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) Sorts 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, whose frequency vary from 1 to 68,252.

Some preliminary work has been done consisting in the automatic labeling of each and every word of the corpus with its grammatical category, 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. 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
  \[ w_{5+i} \leq 0_{5+i} \] and 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 adjacent, i.e. if \( w_i \in \text{ORDENA}_i \) and \( w_j \in \text{ORDENA}_{i+1} \) 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
CONCOURANCES OF THE WORD "EDA" (AGE).

1. 012176020 UN HOMBRECITO BO/CIL Y MA/S PARLANCHÍN QUE EL COMÚN DE LOS NATIVOS DE 1368030006846848
SU EDAD HACÍA PREGUNTAS DISPARATADAS QUE EL VIEJO NO PODÍA CONTESTAR 11280000081010
Y PLE A LO DISPARATADAS NO EXISTEN TOTALEMENTE DE AGUDEZA...
103940010140
2. 017076515 CLÁSICAS A SABER: PRIMERO HAY QUE VIVIR ANTES SE NECESITA HABER 11040000015999
10089668668686
LEÍ/DO TODO CERVANTES ESCRIBÍÓ EL QUIJOTE A UNA EDAD AVANZADA SIN 11870030046800
EXPLICACIONES NO HAY ARTISTAS, Y OTRAS POR EL ESTILO. HASTA LOS
3. 011065263 DIFÍCIL DEL NIÑO, POR OTRA PARTE INECI0 Y LUSITAN HABI/AN LLEGADO YA 104689460350980
A LA EDAD DE IR A LA ESCUELA Y NOSOTROS, QUE CONOCIMOS EN LA 1078704000942
ELECCIÓN DEL INSTITUTO AL QUE DEBERÍAMOS MA/CHARAS ATRIBUIMOS A EAS
4. 020201045 LA GALAXIA, MONOS, ARCHIMONOS, ESTUPIDOS, VÍELES E INOCENTES, CON LA 1000000003040
INOCENCIA DE UNA PUTA DE DIEZ AROS DE EDAD, TA ESTU/PIDOS COMO PARA NO 1500000005210403
DARSE CUENTA DE QUE LOS PRESOS EREN ELLOS Y NO HABÍA MA/S CON TODO Y
5. 020501412 HABLAMOS MUCHO TIEMPO DE NUESTRAS EXPERIENCIAS DURANTE EL CATACLISMO 9 004028448
HASTA QUE EL PROFESOR DIJO QUE A NUESTRA EDAD Y EN CUERPO DE PRIMARIA 1100000008568783
NO PODÍA/NIOS CREER EN SUPERSTICIONES COMO EL RESTO DEL PUEBLO. NI
6. 020606016 PORQUE ASÍ/ ME VEÍA MA/S BONITO Y MAMA/ ESTABA DE ACUERD0 POR ALLA/ 1209800007426
VEÍ/AN LUISA/ COUCHA Y CARMELA/ TRES NIÑAS DEL BARRIO: DE MI EDAD 1268800003630
UNA SOLA SORRIMILLA FLOREADA PARA LAS TRES, PERO LUISA Y CARMELA Eran
7. 020606026 DE CARLOS, EL HERMANO DE LUISA Y CARMELA; CARLOS ERA TAMBIÉN DE MI 13468403009142
EDAD, PERO USABA UNOS ZAPATONES DE SUELAS ENORMES LE GUSTABAN LAS 1183950400560
COCHEAS Y LOS GUITOS, PUES COMO FUI A DECIRLES A MIS AMIGOS CUANDO MI
8. 020807762 HOMRECILLO... 10 1101900191655996
SÍ/ TAMAÑO 34 AROS, YA ESTOY DONDE LA EDAD SE EQUIVA Para LOS 1108000000142
DENAVS, PARA UNO MISMO, HA FLUIDO LA SANGRE INCUMANTEMENTE EN MIS
9. 021040833 ARDÍAN EN SUS FÍMILAS FELICES Y ATERRADAS REMIRO/ SUS ESCOTES SIN 1108000000142
EDAVS SUS ONGILATES SALITRES DE CABELGAURAS/ ES APANTABLE ESPANTO/ NO 9 0288000290
ERA EL POLVO DEL SOL SOBRE EL MANTEL CALIZO/ NI LOS PASES DIMINUTOS
10. 020320438 LOS MAGAJES QUE SADE DAR, LAS ZONAS ERA/GENAS QUE NO HAY DESCUBIERTO 141067800803000
SÍN A MI EDAD POR EL CERDITO INTERCÉS/ EL ENFISMA ENTRE TODAS LAS 150342842800000
MUJERES Y EL FRIJO DE LOS Pechos TAT-TAT/ POR ACA/ Y POR ALLA/ EL
11. 020412907 RAJO DE ESTATUA APPELLIDO DESCONOCIDO... 8 01900000
INTERPRETACIONES POSIBLES DE LA DIFERENCIA DE EDAD:
AJEJU NARCA MAL LUGO/ CON LO DEMUESTRAL ALTER
12. 0206168030 SÍ/ LA ADOLESCENCIA NO PUEDE SER SUPERADA SINO COMO OLVIDO DE SÍ/... 1259000300545
CONO ENTREGA, POR ESO LA ADOLESCENCIA NO ES SO/LO LA EDAD DE LA 13099001916644
SOLIDAD SINO TAMBIÉN/ LA EPoca DE LOS GRANDES AMORES/ DEL HÉRO/S/NO Y
13. 020412907 EL MISMO MOTTAUEN O ACASO EL PADRE LASO CASASO/ UNA EDAD DE ORO QUE... 14680316886883
PODI/A SITUARSE EN EL NEOLÍTICO/ LA IDEA ES DE LAS MÁS INTERESANTES 12909009000090

