"Lingüística de corpus y lingüística descriptiva: el caso de la 'duplicación de objetos'"

José Mª García-Miguel (Universidad de Vigo);
Victoria Vázquez (Universidad de Santiago)

0. Introducción

El objetivo de esta comunicación es mostrar las posibilidades que ofrece la utilización de un corpus analizado sintácticamente para el estudio de un aspecto concreto de la sintaxis del español. El problema que vamos a considerar es el de la coexistencia, en principio opcional, de un pronombre clítico objeto y un constituyente nominal -complemento directo (CDIR) o complemento indirecto (CIND)- correferente y concordado con el átono, como en los ejemplos siguientes:

(1) lamenté haber**lo** dejado **al pobre** donde lo dejé (LAB, 42).

(2) **colegios de monjas** ya **los** tengo en el barrio (MADRID, 206).

(3) me pediste que **le** comunicara **a Agustina** que no ibas a cenar (SUR, 37).

Los estudios gramaticales del español no han llegado a explicar por completo cuál es la razón de esta 'duplicación' ni sus condiciones de uso. Los trabajos más relevantes sobre el fenómeno (Poston, 1953; Barrenechea-Orecchia, 1970) utilizan una metodología cuantitativa y consisten esencialmente en recuentos de frecuencias en función de distintas variables. En estos trabajos se presentan datos sin interpretarlos, y se maneja un número relativamente reducido de ejemplos tratados 'manualmente'[1], lo cual hace que algunas correlaciones entre variables sean estadísticamente poco fiables.

Frente a esta situación, la informatización de corpus del español nos permitirá consultar de una manera rápida y sencilla un conjunto de ejemplos mucho más amplio que el que pudieron manejar los trabajos anteriores.

1. Características del corpus utilizado

En este caso lo que necesitamos es un corpus total o parcialmente analizado sintácticamente, pues lo que buscamos son estructuras específicas en lugar de meras

---

1 Barrenechea-Orecchia (1970), por ejemplo, basan sus cuadros estadísticos en 1.924 cláusulas.

secuencias de palabras o etiquetas. El problema es que la producción a gran escala de corpus analizados sintácticamente es bastante más compleja que el simple almacenamiento de textos o su etiquietado gramatical (cfr., por ejemplo, Aarts & Van den Heuvel (1985); Garside & Leech (1987); Garside(1993))[2].

Para nuestro estudio sobre la duplicación pronominal, hemos utilizado el corpus de la Universidad de Santiago, en el que contamos con un total de 1.500.000 palabras de textos impresos del español contemporáneo. Salvo alguna excepción poco relevante, el corpus incluye textos completos, mejor que fragmentos de tamaño más o menos equivalente para cada texto. El 79% de los textos son de España y el 21% restante de Hispanoamérica. Los textos se han clasificado en cuatro tipos: narrativos (37%), ensayísticos (18%), teatrales (15%), periodísticos (11%), orales (19%).

El corpus de la Universidad de Santiago presenta la particularidad de que fue concebido desde el principio como apoyo a investigaciones sobre la sintaxis de la cláusula en español y, en particular, para el estudio del régimen verbal (valencia). Por ello, se acometió inmediatamente, por procedimientos 'manuales', un análisis sintáctico de todas las cláusulas del corpus. La *Base de datos sintácticos de la Universidad de Santiago* (BDSUS) recoge todos aquellos rasgos que se consideran pertinentes para ese tipo de problemas, pero no otros considerados no relevantes. En concreto, frente a lo que sería un análisis completo en forma de diagrama arbóreo, o corchetes rotulados, han quedado fuera de momento o se consideran sólo marginalmente:

- la estructura de la oración (entendida como complejo de cláusulas)
- la estructura interna de los constituyentes de la cláusula (por ejemplo, la modificación en la frase)
- información morfosintáctica sobre palabras individuales
- los circunstanciales de la cláusula, de los que sólo se ha considerado la posición.

A cambio, esto nos permite contar en el momento de redactar esta comunicación con unas 130.000 cláusulas analizadas (el total previsto es de unas 150.000) y, lo que es más importante, permite ser más detallado en aquellos aspectos que interesan. Para cada cláusula se han anotado datos como los siguientes:

---

2 Hallebeek (1992) presenta una gramática formal del español diseñada con el propósito de analizar el Corpus de Nimega de textos españoles contemporáneos, formado por unas 500.000 palabras

a) Estructura funcional de la cláusula, marcando la presencia, en su caso, de las funciones sintácticas nucleares: sujeto, complemento directo, complemento indirecto, complemento(s) preposicional(es), complemento agente, complemento predicativo.

b) Características inherentes de los elementos funcionales nucleares: categoría sintáctica (y subcategoría), rasgos semánticos ([±animado], [±concreto], [±contable]), determinación, número, marcas formales variables, ...

c) Orden de constituyentes.

d) Datos referentes a la cláusula como conjunto: tipo, función que desempeña, voz, forma verbal, etc.

Este conjunto de datos se ha distribuido en 57 dimensiones para cada una de las cuales se han determinado previamente un conjunto finito de opciones disponibles (que van desde un mínimo de dos opciones -sí /no- hasta un máximo de 57 opciones para el tipo de unidad que desempeña cada función). Para tales opciones se ha utilizado un sistema de claves numéricas que permite jerarquizar los rasgos.

Ya hemos mencionado que el principal objetivo de la BDSUS es el estudio del régimen verbal y la elaboración de un Diccionario de contrucciones verbales con información sobre frecuencias (vid. Rojo, 1992); pero al mismo tiempo proporciona una importante base empírica para el estudio de cualquier problema descriptivo relacionado con la estructura sintáctica nuclear de la cláusula. El diseño de la base de datos la hace particularmente adecuada para el estudio estadístico de correlaciones entre propiedades de diferentes campos (dimensiones). En el caso concreto de la duplicación pronominal podemos comprobar sencillamente si son nulos o no el campo correspondiente a la unidad que desempeña la función CDIR o CIND, y el campo reservado para los clíticos pronominales de cualquiera de esas funciones. Además, podemos buscar correlaciones con rasgos más específicos en esos campo o con otros campo como los reservados para animación, determinación u orden de constituyentes.

