

Un modelo para la perífrasis española y el sistema de pronombres clíticos en HPSG¹

Luis Pineda

Iván Meza

Departamento de Ciencias de la Computación
IIMAS, UNAM

In this paper a grammatical model of the Spanish periphrasis and the system of clitic pronouns in HPSG is presented. The model covers both simple periphrases including, in addition to auxiliary and modal verbs, a single verb with a full semantic content, and also the analysis of complex periphrasis with two content verbs. In the second part of the paper a basic model of pronominal clitic sentences including enclitics and proclitics is presented; finally, in the last part, the interaction of the clitic system and the periphrasis, both simple and complex, is presented. We also report a computational implementation of the model and its validation through the analysis of these phenomena in the DIME Corpus (Villaseñor et. al., 2001), which was developed in the context of the DIME project (Pineda et. al., 2002).

Palabras clave: *verbos auxiliares del español, perífrasis española, pronombres clíticos, HPSG.*

Fecha de recepción del manuscrito: febrero del 2004

Luis Pineda e Iván Meza.

Departamento de Ciencias de la Computación

IIMAS, UNAM, Circuito Escolar, Ciudad Universitaria

04510, México D. F.

correos electrónicos: luis@leibniz.iimas.unam.mx, ivanvladimir@leibniz.iimas.unam.mx

¹ Una versión anterior de parte del contenido de este artículo se encuentra en las memorias del workshop “Multilingual Information Access and Natural Language Processing”, IBERAMIA 2002, Julio Gonzalo, Anselmo Peñas y Antonio Ferrández (Eds.), Sevilla, España, Noviembre de 2002.

En este artículo se presenta un modelo computacional de la perífrasis española y su interacción con el sistema de pronombres clíticos en HPSG. El modelo contempla tanto la perífrasis simple que contiene, además de verbos auxiliares y modales, un solo verbo con contenido conceptual, como la perífrasi compleja que contienen dos verbos con contenidos semánticos; el modelo incluye también el análisis de las perífrasis con copulativos que se presentan en las oraciones atributivas y pasivas. En la segunda parte se presenta un modelo básico del sistema de pronombres clíticos que incluye el análisis de la oración enclítica y proclítica; finalmente, se analiza la interacción del sistema de clíticos con la perífrasis, tanto simple como compleja. Se reporta también la validación del modelo mediante su implementación computacional así como mediante el análisis de estos fenómenos en el corpus DIME (Villaseñor et al., 2001), el cual se desarrolló en el contexto del proyecto DIME (Pineda et. al., 2002).

1. Introducción

A pesar de su importancia, la perífrasis y los pronombres clíticos del español han recibido muy poca atención en los estudios de lingüística computacional; al analizar estos fenómenos se asume con frecuencia que el sistema de verbos auxiliares es similar al del inglés y se modela en dichos términos (e.g. Quirino, 2001). Sin embargo, esta posición conduce a dificultades inmediatamente; mientras que *can* y *have* son auxiliares en inglés, sus contrapartes en español *poder* y *haber* no siempre se analizan como tales. De acuerdo con Gili Gaya (1991), por ejemplo, *poder* no es auxiliar ya que además de conservar su contenido conceptual, es agentivo; *haber*, por su parte, forma parte de la conjugación compuesta y se considera una inflexión. Por otro lado, mientras que el inglés tiene sólo el copulativo *be*, el español tiene los verbos complementarios *ser* y *estar* cuyas sutiles interacciones impactan profundamente la estructura de esta lengua. Otra asimetría importante se revela en expresiones como *te lo digo*, *dime* y *hazlo*, por ejemplo, que son muy comunes en español y que, sin embargo, no tienen contraparte directa en inglés. Más aún, la construcción perífrástica y el sistema de pronombres clíticos interactúan de manera compleja pero sistemática. Por estas razones, estos fenómenos, así como sus formas de interacción, constituyen una parte sumamente importante del núcleo gramatical del español y un modelo que los capture de manera apropiada es indispensable para el procesamiento computacional de esta lengua. En este artículo presentamos un modelo de la perífrasis española, así como su interacción con el sistema de pronombres clíticos, incluyendo el fenómeno conocido como *clitic climbing* (subida de clíticos), tanto simple, cuando los pronombres reemplazan al complemento directo, indirecto, o ambos, de un solo verbo, como compuesto, en el que los pronombres reemplazan complementos de verbos diferentes. Se reporta también la formalización del modelo en HPSG (Pollard y Sag, 1994; Sag y Wasow, 1999), así como su implementación en el ambiente de desarrollo LKB (Copestake, 2002), y su validación a través del análisis de estos fenómenos en el corpus DIME (Villaseñor et. al, 2001), el cual se desarrolló en el contexto del proyecto DIME (Pineda et. al., 2002).

2. La perífrasis española

Iniciamos el estudio de la perífrasis mediante el análisis de los verbos auxiliares del español. Para caracterizar estos verbos partimos de tres nociones generales, como sigue:

- (1) Son palabras gramaticalizadas que tienen una función sintáctica fija en lenguaje; por lo mismo, no tienen contenido conceptual. Verbos auxiliares conllevan normalmente información temporal y aspectual que modifica a la interpretación de los verbos con contenido conceptual que siguen a los auxiliares en la perífrasis.
- (2) En oraciones predicativas aparecen en perífrasis en forma conjugada y llevan como complemento una frase verbal en forma no personal (infinitivo, gerundio y participio). En atributivas tienen función copulativa.

- (3) Son verbos que carecen de agente; por esta razón pueden aparecer tanto con verbos agentivos, como en construcciones impersonales y con verbos unipersonales (e.g. *llover*).

Tomando como punto de partida estas tres nociones gramaticales es posible definir cinco pruebas sintácticas para determinar si un verbo es auxiliar o no en oraciones predicativas. Estas pruebas son:

1. Son verbos conjugados que encabezan una frase verbal y anteceden a uno o más verbos en forma no personal:
(a) *puedes mostrar el catálogo a Juan*
2. Modifican a una frase verbal encabezada por un verbo impersonal:
(a) *puede llover*
(b) **tiene llovido*
3. Se requieren todos los verbos de la perífrasis para preguntar por los complementos del verbo de contenido:
(a) *¿qué puedes mostrar?*
(b) **¿qué puedes?*
4. Se requiere un verbo “comodín” (e.g. *hacer*) para preguntar por la acción denotada por el verbo de contenido:
(a) *¿qué puedes hacer?*
(b) **¿qué puedes?*
5. El verbo de contenido puede presentarse en pasiva pero no el verbo auxiliar.
(a) *El catálogo puede ser mostrado.*
(b) **El catálogo es podido mostrar.*