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.

---

**ORDENA(1:9)**

| Word | Count |
|------|-------|
| 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’.
SUCEDÍA ALLÍ POR EL AÑO DE 1831, CUANDO DON PÉREZ CONDAN UNOS 54 AÑOS DE EDAD Y MUCHOS RIÑONES AUN. TUVO UN IMITADOR NOTABLE, QUE FUE UN BAUTISTERO ANTONIO GONZÁLEZ LEÓN, ORIZABO.* QUIEN DIO A AHORA LA EMPRESA QUE LA TIENEN RIVERTA SE ESTÁ BASTANDO UN DINERO EN EL ESTE SCIAL, BUSCANDO NUEVOS VALORES; MISMO QUE —HASTA QUE SU EDAD SE LOS PERMITA— NO HABRÁN DE SALIR DE ENTRE LOS NIÑOS TERRENOS.

CONSEGUIR DINERO PARA SACAR ADELANTE LA FUNDACIÓN, PRIMERO HABILÓ AL SEÑOR CURA DE DON PEPE Y TAMBO SNACHETA (ESTE SÍ VIEJO Y CULGADO) PROPUSO COLECCIONAR Y Vender.

CABALLOS, 10 EN SAN JOSE Y HABÍA UN MEDIO MILLAR DE HOMBRES EN EDAD DE TOMAR LAS ARMAS Y EIRSE A LA GUERRA, PERO NO TODOS SE CINTRON CON AMIGOS DE CABALLOS Y TIERRAS ANDI HIJOS; LOS MÁS ERAN JÓVENES EN EL VERDOR DE LA EDAD, DE 16 A 30 AÑOS, CON ALGUNA DESTREZA EN EL MANEJO DE ARMAS Y CABALLOS Y SIN DISCIPLINA MILITAR.

ENCUBIERTOS DEL DIABLO: O AL MENOS DOS CAÍLARES INSTRUMENTOS DE SUS AVIÉSOS: UNOS DE OCHO O NUEVE AÑOS HASTA LA DE DIECISIÚN A VEINTE NO EXISTE EL DERECHO DE UN TRABAJO MANUAL PESADO. ESTO ES EXACTO EN LA PERCEPCIÓN DE SUS CALINDADES TANTO MATERIALES COMO FUNCIONALES, ASÍ COMO SU CONVENIENCIA RESPECTO A LA EDAD DE QUIEN LA IBA A USAR OBEVÍA SU CONTENIDO Y MANEJO. SE DIO CUENTA DE SU PESO Y RESISTENCIA ASÍ COMO DESARROLLO DE LA IMAGINACIÓN CREADORA Y ALGUNAS HABILIDADES PARA OFREER CON HERMANSAS. CON LA PLENA REALICACIÓN DE LAS ASPIRACIONES QUE EL HOMBRE HABÍA EN EL TÉRMINO DE LA EDAD MEDIA, Y NO SE DÁNCEA DE QUE EL AÑO 2006 SE DEJO NI LA CULMINACIÓN ROTUNDA Y FELIZ DE UN PERÍODO:

UNOS

NI NUME—AÑOS—DE—

MILLAR—DE—HOMBRES—EN—

EL Termino—DE—

VERDOR—DE—

BEATA—IMAGEN—

HASTA—EN—LA—

PERO—DESDE—

CONVENIENCIA—RESPECTO—

ADECUADO—

QUE—HASTA—QUE—SUI

MUCHOS—RIÑONES—AUN—TENGAN—UN—NIVEL—

MEDIA—ADULTA—

MI—HIJO—

TOMAR—LAS—ARMAS—

ORO—REDIVIVA—SE—

DE—HIERRO—EN—QUE—

OCHO—O—NUEVO—

QUIEN—LA—IBA—

SE—LOS—PERMITA—NO

EDAD

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 occurred 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 omitted from the final output.

- If at some 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 proceeded 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 maintain 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≤30 then the number of concordances selected Q would be
  \[ Q = \frac{F}{2} + 1 \]  
  \[ Q = \frac{F}{4} \]  
- If level is 7 or 3 then
  \[ Q = \frac{F}{3} + 1 \]  \[ F \leq 50 \]
  \[ Q = \frac{F}{5} \]  \[ F > 50 \]
- If level is 8 or 2 then
  \[ Q = \frac{F}{4} + 1 \]  \[ F \leq 70 \]
  \[ Q = \frac{F}{7} \]  \[ F > 70 \]

Finally, if level is 9 or 1 then

\[ Q = \frac{F}{5} + 1 \]  \[ F \leq 50 \]
\[ Q = \frac{F}{10} + 1 \]  \[ F > 50 \]

* 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 word 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 augmented 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 FRECENCIA TOTAL: 370**

| REDUCCIÓN POR LA DERECHA | **EDAD AVANZADA** | **FREC.** | **1** |
|--------------------------|-------------------|-----------|------|
|                          |                   | 188081056 | 5    |

OPORTUNA, Y LOS MEDICAMENTOS ADECUADOS, SUPRIME/NOSE TODA CLASE DE ENSAYOS Y EXPERIMENTOS CON SERES EN EDAD AVANZADA. HACE VOS MUY FERVIENTES POR QUE TALES CONCLUSIONES SE LLEVEN A LA PRA/CÍTICA

|                          | **EDAD DE LA + NM** | **FREC.** | **2** |
|--------------------------|---------------------|-----------|------|
|                          | 469322047           | 50168030  | 3    |

CRIStALITOS DE SISA. LOS RESULTADOS SUGIERN QUE LAS ÚNICAS DIFERENCIAS OBSERVADAS SE EXPLICAN EN Funcion/ON DE LA EDAD DE LA PROTEÍNA PERO QUE NO EXISTEN VARIACIONES ESTRUCTURALES INTRÍNSECAS SI/ LA ADOLESCENCIA NO PUEDE SER SUPERADA SINO COMO OLVIDO DE SI/ como entrega por eso la adolescencia no es solo la edad de la soledad, sino también/IN LA EPOCA DE LOS GRANDES AMORES, DEL HEROISMO Y

|                          | **EDAD DE ORO** | **FREC.** | **4** |
|--------------------------|----------------|-----------|------|
|                          | 054080019       | 335044023 | 5    |

