AVALON, UNA GRAMÁTICA FORMAL BASADA EN CORPUS

AUTORES: Conchita Álvarez Lebredo, Pilar Alvariño Alvariño, Adelaida Gil Martínez, Teresa Romero Quintáns, Mª Paula Santalla del Río, Susana Sotelo Docío.

DIRECCIÓN POSTAL:
Facultade de Filoloxía,
Avda. Burgo das Nacións, s/n
Santiago de Compostela, 15771
Tfno: (981) 563100/575340, ext. 11908
Fax: (981) 574646
e-mail: fesdocio@usc.es

Resumen

En esta exposición se presenta el trabajo que se está llevando a cabo en la Universidad de Santiago de Compostela para desarrollar una gramática formal basada en corpus. El que presentamos aquí es un trabajo asociado al proyecto DoRo, aunque supera ampliamente sus requerimientos y constituye un objetivo independiente en sí mismo. Pretendemos describir el corpus de datos, la concepción general de la gramática en sí misma y, sobre todo, cómo ésta se sirve de aquéllos, con qué fines y, en la medida de lo posible en el estado actual del trabajo, con qué ventajas.

1. LA FUENTE DE DATOS[1]: LA BASE DE DATOS SINTÁCTICOS.

La fuente de datos a la que la gramática formal que presentamos debe su conformación es la *Base de Datos Sintácticos del español actual* (BDS en adelante)[2] elaborada en la Universidad de Santiago de Compostela. La BDS es el resultado de analizar manualmente —con principios de análisis constitutivos y funcionales— el contexto sintáctico de aproximadamente 160.000 formas verbales contenidas en la parte contemporánea del *Archivo de textos hispánicos de la Universidad de Santiago* (ARTHUS). Cada aparición de un verbo en el corpus es recogida en un registro distinto con información detallada acerca de la configuración sintáctica de la secuencia lingüística en la que tal verbo ejerce su influencia directa. En primer lugar, información general acerca del tipo de cláusula del que se trata —independiente, coordinada, completiva con *que*...— función de la cláusula, diátesis —activa, media, pasiva—, modalidad —declarativa, imperativa, interrogativa (total o parcial), exclamativa—, forma flexiva del verbo, número de argumentos y orden en el que se disponen. Seguidamente, información pormenorizada acerca de cada uno de los argumentos—sujeto, complemento directo, ... — identificados para cada forma verbal analizada. Además, junto a rasgos sintácticos propios de cada uno de los argumentos considerados, se incluyen, primero y con un alto grado de especialización el tipo de unidad estructural en que se materializan y, a continuación, ciertos rasgos semánticos de reconocida relevancia sintáctica —determinado/indeterminado, animado/inanimado, contable/incontable —.

2. *EL FORMALISMO AGFL[3], AFFIX GRAMMARS OVER FINITE LATTICES*

AGFL es un formalismo de dos niveles para la descripción de gramáticas libres de contexto. Un primer nivel libre de contexto es extendido con atributos que sirven para ex-

[1] Puede encontrarse más información en http://www.usc.es/~sintx.

[2] La construcción de la BDS ha sido posible gracias a la financiación concedida *por Dirección Xeral de Educación e Ordenación Universitaria* de la *Consellería de Educación* de la *Xunta de Galicia* entre 1989 y 1991 (ref. XUGA 82710088) y la Dirección General de Investigación Científica y Técnica del MEC entre septiembre de 1991 y septiembre de 1994 (ref. PB90-0376).

[3] Asociado a un generador de analizadores (*GEN parser generator*), ha sido desarrollado, y continúa siéndolo, en la Universidad Católica de Nimega por C.H.A. Koster (http://www.cs.kun.nl/agfl).

presar relaciones de concordancia entre constituyentes. Los posibles valores de estos atributos se especifican en el segundo nivel.