La primera de estas pruebas verifica que los verbos auxiliares ocurren en perífrasis (pero hay perífrasis en las cuales no ocurren verbos auxiliares); la segunda garantiza que el verbo auxiliar carece siempre de agente; la tercera verifica que el complemento del verbo auxiliar es una frase verbal, a diferencia de los verbos transitivos que tienen un complemento nominal; la cuarta verifica el vínculo necesario que existe entre el verbo auxiliar y la frase verbal que lo complementa; la última refuerza la no agentividad del verbo auxiliar, ya que si fuera posible pasivizarlo sería agentivo. En estas construcciones, el sujeto del verbo de contenido es también el sujeto de la frase verbal, aunque sea anónimo (e.g. en *puedes mostrar* quien tiene la posibilidad de mostrar es quien muestra). Desde el punto de vista sintáctico los verbos que satisfacen estas propiedades se conocen en la literatura como de subida de sujeto o *subject-raising verbs* (Sag y Wasow, 1999); por lo mismo, aquí se postula que los verbos auxiliares del español son de este tipo sintáctico. De acuerdo con estas pruebas hemos encontrado que los siguientes verbos son auxiliares:

- (a) Con modificador en infinitivo: *poder* (*puedes comer*), *ir a* (*vas a comer*), *venir a* (*vienes a comer*), *volver a* (*vuelves a comer*), *haber de* (*habrás de comer*), *tener que*

(*tienes que comer*), *deber de* (*debes de comer*), *llegar a* (*llegas a comer*), *acabar de* (*acabas de comer*) y *alcanzar a* (*alcanzas a comer*).

- (b) Con modificador en gerundio: *Estar* (*estás comiendo*), *ir* (*vas comiendo*), *venir* (*vienes comiendo*), *seguir* (*sigues comiendo*) y *andar* (*andas comiendo*).
- (c) Con modificador en participio: *ser* (*es querido*), *estar* (*está abierta*) y *haber* (*ha comido*).

Por otro lado, verbos que no satisfacen las cinco condiciones sintácticas no son auxiliares a pesar de que puedan aparecer en perífrasis: *querer* (*quiere comer*), *pasar a* (*paso a comer*) y *haber que* (*hay que comer*).

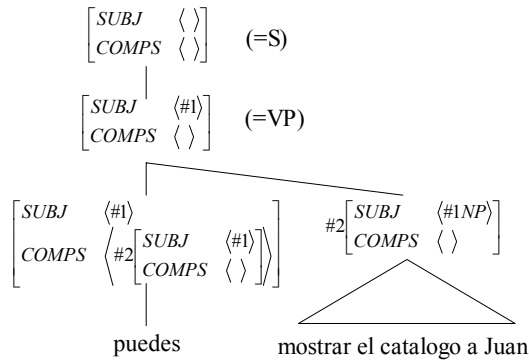
Pasamos ahora a presentar el modelo de verbos auxiliares en HPSG. Como ya se ha dicho, estos verbos son del tipo sintáctico *subject-raising*, y su descripción se presenta en la matriz de atributos y valores (*AVM*) en la figura 1, donde *srv-lxm* corresponde al tipo del verbo; el atributo *SS | LOC | CAT | VAL* se interpreta como la característica sintáctica y semántica del símbolo, con carácter local (en oposición a las dependencias no locales, como las que hay entre los pronombres en las preguntas pronominales y el argumento al que corresponden, por ejemplo), su categoría gramatical, y su capacidad de combinatoria, es decir la especificación de su sujeto y complementos. Lo crucial en esta definición es que el sujeto del verbo auxiliar y el sujeto de la frase verbal, están coindexados, es decir, son el mismo.

Figura 1. Definición léxica de verbos de subida de sujeto

$$\left[\begin{array}{c} \text{srv-lxm} \\ \\ \text{SS | LOC | CAT | VAL} \end{array} \left[\begin{array}{c} \text{SUBJ} \\ \text{COMPS} \end{array} \left\langle \begin{array}{c} \#1 \\ \text{HEAD | CLIT} \quad \text{false} \\ \text{VAL} \end{array} \right\rangle \left[\begin{array}{c} \text{SUBJ} \quad \#1 \\ \text{COMPS} \quad \langle \rangle \end{array} \right] \right] \right]$$

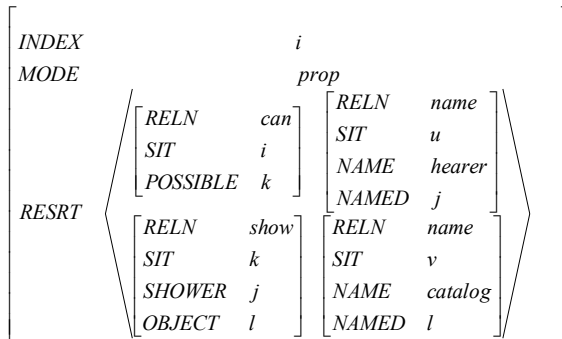
Pasamos ahora al análisis de oraciones con verbos auxiliares. En *puedes mostrar el catálogo a Juan* en la figura 2, *puedes* se combina con su complemento a través de la *Head-Complement Rule* de HPSG (Sag y Wasow, 1999). La frase verbal resultante es objeto de la regla *Subject-PRO-drop Rule* que se ha adicionado a HPSG para capturar los sujetos contenidos en la conjugación del verbo del español.

Figura 2. Oración con verbo auxiliar



La representación semántica resultante de este análisis se muestra en la figura 3. La notación corresponde a la llamada *Minimal Recursion Semantics (MRS)* (Semántica de Recursión Mínima) usada normalmente en HPSG.

Figura 3. Representación semántica no agentiva



En esta figura el atributo *index* se refiere a la situación expresada por la oración en su conjunto (i.e. la situación *i*); el atributo *MOD* indica el modo semántico de la oración, que en este caso es declarativa (*prop*), en oposición a las interrogativas o imperativas, por ejemplo; el atributo *RESTR* contiene una lista de *AVM* con la representación semántica de la oración propiamente; en este caso, que existe una situación *i* en la que una hay relación de poder (*can*) la cual tiene un argumento *k*, el índice de la relación que es posible; *k* se refiere a su vez a la situación de mostrar (*show*) cuyos parámetros son un agente *j* que muestra y un objeto mostrado *l*; a su vez *j* es quien escucha (situación *u*) y *l* es el catálogo (situación *v*).

En el análisis de este tipo de perífrasis es necesario considerar la ambigüedad léxica y estructural que surge cuando las formas léxicas que funcionan como auxiliares conservan su contenido semántico original, lo cual sucede frecuentemente. En este caso, dichos verbos son del tipo sintáctico conocido como *subject-control verb* o verbos con sujeto de control; este tipo es similar al tipo *subject-raising* ya que su sujeto y el de su complemento son también el mismo, sin embargo, los verbos con sujeto de control se realizan de manera agentiva. La descripción del tipo *subject-control verb* se muestra en la figura 4.

Figura 4. definición léxica de verbos de sujeto de control (versión 1)

$$\left[\begin{array}{c} scv-lxm \\ \\ SS|LOC|CAT|VAL \left[\begin{array}{c} SUBJ \quad \#1_i \\ COMPS \left\langle \begin{array}{c} HEAD|CLIT \quad false \\ VAL \quad \left[\begin{array}{c} SUBJ \quad \#1_i \\ COMPS \langle \rangle \end{array} \right] \end{array} \right\rangle \end{array} \right] \end{array} \right]$$

Con esta forma léxica se captura, por ejemplo, la interpretación de *poder* en su sentido de capacidad. La diferencia entre el análisis de *poder* como auxiliar y como verbo de contenido puede apreciarse comparando sus correspondientes representaciones semánticas en las figuras 3 y 5. En el primer caso, la relación de *poder* se establece entre una situación en la que no hay agente y la situación que es posible, mientras que en el segundo (figura 5) la relación de *poder* es una situación que se establece entre el agente que puede, que es también el que muestra, y la capacidad que es podida.