2. La duplicación de complementos en español. Bases teóricas

En contraste con el planteamiento más generalizado, que concibe los clíticos como elementos 'nominales' de características similares a los pronombres tónicos, algunos gramáticos han enfocado el problema de la duplicación desde una perspectiva distinta. Es lo que ocurre con aquellos autores que intentan explicar el hecho como un ejemplo de

"conjugación objetiva" (Lenz, 1920; Heger, 1966; Rothe, 1966; Llorente y Mondéjar, 1974). En esta línea, pero con ciertas matizaciones que llevan a evitar el término "conjugación objetiva" en este caso, se sitúa la consideración de los clíticos como signos de concordancia con los complementos (vid. García-Miguel, 1991). Los clíticos serían entonces constituyentes del predicado -y ya no constituyentes inmediatos de la cláusula- junto a la forma verbal, y su función como morfemas de concordancia con CDIR y CIND sería similar a la de las desinencias de número y persona que señalan en el verbo las características morfológicas del sujeto.

De esta manera nos encontramos con que en el predicado se codifican rasgos gramaticales de determinados constituyentes de la cláusula, en concreto, aquellos que desempeñan las funciones SUJ, CDIR y CIND. Cabe preguntarse, entonces, qué tienen en común estas tres funciones clausales que pueda relacionarse con su capacidad para concordar con el predicado.

En nuestra opinión, la distinción entre funciones centrales y no centrales confiere unidad y sentido a esta concordancia entre el predicado y ciertos actantes. Según el planteamiento que hemos expuesto con detalle en otros trabajos (Vázquez Rozas, 1989; García-Miguel, 1992), las funciones centrales se caracterizan frente a las no centrales por su alto grado de gramaticalización, lo cual se manifiesta en los procedimientos de expresión empleados (morfológicamente poco marcados), en la frecuencia de las funciones centrales en los diversos tipos de cláusulas, y en su caracterización semántica, que escapa a una identificación simplista entre función sintáctica y papel semántico (del tipo Agente, Paciente, Receptor, etc.). La codificación sintáctica de ciertas entidades a través de las funciones centrales constituye un instrumento a disposición del hablante para seleccionar aquellos participantes que desempeñan un papel relativamente más relevante en la predicación, definiendo así la ***perspectiva*** lingüística con que se enfoca el estado de cosas designado por la cláusula, las entidades situadas en primer plano (cfr. el concepto de **'perfil'** de Langacker).

La relevancia de los argumentos representados por las funciones centrales parece ser consecuencia de la interacción de ciertos rasgos inherentes a las propias entidades implicadas (animación, determinación, especificidad), y de su papel semántico-designativo ("case-

roles"). De todos modos, el significado de cada una de las funciones centrales no puede ser definido por referencia directa a estos factores, sino que se establece por oposición entre ellas. Hablamos, pues, ante valores relativos, cuyas manifestaciones concretas varían dependiendo de los contrastes específicos que se dan en cada tipo de esquema clausal.

Este contraste entre funciones centrales se muestra de manera especial en las cláusulas biactanciales SUJ-PRED-CDIR. En este esquema observamos la polarización entre ambas funciones en lo que atañe a su caracterización semántica y pragmático-discursiva. El contraste se articula en torno a dos nociones complejas e interrelacionadas: la **agentividad** y la **topicalidad**. Tanto la agentividad y la topicalidad como los rasgos que las constituyen se conciben como magnitudes graduales, y no como categorías discretas susceptibles de una formulación en términos de oposiciones privativas. Así, la agentividad es un concepto pluridimensional en el que confluyen factores como la animación, el control, la intención, la efectividad, etc., que se entienden habitualmente a modo de jerarquías o continuos. Por su parte, la noción de topicalidad es también gradual e incluye los conceptos de tematicidad, información dada y carga anafórica[3].

Como resultado de la actuación de estas propiedades, el contraste entre las funciones centrales SUJ y CDIR se manifiesta en el carácter prototípico y sintácticamente no marcado de aquellas cláusulas transitivas cuyo SUJ es una entidad animada, definida, específica e informativamente dada, que realiza voluntariamente una acción que afecta directamente a la entidad representada por el CDIR, que muestra unas características claramente divergentes de las del SUJ en lo que se refiere a los rasgos aludidos.

Pero SUJ y CDIR no son las únicas funciones centrales de la cláusula en español. Hay argumentos de índole sintáctico-semántica que nos llevan a incluir también el CIND en la esfera de la 'centralidad'. En consecuencia, el valor lingüístico básico de esta última función ha de definirse en relación con los valores de SUJ y CDIR.

Tanto desde el punto de vista semántico-referencial como desde la perspectiva pragmático-discursiva, el CIND muestra unas características que lo sitúan en una posición

---

3 Utilizamos 'topicalidad' en el sentido de Givón (1983 y 1990), definida como "the relative degree to which one NP is considered 'more old information', 'more presupposed', 'less focussed' or 'less foregrounded' than another" (1976, 186).

intermedia entre SUJ y CDIR en las escalas de agentividad y topicalidad. Resulta así la siguiente jerarquía de funciones centrales en español:

SUJ > CIND > CDIR

Si nuestro planteamiento es correcto, cabe esperar que las diferencias de contenido señaladas entre las funciones centrales se reflejen en los mecanismos de codificación sintáctica de las mismas. Y, ciertamente, por lo que concierne al fenómeno de la concordancia entre el predicado y los actantes, los hechos del español apuntan en la dirección esperada. Es decir, si la concordancia está en correlación con la centralidad, y esta se fundamenta en las nociones de agentividad y topicalidad, es natural que la extensión de la concordancia sea máxima para aquella función que alcance un grado más alto de agentividad y topicalidad -el SUJ-, y descienda a medida que bajamos en estas escalas.

Sin duda la concordancia prácticamente obligatoria entre el predicado y el actante SUJ sitúa a esta función en un lugar destacado frente a las otras dos, y constituye una prueba más del puesto que se le asigna en la jerarquía de centralidad. En cuanto a la ordenación relativa atribuida a CIND y CDIR, tiene su reflejo en la frecuencia de casos de duplicación frente a aquellos otros en que estas funciones están desempeñadas por un constituyente "pleno" no duplicado:

	CDIR		CIND	
Clítico + f. plena	1350	*2,31%*	2063	*63,40%*
F. plena no dupl.	57025	*7,69%*	1191	*36,60%*
TOTAL	58375		3254	

Siguiendo el planteamiento de Givón (1983), el uso de recursos anafóricos débiles (v. gr. el empleo de clíticos sin forma plena) sería también concomitante con un alto grado de topicalidad del participante, y de nuevo aquí los datos nos confirman la mayor topicalidad del CIND (78'78% clítico sólo) frente al CDIR (20,47% clítico sólo).