En el primer nivel, una regla tiene la forma siguiente:

```
izquierda de la regla :
   derecha de la regla.
```

En la izquierda de la regla sólo puede aparecer un elemento. En la derecha de la regla pueden aparecer dos operadores distintos: ",", que indica concatenación de elementos (*miembros*), y ";", que indica disyunción de elementos (*alternativas*). Los elementos que pueden aparecer a la derecha y a la izquierda de la regla se llaman no terminales. Un no terminal es un nombre y una serie facultativa de atributos, expresados de la forma siguiente:

```
no terminal A ( atributo A, atributo B …)
   Clause( clause_type, mood )
```

Los valores de los atributos (no más de treinta y dos) pueden anidarse y se hacen explícitos en el segundo nivel, mediante reglas que toman la forma siguiente:

```
atributo A :: valor a; valor b … atributo B.
atributo B :: valor c; valor d … valor t.
clause_type :: MAIN; SUBORDINATED; EL_SUBORDINATED; QUE; EL_QUE; TOTAL; PARTIAL.
mood :: non_imperative; IMPERATIVE; non_personal_mood.
non_imperative :: INDICATIVE; SUBJUNCTIVE.
non_personal_mood :: PARTICIPLE; GERUND; INFINITIVE
```

En el interior de una regla todos los atributos representados del mismo modo deben obtener el mismo valor. Para expresar el hecho de que el mismo atributo presente en distintos no terminales en la misma regla puede obtener valores distintos para cada uno de los no terminales en cuestión, se añade un número, un índice, distinto a cada uno de los atributos. Los atributos pueden ser resueltos ya en el interior de las reglas en uno o más valores, en este último caso separados por la barra vertical "|". Se pueden establecer prioridades entre las alternativas añadiéndoles un miembro *penalty* a las que se quiera relegar. Un *penalty* tiene la forma:

```
$PENALTY ( n )
```

donde n es cualquier número: cuanto más alto, menos posibilidades tendrá la alternativa penalizada de ser seleccionada frente a sus competidoras. Todo lo que en una línea esté precedido de un sostenido "#" será ignorado por el generador de analizadores. Un ejemplo simple de todo lo anterior lo constituye la regla siguiente:

```
(1)   Clause ( MAIN|SUBORDINATED, non_imperative ):
#***************con dos argumentos expresos:
(1a)  # Orden: sujeto + predicado + predicativo del cd
# Ejs.: "yo la considero insoportable"
HIDDEN SUBJECT ( +FIRST,  const_category1, person, gender, number ),
HIDDEN PREDICATE ( -FIRST,  SDOPR, ACT|MIDD, person,
number, non_imperative, ACC, person_one, gender_one,
number_one ),
HIDDEN PREDICATIVE COMPLEMENT ( -FIRST,  const_category2,
gender_one, number_one );
```

La alternativa (1a) de la regla (1) da cuenta de la constitución de una cláusula a partir de sujeto, predicado y predicativo del complemento directo en ese orden. Los atributos *person, gender* y *number* del predicado deben coincidir con los del sujeto; los atributos *gender_one* y *number_one* del predicado, que contienen información acerca del clítico que reproduce al complemento directo, deben coincidir con los del predicativo. Los atributos *const_category* del sujeto y el predicativo, que contienen información acerca de la categoría en que se materializan tales argumentos, toman valores (distintos o iguales) de modo independiente gracias a los índices 1 y 2 que los separan.

3. *UN LEXICÓN[4] DE FORMAS FLEXIVAS PARA AVALON*

3.1 LEXICÓN DE SUSTANTIVOS Y ADJETIVOS

El lexicón de sustantivos y adjetivos se distribuye en 54 ficheros con un total de 80116 lemas, agrupados según sus características morfológicas; cada uno de ellos contiene una cabecera con la información morfológica asociada a todas las entradas incluidas, y que comprende categoría, género, número, modelo de flexión, etc. Los datos van organizados en campos divididos por una barra inclinada (\) y, lógicamente, varían dependiendo de la categoría (sustantivo o adjetivo, en este caso).

```
\N\COMMON\FEM\_\= [es] \\\_\
abdicación
aberración
   ...
```

```
\ADJ\QUAL\_\_\o/a [s] \POS\_\
 desamparado
   ...
\ADJ\QUAL\MASC|FEM\_\= [s] \POS\?\_\
*breve
   ...
```

Figs. 1-2: ficheros originales a partir de los que genera el lexicón de formas flexionadas)