Figura 5. Representación semántica agentiva

$$\left[\begin{array}{c} INDEX \\ MODE \\ \\ RESRT \end{array} \left\langle \begin{array}{c} \left[\begin{array}{cc} RELN & can \\ SIT & i \\ CAPABLE & j \\ CAPACITY & k \\ RELN & show \\ SIT & k \\ SHOWER & j \\ OBJECT & l \end{array} \right] \left[\begin{array}{cc} RELN & name \\ SIT & u \\ NAME & hearer \\ NAMED & j \\ RELN & name \\ SIT & v \\ NAME & catalog \\ NAMED & l \end{array} \right] \end{array} \right\rangle \right]$$

Al igual que *poder*, varias formas que se usan como auxiliares pueden aparecer con su contenido semántico original, y en este caso son también verbos de sujeto de control. Además de estas formas, los llamados verbos modales como *querer*, *deber*, *osar*, *pensar*, *esperar*, *lograr* y *saber* ocurren en perífrasis como verbos de sujeto de control que requieren un agente que lleve a cabo la intención o acción enunciada por el verbo. Los requerimientos de subcategorización de estos últimos verbos son un problema muy complejo que queda fuera del alcance de este trabajo; sin embargo, podemos decir de manera general que hay tres casos principales: cuando se siguen de un complemento nominal y funcionan como simples transitivos (e.g. *quiere los juguetes*), cuando tienen como complemento una frase verbal infinitiva (e.g. *quiere comer*) y cuando el complemento es una frase relativa (e.g. *quiere que le den de comer*). En este trabajo definimos los verbos de control para los primeros dos casos, y dejamos el estudio de las relativas para investigaciones futuras. En particular, la definición de la Figura 4 corresponde al caso de los verbos de control que tienen como complemento una frase infinitiva.

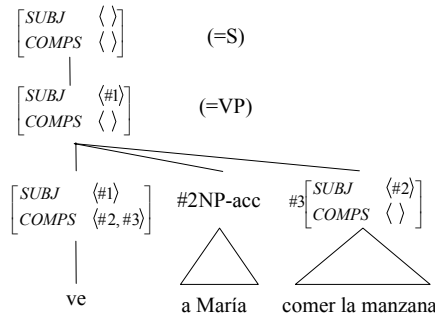
Además de los verbos auxiliares y modales que están en mayor o menor medida gramaticalizados, la perífrasis contiene cuando menos un verbo con contenido semántico, el cual va normalmente al final de la misma. Sin embargo, hay perífrasis complejas en las que hay dos verbos de contenido, los cuales van también normalmente al final de la perífrasis, seguidos de sus complementos nominales, aunque estos pueden ser compartidos y aparecer en diferentes órdenes. Estas construcciones se dan cuando ocurren los verbos de tipo *object-control* o de objeto de control que tienen dos complementos: una frase nominal (su objeto directo) y una frase verbal encabezada por el segundo verbo de contenido; adicionalmente, el objeto directo del verbo de objeto de control se comparte con el sujeto del segundo verbo. Para este tipo de perífrasis hemos identificado tres grupos principales de verbos: de los sentidos, de la voluntad y el factitivo *haber*.

En el primer caso se encuentran, por ejemplo, los verbos *ver* y *oir*. En *ve a María comer la manzana*, el objeto directo de *ver* es también el sujeto de *comer*; del mismo modo, en *oye cantar a los pájaros*, el objeto directo de *oye* es también el sujeto de *cantar*. En los verbos de voluntad, como *mandar* y *permitir*, la relación es similar; en *manda a los niños a dormir* el niño a quien se manda es también quien duerme, y en *permite jugar a los niños* quien permite no es quien juega, pero los niños son el objeto directo de *permite* y el sujeto de *jugar*. El caso de *hacer* es más complejo desde el punto de vista semántico, pero sintácticamente su comportamiento es similar; en *hace reír a la gente*, el objeto directo de *hacer* es *a la gente* quien es además el sujeto de *reír*, es decir, quien ríe. En un análisis alternativo enfocado semánticamente el verbo *hacer* tiene tan sólo un complemento, que es la frase verbal que le sigue y denota un evento o un proceso; es tal vez más intuitivo el análisis en el que el payaso hace a la gente reír, que el análisis en el que el payaso hace a la gente, que ríe. Sin embargo, en el presente trabajo adoptamos por uniformidad y simplicidad el enfoque sintáctico de los verbos de objeto de control y dejamos el análisis alternativo para investigaciones posteriores. La definición de los verbos de objeto de control se muestra en la figura 6 y el análisis de *ve a María comer la manzana* en la figura 7.

Figura 6. Definición léxica de los verbos de objeto de control (versión 1)

$$\left[\begin{array}{l} ocv-lxm \\ SS | LOC | CAT | VAL \\ \left[\begin{array}{l} SUBJ \\ COMPS \end{array} \left\langle \begin{array}{l} \#1 \\ \#2 \& NP-acc_i, \end{array} \right\rangle, \left[\begin{array}{l} SUBJ \\ COMPS \end{array} \left\langle \begin{array}{l} \#2_i \\ \langle \rangle \end{array} \right\rangle \right] \end{array} \right] \right]$$

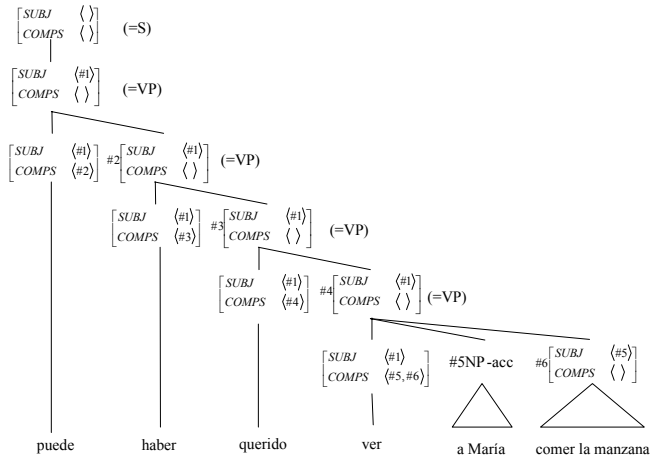
Figura 7. Oración con verbos de objeto de control



Los verbos de objeto de control pueden también ocurrir en perífrasis con verbos auxiliares y modales; en este caso el verbo de control va necesariamente en forma no personal, ya que el verbo conjugado es siempre el que va al frente en la perífrasis. Un ejemplo de este tipo de oración es *puede haber querido ver a María comer la manzana*. En este caso la perífrasis contiene cinco verbos: dos auxiliares, un modal y dos verbos de contenido semántico. Sin embargo, con las definiciones de los verbos de subida de sujeto, de sujeto de control y de objeto de control, el análisis de este tipo de oraciones es simple y directo, como se muestra en la figura 8.

