IGNACIO DEL CAMPO – ISABEL GONZALES – Mª TERESA MOLINA – FRANCISCO MARCOS

THE AUTOMATIC SYNTACTIC ANALYSIS AS AN AID IN DICTIONARY MAKING

In this paper we are going to explain some of the problems we have found in our attempts to mechanize the *Historical Dictionary* (D.H.) of the Spanish language. Our experimental project is a collaboration of the Royal Spanish Academy (R.A.E.) and the Computer Center of the University of Madrid (CCUM).

There are, indeed, many procedures for making concordances of a text, and those procedures are, in general, very successful ones. We are not going to be concerned with concordance making; nevertheless, it seems obvious to declare that a good concordance system must be in the basis of our researches. We must suppose that in our concordances we get syntactically limited utterances, i.e. that in our concordances we do not have to deal with words belonging to sentences whose verbs are not included in the text given in the concordance. By now, the best method of getting this kind of concordances is to limit them by full stops. So, we get our text divided by full stops or semicolons and we analyze as many sentences as verbs could be included between those punctuation marks. But, and we must say it quickly, we are not dealing with complex sentences yet, on the contrary, we are analyzing rather simple structures of

<div align="center">Subject   Verb   Direct Object</div>

which means really that we are occupying ourselves with problems of determiners, inflectional endings, agreement or concord, etc., instead of dealing with word order problems. We are making some steps in semantic analysis too.

Nobody shall expect, then, from our so clearly limitated work, any marvellous discovery. We have just proved that our IBM 7090, assisted by an IBM 1401 is able to analyze so simple a phrase as each of the components of the set of experiments you will see.

11

We have written our programs in SNOBOL, a language specially indicated for the management of linguistic structures. One of the many problems of our SNOBOL compiler is its inadequacy to give the time used by the computer in performing its task, but we have calculated that for all the lecture, analysis and listing we show, it must be about three minutes.

Our error message is HA FALLADO, which indicates that something has gone wrong since the beginning. Other error messages are *LA PALABRA " PALABRA " NO ESTA EN NUESTRA TABLA*, which means that we are using a word that does not exist in our lexicon (Table 1), *ESTA FRASE NO FORMA ORACION*, i.e. we

```
TABLA1 (LEXICO)
* * * * * * * * * * * * * *
LAS=DET,FEM,PLU,
EL=DET,MASC,SING,
LA=DET,FEM,SING,
PERRO=SUST,MASC,SING,ANIMADO,
GATO=SUST,MASC,SING,ANIMADO,
PEDRO=SUST,MASC,SING,ANIMADO,INTELIGENTE,
PAN=SUST,MASC,SING,SOLIDO,
LECHE=SUST,FEM,SING,LIQUIDO,
FLORES=SUST,FEM,PLU,SOLIDO,NATURAL,
COME=VT+AUX,SUJANIMADO,COMSOLIDO,
BEBE=VT+AUX,SUJANIMADO,COMLIQUIDO,
RIEGA=VT+AUX,SUJINTELIGENTE,SUJANIMADO,COMNATURAL,
```

have got a set of words without grammatical organization. As grammatical organization we understand our Table 2, " grammar ". *EL*

```
TABLA2 (GRAMATICA)
* * * * * * * * * * * * * * * * *
DET SUST=SN
SUST=SN
VT+AUX=GV
GV SN=SV
SN SP=C
SV=SP
```

SUSTANTIVO NO CONCUERDA CON EL ARTICULO EN EL SUJETO (or *EN EL COMPLEMENTO*) " the substantive does not agree with the article in the subject (or in the object) "; *LA PALABRA " PALABRA " NO ES ARTICULO, SUSTANTIVO NI VERBO*, the word we are dealing with is not in our grammar, which only includes

articles, substantives, and verbs; *NO HAY COHERENCIA ENTRE SUJETO Y VERBO* " there is no semantical agreement between subject and verb ", i.e., the features of the subject are not those demanded by the verb; *LA PALABRA " PALABRA " ESTA MAL COLOCADA*, meaning that we have a word out of place (referring always to our grammar); *EL SUSTANTIVO NO CONCUERDA CON EL ARTICULO* " there is no equivalent among the grammatical features of the article and those of the substantive ", and *NO HAY COHERENCIA ENTRE COMPLEMENTO Y VERBO* " the semantic features of the object are not those demanded by the verb ".