La herramienta *LexTool*[5] se encarga de convertir estos ficheros al formato de *mmorph*[6], así como de transformar la salida con las formas flexionadas en el archivo final (fig. 4). En una primera fase, cada lema es asociado con la información presente en la cabecera; esta información es utilizada para producir la raíz correspondiente a dicho lema e insertar ambos en el fichero fuente que *mmorph* utiliza para generar las formas flexionadas de acuerdo con las especificaciones que prevé la gramática en el formalismo *mmorph* para cada grupo.

```
;; \ADJ\QUAL\_\_\o/a [s] \POS\_\
   Adj ( gdtype=o|a adjtype=QUAL
deg=POS mente=N ) "desamparad" = "de-
samparado"
```

(fig 3: muestra del fichero fuente de mmorph con el léxico)

```
NounSt ( COMMON, FEM, SING ): "aberra-
ción". #"aberración"
NounSt ( COMMON, FEM, PLU ): "aberra-
ciones". #"aberración"
AdjSt ( QUAL, POS, MASC|FEM|NEUT, SING
): "breve". #"breve"
AdjSt ( QUAL, POS, MASC|FEM, PLU ):
"breves". #"breve"
AdvSt ( POS ): "brevemente". #"breve-
mente"
```

(fig. 4: salida final producida por LexTool en formato AGFL)

Además de la generación de sustantivos y adjetivos, aquellas entradas marcadas con un asterisco (fig. 2) dan lugar también al correspondiente adverbio en *-mente* formado a partir del femenino singular. El lexicón de sustantivos y adjetivos que utilizamos contiene información extraída en buena medida de diccionarios, que, por sus características específicas (exhaustividad, información altamente estructurada), se han convertido en los últimos años en una fuente importante para la elaboración de lexicones computacionales

El objetivo primordial de los diccionarios es dar cuenta del léxico de una lengua; la exhaustividad con que esto se lleva a cabo hace que se incluyan a un mismo nivel formas de uso común con otras claramente desusadas o limitadas a un espacio geográfico o ámbito concreto, lo que puede dar lugar a ambigüedades no siempre fáciles de resolver. Pero por otra parte, y pese a la exhaustividad que los caracteriza, los diccionarios por lo general no siempre ofrecen una información lo suficientemente actualizada o, al menos, no en la misma medida que los *corpus*; esto, junto a la sobrecarga de información de que hablábamos antes, puede producir problemas que repercuten directamente en la eficiencia de la aplicación a la que va destinado el lexicón.

La estrategia encaminada a solventar estas cuestiones y a adaptar los datos con los que ya contamos se asienta básicamente

---

[4] Hemos de expresar aquí nuestro agradecimiento por la valiosísima contribución prestada en este punto por el grupo de investigación de la Universitat de Barcelona dirigido por M.ª Antònia Martí, que nos ha proporcionado el amplio conjunto de lemas que constituyó el núcleo inicial a partir del cual hemos podido desarrollar el trabajo que describimos en este apartado.

[5] Desarrollada en el seno del grupo utilizando el lenguaje Tcl bajo SunOs. En un futuro está prevista la construcción de una interfaz gráfica en Tcl/Tk que reúna en un sistema de menús las herramientas de generación que ahora se agrupan bajo la denominación genérica de LexTool.

[6] Herramienta de análisis morfológico desarrollada por ISSCO dentro del marco del proyecto Multext,http://www.lpl.univ-aix.fr/projects/multext)

sobre dos pilares: la incorporación de información procedente de *corpus* y el diseño de un sistema de delimitación de grandes grupos entre las entradas a través de marcas. Así, por un lado cubrimos posibles omisiones mediante cruces con un corpus de formas asociadas a su frecuencia de aparición, incorporando al lexicón aquellas palabras ausentes con una frecuencia superior a 10. Por otra parte, establecimos un sistema de marcas que nos permiten en todo momento la generación de varios tipos de lexicones dependiendo de las necesidades de la gramática. Este sistema proporciona una gran flexibilidad, ya que permite disponer de un lexicón orientado específicamente hacia la gramática formal al mismo tiempo que conservamos toda la información originaria intacta en otro lexicón, cuya organización dependerá de criterios lexicográficos, y no de la aplicación a la que está destinado.