Dado que nuestro análisis persigue una explicación coherente de los diversos aspectos de la duplicación, no es suficiente con señalar la existencia de una correlación general entre nociones semánticas y pragmáticas y la diferente extensión global del fenómeno con CIND y CDIR. Para que nuestra propuesta sea aceptable, los parámetros que manejamos

han de dar cuenta asimismo de las variaciones que se observan en el empleo de la duplicación con cada una de estas funciones sintácticas.

Desde una perspectiva general, este ejemplo de falta de uniformidad en los mecanismos de expresión de las funciones clausales es una muestra de lo que G. Lazard (1984) llama "variación actancial", que consiste en la existencia de posibilidades alternativas para la codificación sintáctica de los argumentos de una predicación. Los fenómenos que se incluyen bajo este epígrafe han interesado de manera especial a la tipología lingüística, marco en el que se ha mostrado que los factores que condicionan tal variación tienen pertinencia interlingüística. El estudio comparativo de los hechos de variación actancial en las lenguas particulares permite comprobar la existencia de tendencias universales en los correlatos semánticos y pragmáticos que fundamentan las alteraciones sintácticas.

Por lo que concierne al español y al ámbito concreto de la codificación sintáctica de los objetos, pueden explicarse como hechos de variación actancial tanto el uso de la preposición *a* ante CDIR como ciertos casos de leísmo, además, claro está, de la cuestión que nos ocupa aquí.

Un repaso a los rasgos de contenido que se han propuesto para explicar las diversas manifestaciones -intralingüísticas e interlingüísticas- de la variación actancial revela de inmediato evidentes coincidencias con las propiedades en que, según nuestra opinión, se fundamenta el contraste entre participantes centrales de la cláusula. En el caso específico de la duplicación, nos encontramos, pues, con que las mismas nociones que permiten dar cuenta de las diferencias entre funciones centrales, y de las extensión global de la concordancia con cada una de ellas, proporciona el fundamento semántico y pragmático de la distribución interna de la duplicación con CDIR y CIND.

3. La variación de la duplicación en el corpus

En un cierto corpus puede resultar imposible averiguar de un modo directo si efectivamente resultan pertinentes la topicalidad y la agentividad de CDIR y CIND en cuanto a sus posibilidades de aparecer duplicados pronominalmente, y sin embargo podemos llegar a contrastar esta hipótesis (en alguna medida) si nuestra base de datos ofrece información

acerca de rasgos como el orden de los constituyentes clausales, el tipo de unidad de los constituyentes, la animación, la determinación, el empleo de recursos anafóricos débiles, etc., que se relacionan **indirectamente** con las nociones de topicalidad y agentividad.

Nuestra exploración del corpus nos ha mostrado los factores que se detallan en los cuadros como relevantes en el fenómeno de la duplicación pronominal[4]. Su relación con los valores semántico-pragmáticos de los participantes centrales la comentamos a continuación.

En el cuadro 2 tenemos las frecuencias y porcentajes generales de duplicación del CDIR, desglosados en cinco factores:

1) El **orden** es un indicio de tematicidad. La posición inicial es lo que marca al tema en español. Como puede comprobarse, los CDIR antepuestos constituyen la excepción al principio general de que el CDIR normalmente no duplica. La tematicidad implica topicalidad y es también una característica del sujeto; por lo que la duplicación marca la proximidad semántico-pragmática del CDIR con el sujeto y, al mismo tiempo, contribuye a distinguirlo de él.

2) La **animación**. Las entidades animadas son capaces de actuar autónomamente y, también, son cognitivamente las entidades más salientes en el universo del discurso. Esto justifica el mayor porcentaje de duplicación de los CDIR de referente animado, tanto por su mayor agentividad (potencial) como por su mayor topicalidad.

3) La **determinación**. Los porcentajes más altos de duplicación los encontramos con las frases determinadas; es decir, cuyo referente es identificable en el contexto previo. Resulta evidente la conexión entre identificabilidad y el concepto de topicalidad tal como lo hemos entendido en el apartado anterior.

4) El **tipo de unidad**. En este parámetro se combinan los dos factores anteriores. Es casi obligatoria la duplicación de pronombres personales, inherentemente definidos y casi siempre animados. Los pocos casos que hemos encontrado de cláusulas duplicadas presentan siempre información que se da por supuesta:

---

4 Se observará que las frecuencias totales varían dependiendo del factor examinado. Esto se explica por la propia configuración de los datos en la BDSUS. Así, por ejemplo, las diferencias en los totales correspondientes a los rasgos *'animación'* y *'determinación'* se entienden si tenemos en cuenta que las cláusulas no se marcan con respecto al factor determinación, pero sí se incluyen entre los elementos inanimados. También las frecuencias totales resultandes de los rasgos *'orden'* y *'animación'* son distintas, dado que nuestra BDS no registra en el orden de constituyentes elementos tales como relativos, interrogativos o exclamativos.

(4) Inf. A.- Y sí cunde, desde luego
Inf. B.- ¡Hombre que...! ¡Ya **lo** creo **que cunde**! (MADRID, 427)

5) El **registro**. Comprobamos porcentajes ligeramente más altos de duplicación en la lengua hablada que en la escrita. La primera, más expresiva, presenta una mayor ligazón contextual en sus estructuras. La mayor duplicación en lengua hablada quizá muestra que el fenómeno está en progresión[5].

En el cuadro 3 se presentan las frecuencias y porcentajes de duplicación del CIND. Los factores examinados coinciden en parte con los vistos a propósito del CDIR (cuadro 1), y de nuevo los datos resultantes vienen a confirmar la hipótesis apuntada en nuestra propuesta de análisis. La mayor presencia de formas duplicadas en el caso del CIND que en el del CDIR se explica por su mayor topicalidad y agentividad potencial, rasgos que lo aproximan al sujeto.

