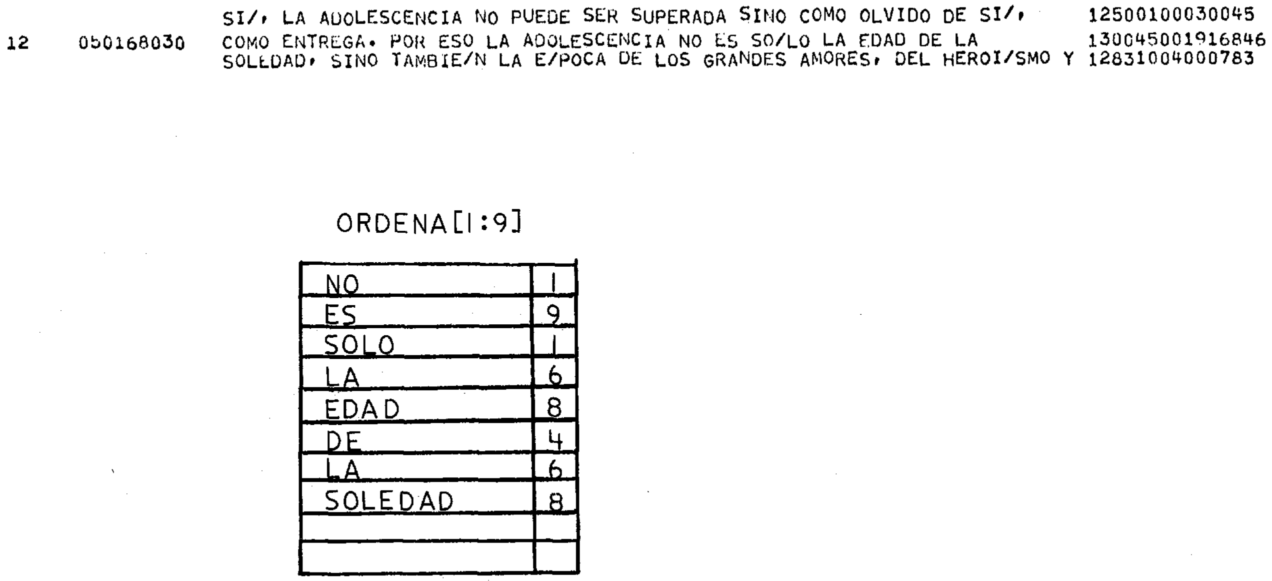
Figure No. 1  File CONCUERDA, where the concordances of the word W in question are stored.

particular context came from, making in this way possible the retrieval operation.

-A node has as many branches as different words are found to be direct descendants to that word, with the same grammatical category through the whole set of concordances.

The process repeats itself until the last concordance has been processed.

Figure No. 3 shows, for a set of 14 concordances, the left and right trees generated.



```
                 SI/, LA ADOLESCENCIA NO PUEDE SER SUPERADA SINO COMO OLVIDO DE SI/,      12500100030045
12   050168030   COMO ENTREGA. POR ESO LA ADOLESCENCIA NO ES SO/LO LA EDAD DE LA          130045001916846
                 SOLEDAD, SINO TAMBIE/N LA E/POCA DE LOS GRANDES AMORES, DEL HEROI/SMO Y  12831004000783
```

ORDENA[I:9]

| | |
|---|---|
| NO | 1 |
| ES | 9 |
| SOLO | 1 |
| LA | 6 |
| EDAD | 8 |
| DE | 4 |
| LA | 6 |
| SOLEDAD | 8 |
| | |
| | |