3.2 LEXICÓN DE VERBOS

El programa encargado de su generación[7], *mkagfl*, toma como punto de partida un fichero con 1286 lemas verbales acompañados de información sobre subcategorización sintáctica, aunque esta última no será procesada hasta el final (vid. 4). El procedimiento es en parte similar al llevado a cabo con el lexicón de sustantivos y adjetivos. En una primera fase de la generación, el módulo *mklista* crea una lista de lemas asociados a su correspondiente modelo morfológico, que extrae de una base de datos con 12000 verbos. Esta información es utilizada para producir las raíces e insertarlas –junto con su lema– en el fichero de entrada de léxico[8] que utiliza *mmorph*.

```
; 001-01 "am" = "amar" Verb[ infl=stem pres_s1=tl
pres_s23p3=tl pres_p1=tl pres_p2_inf=tl pres_p12_inf=nil
impf_type=tl past_s1=tl past_s82p2=tl past_s2p12=nil
past_s3=tl past_p3=tl futcond_type=tl pressubj_s123p3=tl
pressubj_p12=tl impfsubj_type=tl ger_type=tl
part_type=tl imper_s2=tl imper_p2=tl impcor_sing=tl
impcor_plu=tl ]"dej" = "dejar"
```

(fig 5: muestra del fichero fuente de *mmorph*)

[7] Herramienta que controla varios módulos distintos; cada uno de ellos tiene asignada una tarea, y puede ser ejecutado independientemente del resto.

[8] La formalización de la morfología verbal parte de la desarrollada en los recursos lingüísticos liberados para el español por el proyecto Multext (vid. nota 6).

A partir de estas raíces, y de acuerdo con las reglas morfológicas descritas en su gramática, *mmorph* genera, para cada uno de los lemas presentes en el fichero original, el conjunto de sus formas flexivas; la salida es transformada en cláusulas AGFL con información sobre número, persona, tiempo y modo. Los atributos que contienen esquemas sintácticos (*mvtype*) y voz (*voice*) permanecen sin recibir valor alguno hasta la fase final, tal como describimos en el apartado 4.

```
VerbSt ( mvtype, voice, FIRST, SING, PRESENT,
INDICATIVE ):
   "dejo". #"dejar"
VerbSt ( mvtype, voice, SECOND, SING, PAST,
INDICATIVE ):
   "dejaste". #"dejar"
VerbStPART ( mvtype, voice, PARTICIPLE, MASC, PLU
): "dejados". #"dejar"
```

(fig. 6: ejemplos de cláusulas AGFL para formas verbales)

En este proceso se generan tan solo las formas simples de los verbos de significado pleno y auxiliares (*haber*, *ser* y perifrásticos). Los tiempos compuestos y las construcciones perifrásticas son tratados en el interior de AVALON como parte de lo que se considera la frase verbal.

Por otro lado, a la hora de generar formas verbales flexionadas es necesario tener en cuenta el fenómeno de la enclisis pronominal, que puede provocar cambios ortográficos manifestados a través de la acentuación gráfica, de modificaciones en la última desinencia o ambos a un tiempo. Para ello, además de los no terminales *VerbSt* y *VerbStPART* vistos anteriormente, existe un tercer tipo exclusivo de aquellas formas verbales que reciben pronombres enclíticos. A diferencia de las anteriores, no presentan como elementos terminales del lexicón formas verbales completas, sino la variante ortográfica sin clíticos; el número de pronombres es controlado por un nuevo atributo, cuyos valores posibles pueden ser JUST_ONE, ONE o TWO según reciba un único pronombre o uno o más.

4. *DE LA BDS AL LEXICÓN VERBAL PARA AVALON*

De BDS, al procesar todos los datos que hay para cada verbo en concreto, obtenemos, por ejemplo, el tipo de información sumaria siguiente:

DEJAR[1584 casos; 37 esquemas; 184 subesquemas]

Act.	SDPD	625	39.45 %
Act.	SD	375	23.67 %
Act.	SDAD	162	10.23 %
CPrn.	SPS	154	9.72 %
Act.	SDI	119	7.51 %
CPrn.	SSP	17	1.07 %
Act.	SDADPD	14	0.88 %
CPrn.	SD	14	0.88 %
Act.	S	12	0.76 %
Act.	SDIPD	12	0.76 %
CPrn.	SDPD	12	0.76 %
Act.	SDPS	11	0.69 %
CPrn.	S	10	0.63 %
Act.	SDIAD	7	0.44 %
CPrn.	SDAD	7	0.44 %

Tabla 1. Esquemas[9] de DEJAR en BDS.