Hemos incluido también en el cuadro 3 dos factores suplementarios que podrían resultar relevantes, teniendo en cuenta lo señalado anteriormente acerca del contraste entre las funciones centrales. Así, en los esquemas ditransitivos (SUJ-PRED-CDIR-CIND) se observa una polarización entre los objetos que hace que las posibilidades de duplicación del CIND aumenten en aquellos casos en que el CDIR se codifica como un elemento de topicalidad (relativamente) alta -en forma de clítico solo o constituyente pleno duplicado-; por el contrario, la duplicación del CIND desciende cuando las diferencias de topicalidad potencial de CIND y CDIR son suficientes para manifestar el contraste entre ambas funciones centrales, de ahí que, por ejemplo, la duplicación del CIND sea menor en aquellos casos en que el CDIR es un elemento sin determinante alguno, es decir, con un topicalidad potencial muy baja.

---

5 La progresión de la duplicación se comprueba también si comparamos nuestros datos con los que ofrece Rini (1991) para el español medieval y clásico.

Cuadro 1. Factores que afectan a la duplicación de CDIR

	DUPL	NO DUPL	% *DUPL*
1. **Orden**			
Antepuesto	663	882	*42,91%*
Pospuesto	593	49874	*1,18%*
2. **Animación**			
Animado	523	4199	*11,08%*
Inanimado	775	51717	*1,48%*
3. **Determinación**			
Determinado definido	960	22541	*4,08%*
Determinado indefinido	106	10068	*1,04%*
Sin determinante	10	6199	*0,16%*
4. **Tipo de unidad**			
Pronombres personales	297	1	*99,66%*
FN y otros pronombres	1023	45322	*2,21%*
Cláusulas	13	11629	*0,11%*
5. **Registro**			
Lengua oral	345	7894	*4,19%*
Lengua escrita	1005	49131	*2,00%*

Cuadro 2. Duplicación de CDIR (Combinación de registro y posición con animación y determinación)

	L. ORAL			**L.ESCRITA**		
	DUPL	NO DUPL	% *DUPL*	NO DUPLDUPL	% *DUPL*	
A) CDIR ANTEPUESTO	187	153	*55%*	476	729	*43,08%*
1. **Animación**						
Animado	34	8	*80,95%*	177	38	*82,33%*
Inanimado	153	145	*51,34%*	299	691	*30,20%*
2. **Determinación**						
Determinado definido	164	32	*83,67%*	404	161	*71,50%*
Determinado indefinido	14	31	*31,11%*	38	160	*19,19%*
Sin determinante	2	20	*9,09%*	8	65	*10,96%*
B) CDIR POSPUESTO	106	6632	*1,57%*	487	43242	*1,11%*
1. **Animación**						
Animado	60	460	*11,54%*	272	3313	*7,59%*
Inanimado	46	6172	*0,74%*	215	39929	*0,54%*
2. **Determinación**						
Determinado definido	77	2098	*3,54%*	315	20250	*1,53%*
Determinado indefinido	12	1798	*0,66%*	42	8079	*0,52%*
Sin determinante	0	906	*0%*	0	5208	*0 %*

Cuadro 3. Factores que afectan a la duplicación de CIND

	DUPL	NO DUPL	% *DUPL*
1. Orden			
Antepuesto	1032	30	*97,18%*
Pospuesto	983	1079	*47,67%*
2. Animación			
Animado	1878	709	*72,59%*
Inanimado	185	482	*27,74%*
3. Determinación			
Determinado definido	1876	1030	*64,56%*
Determinado indefinido	153	128	*54,45%*
Sin determinante	3	19	*13,64%*
4. Tipo de unidad			
Pronombres personales	814	26	*96,91%*
FN y otros pronombres	1248	1156	*51,97%*
Cláusulas	1	9	*10,00%*
5. Registro			
Lengua oral	509	106	*82,76%*
Lengua escrita	1554	1085	*58,89%*
6. Forma CDIR			
CDIR clítico sólo	118	27	*81,38%*
CDIR duplicado	21	3	*87,5%*
CDIR no duplicado	697	803	*46,47%*
7. Determinación CDIR			
CDIR definido	288	326	*46,91%*
CDIR indefinido	173	182	*48,73%*
CDIR sin determinante	112	225	*33,23%*

4. Límites y perspectivas de la utilización del corpus.

De los datos del corpus que acabamos de exponer parece deducirse que no existen condiciones necesarias y suficientes para la elección de la construcción con 'duplicación' frente a la no duplicada. Las variaciones de frecuencia nos llevan, en el terreno de la lingüística descriptiva, al concepto de regla variable (desarrollado inicialmente por la sociolingüística) y también al concepto de prototipo (y desviaciones y extensiones a partir del prototipo) desarrollado por la lingüística cognitiva. En el campo de la lingüística computacional, tenemos el problema del manejo de reglas de aplicación variable. Las soluciones más inmediatas distorsionan ambas el uso lingüístico, que a veces debe ser

Cuadro 3. Factores que afectan a la duplicación de CIND

	DUPL	NO DUPL	% *DUPL*
1. **Orden**			
Antepuesto	1032	30	*97,18%*
Pospuesto	983	1079	*47,67%*
2. **Animación**			
Animado	1878	709	*72,59%*
Inanimado	185	482	*27,74%*
3. **Determinación**			
Determinado definido	1876	1030	*64,56%*
Determinado indefinido	153	128	*54,45%*
Sin determinante	3	19	*13,64%*
4. **Tipo de unidad**			
Pronombres personales	814	26	*96,91%*
FN y otros pronombres	1248	1156	*51,97%*
Cláusulas	1	9	*10,00%*
5. **Registro**			
Lengua oral	509	106	*82,76%*
Lengua escrita	1554	1085	*58,89%*
6. **Forma CDIR**			
CDIR clítico sólo	118	27	*81,38%*
CDIR duplicado	21	3	*87,5%*
CDIR no duplicado	697	803	*46,47%*
7. **Determinación CDIR**			
CDIR definido	288	326	*46,91%*
CDIR indefinido	173	182	*48,73%*
CDIR sin determinante	112	225	*33,23%*