Our success message is *LA FRASE ESTA BIEN CONSTRUIDA* (" the sentence is a grammatical one "); if there has been a semantic disagreement in the analysis, but not a grammatical one we get *A PESAR DE ELLO LA FRASE ESTA BIEN CONSTRUIDA* (in spite of our signalled semantic disagreement the sentence is well built). With this innovation we are trying to research on the domain of apparent incoherences like metaphors. For instance, we establish that the verb HABLAR " to speak " requires the feature $+ HUMAN$ in the subject, so if we get *EL PERRO HABLO* " the dog spoke " our message got will be " there is no semantic agreement between subject and verb ": " dog " is $- HUMAN$; but, in spite of that, we do not stop our analysis and at the end we obtain " in spite of that the sentence is well built ", which assures us that the sentence is grammatical to a lesser degree than another one with total agreement.

Our grammar is like this:

$$S > SN \ SP$$

(we admit a difference between SP and SV, but it is irrelevant at this step of our job, so we make $SP \equiv SV$)

$$SN > (DET) \ SUBST$$
$$SV > GV[\text{verbal group}] \ (SN)$$
$$GV > VT \text{-} AUX$$

The computer makes substitutions beginning at the left side; if there is no agreement it emits the corresponding error message, and if this disagreement is of a grammatical kind it stops.

```
                                          ┌─────────────┐
                                          │    START    │
                                          └──────┬──────┘
                                                 │
                                                 ▼
   ┌─────────┐                            ╱────────────╲
   │  PRINT  │                           ╱    READ      ╲
   ╲─────────╱                          ╱    TABLES      ╲