No toda la riqueza de esquemas (combinaciones de argumentos y voz en las que entra un verbo) ofrecida por la BDS es aprovechada en el lexicón para AVALON. Hemos diseñado programas adicionales que operan reducciones sobre los datos argumentales más ricos de la BDS. Por ejemplo, los datos relacionados con la voz pasiva y las cláusulas impersonales con *se* no son tenidos en cuenta (ambas construcciones son tratadas en la propia gramática a partir de los esquemas activos); todos los complementos preposicionales son tratados del mismo modo, sin diferenciar tipos entre ellos, y sólo es posible identificar uno. Aun considerando estas reducciones, la variedad de esquemas que ofrecía la BDS era demasiado amplia para ser abarcada en la gramática, con lo cual impusimos un número finito de esquemas, lo dio como resultado un conjunto de valores para el atributo *mvtype* del verbo y la función sintáctica que éste constituye, el PREDICATE. Todos los esquemas que surgen de la BDS distintos de estos son, por ahora, marginados del lexicón verbal para AVALON.

```
mvtype[10] :: P; S; DO; SDO; SPC; SIO; SPR;
SDOIO; SDOPC; SIOPC; SDOPR; SDOIOPC; PC.
```

[9] La combinación de argumentos en los esquemas es representada por las letras S (sujeto), D (complemento directo), I (complemento indirecto), SP (suplemento), AD (adverbial), MD (modal), PR (otros preposicionales), estos 4 últimos son tipos de complementos preposicionales, lógicamente excluyentes entre ellos en las dos posiciones consecutivas que les están reservadas, A (agente), PS (predicativo del sujeto), PD (predicativo del complemento directo), PO (predicativo de otros constituyentes), estos tres últimos son tipos de predicativos, también excluyentes entre ellos, en la posición que ocupan.

Tratando de evitar las frecuencias de verbos o esquemas demasiado bajas y los errores que hayan podido introducirse en la BDS debido al análisis manual, sólo son incluidos en el lexicón para AVALON los esquemas con una frecuencia absoluta superior a 15, o los esquemas con una frecuencia percentualmente superior al 10% en verbos con una frecuencia absoluta superior a 10. En total, 1286 verbos distintos en 2350 esquemas verbales. Para cada forma generada con los valores abiertos *mvtype* y *voice*, el módulo *mvtype* de *mkagfl* produce un número de entradas para cada forma verbal, incluyendo el valor indicado para el posible régimen preposicional allí donde sea necesario.

5. *AVALON*

La gramática formal describe dos grandes tipos de unidades: la cláusula y la frase. Por *cláusula* entendemos una categoría gramatical caracterizada por la presencia de un predicado en torno al cual se articulan una serie de constituyentes sintácticos que desempeñan las funciones de sujeto, complemento directo, complemento indirecto, complemento preposicional, agente, predicativo o circunstancial. *Frase* es toda estructura lingüística construida en torno a un nombre, adjetivo, pronombre, verbo o adverbio para desempeñar una de las funciones sintácticas clausales o para modificar a otra frase.

5.1. LA FRASE

Sobre la estructura constitutiva interna de la *frase*, la BDS no nos ofrece datos ya interpretados. No obstante, donde las características del fenómeno concreto lo permitían —se podían hacer sobre él búsquedas razonablemente ponderadas y limitadas—, hemos acudido directamente a ARTHUS (a su parte contemporánea), para estudiar mediante concordancias cómo se comporta el fenómeno en cuestión. La preocupación fundamental es dar cuenta esencialmente de lo que con más frecuencia se da en la len-

[10] P hace referencia al esquema constituido exclusivamente por un *predicate*; en el resto de los esquemas se omite la referencia al *predicate*. En estos últimos, el resto de los caracteres deben interpretarse del modo siguiente: S, *subject*, DO, *direct object*, PC, *prepositional complement*, IO, *indirect object*, PR, *predicative*.

gua, y hacerlo de manera que se puedan añadir fácilmente nuevas especificaciones para el fenómeno tratado en cada caso. Junto a ello, nuestro propósito en AVALON es identificar frases y límites de constituyentes sintácticos clausales. Para ello hemos de abrir el camino a la actuación de la *recursividad* en todos los puntos de la gramática en los que de hecho sea posible y, al mismo tiempo, aprovechar al máximo todas las pistas que la lengua dé acerca de qué ítems pueden sucederse efectivamente en el interior de un único constituyente sintáctico.