Figure No. 2 Table ORDENA is obtained from a given concordance. Note that ORDENA(9) is void, since there is a comma (,) after the word 'soledad'

```
                  SUCEDI/A ALLA/ POR EL A+O DE 1881, CUANDO DON¬ PEPE@ TENI/A UNOS 55       13914684C10B06C
90     323023064  A+OS DE EDAD Y MUCHOS RI+ONES AU/N. TUVO UN IMITADOR NOTABLE, QUE FUE     1384832010630 9
                  UN BANDERILLERO LLAMADO ANTONIO@ GONZA/LEZ@ EL¬ ORIZABE+O, QUIEN DIO A    10680B868594

                  AHORA, LA EMPRESA QUE LA TIENE RENTADA, SE ESTA/ GASTANDO UN DINERAL EN  13100 0 59 684
91     324034073  ESTE SERIAL, BUSCANDO NUEVOS VALORES, MISMOS QUE - HASTA QUE SU EDAD SE  122800 343285
                  LOS PERMITA - NO HABRA/N DE SALIR DE ENTRE LOS NI+OS TOREROS.             1159104046683

                  CONSEGUIR DINERO PARA SACAR ADELANTE LA FUNDACIO/N. PRIMERO HABLO/ EL     100       96
92     333060012  SE+OR CURA QUE ENTONCES NO TENI/A NI TREINTA A+OS DE EDAD. LUEGO DON      1380011038848010
                  TOMA/S@ SA/NCHEZ@ (ESTE SI/ VIEJO Y COLUDO) PROPUSO COLECTAS Y RIFAS.     1185219300 3

                  CABALLOS.             1 0
93     333148021  EN SAN* JOSE/@ HABI/A MEDIO MILLAR DE HOMBRES EN EDAD DE TOMAR LAS        134880084843496
                  ARMAS E IRSE A LA GUERRA, PERO NO TODOS SE SINTIERON CON A/NIMOS DE       148304031 59494

                  CASADOS Y TENI/AN HIJOS. LOS MA/S ERAN JO/VENES EN EL VERDOR DE LA        1383980 46846
94     333148034  EDAD, DE 16 A 30 A+OS, CON ALGUNA DESTREZA EN EL MANEJODE ARMAS Y        1584C4C040 468483
                  CABALLOS Y SIN DISCIPLINA MILITAR.                   5 03400

                  ENCUBIERTOS DEL DIABLO, O AL MENOS DO/CILES INSTRUMENTOS DE SUS AVIESOS  110783700 428
95     335044023  DESIGNIOS, LA BEATA IMAGEN DE LA EDAD DE ORO REDIVIVA SE TRANSMUTO/, AL  130088468438597
                  CONJURO DEL DESENGA+O, EN EDAD DE HIERRO EN QUE DOMINABA LA CRECIENTE     12878484040 

                  DESIGNIOS, LA BEATA IMAGEN DE LA EDAD DE ORO REDIVIVA SE TRANSMUTO/, AL  1300 46840 597
96     335044024  CONJURO DEL DESENGA+O, EN EDAD DE HIERRO EN QUE DOMINABA LA CRECIENTE     1287848404500
                  CONVICCIO/N DE QUE ESOS DESNUDOS HIJOS DEL OCE/ANO¬ FORMABAN PARTE DEL    110402807C907

                  INDI/GENAS, COMO ES AU/N, EN PARTE ESTE/RIL, SINO QUE REALIZARI/A SU      1101914033 2
97     342257019  PROGRESIVA EDUCACIO/N EN LA ADOLESCENCIA Y HASTA EN LA EDAD ADULTA'.     118040 34 685
                  EN EL PLAN DEFINITIVAMENTE REGENERADOR DICTADO EN EL LLANO DEL RODEO¬     1146810046878

                  JURA/IS Y YO PIERDO UN ALUMNO.                6 935968
98     343065005  PERO DESDE LA EDAD DE OCHO O NUEVE A+OS HASTA LA DE DIECINUEVE O VEINTE 153468483984040830
                  NO EXISTE EL DESEO DE UN TRABAJO MANUAL PESADO. ESTO ESEXACTO EN LA      1410684680 59040

                  PERCIBIR SUS CUALIDADES TANTO MATERIALES COMO FUNCIONALES, ASI/ COMO      9 0280   1
99     344096030  SU CONVENIENCIA RESPECTO A LA EDAD DE QUIEN LA IBA A USAR; OBSERVO/ SU   1428146845594092
                  CONTENIDO Y MANEJO, SE DIO CUENTA DE SU PESO Y RESISTENCIA ASI/ COMO DE  1483059842830104

                  DESARROLLO DE LA IMAGINACIO/N CREADORA YALGUNAS HABILIDADES PARA OPERAR  9 8409
100    344096036  CON HERRAMIENTAS SENCILLAS; ES, ADEMA/S, ADECUADO A LA EDAD DE MI HIJO.  12400918468428
                  INSTRUCCIONES. LA PRESENTE ESCALA CONTIENE OCHO ASPECTOS ESENCIALES EN   9 00      4

                  LE ES FA/CIL HACER AMISTADES: 'ME ES BASTANTE FA/CIL HACERLAS Y ME        125900 590  35
101    345322048  GUSTA QUE SEAN ALEGRES, DE MI EDAD Y TENGAN UN NIVEL CULTURAL POCO MA/S  14909842839680
                  O MENOS COMO EL MI/O.' Y FRENTE A UN GRUPO DE NI+OS: 'ME DA GUSTO VER    16300653346848598O

                  TERMINAR LA CARRERA DE MEDICINA.                   5 0  4
102    345350029  CASO 2. ALUMNO DE 19 A+OS DE EDAD; SEXO MASCULINO. PROCEDENTE DE LA       130C84C8400 4
                  ESCUELA¬ DE ARQUITECTURA DE UNA UNIVERSIDAD DE PROVINCIA.                 8 04046840

                  IDENTIFICA CON LA PLENA REALIZACIO/N DE LAS ASPIRACIONES QUE EL HOMBRE   1 940 4   68
103    346037019  TIENE DESDE EL TE/RMINO DE LA EDAD¬ MEDIA¬, Y NO SE DANCUENTA DE QUE     150468468 3159840
                  EL A+O 200  PIEDE NO SE LA CULMINACIO/N ROTUNDA Y FELIZDE UN PERIODO     1468C0153  3 468
```