        │                               ╲────────────────╱
        ▼                                      │
     ╭─────╮                                   ▼
     │STOP │                            ╱────────────╲
     ╰─────╯                           ╱    READ      ╲
                                      ╱    PHRASE      ╲
                                      ╲────────────────╱
                                             │
                                             ▼
                                      ╱────────────╲
                                     ╱    WRITE      ╲
                                    ╱    PHRASE       ╲
                                    ╲─────────────────╱
                                             │
                                             ▼
                              ┌───────────────┐         ╱───────────╲
                              │   TAKE ONE    │────────▶│  ERROR 1   ╲
                              │     WORD      │         ╲────────────╱
                              └───────┬───────┘
                                      │
                                      ▼
                              ┌───────────────┐         ╱───────────╲
                              │     LOOK      │────────▶│  ERROR 8   ╲
                              │   TABLE 1     │         ╲────────────╱
                              └───────┬───────┘
                                      │
                                      ▼
                              ┌───────────────┐
                              │ GRAMMATICAL   │       ╭───╮
                              │  VALUE IN     │──────▶│ A │
                              │ SUBSTITUTION  │       ╰───╯
                              └───────────────┘

                              ┌───────────────┐       ╭───╮
                              │    WRITE      │◀──────│ B │
                              │    PHRASE     │       ╰───╯
                              └───────┬───────┘
                                      │
                                      ▼
                              ┌───────────────┐
                NO            │     LOOK      │
                              │   TABLE 2     │
                              └───────┬───────┘
                                      │
                                      ▼
  YES        ╱─────────╲        ╱──────────────╲
        ╱─────────────╲    YES ╱   IS THERE      ╲   NO
       ╱    PHRASE      ╲◀────╱ ANY SUBSTITUTION  ╲──────▶
       ╲       ≅         ╱◀───╲     RULE ?         ╱
        ╲  SENTENCE     ╱      ╲──────────────────╱
         ╲     ?       ╱
          ╲───────────╱
               ▲
               │
         ╱───────────╲
        ╱   WRITE     ╲
       ╱   PHRASE      ╲
       ╲───────────────╱
```
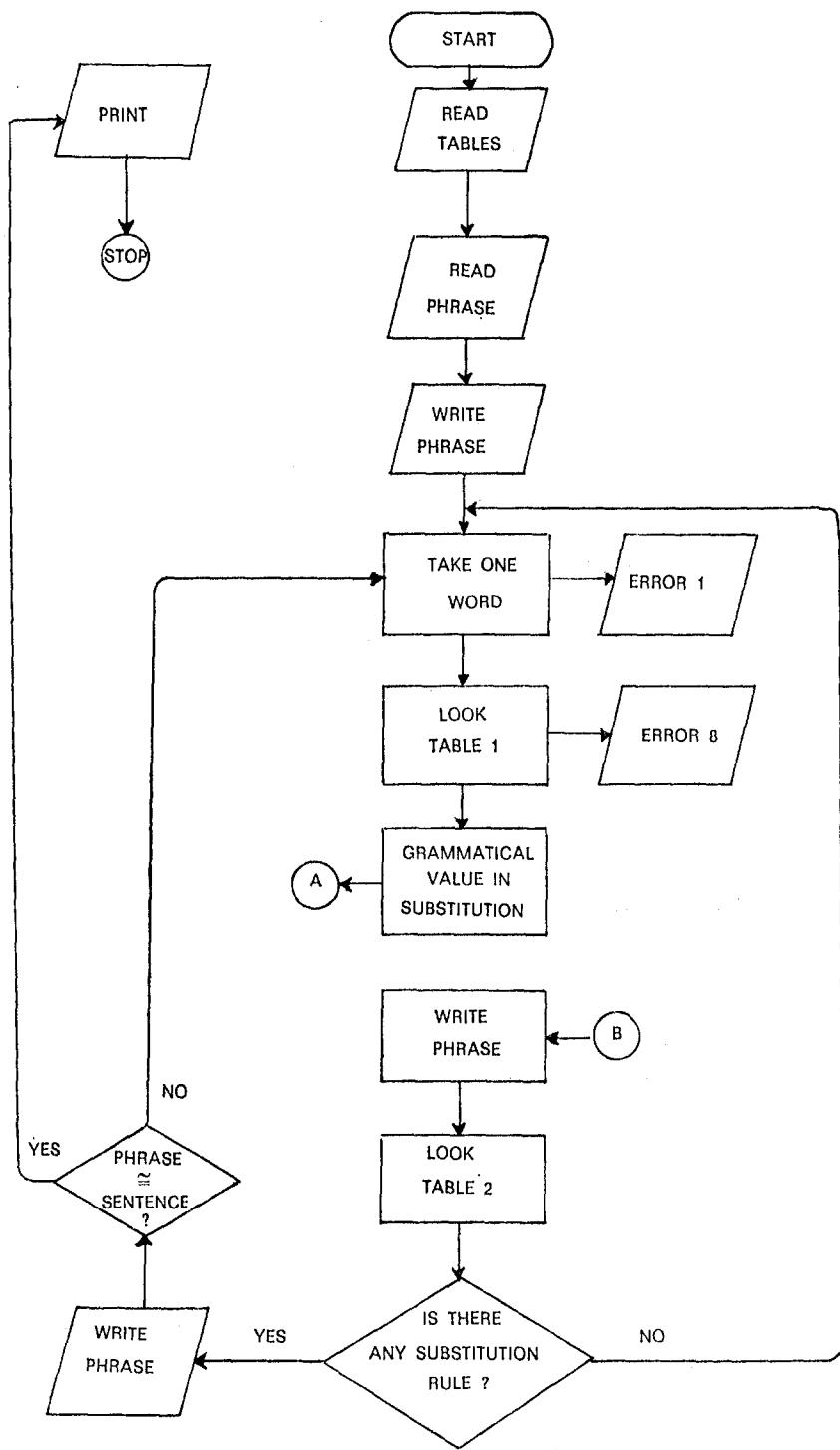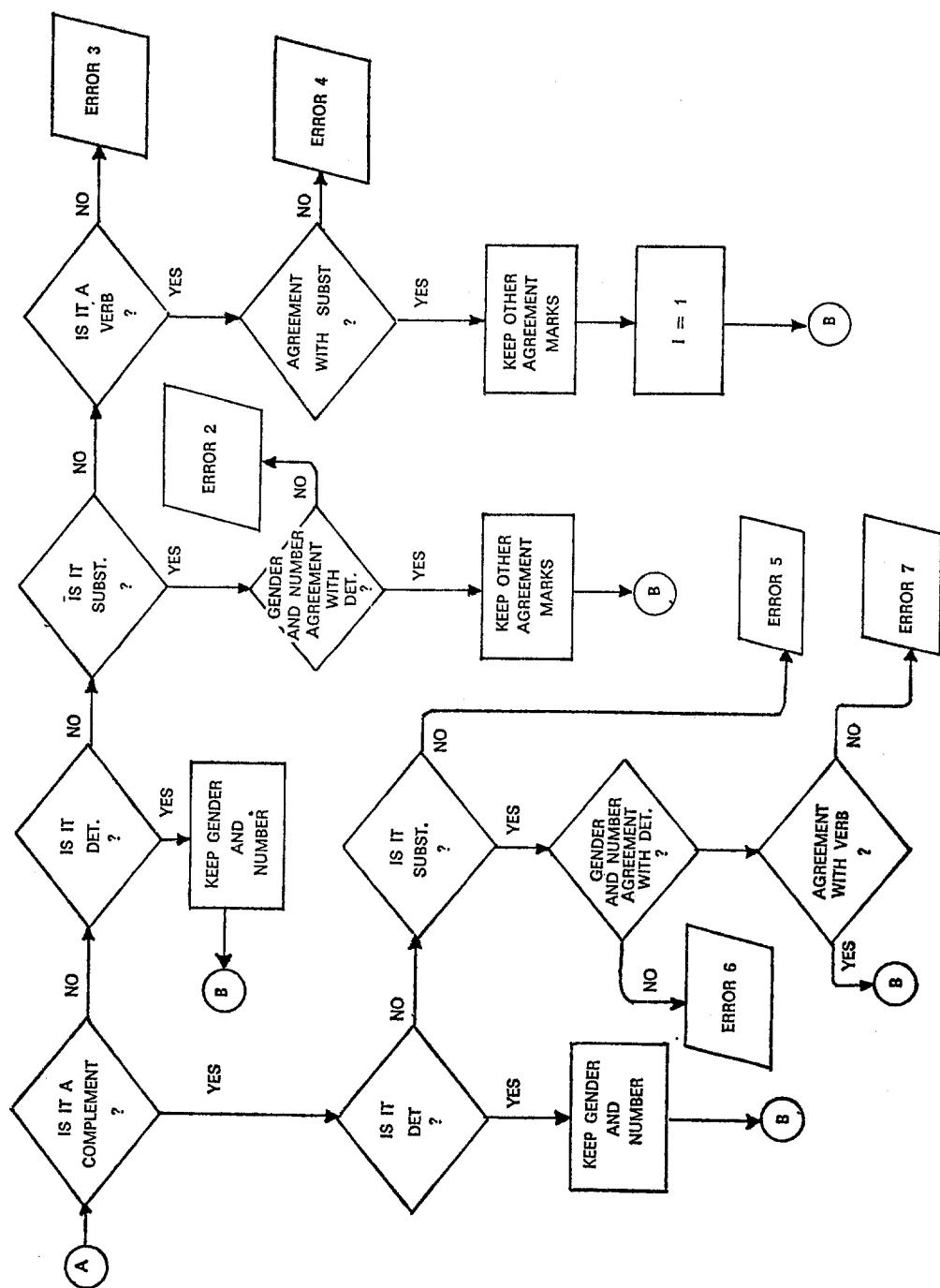
A part of our lexicon is included in the listing, in which it may be seen that the first feature belongs to the grammatical analysis, and the rest to the semantic one. So PEDRO is a substantive, masculine, singular, on the grammatical side, and animated, intelligent, on the semantic side. The verb COME "eats" is a transitive verb, which needs an animated subject and a solid object. It seems fair to declare that till now we were much more concerned about syntactic problems than about those of morphology. So, we operated with verbs in the third person singular. Now we are trying to build a morphology which will permit us to apply our analysis to a broader field.