5.2. La CLÁUSULA

5.2.1. Tipos de cláusulas. Los datos que ofrece la BDS

En AVALON, las reglas de reescritura de la unidad cláusula constan de una serie de alternativas cada una de las cuales describe un orden diferente de los constituyentes sintácticos entre sí y respecto al PREDICADO. Esquema sintáctico y orden no son equivalentes tal como los concebimos en la gramática. El esquema se refiere al tipo de constituyentes que se puede combinar con un predicado dado, y el orden a la posición relativa de los argumentos plenos, es decir, sin tener en cuenta pronombres clíticos y sujetos implícitos. Un mismo esquema, como por ejemplo SDOPR (sujeto, predicado, complemento directo, predicativo del complemento directo) puede concretarse en órdenes que implican diferente número de argumentos explícitos:

En nuestra gramática nos interesa recoger todas las ordenaciones posibles o, al menos, las más frecuentes, de los elementos integrantes de la cláusula. Pero sólo las que se dan en el discurso real, ya que un exceso de alternativas de reescritura sólo contribuye a complicar la gramática, provocando ambigüedades innecesarias y favoreciendo la aparición de análisis erróneos. El problema es que no existen estudios detallados del orden en español. La intuición del lingüista tampoco basta. Por ello, el corpus es el instrumento ideal para extraer la información de orden. Nos permite manejar datos reales, de la lengua viva. Nos proporciona tanto los datos de las combinaciones efectivas de los constituyentes sintácticos como su frecuencia.

Un análisis de los textos del corpus (a través de su versión interpretada, la BDS) nos ha permitido descubrir diferentes comportamientos en el orden de los constituyentes clausales según el tipo de cláusulas que estemos considerando. Esto nos ha inducido a establecer varios grupos en el interior de los cuales reina una cierta homogeneidad en las posibilidades de ordenación. Son los siguientes:

- Grupo general: agrupa las cláusulas declarativas con verbo en forma personal, tanto las "principales" de la gramática tradicional, como las precedidas de la conjunción *que*; las interrogativas totales directas e indirectas; las claúsulas miembros de oración bipolar (concesivas, condicionales, consecutivas, etc.).
- Exhortativas o imperativas
- Cláusulas no personales (de infinitivo y gerundio) no relativas ni interrogativas
- Interrogativas parciales (directas o indirectas)
- Interrogativas en infinitivo
- Relativas
- Relativas en infinitivo
- Participio

Cada uno de estos grandes grupos se corresponde con una regla de reescritura de la gramática formal. En la parte derecha de cada regla aparecen tantas alternativas como posibilidades de ordenación pero sólo están activas las posibilidades de orden que aparecen ilustradas con ejemplos en el corpus. A ellas añadimos una serie de posibilidades no confirmadas pospuestas al símbolo de comentario "#" referido en el apartado 2, pero que quizá pudieran aparecer en otro conjunto de textos. No podemos olvidar que el corpus que manejamos y que sirve de fuente de información a la BDS es un corpus no demasiado amplio. Aunque nos proporciona datos muy valiosos y nos muestra las tendencias generales del idioma, no podemos asegurar que todo lo que está ausente en ARTHUS es imposible o agramatical.

Las consultas a la BDS no se han realizado de una forma aleatoria, sino organizada. En primer lugar, se han extraído los datos de orden y de frecuencia teniendo en cuenta los 6 grandes grupos de cláusulas de que hemos hablado. Dentro de cada grupo, hemos separado los casos de voz activa y media por un lado, y los de pasiva por otro.