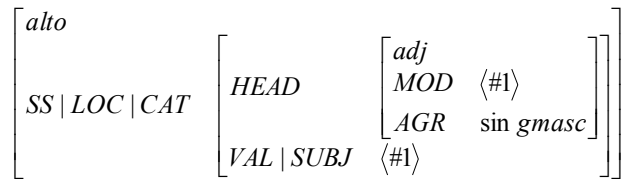
Concluimos el análisis de la perífrasis española con los copulativos *ser* y *estar* en las oraciones atributivas y pasivas. Estos verbos son también del tipo *subject-raising* pero a diferencia de las oraciones con predicación verbal, en las atributivas lo que se comparte es el sujeto que posee la predicación nominal (i.e. la propiedad adjetiva o nominal especificada como el atributo) con el sujeto de la oración. Como uno de sus sellos distintivos el español tiene dos auxiliares para esta función: *ser* y *estar*. Intuitivamente, *ser* se utiliza para hacer atribuciones permanentes (*el cielo es infinito* pero no *el cielo está infinito*) mientras que *estar* se utiliza para hacer atribuciones transitorias (*La fruta está madura* pero no *la fruta es madura*). Sin embargo, esta distinción no es suficiente: tanto *el niño es alto* como *el niño está alto* son oraciones gramaticales e idiomáticas, pero su significado es diferente. Más aún, no se puede decir *Juan es muerto*, a pesar de que nada puede ser más permanente que dicha condición, y se tiene que decir *Juan está muerto*.

Figura 8. Análisis de la oración con perífrasis compleja



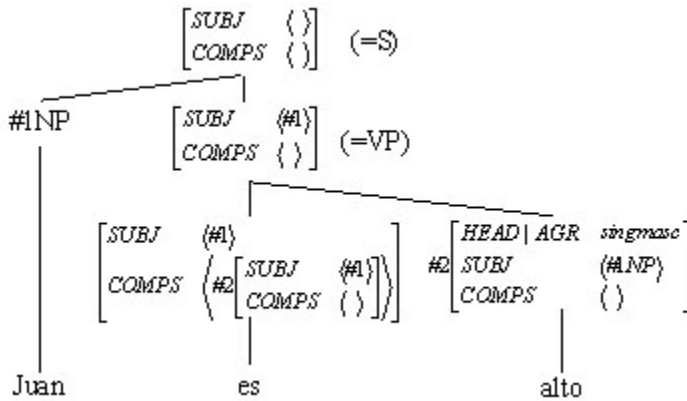
El uso de estos verbos se aclara cuando se toma en cuenta su carácter aspectual; mientras que *ser* es un verbo imperfectivo, *estar* es perfectivo, y estos auxiliares comunican este aspecto a sus atributos. Si los atributos que son permanentes de sus sujetos son compatibles con el carácter imperfectivo del verbo *ser*, oraciones como *el cielo es infinito* son aceptables, pero si el adjetivo representa una propiedad que el sujeto llega a tener como resultado de un proceso, su significado no es compatible con el sentido imperfectivo de *ser* y oraciones como *Juan es muerto* no lo son; de manera análoga el sentido de la propiedad que se tiene o que se llega a tener debe ser compatible con el sentido perfectivo de *estar*, por lo que *Juan está muerto* es aceptable pero *el cielo está infinito* no lo es.

Presentamos ahora el modelo de las atributivas en HPSG. Como ya se ha dicho *ser* y *estar* son del tipo *srv-lxm*. Por su parte, la entrada léxica del atributo marca que éste modifica a su propio sujeto, como se muestra en la entrada léxica de *alto* en la figura 9 donde el sujeto del adjetivo (i.e. quien tiene la propiedad de ser alto) está coindexado con la frase nominal que dicho adjetivo modifica, es decir con el valor del atributo MOD del HEAD de esta entrada léxica.

Figura 9. Entrada léxica para *alto*

El análisis de la oración atributiva procede combinando el auxiliar con el atributo mediante la *HCR* para formar la predicación nominal, y mediante la *HSR* para formar la oración completa, como se muestra en la figura 10.

Figura 10. Estructura para oración atributiva



Los análisis sintácticos de las atributivas con *estar* son similares y la diferencia de significados se representa en la semántica. Para este efecto, si el atributo representa una propiedad que tiene que ser permanente necesariamente (“imperfectiva”) así se marca en la semántica de la entrada léxica del adjetivo; del mismo modo, si es una propiedad que necesariamente se llega a tener después de un devenir (“perfectiva”) así se marca también en la entrada léxica correspondiente. Sin embargo, si la propiedad puede tomar ambos valores (*alto, guapa*), su valor aspectual queda indeterminado en el lexicon y el valor de este atributo lo determina el copulativo en el proceso de combinación. La semántica de nuestro ejemplo se ilustra en la figura 11.

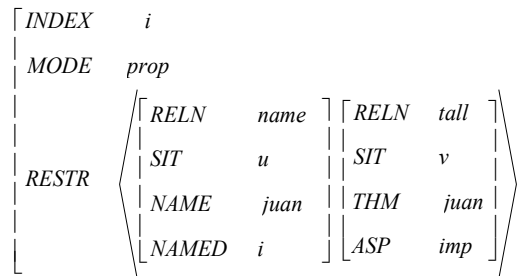


Figura 11. Semántica de atributiva con *ser*

Cuando los auxiliares *ser* y *estar* aparecen en pasivas, comunican a sus predicaciones su carácter aspectual de la misma manera que en las atributivas; por lo mismo, la pasiva con *estar* se usa con verbos perfectivos mientras que la pasiva con *ser* requiere verbos imperfectivos: se puede decir *María es querida* pero no *María está querida*; asimismo, se puede decir *la puerta está abierta*, pero no *la puerta es abierta* a menos que el significado de esta última oración se complete con un modificador que complete el sentido de un proceso reiterativo como en *la puerta es abierta todas las mañanas*. Con esta observación concluimos el estudio de los verbos auxiliares.

3. El sistema de pronombres clíticos

Las expresiones perifrásticas mostradas en la sección anterior contienen frecuentemente pronombres clíticos en lugar de los argumentos explícitos del verbo de contenido. Por ejemplo *puedes mostrar el catálogo a Juan* puede realizarse de manera mucho más directa como *se lo puedes mostrar* o *puedes mostrárselo*. Más aún, este tipo de pronombres ocurre también con mucha frecuencia en construcciones en las que no intervienen verbos auxiliares. En general, las partículas clíticas han sido objeto de una gran cantidad de estudios lingüísticos (e.g. véase Nevis et. al, 1994); sin embargo, a pesar de su importancia y la atención recibida, el nivel de representación de los pronombres clíticos es sujeto de controversia. Por un lado hay evidencia de que son fenómenos de carácter estrictamente morfológico, mientras que por otro se arguye que los pronombres clíticos son objetos léxicos independientes y por lo mismo pertenecen al nivel de representación sintáctico.