We reproduce here a listing of one of our experiments.

```
JOB    17    CCUMOO MTM B003130 GRAMATICA        SNOBOL    5  1
TIME OF DAY      00 HR 00 MNS  SYS CUMUTV 00001 JOB  00/00/00
          $EXECUTE        SNOBOL


Z  DEFINE('BLANCOS(K)','Z1') /(T)                                    1
Z1 BLANCOS =                                                         2
F1 JJ = 'O'                                                          3
F9 BLANCOS = BLANCOS ' '                                             4
   JJ = JJ + '1'                                                     5
   .EQ(JJ,'80') /F(U)                                                6
   SYSPCT = ' HA FALLADO' /(RETURN)                                  7
U EQUALS(JJ,K) /F(F9)S(RETURN)                                       8
T DEFINE('ASTE(L)','V1') /(R)                                        9
V1 ASTE =                                                            10
FF1 K   ='0'                                                         11
FF9 ASTE = ASTE '*'                                                  12
   K = K + '1'                                                       13
   .EQ(K,'60') /F(UU)                                                14
   SYSPOT = 'HA FALLADO' /(RETURN)                                   15
UU EQUALS(K,L) /(FF9)S(RETURN)                                       16
R DEFINE('LINEAS(VA)','R1') /(W)                                     17
R1 II = '1'                                                          18
R2 SYSPOT =                                                          19
   EQUALS(II,VA) /S(RETURN)                                          20
   II = II + '1' /(R2)                                               21
W TABLA =                                                            22
   CUENTA = '0'                                                      23
Q SYSPIT *TEXTO/'72'*                                                24
   TEXTO = TRIM(TEXTO)                                               25
   TEXTO 'ENDTABLA1' /S(AE)                                          26
   TABLA1 = TABLA1 TEXTO /(Q)                                        27
AE SYSPOT = 'TABLA1 (LEXICO)'                                        28
   CUENTA = CUENTA + '1'                                             29
   SYSPOT = ASTE('15')                                               30
   CUENTA = CUENTA + '1'                                             31
```

```
AF TABLA1 *ENTE* '.' = /F(AG)                                          32
   SYSPOT = ENTE                                                       33
   CUENTA = CUENTA + '1'                                               34
   TA = TA ENTE '.' /(AF)                                              35
AG TABLA1 = TA                                                         36
   SYSPOT =                                                            37
   CUENTA = CUENTA + '1'                                               38
   SYSPOT =                                                            39
   CUENTA = CUENTA + '1'                                               40
     SYSPOT =                                                          41
   CUENTA = CUENTA + '1'                                               42
   TABLA2 = 'DET SUST=SN.SUST=SN.VT+AUX=GV.GV SN=SV.SN SP=O.
     SV=SP.                                                            43
   SYSPOT = 'TABLA2 (GRAMATICA)'                                       44
   CUENTA = CUENTA + '1'                                               45
   SYSPOT = ASTE('18')                                                 46
   CUENTA = CUENTA + '1'                                               47
AI TABLA2 *ANTE* '.' = /F(AA)                                          48
   CUENTA = CUENTA + '1'                                               49
   SYSPOT = ANTE /(AI)                                                 50
AA SYSPIT *FRASE/'72'*                                                 51
   NN = '65' - CUENTA                                                  52
   SYSPOT = LINEAS(NN)                                                 53
   CUENTA = '0'                                                        54
   FRASE 'FINDETARJETAS' /S(END)                                       55
   FRASE = TRIM(FRASE)                                                 56
   FRASE = FRASE ' '                                                   57
   SYSPOT = BLANCOS('48') 'ANALISIS GRAMATICAL DE LA FRASE'            58
   CUENTA = CUENTA + '1'                                               59
   SYSPOT = BLANCOS('48') ASTE('31')                                   60
   CUENTA = CUENTA + '1'                                               61
   SYSPOT =                                                            62
   CUENTA = CUENTA + '1'                                               63
   I = '0'                                                             64
   J = '0'                                                             65
   K = '0'                                                             66
   N = SIZE(FRASE) / '2'                                               67