4. Límites y perspectivas de la utilización del corpus.

De los datos del corpus que acabamos de exponer parece deducirse que no existen condiciones necesarias y suficientes para la elección de la construcción con 'duplicación' frente a la no duplicada. Las variaciones de frecuencia nos llevan, en el terreno de la lingüística descriptiva, al concepto de regla variable (desarrollado inicialmente por la sociolingüística) y también al concepto de prototipo (y desviaciones y extensiones a partir del prototipo) desarrollado por la lingüística cognitiva. En el campo de la lingüística computacional, tenemos el problema del manejo de reglas de aplicación variable. Las soluciones más inmediatas distorsionan ambas el uso lingüístico, que a veces debe ser

sacrificado en favor de la efiuciencia de los programas. Bien podemos considerar la regla como simplemente opcional, descartando la tendencias que se reflejan en las diferencias de frecuencia (quizá sea ésta la solución preferible en el análisis automatizado); bien podemos formular reglas más restrictivas que permitan sólo las opciones más frecuentes (solución quizá preferible en la generación)

También nos ha parecido interesante plantearnos hasta dónde podíamos llegar acudiendo a los rasgos previstos en una base multiuso cuando se estudia un problema concreto como el de la duplicación.

Considerando tanto propiedades que nosotros mismos hemos comprobado en algunos ejemplos procedentes de este corpus o de otras fuentes como propiedades citadas por otros autores que han prestado atención a la duplicación pronominal, estas son algunas de las insuficiencias que hemos encontrado en nuestra BDS para la obtención automática de datos estadísticos sobre la duplicación pronominal[6]:

a) Los rasgos formales no son siempre lo suficientemente específicos. Por ejemplo, la mayor parte de los CDIR duplicados pospuestos en los textos del español peninsular corresponden, además de a pronombres personales, a la forma *todo*:

(5) Ya te **lo** conté **todo**(SON, 233)

En el diseño de la BDS no se consideró necesario separar tal palabra de otras frases nominales añadiéndola, como opción independiente, a las 54 opciones ya previstas entre los tipos de unidad que desempeñan una función. De haberlo hecho, nada garantiza que las opciones resultantes fueran suficientes para el estudio de este u otros problemas.

b) Faltan rasgos semánticos, no directamente ligados a diferencias formales, que pueden resultar relevantes. Por ejemplo, en la cláusula siguientes no hay duplicación, en contra de la tendencia general, de un CIND animado y determinado. La razón parece estar, siguiendo a Suñer (1988), en que se trata de una expresión definida genérica:

(6) No sé decir que no **a las mujeres guapas** (LAB, 104)

---

6 Existen, por supuesto, otras muchas posibilidades en nuestra BDS que no hemos examinado por considerarlas menos relevantes que las consideradas. Creemos que podrían obtenerse algunas correlaciones interesantes al estudiar el uso de la duplicación en cláusulas declarativas vs. interrogativas, afirmativas vs. negativas, activas vs. construcciones pronominales, distintos esquemas funcionales de la cláusula, etc.

sacrificado en favor de la efiuciencia de los programas. Bien podemos considerar la regla como simplemente opcional, descartando la tendencias que se reflejan en las diferencias de frecuencia (quizá sea ésta la solución preferible en el análisis automatizado); bien podemos formular reglas más restrictivas que permitan sólo las opciones más frecuentes (solución quizá preferible en la generación)

También nos ha parecido interesante plantearnos hasta dónde podíamos llegar acudiendo a los rasgos previstos en una base multiuso cuando se estudia un problema concreto como el de la duplicación.

Considerando tanto propiedades que nosotros mismos hemos comprobado en algunos ejemplos procedentes de este corpus o de otras fuentes como propiedades citadas por otros autores que han prestado atención a la duplicación pronominal, estas son algunas de las insuficiencias que hemos encontrado en nuestra BDS para la obtención automática de datos estadísticos sobre la duplicación pronominal[6]:

a) Los rasgos formales no son siempre lo suficientemente específicos. Por ejemplo, la mayor parte de los CDIR duplicados pospuestos en los textos del español peninsular corresponden, además de a pronombres personales, a la forma *todo*:

(5) Ya te **lo** conté **todo**(SON, 233)

En el diseño de la BDS no se consideró necesario separar tal palabra de otras frases nominales añadiéndola, como opción independiente, a las 54 opciones ya previstas entre los tipos de unidad que desempeñan una función. De haberlo hecho, nada garantiza que las opciones resultantes fueran suficientes para el estudio de este u otros problemas.

b) Faltan rasgos semánticos, no directamente ligados a diferencias formales, que pueden resultar relevantes. Por ejemplo, en la cláusula siguientes no hay duplicación, en contra de la tendencia general, de un CIND animado y determinado. La razón parece estar, siguiendo a Suñer (1988), en que se trata de una expresión definida genérica:

(6) No sé decir que no **a las mujeres guapas** (LAB, 104)

---

6 Existen, por supuesto, otras muchas posibilidades en nuestra BDS que no hemos examinado por considerarlas menos relevantes que las consideradas. Creemos que podrían obtenerse algunas correlaciones interesantes al estudiar el uso de la duplicación en cláusulas declarativas vs. interrogativas, afirmativas vs. negativas, activas vs. construcciones pronominales, distintos esquemas funcionales de la cláusula, etc.

No parece previsible que en un corpus amplio se incluya información sobre la [illegible]ionalidad de las expresiones nominales, si tenemos en cuenta la dificultad de asignar [illegible] tipo de rasgos tanto por procedimientos manuales como sobre todo (dada la ausencia de [illegible]nifestaciones formales claras) por procedimientos automáticos.

c) Faltan indicaciones sobre entonación y estructura informativa. Este factor, del que [illegible]pende la distinción *dado / nuevo* es muy relevante en la posibilidad de duplicación (vid. [illegible]ucher, 1956; Silva-Corvalán, 1983). Mientras que, por ejemplo, la anteposición de CDIR [illegible]terminado suele implicar su duplicación, ésta no se produce si la entonación lo marca como [illegible] informativo del mensaje. Compárense:

(7) a. // **muchos disgustos** le proporcionaba aquella criatura //
b. // muchos disgustos se *los* proporcionaba **aquella criatura** //

Evidentemente, incorporar a un corpus valores informativos como los de *dado* y *nuevo* supone un trabajo laborioso y cargado de conjeturas. En principio, parece más objetivo (y, teóricamente, susceptible de elaboración computacional) un sistema de transcripción de la entonación de los textos hablados; pero no es el caso de un corpus basado en textos escritos o transcritos.