Aquí se presenta un modelo en HPSG que toma en cuenta ambas tendencias y se propone que las secuencias de pronombres clíticos constituyen unidades morfológicas indisolubles, independientemente de que aparezcan en forma proclítica o enclítica. Sin embargo, estos grupos pueden ser sujetos de una operación sintáctica, y pertenecen también a este último nivel de representación gramatical. El modelo permite analizar oraciones incluyendo pronombres clíticos atómicos o en grupos y en forma enclítica y proclítica. Los elementos esenciales del modelo son:

- Las entradas léxicas de los verbos que pueden recibir pronombres clíticos como argumentos.
- Las entradas léxicas de las *palabras clíticas* que corresponden a las secuencias posibles de pronombres clíticos en el lenguaje.
- Las reglas léxicas que modifican la valencia de los verbos en el proceso de clitización.
- Las reglas léxicas que realizan a los pronombres cuando estos se presentan de forma enclítica.
- Una regla gramatical para combinar una secuencia de uno o más pronombres clíticos con la frase verbal cuando estos se presentan de forma proclítica.

A continuación se muestran los elementos del modelo mediante el análisis de las formas clíticas de *muestras el catálogo a Juan: se lo muestras y muéstraselo*. La entrada léxica de *muestras* tiene su sujeto y dos complementos como se muestra en la figura 12.

Figura 12. Entrada léxica de *muestras*

$$\left[\begin{array}{l} \textit{muestras} \\ SS | LOC | CAT \end{array} \left[\begin{array}{l} HEAD \quad \textit{verb} \\ VAL \quad \left[\begin{array}{l} SUBJ \quad \langle NP \rangle \\ COMPS \quad \langle NP, NP \rangle \end{array} \right] \end{array} \right] \right]$$

En esta forma básica la palabra *muestras* debe ser seguida por sus complementos acusativo y dativo (qué y a quién); sin embargo, cuando ocurre en construcciones con pronombres proclíticos la forma en la figura 12 se modifica como se indica en la figura 13.

Figura 13. Forma clitzada de *muestras*

$$\left[\begin{array}{l} \textit{muestras} \\ SS | LOC | CAT \end{array} \left[\begin{array}{l} HEAD \quad \textit{verb} \\ VAL \quad \left[\begin{array}{l} SUBJ \quad \langle NP \rangle \\ COMPS \quad \langle \rangle \\ CL-LIST \quad \langle NP, NP \rangle \end{array} \right] \end{array} \right] \right]$$

Esta transformación se logra mediante una regla léxica llamada de *clitización* cuyo efecto es consumir a los complementos del verbo en el rasgo *COMPS* e incluirlos en lista de clíticos en el rasgo *CL-LIST*; de este modo el verbo se transforma en una frase verbal que pospone la combinación con sus argumentos. Por otro lado, la secuencia de pronombres “*se lo*” se considera una unidad morfológica indisoluble y se combina con el verbo en una sola operación mediante la regla de *Head Proclitic Rule (HPR)*. En la figura 13, el atributo *CL-LIST* contiene los argumentos del verbo clitzado en el orden requerido para su combinación mediante la regla de combinación proclítica. La entrada léxica de la secuencia de clíticos se muestra en la figura 14 y la regla *HPR* en la figura 15; el análisis sintáctico de nuestra oración de ejemplo se muestra en la figura 16. El último paso del análisis para la combinación del sujeto con la frase verbal en la figura 16 se logra mediante la aplicación de la regla *Head-PRO Drop Rule*, como ya se ha mencionado.

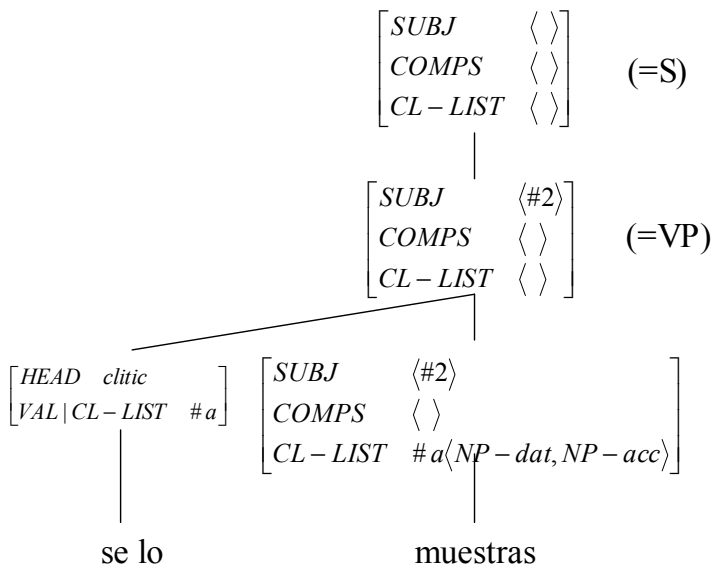
Figura 14. Entrada léxica de compuesto clítico

$$\left[\begin{array}{l} HEAD | CLIT \quad true \\ VAL | CL-LIST \quad \langle \rangle \end{array} \right], \left[\begin{array}{l} HEAD \quad clitic \\ VAL | CL-LIST \quad \#1 \end{array} \right] H \left[\begin{array}{l} VAL \left[\begin{array}{l} SUBJ \quad \langle I \rangle \\ COMPS \quad \langle \rangle \\ CL-LIST \quad \#1 \end{array} \right] \end{array} \right]$$

Figura 15. Head Proclitic Rule

$$\left[\begin{array}{l} se \quad lo \\ SS | LOC \left[\begin{array}{l} CAT \left[\begin{array}{l} HEAD \quad clitic \\ VAL | CL-LIST \quad \langle NP-dat_i, NP-acc_j \rangle \end{array} \right] \\ CONT | RESTR \quad \{ [individual-rel]_i, [male-rel]_j \} \end{array} \right] \end{array} \right]$$

Figura 16. Estructura de oración proclítica



Pasamos ahora a mostrar el análisis para la forma enclítica *muéstramelo*. En este caso el verbo de contenido está en forma no personal (i.e., imperativo) y los pronombres *se* y *lo* son una sola forma léxica. La entrada básica del verbo *mostrar* se modifica mediante una regla léxica para consumir sus dos argumentos y formar una frase verbal saturada en una sola operación. En la figura 17 se muestra dicha regla para clitizar el verbo *mostrar* con dos argumentos enclíticos. La forma resultante se puede combinar con el sujeto, en este caso omitido, de manera similar a la construcción en la figura 16.

Figura 17. Regla de clitización enclítica

$$\left[\begin{array}{ll} \textit{mostrar} & \\ \textit{ORTH} & \#1 \\ \textit{VAL} \mid \textit{CL} - \textit{LIST} & \langle \rangle \\ \textit{ARG} - \textit{ST} & \langle \#2, [], [] \rangle \end{array} \right] \mapsto \left[\begin{array}{ll} \textit{ORTH} & \textit{muestramelo} \\ \textit{VAL} \mid \textit{CL} - \textit{LIST} & \langle \rangle \\ \textit{ARG} - \textit{ST} & \langle \#2 \rangle \end{array} \right]$$

La semántica resultante del análisis de la oración enclítica se muestra en la figura 18.