   M = '63' - N                                                        68
   FRASE1 = FRASE                                                      69
   SYSPOT = BLANCO(M) FRASE                                            70
   CUENTA = CUENTA + '1'                                               71
   SYSPOT =                                                            72
   CUENTA = CUENTA + '1'                                               73
   SYSPOT =                                                            74
   CUENTA = CUENTA + '1'                                               75
   SYSPOT =                                                            76
   CUENTA = CUENTA + '1'                                               77
J TABLA2 = 'DET SUST=SN.SUST=SN.VT+AUX=GV.GV SN=SV.SN SP=O.
     SV=SP.                                                            78
   FRASE1 *PALABRA* ' ' = /F(E1)                                       79
   TABLA1 PALABRA '=' *RESTO* '.' /F(E8)                               80
   RESTO *CLASE* ',' *DEMAS*                                           81
   FRASE PALABRA = CLASE                                               82
```

```
  .EQ(I,'1')  /S(A)                                                      83
  CLASE 'DET'  /F(B)                                                     84
  J = '1'                                                                85
  DEMAS *GENERO* ',' *NUMERO* ',' = /(C)                                 86
B CLASE 'SUST' /F(D)                                                     87
  DEMAS *GEN1* ',' *NUM1* ',' *GESTO*                                    88
  .EQ(J,'1') /F(C)                                                       89
  J = '0'                                                                90
  EQUALS(GENERO,GEN1) /F(E2)                                             91
  EQUAL(NUMERO,NUM1) /F(E2)S(C)                                          92
D CLASE 'VT+AUX' /F(E3)                                                  93
  SIG = DEMAS                                                            94
E SIG 'SUJ' *LT* ',' =  V  /F(F)                                         95
  GESTO LT /S(E)F(E4)                                                    96
F I = '1' /(C)                                                           97
A CLASE 'DET' /F(G)                                                      98
  J = '1'                                                                99
  DEMAS *GENERO* ',' *NUMERO* ',' = /(C)                                100
G CLASE 'SUST' /F(E5)                                                   101
  DEMAS *GEN1* ',' *NUM1* ',' *RESTO*                                   102
  .EQ(J,'1') /F(H)                                                      103
  J = '0'                                                               104
  EQUALS(GENERO,GEN1) /F(E6)                                            105
  EQUALS(NUMERO,NUM1) /F(E6)                                            106
H SIG 'COM' *LT* ',' = /F(C)                                            107
  RESTO LT /S(H)F(E7)                                                   108
C N = SIZE(FRASE) / '2'                                                 109
  M = '63' - N                                                          110
  SYSPOT = BLANCOS(M) FRASE                                             111
  CUENTA = CUENTA + '1'                                                 112
  SYSPOT =                                                              113
  CUENTA = CUENTA + '1'                                                 114
I1 TABLA2 = 'DET SUST=SN.SUST=SN.VT+AUX=GV.GV SN=SV.SN SP=O.
   SV=SP '                                                              115
I TABLA2 *STR* '=' *CADENA* '.' = /F(J)                                 116
  FRASE STR = CADENA /F(I)                                              117
  N = SIZE(FRASE) / '2'                                                 118
  M = '63' - N                                                          119
  SYSPOT = BLANCOS(M) FRASE                                             120
  CUENTA = CUENTA + '1'                                                 121
  SYSPOT =                                                              122
  CUENTA = CUENTA + '1'                                                 123
  EQUALS(FRASE,'O ') /S(EXITO)F(I1)                                     124
EXITO SYSPOT =                                                          125
  CUENTA = CUENTA + '1'                                                 126
  .EQ(K,'1') /S(AC)                                                     127
  SYSPOT = '***LA FRASE ESTA BIEN CONSTRUIDA***'                        128
  CUENTA = CUENTA + '1'                                                 129
  CUENTA = CUENTA + '1'                                                 130
  SYSPOT = /(AA)                                                        131
E1 SYSPOT =                                                             132
  CUENTA = CUENTA + '1'                                                 133
  SYSPOT = '*ESTA FRASE NO FORMA ORACION*'                              134
```