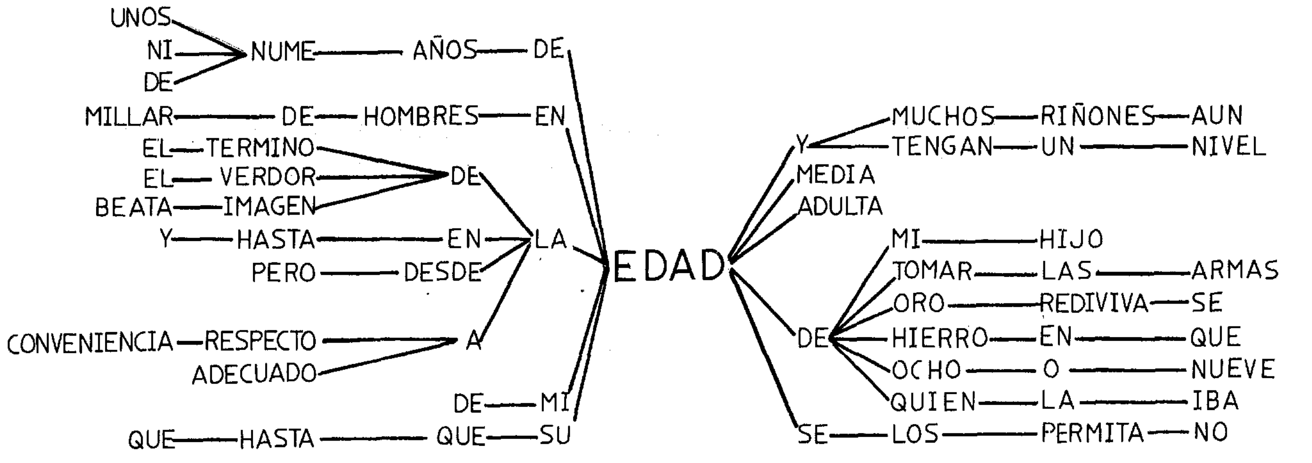


Figure No. 3 Left and right trees generated from a set of 14 concordances.

### 3.3 The algorithm to select significant concordances.

Once the tree is fully constructed it is proceeded to make the actual reduction.

There are some facts to be considered beforehand:

- The more words repeated exactly in the same context, the greater is the probability that the meaning of the word W in that context is the same.
- A set of words repeated a small number of times may be more significant than another one repeated a larger number of times since there are not so many different meanings or grammatical functions of a word W followed by the same set of words.

Next, it will be described the procedure:

In order to analyse the tree, a leftmost path is followed.