Estas "carencias" nos exigen la continua incorporación al corpus de información cada vez más detallada, que debería llevarnos al menos a incluir el análisis sintáctico completo de todas las secuencias y a ser lo más específicos posible en la anotación de rasgos gramaticales. Sería deseable también, aunque más difícil de realizar, la inclusión de valores sintáctico-semánticos e informativos. Evidentemente, cuanto mayor sea la información contenida en una base de datos más fácil resultará proporcionar apoyo empírico para nuestros estudios gramaticales.

Aun en el supuesto de que dispusiéramos sin limitaciones de la información que acabamos de reseñar, está claro que siempre nos quedará la labor de interpretar los datos o de emitir hipótesis sobre ellos. En este punto, debemos destacar que los cuadros expuestos en apartados anteriores revelan correlaciones estadísticas más o menos fuertes, pero no condiciones. La anteposición al predicado, la animación, la determinación y similares son sólo un reflejo indirecto de los principios funcionales que motivan en español la duplicación (y, sin duda, también fenómenos similares de otras lenguas). Nosotros hemos situado esa

No parece previsible que en un corpus amplio se incluya información sobre la referencialidad de las expresiones nominales, si tenemos en cuenta la dificultad de asignar este tipo de rasgos tanto por procedimientos manuales como sobre todo (dada la ausencia de manifestaciones formales claras) por procedimientos automáticos.

c) Faltan indicaciones sobre entonación y estructura informativa. Este factor, del que depende la distinción *dado / nuevo* es muy relevante en la posibilidad de duplicación (vid. Hatcher, 1956; Silva-Corvalán, 1983). Mientras que, por ejemplo, la anteposición de CDIR determinado suele implicar su duplicación, ésta no se produce si la entonación lo marca como Foco informativo del mensaje. Compárense:

(7) a. // **muchos disgustos** le proporcionaba aquella criatura //
b. // muchos disgustos se *los* proporcionaba **aquella criatura** //

Evidentemente, incorporar a un corpus valores informativos como los de *dado* y *nuevo* supone un trabajo laborioso y cargado de conjeturas. En principio, parece más objetivo (y, teóricamente, susceptible de elaboración computacional) un sistema de transcripción de la entonación de los textos hablados; pero no es el caso de un corpus basado en textos escritos o transcritos.

Estas "carencias" nos exigen la continua incorporación al corpus de información cada vez más detallada, que debería llevarnos al menos a incluir el análisis sintáctico completo de todas las secuencias y a ser lo más específicos posible en la anotación de rasgos gramaticales. Sería deseable también, aunque más difícil de realizar, la inclusión de valores sintáctico-semánticos e informativos. Evidentemente, cuanto mayor sea la información contenida en una base de datos más fácil resultará proporcionar apoyo empírico para nuestros estudios gramaticales.

Aun en el supuesto de que dispusiéramos sin limitaciones de la información que acabamos de reseñar, está claro que siempre nos quedará la labor de interpretar los datos o de emitir hipótesis sobre ellos. En este punto, debemos destacar que los cuadros expuestos en apartados anteriores revelan correlaciones estadísticas más o menos fuertes, pero no condiciones. La anteposición al predicado, la animación, la determinación y similares son sólo un reflejo indirecto de los principios funcionales que motivan en español la duplicación (y, sin duda, también fenómenos similares de otras lenguas). Nosotros hemos situado esa

motivación funcional en el carácter de participantes centrales de SUJ, CDIR y CIND y en el contraste significativo que se establece entre ellos, lo cual lleva a marcar en el predicado las entidades situadas en perspectiva, las más esperadas en la conceptualización de la situación, las más próximas en sus valores semánticos y pragmático-discursivos al sujeto. Estas propiedades se advierten al interpretar los textos; pero por medios "objetivos" sólo pueden comprobarse indirectamente. No obstante, unos medios de comprobación son más indirectos que otros. En lo que más directamente influye el peso relativo de un participante en un proceso es, sin duda, en su relación en el resto del texto. En un texto se habla más de los participantes principales que de los secundarios. T. Givón (1983), y seguidores, ha utilizado este principio para proponer un método estadístico para medir en un texto la topicalidad relativa de las entidades. Utiliza como criterios la 'distancia referencial' (en número de cláusulas con respecto a la mención anterior del mismo referente) y la 'persistencia' (en cuanto a la mención del mismo referente en las cláusulas subsiguientes). Que sepamos, el método ha sido aplicado hasta ahora a una cantidad de textos relativamente breve. Su aplicación a corpus más amplios permitiría afinar el método y sus resultados, especialmente en el estudio de las diferencias de registro y de la interacción con otras variables. Con ello, saltamos de la estructura de la cláusula, en la que nos hemos estado moviendo hasta ahora, a la estructura del discurso. Resultan necesarios sistemas computacionales de análisis textual (ya no es poca cosa que se puedan identificar correctamente los antecedentes de variables referenciales no ligadas) que sin duda supondrán un importante apoyo para nuestra comprensión del funcionamiento del lenguaje.

motivación funcional en el carácter de participantes centrales de SUJ, CDIR y CIND y en el contraste significativo que se establece entre ellos, lo cual lleva a marcar en el predicado las entidades situadas en perspectiva, las más esperadas en la conceptualización de la situación, las más próximas en sus valores semánticos y pragmático-discursivos al sujeto. Estas propiedades se advierten al interpretar los textos; pero por medios "objetivos" sólo pueden comprobarse indirectamente. No obstante, unos medios de comprobación son más indirectos que otros. En lo que más directamente influye el peso relativo de un participante en un proceso es, sin duda, en su relación en el resto del texto. En un texto se habla más de los participantes principales que de los secundarios. T. Givón (1983), y seguidores, ha utilizado este principio para proponer un método estadístico para medir en un texto la topicalidad relativa de las entidades. Utiliza como criterios la 'distancia referencial' (en número de cláusulas con respecto a la mención anterior del mismo referente) y la 'persistencia' (en cuanto a la mención del mismo referente en las cláusulas subsiguientes). Que sepamos, el método ha sido aplicado hasta ahora a una cantidad de textos relativamente breve. Su aplicación a corpus más amplios permitiría afinar el método y sus resultados, especialmente en el estudio de las diferencias de registro y de la interacción con otras variables. Con ello, saltamos de la estructura de la cláusula, en la que nos hemos estado moviendo hasta ahora, a la estructura del discurso. Resultan necesarios sistemas computacionales de análisis textual (ya no es poca cosa que se puedan identificar correctamente los antecedentes de variables referenciales no ligadas) que sin duda supondrán un importante apoyo para nuestra comprensión del funcionamiento del lenguaje.