AUTOMATIC SYNTACTIC ANALYSIS IN DICTIONARY MAKING   169

```
    CUENTA = CUENTA + '1'                                      135
    CUENTA = CUENTA + '1'                                      136
    SYSPOT = /(AA)                                             137
E2 SYSPOT =                                                    138
    CUENTA = CUENTA + '1'                                      139
    SYSPOT= '*EL SUSTANTIVO NO CONCUERDA CON EL ARTICULO EN EL
    SUJETO*                                                    140
    CUENTA = CUENTA + '1'                                      141
    CUENTA = CUENTA + '1'                                      142
    SYSPOT = /(AA)                                             143
    CUENTA = CUENTA + '1'                                      144
E3 SYSPOT = '*LA PALABRA' ' ' PALABRA ' ' 'NO ES ART,SUST, NI VER *'  145
. /(AA)                                                        145
E4 SYSPOT =                                                    146
    CUENTA = CUENTA + '1'                                      147
    SYSPOT = '*NO HAY COHERENCIA ENTRE SUJETO Y VERBO*'        148
    CUENTA = CUENTA + '1'                                      149
    CUENTA = CUENTA + '1'                                      150
    SYSPOT = /(AB)                                             151
E5 SYSPOT =                                                    152
    CUENTA = CUENTA + '1'                                      153
    SYSPOT = '*LA PALABRA ' PALABRA ' ESTA MAL COLOCADA*'      154
    CUENTA = CUENTA + '1'                                      155
    CUENTA = CUENTA + '1'                                      156
    SYSPOT = /(AA)                                             157
E6 SYSPOT =                                                    158
    CUENTA = CUENTA + '1'                                      159
    A = ' EN EL COMPLEMENTO*'                                  160
    SYSPOT = '*EL SUSTANTIVO NO CONCUERDA CON EL ARTCULO ' A   161
    CUENTA = CUENTA + '1'                                      162
    CUENTA = CUENTA + '1'                                      163
    SYSPOT = /(AA)                                             164
E7 SYSPOT =                                                    165
    CUENTA = CUENTA + '1'                                      166
    SYSPOT = '*NO HAY COHERENCIA ENTRE COMPLEMENTO Y VERBO*'   167
    CUENTA = CUENTA + '1'                                      168
    CUENTA = CUENTA + '1'                                      169
    SYSPOT =/(AD)                                         .    170
E8 SYSPOT =                                                    171
    CUENTA = CUENTA + '1'                                      172
    SYSPOT = '*LA PALABRA ' PALABRA ' NO ESTA EN NUESTRA TABLA*'  173
    CUENTA = CUENTA + '1'                                      174
    CUENTA = CUENTA + '1'                                      175
    SYSPOT = /(AA)                                             176
AB K = '1' /(F)                                                177
AD K = '1' /(C)                                                178
AC SYSPOT = '*A PESAR DE ELLO LA FRASE ESTA BIEN CONSTRUIDA*'  179
    CUENTA = CUENTA + '1'                                      180
    CUENTA = CUENTA + '1'                                      181
    SYSPOT = /(AA)                                             182
END Z                                                          183
```