- A 6th level branch of the tree is first analysed (Remember that the root is in level 5, and that the tree to the right of W is being analysed). If the frequency is greater than 1, then its leftmost direct descendant is analysed in the same way.
- If a 9th level rode is reached in this form, and the frequency $n > 1$, it means that the words W followed by these four words ocurred a times in n different concordances. As it was said before there is a good probability that the meaning of the word W in this particular context is the same in all of the n concordances; hence, by talking only one or two of them, by means of a random function, we obtain a significant concordance, and the ( n - 1) or (n-2) left can be safely omited from the final output.
- If at same intermediate level it is found that the frequency of the word associated to that node is 1, then the analysis of such branch would have to be stopped; however, it was thought that a possible way to reduce was not by identical words but by the same grammatical category. It is preceeded then to find all direct descendants of its own direct ascendant with the same frequency and grammatical category, and then the number of these concordances is reduced.

It is clear that the process takes into account that as the level of reduction is closer to 5, then the context is less significant; hence a larger number of concordances have to be chosen to mantain the required quality information.

After some study and many trials it was empirically decided by our team of linguists* that a reasonable pattern of reduction was the following:

- If the level of reduction is 4 or 6 and the frequency $F \leq 30$ then the number of concordances selected Q would be
Q=F//2 + 1 and
Q=F//4 if $F > 30$.
- If level is 7 or 3 then
Q = F//3 + 1 for $F \leq 50$
Q = F//5 for $F > 50$
- If level is 8 or 2 then
Q = F//4+1 for $F \leq 70$
Q = F//7 for $F > 70$

Finally, if level is 9 or 1 then
Q = F//5 + 1 for $F \leq 50$
Q = F//10 + 1 for $F > 50$

* At this point, we would like to thank in particular to Paulette Levy for her valuable discussions and interesting suggestions.

It has to be mentioned here, that this pattern of reduction may be changed according to the wprd analysed, as to obtain the best results each time.

When it is already Known the number of concordances that will be chosen ( Q out of F) it is proceeded to select them again, by means of a random function, and each one of them is marked as such, to avoid any one of them be selected twice or more times.

3.4 Output.

The final output is presented indicating the group of words repeated the grammatical category of the last word - when applicable - and the frequency. Next, the Q concordances chosen are listed below.

Figure No 4 shows the form in which the output is presented.

## IV The Computational System.

The system was implemented in the University of Norway version of ALGOL 60 NUALGOL for a UNIVAC 1106 computer of the "Centro de Procesamiento Arturo Rosenblueth" of the Secretaría de Educación Pública (Ministry of Education), with 262K words of 36 bites of central memory and 8,000,000 of characters in disc.

### 4.1 Data Storage.

We made use of 3 files:

a) File CONCUERDA, where the whole set of concordances of the word W was stored, and it was described above.

b) Files ARBOL and CONCORD; these two files are supposed to contain the information obtained while generating the right and left trees.

ARBOL: Each node of the tree is stored in a line composed of 72 characters, distributed in the following way:

7 for its own address in file ARBOL

1 for the level

24 for the word

2 for the grammatical category

3 for the length of the word

4 for the frequency

7 for the address of its direct ascendant

7 for the address of the next direct descendant of its own direct ascendant (i.e. like next brother)

7 for the address of the first direct descendant

4 for the number of direct descendants (i.e. No of branches emerging from it) and

6 for the address in file CONCORD where it is stored the number of the concordance where it comes from.

From the computational point of view, each one of the trees is generated in the following way:

- The root, whose node associated is the word W is in a prefixed address, and it will be present in every concordance. This word is taken from ORDENA ( 5 )
- The next word in ORDENA will be stored by means of a hash function, and it is decided to be the same node as one previously stored, if and only if the word, its grammatical category, level and direct ascendant are exactly the same, in such case the frequency is aumented by one and in file CONCORD is stored the number of this concordance in addition to the previous one.

```
                 CONCORDANCIAS REDUCIDAS DE LA PALABRA ** EDAD                       **  CON FRECUENCIA TOTAL * 379

REDUCCION POR LA DERECHA:

      EDAD AVANZADA   FREC= 3

                                     OPORTUNA, Y LOS MEDICAMENTOS ADECUADOS, SUPRIMIE/NDOSE TODA CLASE DE
             188081055       1.-     ENSAYOS Y EXPERIMENTOS CON SERES EN EDAD AVANZADA. HACEMOS VOTOS MUY
                                     FERVIENTES POR QUE TALES CONCLUSIONES SE LLEVEN A LA PRA/CTICA

      EDAD DE LA + NM    FREC= 5

                                     CRISTALITOS DE SISA. LOS RESULTADOS SUGIEREN QUE LAS U/NICAS
             469322047       2.-     DIFERENCIAS OBSERVADAS SE EXPLICAN EN FUNCIO/N DE LA EDAD DE LA
                                     PROTEI/NA PERO QUE NO EXISTEN VARIACIONES ESTRUCTURALES INTRI/NSECAS

                                     SI/, LA ADOLESCENCIA NO PUEDE SER SUPERADA SINO COMO OLVIDO DE SI/,
             050168030       3.-     COMO ENTREGA. POR ESO LA ADOLESCENCIA NO ES SO/LO LA EDAD DE LA
                                     SOLEDAD, SINO TAMBIE/N LA E/POCA DE LOS GRANDES AMORES, DEL HEROI/SMO Y

      EDAD DE ORO   FREC= 3

                                     SIMBO/LICA; ESTO LE PERMITE TAMBIE/N ENCONTRAR - COMO ROUSSEAU@, COMO
             054080014       4.-     EL MISMO MONTAIGNE@ O ACASO EL PADRE LAS@ CASAS@ - UNA EDAD DE ORO QUE
                                     PODRI/A SITUARSE EN EL NEOLI/TICO@ (LA IDEA ES DE LAS MA/S INTERESANTES

                                     ENCUBIERTOS DEL DIABLO, O AL MENOS DO/CILES INSTRUMENTOS DE SUS AVIESOS
             335044023       5.-     DESIGNIOS, LA BEATA IMAGEN DE LA EDAD DE ORO REDIVIVA SE TRANSMUTO/, AL
                                     CONJURO DEL DESENGA+O, EN EDAD DE HIERRO EN QUE DOMINABA LA CRECIENTE

      EDAD DE LOS + NM    FREC= 5

                                     SILENCIOSOS.
             408212010       6.-     LA EDAD DE LOS PECES SE PUEDE DETERMINAR EN MUCHOS CASOS CONTANDO EL
                                     NU/MERO DE ANILLOS DE LAS ESCAMAS, LOS CUALES REPRESENTAN ZONAS DE

                                     CASANDRA@
             107010049       7.-     (UN POCO PEDANTE) SI QUIEREN DECIRLO ASI/. BUENO, CIERTAMENTE LA EDAD
                                     DE LOS YELMOS BRILLANTES COMO ESPEJOS NO ES E/STA (EN CRESCENDO BRIOSO)

      EDAD DE NUME A+OS   FREC= 3

                                     ANISOMETROPI/A PUEDE SER DISMINUIDO ENORMEMENTE POR UN PEDIATRA ALERTA
             472302007       8.-     O UN ME/DICO GENERAL QUE EXAMINE LA AGUDEZA VISUAL A LA EDAD DE 4 A+OS.
                                     SE PUEDE SOLICITAR LA AYUDA DE LA MADRE; A ELLA SE LEPUEDE DAR UNA
```

Figure No. 4  Final Output of the selected concordances of the word EDAD (AGE).