SIMBO/LICA: ESTO LE PERMITE TAMBIÉN ENCONTRAR - COMO ROUSSEAU, COMO EL MISMO MONTAINEGRO O ACASO EL PADRE LASO CASADO - UNA EDAD DE ORO QUE PUDRI/A SITUARSE EN EL NEOLÍTICO. (LA IDEA ES DE LAS MÁS INTERESANTES EXUEBTEROS DEL DIABLO, O AL MENOS DO/CILES INSTRUMENTOS DE SUS AVIESOS DESIGNOS. LA BEATA IMAGEN DE LA EDAD DE ORO RETÍVIA SE TRANSFORMA/, AL CONJUNTO DEL DESINGA/O, EN EDAD DE HIERRO EN QUE DOMINABA LA CRESCIENTE

|                          | **EDAD DE LOS + NM** | **FREC.** | **5** |
|--------------------------|----------------------|-----------|------|
|                          | 408212010            | 107010049 | 6    |

SILENCIOSOS.

|                          | **EDAD DE NUEVA APOS** | **FREC.** | **7** |
|--------------------------|------------------------|-----------|------|
|                          | 472302007              |           | 6    |

ANISOMETRÍA PUEDE SER DISMINUIDO ENORMEMENTE POR UN PEDIATRA ALERTA O UN MÉDICO GENERAL QUE EXAMINE LA AGUDEZA VISUAL A LA EDAD DE 4 APOS. SE PUEDE SOLICITAR LA AYUDA DE LA MADRE A ELLA SE LE PUEDE 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 0647
31601 3Y 32712 36576 36552 1 3228
31620 4MI 33160 33100 4 69
31632 2MÚQUERADAMENTE 36250 2046
31644 2MÚQUES 36246 3579
31656 3Y 36008 32496 37344 4 397
31666 3T 36268 5232 1 419
31669 3UE 35372 4674
31692 2GUE 31056 4749
31716 350/L0 34008 32208 36464 4 141
31728 4ESA 31980 31920 219
31740 2ALU/N 35400 633
31752 2PROBABILMENTE 35644 35028 1563
31760 2CON 31980 31920 219
31782 2CON 35400 633
31799 2CON 31980 31920 219
31802 3A 31920 31968 32016 2 507
31813 3A 31856 32016 2 507
31824 4TU 31728 31116 2 498
31836 4TAL 32352 32380 1 2802
31848 4FOCA 32076 32148 1 540
31866 2INVERSAMENTE 32424 32580 1749
31900 2PAHA 31980 2067
31912 3A 34500 5572 5 15
31944 2COYA 34008 30828 3568 20 27
31956 3A 35252 31332 1 54
31968 3A 31080 31500 1 114
31980 3GUE 34008 32112 31800 2 216
31992 4TUGA 32324 32352 1 366
32004 3A 32176 32496 1 537
32016 3GUE 31080 31500 1 114
32020 3A 33060 1740
32040 3CIERTAMENTE 34008 35604 411
32052 4ESTA 32352 32184 5 1980
32064 2POH 35604 2065
32076 4COYA 32556 648
32088 3UE 31856 32016 2 507
32100 3DE 31968 31344 4 72
32112 3DE 34008 32040 37320 19 345
32124 4OTHA 31080 31500 1 114
32136 3UE 31216 34204 1 304
32148 3UE 31080 31500 1 114
32160 2HASTA 32016 1131
32172 2HASTA 33060 1740
32184 3DE 32556 648
32196 2CONFORME 35626 2346
32208 3EN 31980 35256 8 6
32220 3TA 31968 31344 4 72
32232 3EN 31856 32112 31848 1 381
32244 3EN 3124 34204 1 304
32256 4H0 31080 31500 1 114
32260 4ESTE 32016 1131
32280 3A 32208 34204 1 304
32292 3PERO 32556 648
32304 3A 32712 31656 2577
32316 3LE 32016 1131
32328 3EXACTAMENTE 33060 1740
32340 3A 32556 648

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 Lexicografia, in Investigaciones Linguisticas en Lexicografia, Jornadas 89 El Colegio de México, 1979

2.- Isabel Garcia Hidalgo: La Formalización del Analizador Gramatical del DEM y Luis Fernando Lara y Roberto Ham Chande: Base Estadistica del DEM in Investigaciones lingüisticas en Lexicografia. Jornadas 89 El Colegio de México, 1979.

3.- Centre du Trésor de la Langue Francaise: Le Traitement des Groupes Binaries. Cahiers de Lexicologie. 17 - 1940 - II