Figura 18. Semántica de frase verbal enclítica

$$\left[\begin{array}{l} \textit{RESTR} \\ \left\langle \begin{array}{l} \left[\begin{array}{ll} \textit{RELN} & \textit{show} \\ \textit{WHO} & i \\ \textit{WHAT} & j \\ \textit{TO_WHOM} & k \end{array} \right] \\ \left[\begin{array}{ll} \textit{RELN} & \textit{third-person} \\ \textit{SIT} & k \end{array} \right] \end{array} \right\rangle \left[\begin{array}{ll} \textit{RELN} & \textit{object-mas} \\ \textit{SIT} & j \end{array} \right] \end{array} \right]$$

Una característica sumamente distintiva de los sistemas de clíticos en diferentes lenguas es el orden estricto pero idiosincrásico que se da entre los pronombres; en español el orden es: *se, te, me, le, lo, la*, donde *lo, la* y *le* nunca aparecen juntos. Por ejemplo *se lo das*, pero no *lo se das* o *me lo dices* pero no *lo me dices*, etc. Adicionalmente este orden se conserva independientemente de que los pronombres se presenten de manera pro o enclítica; por ejemplo: *dáselo* pero no *dalose*, *dímelo* pero no *dilome*. Asimismo, en español todos los pronombres son proclíticos o enclíticos: *se lo das* o *dáselo* pero no *se daslo*. El orden en el caso proclítico se captura en la definición de la entrada léxica de la palabra clítica como se muestra en la figura 14; los diferentes ordenes se capturan definiendo una palabra clítica para cada orden posible. Por su parte, la regla de combinación proclítica contiene los argumentos clitizados en el mismo orden, ya que la palabra clítica y la lista de argumentos de la frase verbal están coindexados, como se puede

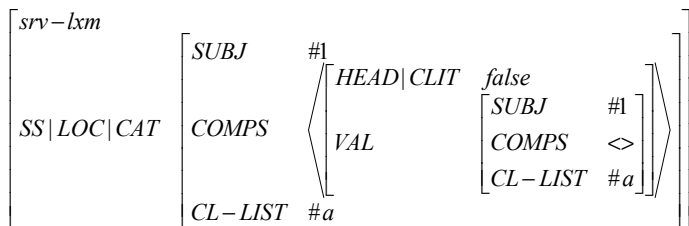
apreciar en la figura 15. En el caso enclítico, por su parte, el orden se captura de forma explícita en la regla de cliticación en la figura 17, la cual asigna el *me* dativo y el *lo* acusativo en la semántica correspondiente, como se muestra en la figura 18. En general, las palabras léxicas con los diferentes órdenes posibles corresponden a las formas posibles de realización enclítica, por lo que el diccionario clítico es común, en última instancia, para la realización tanto de enclíticos como proclíticos.

Concluyendo, el presente análisis postula la existencia de un conjunto de secuencias de pronombres clíticos que se consideran como unidades léxicas indisoluble: “*se*”, “*te*”, “*me*”, “*le*”, “*lo*”, “*la*”, “*se te*”, “*se me*”, “*se lo*”, “*se la*”, “*te me*”, “*te lo*”, “*te la*”, “*me le*”, “*me lo*”, “*me la*”, etc., un conjunto de reglas léxicas para cliticar los verbos y para formar enclíticos y una regla de combinación sintáctica para formar complejos clíticos. A pesar de que esta lista de palabras clíticas es parcial y no incluye los órdenes de clíticos posibles, se puede apreciar que el número de estas construcciones es finito y bajo; para el español de México, por ejemplo, el número de combinaciones incluyendo los diferentes casos, número, género y persona posibles, no pasan de 60. Por lo mismo, este mecanismo permite modelar con un costo computacional aceptable, centrado en el lexicón, las complejidades del fenómeno de clíticos debidas a su orden rígido e idiosincrático.

4. Interacción entre sistemas

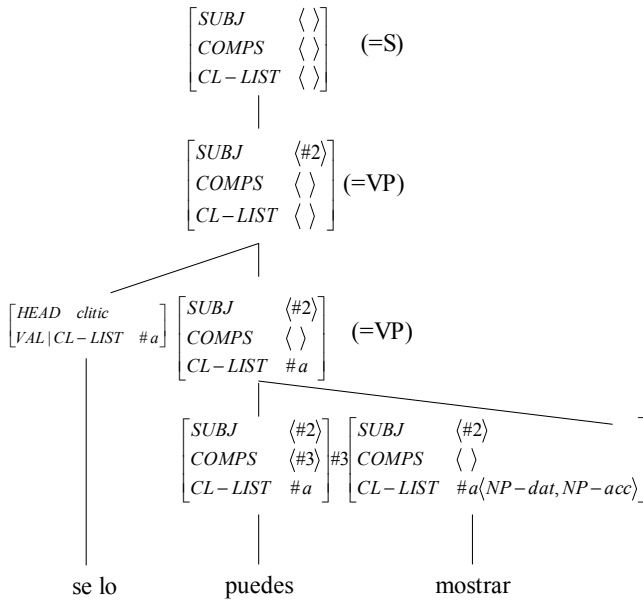
Los sistemas de verbos auxiliares y pronombres clíticos no son independientes; en español es posible que exista un verbo auxiliar o una frase verbal entre dichos pronombres y el verbo de contenido de la oración. Este fenómeno se conoce como *clitic climbing* o *subida de clíticos* y ha sido sujeto a numerosos estudios (e.g. Monachesi 1999, Miller y Sag, 1995). Para modelar este tipo de perífrasis es necesario extender la definición de los verbos que toman como complemento una frase verbal, ya sean estos *subject-raising* o *subject-control*, permitiendo que no sólo el sujeto del auxiliar y su complemento se compartan, sino también que la lista de clíticos del auxiliar se comparta con la de su complemento, como se muestra en la figura 19, que actualiza a la definición de la figura 1. La definición de los verbos de sujeto de control en la figura 4 se actualiza de manera similar.

Figura 19. Entrada léxica de verbos de subida de sujeto (versión 2)



Esta definición permite que la lista de clíticos pase hacia “arriba” en el proceso de clitic climbing. En el proceso de combinación sintáctica y al igual que el resto de los atributos de valencia (i.e. el sujeto y los complementos verbales) el valor de estos atributos del nodo madre de la estructura sintáctica es el mismo que el de la hija “cabeza”, a menos que la regla de combiación sintáctica defina explícitamente el valor de este atributo (i.e. el principio de valencia en HPSG). Para el caso de verbos no clítizados la lista de clíticos es simplemente vacía, por lo que el mecanismo funciona de manera general. A continuación ilustramos el modelo con el análisis de la oración *se lo puedes mostrar*, donde el auxiliar *puedes* toma como complemento la forma clítizada de *mostrar*. Esta combinación se logra mediante la aplicación de la *HCR* y la estructura resultante se combina con la forma clítica “*se lo*” mediante la aplicación de la *HPR*; finalmente el análisis de la oración completa se logra mediante el uso de *Subject-PRO Drop Rule* al igual que en los ejemplos anteriores. El árbol sintáctico se muestra en la figura 20.