PROCEDENCIA DE LOS EJEMPLOS CITADOS

LAB: Eduardo Mendoza, *El laberinto de las aceitunas*, Seix Barral, Barcelona, 1982.

MADRID: M. Esgueva y M. Cantarero (eds.), *El habla de la ciudad de Madrid. Materiales para su estudio*, C.S.I.C., Madrid, 1981.

SON: José Luis Sampedro, *La sonrisa etrusca*, Alfaguara, Madrid, 1985.

SUR: Adelaida García Morales, *El Sur* seguido de *Bene*, Anagrama, Madrid, 1985.

REFERENCIAS BIBLIOGRAFICAS

Aarts, J. & Th. van den Heuvel (1985): "Computational tools for the syntactic analysis of corpora", Linguistics, 23, pp. 303-335.

Barrenechea, A.M. y T. Orecchia (1970): "La duplicación de objetos directos e indirectos en el español hablado en Buenos Aires", *Romance Philology*, 24/1, 58-83.

García-Miguel, J.M. (1991): "La duplicación de complemento directo e indirecto como concordancia", *Verba*, 18, 375-401.

García-Miguel, J.M. (1992): *Aspectos de la estructura de la cláusula: Transitividad y complementación preposicional en español*, Tesis doctoral, Universidad de Santiago.

Garside, R. (1993): "The Large-Scale Production of Syntactically Analysed Corpora", *Literary and Linguistic Computing*, 8/1, 39-46.

Garside, R. & Leech, F. (1987): "The UCREL probabilistic parsing system", en Garside et alt. (eds): *The Computational Analysis of English. A corpus based approach*, Longman, Londres, pp. 66-81.

Givón, T. (ed.)(1983): *Topic Continuity in Discourse. A Quantitative Cross-language Study*, J. Benjamins, Amsterdam.

Givón, T. (1990): *Syntax: A Functional-Typological Introduction*, vol. II, J. Benjamins, Amsterdam.

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

Hatcher, A.M. (1956): "On the inverted object in Spanish", *Modern Language Notes*, 71, 362-373.

Heger, K. (1966): "La conjugaison objective en français et en espagnol", *Langages*, 3, 19-39.

Lazard, G. (1984): "Actance Variations and Categories of the Object", en Plank (ed.):*Objects. Towards a Theory of Grammatical Relations*, Academic Press, Londres, pp. 269-292.

Lenz, R. (1920): *La oración y sus partes. Estudios de gramática general y castellana*, Centro de estudios históricos, Madrid, 1935^3.

Llorente, A. y J. Mondéjar (1974): "La conjugación objetiva en español", *R.S.E.L.*, 4/1, 1-60.

Poston, L. (1953): "The redundant object pronoun in contemporary Spanish", *Hispania*, 36, 263-272.

Rini, J. (1991): "The Redundant Indirect Object Constructions in Spanish: A New Perspective", *Romance Philology*, 45/2, 269-286.

Rojo, G. (1992): "El futuro *Diccionario de construcciones verbales del español actual*", en Martín Vide, C. (ed.): *Lenguajes naturales y lenguajes formales*, VIII, Universitat de Barcelona, 1992, pp. 41-50.

Rothe, W. (1966): "Romanische Objektkonjugation", *Romanische Forschungen*, 78, 530-547.

Silva-Corvalán, C. (1983): "On the interaction of word order and intonation. Some OV constructions in Spanish", en F. Klein-Andreu (ed.): *Discourse Perspectives on Syntax*, Academic Press, Londres, 117-140.

Suñer, M. (1988): "The role of agreement in clitic-doubled constructions", *Natural Language and Linguistic Theory*, 6, 391-434.

Vázquez Rozas, V. (1989): *El complemento indirecto en español*, Tesis doctoral, Universidad de Santiago.

PROCEDENCIA DE LOS EJEMPLOS CITADOS

LAB: Eduardo Mendoza, *El laberinto de las aceitunas*, Seix Barral, Barcelona, 1982.

MADRID: M. Esgueva y M. Cantarero (eds.), *El habla de la ciudad de Madrid. Materiales para su estudio*, C.S.I.C., Madrid, 1981.

SON: José Luis Sampedro, *La sonrisa etrusca*, Alfaguara, Madrid, 1985.

SUR: Adelaida García Morales, *El Sur* seguido de *Bene*, Anagrama, Madrid, 1985.

REFERENCIAS BIBLIOGRAFICAS

Aarts, J. & Th. van den Heuvel (1985): "Computational tools for the syntactic analysis of corpora", Linguistics, 23, pp. 303-335.

Barrenechea, A.M. y T. Orecchia (1970): "La duplicación de objetos directos e indirectos en el español hablado en Buenos Aires", *Romance Philology*, 24/1, 58-83.

García-Miguel, J.M. (1991): "La duplicación de complemento directo e indirecto como concordancia", *Verba*, 18, 375-401.

García-Miguel, J.M. (1992): *Aspectos de la estructura de la cláusula: Transitividad y complementación preposicional en español*, Tesis doctoral, Universidad de Santiago.

Garside, R. (1993): "The Large-Scale Production of Syntactically Analysed Corpora", *Literary and Linguistic Computing*, 8/1, 39-46.

Garside, R. & Leech, F. (1987): "The UCREL probabilistic parsing system", en Garside et alt. (eds): *The Computational Analysis of English. A corpus based approach*, Longman, Londres, pp. 66-81.

Givón, T. (ed.)(1983): *Topic Continuity in Discourse. A Quantitative Cross-language Study*, J. Benjamins, Amsterdam.

Givón, T. (1990): *Syntax: A Functional-Typological Introduction*, vol. II, J. Benjamins, Amsterdam.

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

Hatcher, A.M. (1956): "On the inverted object in Spanish", *Modern Language Notes*, 71, 362-373.