Por otra parte, se ha comprobado que las posibilidades de orden varían dependiendo de los clíticos que acompañan al predicado, permitiéndonos establecer cuatro subgrupos dentro de los casos de la voz activa: a) cláusulas sin clíticos funcionales (*comió una manzana; se comió una manzana*); b) cláusulas con clítico dativo (*le comió sus manzanas)*; c) con clítico acusativo (*la comió; me la comí*); d) con dos clíticos, uno acusativo y otro dativo (*nos la dio*). A su vez, dentro del grupo de pasiva, es posible distinguir casos sin clíticos (*fue visitado*) y casos con dativo (*le fue entregado el premio anoche*). Toda esta información nos la ofrece la BDS.

Como ejemplo ilustrativo, la Tabla 2 recoge los datos del corpus referidos a las combinaciones de los constituyentes sujeto, predicado y predicativo del complemento directo, cuando el predicado va acompañado de un clítico acusativo, en voz activa o media y para las cláusulas que hemos llamado "de tipo general".

ESQUEMA	ORDEN	Nº de casos	Frecuencia (sobre el esquema)
SDOPR	PVS[11]	8	(sobre 1196: 0.67)
SDOPR	SPV	1	(sobre 1196: 0.08)
SDOPR	SVP	283	(sobre 1196: 23.66)
SDOPR	VPS	10	(sobre 1196: 0.84)
SDOPR	VSP	11	(sobre 1196: 0.92)

Tabla 2. Extracto de los datos de orden del esquema SDOPR.

De este modo comprobamos, entre otras cosas, que el orden PSV no aparece en el corpus y que el orden SPV no es significativo, por lo que estas alternativas deberían estar inactivas en la gramática. El corpus también nos facilita información sobre las cláusulas que contienen elementos de orden fijo: relativas, interrogativas, exclamativas parciales. En estos tipos de cláusula, las posibilidades de orden varían dependiendo de la función sintáctica del elemento inicial, que, al tener una posición fija reduce el número de alternativas a considerar.Todo este grado de detalle en la información de la BDS y que se refleja directamente en AVALON repercute en el nivel de acierto del analizador, ya que no trabaja con restricciones de selección.

5.2.2 CRITERIOS DE DISPOSICIÓN DE LAS ALTERNATIVAS CLAUSALES

En la organización de las alternativas seguimos unos criterios fijos que exponemos a continuación. En primer lugar, colocamos las alternativas referidas a la voz activa y media. Cierran la regla las posibilidades con voz pasiva. Dentro de cada uno de estos dos apartados, seguimos un orden decreciente, es decir, las alternativas con mayor número de constituyentes sintácticos preceden a las más cortas. Primero colocamos las compuestas por 4 constituyentes más un predicado, seguidamente las de 3 más predicado, y así sucesivamente.

Dado que manejamos una opción de ejecución del sistema para que ofrezca un único análisis de cada secuencia, el orden decreciente desempeña un papel crucial. A igual número de penalizaciones, el primer análisis encontrado es el ofrecido. Cada argumento (SUJETO, COMPLEMENTO DIRECTO, COMPLEMENTO INDIRECTO...) se reescribe seguido opcionalmente de uno o varios CIRCUNSTANCIALES. El hecho de que las alternativas más largas estén situadas antes provoca que el *parser* prefiera los análisis con argumentos exigidos por el verbo antes que los análisis con circunstánciales, siempre de acuerdo con los datos de subcategorización del predicado. Así, por ejemplo: *El hombre pensaba en su casa,* será analizada como SUJETO + PREDICADO + COMPLEMENTO PREPOSICIONAL, y no como SUJETO + (PREDICADO + CIRCUNSTANCIAL) porque la primera opción implica mayor número de argumentos sintácticos y el verbo 'pensar' admite la construcción con un COMPLEMENTO PREPOSICIONAL precedido de *en.* En cambio, *El hombre pensaba mientras estaba en su casa* será analizado como SUJETO + (PREDICADO + CIRCUNSTANCIAL), ya que la cláusula adverbial *mientras estaba...* no puede ser interpretada de otro modo.