Figura 20. Oración con subida de clíticos



Con los elementos dados hasta este momento es posible analizar estructuras sintácticas con subida de clíticos en perífrasis con verbos de subida de sujeto y sujeto de control; sin embargo, este modelo básico no contempla las construcciones con los llamados verbos de objeto de control como *ver*, *mandar*, *hacer*, etc., que tienen como complementos una frase nominal (i.e. su objeto directo) y una frase verbal infinitiva,

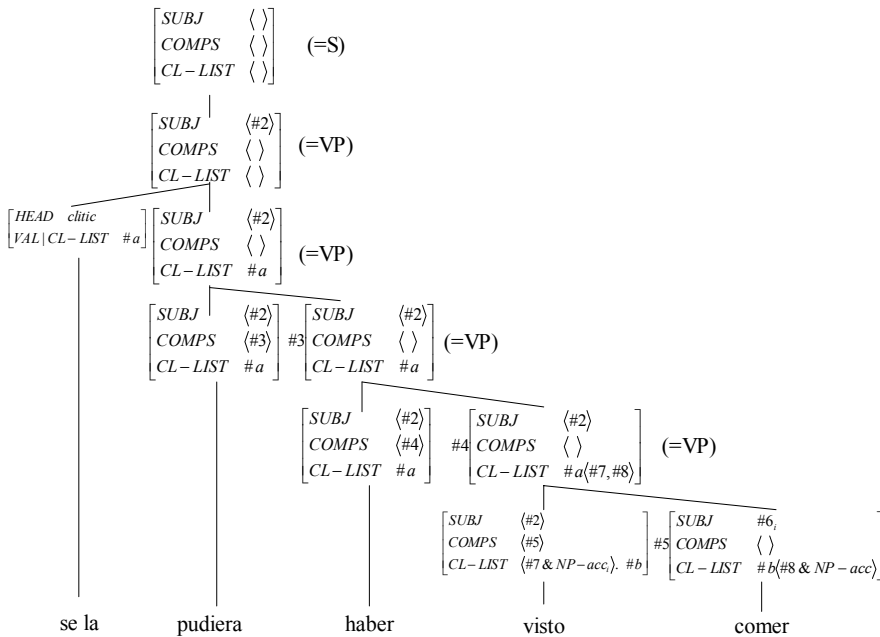
como *ver* en *puede ver a Juan comer la manzana*, *puede ver a Juan comerse la manzana* y su forma clítica *se la puede ver comer*. Más aún, esta última forma es ambigua ya que si corresponde a una cliticación de la primera o la segunda oración, el pronombre *se* es acusativo o dativo respectivamente, y el significado de la oración clítica es diferente. Esto se puede apreciar si tomamos en cuenta las formas clíticas *puede verlo comerla* y *puede verlo comérsela*, donde *lo* y *la* son acusativos ya que substituyen a los complementos directos de los verbos respectivos, mientras que *se* es un pronombre dativo que reemplaza a un segundo complemento opcional del verbo *comer*, e indica que su referente es el beneficiario de la acción de comer (i.e. dativo ético).

Analizamos primero el caso en que *se la puede ver comer* es una forma clítica de *puede verlo comerla*. En este caso los dos pronombres se frontalizan formando la secuencia “*se la*” al tiempo que *se* reemplaza a *lo*, ya que la secuencia “*lo la*”, que daría origen a la forma no gramatical *lo la puede ver comer*, no puede darse en español; en este caso el pronombre *se*, en acusativo, es una repetición del *le* personal (derivado del *illi* Latino), dándose simultáneamente un caso de leísmo, en el que la forma *le* y su duplicación *se*, normalmente dativos, se realizan como acusativos en vez de la forma normal *lo*. Para capturar este fenómeno definimos la operación de *composición de clíticos*. Esta oración toma dos cliticaciones de dos verbos diferentes y las compone en una sólo cliticación que se cancela con una secuencia de pronombres clíticos. En nuestro ejemplo, las cliticaciones de los complementos directos de *ver* y *comer* se combinan en una sola cliticación abstracta (i.e. de la frase verbal *ver comer*) la cual se cancela con la palabra clítica “*se la*”, con ambos pronombres en acusativo, en una sola operación. La operación de composición se define en la entrada léxica de los verbos de objeto de control, donde el atributo de la lista de clíticos definido en la valencia de estos verbos es la suma de la cliticación de su objeto directo, en caso que éste se encuentre cliticado, y los clíticos de su segundo complemento. Esta definición se muestra en la figura 21 y corresponde a la actualización la definición en la figura 6. Con estas definiciones podemos analizar la oración de ejemplo, como se muestra en la figura 22.

Figura 21. Definición de verbos de objeto de control (versión 2)

$$\left[\begin{array}{l} ocv - lxm \\ \\ SS | LOC | CAT | VAL \end{array} \left[\begin{array}{l} SUBJ \quad \#1 \\ COMPS \quad \left\langle \begin{array}{l} \#2i, \left[\begin{array}{l} SUBJ \quad \#2i \\ COMPS \quad \langle \rangle \\ CL - LIST \quad \#a \end{array} \right] \end{array} \right\rangle \\ CL - LIST \quad \langle \dots \rangle. \#a \end{array} \right] \right]$$

Figura 22. Oración con composición de clíticos



Pasamos ahora al análisis de la segunda interpretación de la oración en que *se la puede ver comer* es una forma clítica de *puede ver a Juan comersela la manzana*; como ya se ha dicho, en esta oración *se* es un dativo ético que indica quien es el beneficiario de la acción de comer. La estructura de la oración puede ser mejor apreciada si se toman en cuenta las formas clíticas intermedias *puede ver a Juan comersela* y *puede verlo comersela*. En particular, si no se clitiza el complemento directo de *ver*, no puede haber ambigüedad respecto a cuál es el argumento reemplazado por *se*. Adicionalmente, la referencia del dativo ético, es decir Juan, es no sólo el objeto directo de *ver* (i.e. es a quien puede ver) sino también el sujeto de comer (es quien come), por lo que tres argumentos de la oración tienen al mismo referente y están coindexados.

Desde nuestro punto de vista, resulta sumamente interesante el hecho de que la forma *puede verlo comersela* es equivalente en significado a *se la puede ver comer* a pesar de que en ésta última no se realiza el pronombre *lo*. Más aún, notamos que en esta última forma sólo se realiza una instancia del pronombre *se* que representa a los tres argumentos correferenciales. En este caso la secuencia "*se lo*" en la forma proclítica de la oración reemplaza a la secuencia "*lo + sela*" con la ecuación clítica *se = lo + se* que se permite si ambos pronombres están coindexados. Para modelar este fenómeno

definimos la operación de subsumción, que permite combinar dos cliticaciones contiguas subsumiendo una en la otra si los pronombres correspondientes están coindexados. En nuestro ejemplo, el pronombre *lo* se subsume en el pronombre *se*, prevaleciendo el caso dativo sobre el acusativo (i.e. el *lo* acusativo queda subsumido dentro del *se* dativo). Finalmente, la palabra clítica resultante de la operación de subsumción se cancela con la realización explícita de los pronombres al frente de la perífrasis. La operación de subsumción clítica se define mediante una regla léxica adicional que modifica la valencia del verbo de objeto de control; esta regla se muestra en la figura 23.