SUCCESSFUL COMPILATION

ANALISIS GRAMATICAL DE LA FRASE
****

PEDRO RIEGA LAS FLORES

SUST RIEGA LAS FLORES

SN RIEGA LAS FLORES

SN VT+AUX LAS FLORES

SN GV LAS FLORES

SN GV DET FLORES

SN GV DET SUST

SN GV SN

SN SV

SN SP

O

LA FRASE ESTA BIEN CONSTRUIDA

ANALISIS GRAMATICAL DE LA FRASE
****
EL PERRO COME LA LECHE

DET PERRO COME LA LECHE
DET SUST COME LA LECHE
SN COME LA LECHE
SN VT+AUX LA LECHE
SN GV LA LECHE
SN GV DET LECHE

SN GV DET SUST
SN GV SN
SN SV
SN SP
O

NO HAY COHERENCIA ENTRE COMPLEMENTO Y VERBO

A PESAR DE ELLO LA FRASE ESTA BIEN CONSTRUIDA

ANALISIS GRAMATICAL DE LA FRASE
****
PEDRO RIEGA FLORES

SUST RIEGA FLORES

SN RIEGA FLORES

SN VT+AUX FLORES

SN GV FLORES

SN GV SUST

SN GV SN

SN SV

SN SP

O

LA FRASE ESTA BIEN CONSTRUIDA

ANALISIS GRAMATICAL DE LA FRASE
****
PEDRO BEBE VINO

SUST BEBE VINO

SN BEBE VINO

SN VT+AUX VINO

SN GV VINO

LA PALABRA VINO NO ESTA EN NUESTRA TABLA

ANALISIS GRAMATICAL DE LA FRASE
****
EL GATO BEBE LECHE

DET GATO BEBE LECHE
DET SUST BEBE LECHE
SN BEBE LECHE
SN VT+AUX LECHE
SN GV LECHE
SN GV SUST
SN GV SN
SN SV
SN SP
O

•••LA FRASE ESTA BIEN CONSTRUIDA•••

ANALISIS GRAMATICAL DE LA FRASE
****
PEDRO COME EL PAN

SUST COME EL PAN
SN COME EL PAN
SN VT+AUX EL PAN
SN GV EL PAN
SN GV DET PAN
SN GV DET SUST
SN GV SN
SN SV
SN SP
O

•••LA FRASE ESTA BIEN CONSTRUIDA•••

ANALISIS GRAMATICAL DE LA FRASE
****
EL PERRO COME PAN

DET PERRO COME PAN
DET SUST COME PAN
SN COME PAN
SN VT+AUX PAN
SN GV PAN
SN GV SUST
SN GV SN
SN SV
SN SP
O

LA FRASE ESTA BIEN CONSTRUIDA

ANALISIS GRAMATICAL DE LA FRASE
****
EL GATO COME EL PAN

DET GATO COME EL PAN
DET SUST COME EL PAN
SN COME EL PAN
SN VT+AUX EL PAN
SN GV EL PAN
SN GV DET PAN
SN GV DET SUST
SN GV SN
SN SV
SN SP
O

LA FRASE ESTA BIEN CONSTRUIDA

ANALISIS GRAMATICAL DE LA FRASE
****

EL PERRO RIEGA LAS FLORES

DET PERRO RIEGA LAS FLORES
DET SUST RIEGA LAS FLORES
SN RIEGA LAS FLORES

SN VT+AUX LAS FLORES
SN GV LAS FLORES
SN GV DET FLORES
SN GV DET SUST
SN GV SN
SN SV
SN SP
O

NO HAY COHERENCIA ENTRE SUJETO Y VERBO

A PESAR DE ELLO LA FRASE ESTA BIEN CONSTRUIDA

ANALISIS GRAMATICAL DE LA FRASE
****

EL GATO BEBE LAS LECHE

DET GATO BEBE LAS LECHE
DET SUST BEBE LAS LECHE
SN BEBE LAS LECHE
SN VT+AUX LAS LECHE
SN GV LAS LECHE
SN GV DET LECHE

EL SUSTANTIVO NO CONCUERDA CON EL ARTICULO EN EL COMPLEMENTO

ANALISIS GRAMATICAL DE LA FRASE
****

LA GATO BEBE LA LECHE

DET GATO BEBE LA LECHE

EL SUSTANTIVO NO CONCUERDA CON EL ARTICULO EN EL SUJETO

Even if we are at this moment at the preliminary steps, we hope that in the near future we shall be able to analyze more complicated sentences, helping this way to lemmatization by distinguishing, for instance, *CANTO* substantive, " song ", from *CANTO*, verb, " I sing ", or the two possibilities of *ESPERABA*, " I hoped ", " he hoped ".