TREE STRUCTURE GENERATED FOR WORD *EDAD* (AGE).

```
31596  2DE                       PR2  1  30936                           4647
31608  3Y                        CO1  1  32712  36576  36552  1   3228
31620  4MI                       AJ2  9  12     33168  32100  4   69
31632  2MODERADAMENTE            AV13 1  36360                           2046
31644  2PUES                     CO4  1  32436                           3879
31656  3Y                        CO1  1  34008  32496  37344  1   3987
31668  3Y                        CO1  1  32688         5232   1   4149
31680  2DE                       PR2  2  33372                           4674
31692  2DE                       PR2  1  31056                           4749
31716  3SO/LO                    AV5  1  34008  32208  34644  1   141
31728  4ESA                      AJ3  5  12     31896  32232  4   507
31740  2ALGU/N                   AJ6  2  36312                           2685
31752  2PROBABLEMENTE            AV13 1  31944  35028                    1563
31800  2CON                      PR3  1  31980  31920                    219
31812  2CON                      PR3  1  35400                           633
31824  2CON                      PR3  1  35664  33192                    1680
31836  4SU                       AJ2  16 12     34344  32088  6   3
31848  4TU                       AJ2  2  12     31728  31116  2   498
31860  4TAL                      AJ3  1  12     32352  32280  1   2802
31896  4POCA                     AJ4  2  12     32076  32148  1   540
31908  2INVERSAMENTE             AV12 1  32424  32580                    1749
31920  2PARA                     PR4  1  31980                           2607
31932  3A                        PR1  10 34344  5556   5472   5   15
31944  3A                        PR1  25 34008  30828  3468   20  27
31956  3A                        PR1  1  32832         31332  1   54
31968  3A                        PR1  1  31620  32880  31560  1   114
31980  3QUE                      CO3  2  34008  32112  31800  2   216
31992  4TODA                     AJ4  1  12     32124  32532  1   366
32004  3A                        PR1  2  31728  36324  35496  1   537
32016  3QUE                      CO3  1  31836  32568  32160  1   1128
32028  3A                        PR1  1  33060                           1740
32040  3CIERTAMENTE              AV11 1  34008  35604                    411
32052  4ESTA                     AJ4  6  12     33252  32184  5   1980
32064  2POR                      PR3  2  30504                           2625
32076  4CUYA                     AJ4  1  12     32556                    648
32088  3DE                       PR2  9  31836  32016  35256  8   6
32100  3DE                       PR2  6  31620  31968  31344  4   72
32112  3DE                       PR2  21 34008  32040  37320  19  345
32124  4OTRA                     AJ4  1  12     31848  32136  1   381
32136  3DE                       PR2  1  32124         34524  1   384
32148  3DE                       PR2  2  31896         30996  2   543
32160  2HASTA                    PR5  1  32016                           1131
32172  2HASTA                    PR5  1  32208  34836                    1215
32184  3DE                       PR2  1  32052  32304  35124  1   1983
32196  2CONFORME                 CO8  1  35628                           2346
32208  3EN                       PR2  15 34008  31980  35340  11  189
32220  6YA                       AV2  2  24     34080  39324  1   222
32232  3EN                       PR2  1  31728  32004  35976  1   510
32244  3EN                       PR2  1  34344  32796  35892  1   1809
32256  6NO                       AV2  2  24     39264  37884  2   651
32268  4ESTE                     AJ4  2  12     31860  36156  1   2787
32280  3A                        PR1  1  31860         36168  1   2805
32292  3PERO                     CO4  2  32712  31608                    2877
32304  3A                        PR1  2  32052  32856  36804  1   3000
32316  2LE                       PN2  1  36900                           3159
32328  3EXACTAMENTE              AV11 1  34008  30468  36516  1   3165
32340  3A                        PR1  1  31848         5220   1   3192
```

Figure No. 5 File ARBOL, where the tree structure is generated.

- Otherwise it will be a new rode.
  Figure No 5 shows part of file ARBOL, EDAD (AGE) is being processed.

## V Results And Applications.

The first results were very encouraging, since for those words with medium number of concordances - say up to 600 - we were able to reduce the number between 30% and 40%, according to the word in question.

No lost information was reported (by comparing the original set of concordances with the reduced version)

It is expected that for words with higher frequency, the method here described will be more efficient.

However, from the computational point of view, there are still some difficulties, since the generation of each tree is very time consuming as the frequency of the word in question increases. We are still working to optimize it.

The most important application besides the original main objectives, is that by this method it is possible to find expressions and patterns of language repeated and used consistently.

## VI References

1.- Roberto Ham Chande: Del 1 al 100 en Lexicografía, in Investigaciones Linguísticas en Lexicografía, Jornadas 89 El Colegio de México, 1979

2.- Isabel García Hidalgo: La Formalización del Analizador Gramatical del DEM y
Luis Fernando Lara y Roberto Ham Chande: Base Estadística del DEM in Investigaciones linguísticas en Lexicografía. Jornadas 89 El Colegio de Mêxico, 1979.

3.- G.Gorcy, R.Martin, J.Maucourt, R.Vienney
Centre du Trésor de la Langue Francaise: Le Traitement des Groupes Binaries. Cahiers de Lexicologie. 17 - 1970 - II