A igual número de argumentos, seguimos un orden de frecuencia. Las alternativas más frecuentes preceden a las menos usuales. De este modo, en el caso de que en el

[11] La V (de verbo) representa al predicado en el orden, mientras que la P se reserva para el complemento predicativo.

input hubiese una ambigüedad teórica, el *parser* resuelve el conflicto ofreciendo el análisis más frecuente de acuerdo con los datos del corpus. Aunque caben las dos interpretaciones, el sistema escoge la primera porque la hemos colocado antes basándonos en los datos de frecuencia de la BDS:

AVALON da prioridad a las cláusulas subordinadas más largas sobre las que contienen menos argumentos funcionales. Esto lo conseguimos añadiendo progresivamente más *penalties* a medida que desciende el número de argumentos en las reglas de reescritura de las cláusulas subordinadas. De esta forma, la secuencia *quiere comer un pollo* será analizada como:

> PREDICADO (*quiere*) + COMPLEMENTO DIRECTO (*comer un pollo*)

Y no como

> PREDICADO (*quiere*) + COMPLEMENTO DIRECTO (*comer*) + SUJETO (*un pollo*),

a pesar de que esta última alternativa posee más argumentos (2 + predicado) y por tanto, está situada antes en la parte derecha de la regla. La razón es que nuestra gramática da más peso al hecho de que la subordinada sea más larga. Nótese que el análisis preferido por el sistema contiene una subordinada de infinitivo más larga:

PREDICADO (comer) + COMPLEMENTO DIRECTO (un pollo)

que la del análisis desechado: PREDICADO (comer)

6. *Final*

En primer lugar, AVALON es una gramática íntimamente asociada a un corpus textual, ARTHUS (sección contemporánea) que, tal como es concebida, resulta de la tensión entre dos principios contrapuestos. De una parte, la dependencia de los datos derivados del corpus, que constituye un principio reductor, sobre todo cuando las frecuencias altas se convierten en factores decisivos para determinar lo que va a entrar en la zona de cobertura de la gramática. De otra, la extensibilidad, que compensa al anterior en tanto que deja el camino abierto a ampliaciones fáciles a medida que vaya siendo necesario dar cuenta de más especificaciones para fenómenos lingüísticos concretos.

En segundo término, la concepción derivada de corpus con la que elaboramos AVALON pretende que, al ejecutar el parser a ella asociado de manera que sólo produzca un análisis, éste tenga las más altas probabilidades de ser el correcto.

7. *BIBLIOGRAFÍA.*

Contreras, H. (1978): *El orden de palabras en español*. Madrid: Ediciones Cátedra.

Diccionario Actual de la Lengua Española, DALE (1990): Barcelona: Bibliograf.

García-Miguel, J.M.ª (1994): "Corpus de textos analizados sintácticamente", en J. Gómez Guinovart (ed.), *Aplicaciones lingüísticas de la informática*, Santiago de Compostela: Tórculo, 19-34.

Hallebeek, J. (1992): *A Formal Approach to Spanish Syntax*, Amsterdam: Rodopi.

Koster, C.H.A. (1991): "Affix Grammars for Natural Languages", en *Attribute Grammars, Applications and Systems, International Summer School SAGA*. Lecture Notes in Computer Science, vol. 545, pp. 358-373. Praga: Springer-Verlag.

López Meirama, B. (1997a): *La posición del sujeto en la cláusula monoactancial en español*. Lalia, Series Maior, nº 7. Universidad de Santiago de Compostela.

López Meirama, B. (1997b): "Los verbos "ergativos" y la posición del sujeto en las cláusulas monoactanciales en castellano". *Estudios de Lingüística General: conferencias y trabajos presentados en el II Congreso Nacional de Lingüística General*, Granada, 25 al 27 de marzo de 1996. Edición de Francisco Fdez. García. Granada.

Oostdijk, N. (1996). *Corpus Linguistics and the Automatic Analysis of English*. Amsterdam: Rodopi.

Rojo, G. (1978). *Cláusulas y oraciones*. Anexo 14 de *Verba*. Universidad de Santiago de Compostela.

Rojo, G., T. Jiménez Juliá (1989): *Fundamentos del análisis sintáctico funcional*, Lalia, 2, Universidad de Santiago de Compostela.

Rojo, G.: "La base de datos sintácticos del español actual", Español actual, 59, 1993, 15-20 (pero 1995).