Heger, K. (1966): "La conjugaison objective en français et en espagnol", *Langages*, 3, 19-39.

Lazard, G. (1984): "Actance Variations and Categories of the Object", en Plank (ed.):*Objects. Towards a Theory of Grammatical Relations*, Academic Press, Londres, pp. 269-292.

Lenz, R. (1920): *La oración y sus partes. Estudios de gramática general y castellana*, Centro de estudios históricos, Madrid, 1935[3].

Llorente, A. y J. Mondéjar (1974): "La conjugación objetiva en español", *R.S.E.L.*, 4/1, 1-60.

Poston, L. (1953): "The redundant object pronoun in contemporary Spanish", *Hispania*, 36, 263-272.

Rini, J. (1991): "The Redundant Indirect Object Constructions in Spanish: A New Perspective", *Romance Philology*, 45/2, 269-286.

Rojo, G. (1992): "El futuro *Diccionario de construcciones verbales del español actual*", en Martín Vide, C. (ed.): *Lenguajes naturales y lenguajes formales*, VIII, Universitat de Barcelona, 1992, pp. 41-50.

Rothe, W. (1966): "Romanische Objektkonjugation", *Romanische Forschungen*, 78, 530-547.

Silva-Corvalán, C. (1983): "On the interaction of word order and intonation. Some OV constructions in Spanish", en F. Klein-Andreu (ed.): *Discourse Perspectives on Syntax*, Academic Press, Londres, 117-140.

Suñer, M. (1988): "The role of agreement in clitic-doubled constructions", *Natural Language and Linguistic Theory*, 6, 391-434.

Vázquez Rozas, V. (1989): *El complemento indirecto en español*, Tesis doctoral, Universidad de Santiago.

Cuadro 1. Totales de duplicación y no duplicación de CDIR y CIND

	CDIR		CIND	
Clítico + F. plena	1350	*2,31%*	2063	*63,40%*
F. plena no dupl.	57025	*7,69%*	1191	*36,60%*
TOTAL	58375		3254	

Cuadro 2. Factores que afectan a la duplicación de CDIR

	DUPL	NO DUPL	% DUPL
1. **Orden**			
Antepuesto	663	882	*42,91%*
Pospuesto	593	49874	*1,18%*
2. **Animación**			
Animado	523	4199	*11,08%*
Inanimado	775	51717	*1,48%*
3. **Determinación**			
Determinado definido	960	22541	*4,08%*
Determinado indefinido	106	10068	*1,04%*
Sin determinante	10	6199	*0,16%*
4. **Tipo de unidad**			
Pronombres personales	297	1	*99,66%*
FN y otros pronombres	1023	45322	*2,21%*
Cláusulas	13	11629	*0,11%*
5. **Registro**			
Lengua oral	345	7894	*4,19%*
Lengua escrita	1005	49131	*2,00%*

Cuadro 3. Factores que afectan a la duplicación de CIND

	DUPL	NO DUPL	% DUPL
1. **Orden**			
Antepuesto	1032	30	*97,18%*
Pospuesto	983	1079	*47,67%*
2. **Animación**			
Animado	1878	709	*72,59%*
Inanimado	185	482	*27,74%*
3. **Determinación**			
Determinado definido	1876	1030	*64,56%*
Determinado indefinido	153	128	*54,45%*
Sin determinante	3	19	*13,64%*
4. **Tipo de unidad**			
Pronombres personales	814	26	*96,91%*
FN y otros pronombres	1248	1156	*51,97%*
Cláusulas	1	9	*10,00%*
5. **Registro**			
Lengua oral	509	106	*82,76%*
Lengua escrita	1554	1085	*58,89%*
6. **Forma CDIR**			
CDIR clítico sólo	118	27	*81,38%*
CDIR duplicado	21	3	*87,5%*
CDIR no duplicado	697	803	*46,47%*
7. **Determinación CDIR**			
CDIR definido	288	326	*46,91%*
CDIR indefinido	173	182	*48,73%*
CDIR sin determinante	112	225	*33,23%*

Cuadro 1. Totales de duplicación y no duplicación de CDIR y CIND

	CDIR		CIND	
Clítico + F. plena	1350	*2,31%*	2063	*63,40%*
F. plena no dupl.	57025	*7,69%*	1191	*36,60%*
TOTAL	58375		3254	

Cuadro 2. Factores que afectan a la duplicación de CDIR

	DUPL	NO DUPL	% DUPL
1. **Orden**			
Antepuesto	663	882	*42,91%*
Pospuesto	593	49874	*1,18%*
2. **Animación**			
Animado	523	4199	*11,08%*
Inanimado	775	51717	*1,48%*
3. **Determinación**			
Determinado definido	960	22541	*4,08%*
Determinado indefinido	106	10068	*1,04%*
Sin determinante	10	6199	*0,16%*
4. **Tipo de unidad**			
Pronombres personales	297	1	*99,66%*
FN y otros pronombres	1023	45322	*2,21%*
Cláusulas	13	11629	*0,11%*
5. **Registro**			
Lengua oral	345	7894	*4,19%*
Lengua escrita	1005	49131	*2,00%*

Cuadro 3. Factores que afectan a la duplicación de CIND

	DUPL	NO DUPL	% DUPL
1. **Orden**			
Antepuesto	1032	30	*97,18%*
Pospuesto	983	1079	*47,67%*
2. **Animación**			
Animado	1878	709	*72,59%*
Inanimado	185	482	*27,74%*
3. **Determinación**			
Determinado definido	1876	1030	*64,56%*
Determinado indefinido	153	128	*54,45%*
Sin determinante	3	19	*13,64%*
4. **Tipo de unidad**			
Pronombres personales	814	26	*96,91%*
FN y otros pronombres	1248	1156	*51,97%*
Cláusulas	1	9	*10,00%*
5. **Registro**			
Lengua oral	509	106	*82,76%*
Lengua escrita	1554	1085	*58,89%*
6. **Forma CDIR**			
CDIR clítico sólo	118	27	*81,38%*
CDIR duplicado	21	3	*87,5%*
CDIR no duplicado	697	803	*46,47%*
7. **Determinación CDIR**			
CDIR definido	288	326	*46,91%*
CDIR indefinido	173	182	*48,73%*
CDIR sin determinante	112	225	*33,23%*