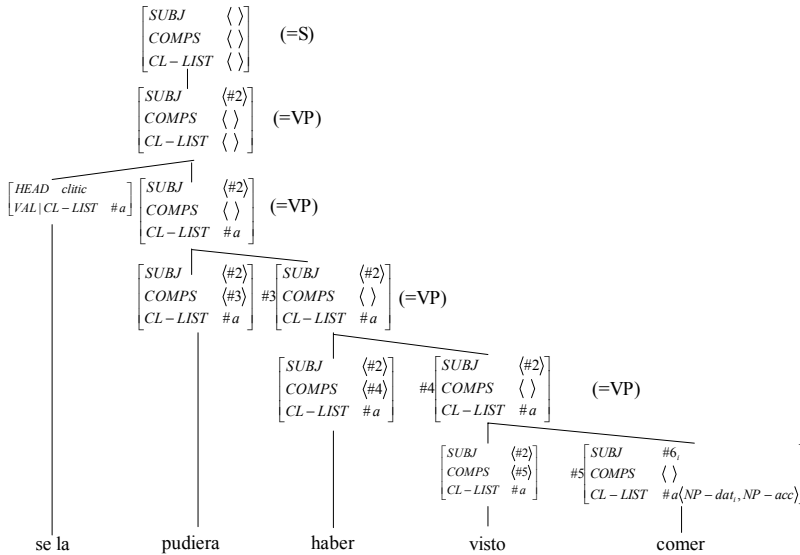
Figura 23. Regla léxica para la operación de subsumción

$$\left[\begin{array}{ll} ocv-lxm \\ ORTH & \#1 \\ HEAD & \#2 \\ ARG-ST & \langle \#3, [_i], \#4 \rangle \end{array} \right] \mapsto \left[\begin{array}{ll} ORTH & \#1 \\ HEAD & \#2 \\ VAL|CL-LIST & \#a \\ ARG-ST & \left\langle \#3_i, \#4 \& \left[\begin{array}{ll} SUBJ & \langle [_i] \rangle \\ CL-LIST & \#a \& \langle [_i, \dots] \rangle \end{array} \right] \right\rangle \end{array} \right]$$

Esta regla especifica que si el complemento directo del verbo de objeto de control está cliticado, y as su vez está coindexado con el primer elemento de la lista de clíticos del segundo verbo, la lista de clíticos de la frase verbal formada por los dos verbos (i.e. *ver comer*) es simplemente la lista de clíticos del segundo verbo; esta regla define también que el argumento subsumido está coindexado con el sujeto y el segundo complemento del verbo de objeto de control. Con esta definición es posible analizar a la segunda interpretación de *se la puede ver comer*, como se muestra en la figura 24.

Una alternativa para analizar la oración sería utilizar la operación de composición de clíticos definida arriba; sin embargo, dicha composición produciría en este caso la palabra clítica “*lo se la*” que no ocurre en español. En general, de acuerdo con nuestro análisis, las operaciones de composición y subsumción de clíticos tienen motivaciones independientes, por lo que ambas son necesarias para completar el análisis de la perífrasis compleja con pronombres clíticos. Con esta observación concluimos nuestro análisis de la perífrasis española y su interacción con el sistema de pronombres clíticos. Un análisis más elaborado de estos fenómenos se presenta en (Pineda y Meza, 2003).

Figura 24. Oración con subsumión de clíticos.



5. Implementación y verificación

Ambos sistemas así como su interacción han sido verificados empíricamente con expresiones del Corpus DIME (Villaseñor et. al., 2001). Asimismo, el modelo ha sido implementado en el ambiente de desarrollo de HPSG, Linguistic Knowledge Building (LKB) (Copestake 2002), partiendo para este efecto de la gramática del español desarrollada por Quirinos (2001), que corresponde en términos generales a la gramática básica del inglés de Sag y Wasow (1999).

El corpus DIME está compuesto por 6121 elocuciones, de las cuales aproximadamente un 55% no admiten un análisis gramatical exhaustivo (e.g. interjecciones, actos del habla expresados en una palabra o frase hecha, oraciones elípticas, etc.). De las expresiones parseables un total de 748 incluyen clíticos, de las cuales 688 incluyen clíticos propiamente, y el resto corresponde a otro tipo de oraciones con *se* (e.g. pasiva refleja, impersonal activa, reflexivas, reciprocas, pseudo reflejas, etc.). Del total de las oraciones que contienen clíticos propiamente, 677 contienen un solo pronombre clítico y tan sólo 11 llevan dos. No hay ningún caso de clíticos con verbos de control o auxiliares clitizados, y todos los casos que se presentan en el corpus reciben un análisis adecuado con nuestro modelo básico. Como dato singular en todo el corpus no se presenta ninguna oración en voz pasiva, y siempre que se requiere este recurso expresivo se usa la forma pasiva refleja (39 oraciones). El análisis de el resto de las oraciones con *se* se presenta en (Pineda y Meza, 2003).

6. Conclusiones

En este artículo se presenta un modelo computacional de la perífrasis y los pronombres clíticos del español. Estos sistemas, en conjunto, constituyen un núcleo sintáctico esencial para el procesamiento computacional de esta lengua. En particular se destacan los siguientes resultados: Los verbos auxiliares del español son del tipo de subida de sujeto, y el español es rico en estas formas; sin embargo, los verbos auxiliares aparecen también con su sentido original por lo que surgen numerosas ambigüedades que deben ser resueltas. En particular, en su forma agentiva estos verbos son del tipo de sujeto de control. Nuestro modelo provee un análisis adecuado para estos fenómenos. Adicionalmente, en perífrasis complejas aparecen los verbos llamados de objeto de control. Con relación al sistema de pronombres clíticos, se presenta una solución innovativa en la que se adopta una posición intermedia en la controversia acerca del nivel de representación de este fenómeno. Por un lado se considera a las secuencias de uno o más pronombres clíticos como estructuras léxicas indisolubles que se representan en la morfología; por el otro, el modelo aquí propuesto utiliza una regla sintáctica: la *Head-Proclitic Rule*, por lo que el sistema de pronombres clíticos también forma parte de este nivel de representación gramatical. Sin embargo, la regla de combinación proclítica requiere que la palabra clítica preceda de manera inmediata a la frase verbal por lo que el resultado de esta combinación se presenta al observador externo como una unidad morfológica indisoluble, atendiendo al criterio que considera al fenómeno como morfológico: aunque oculta a la información de entrada y salida del proceso, es esta regla sintáctica la que da cuenta de la estructura de la oración proclítica. Se destaca también el uso de las operaciones de composición y subsumisión de clíticos que permiten combinar argumentos clíticos de verbos diferentes en un sólo cliticación de carácter abstracto, que corresponde a la cliticación de una frase verbal, y que se combina con una secuencia del lexicon clítico. La definición de estas operaciones permite analizar la perífrasis compleja con pronombres clíticos. Finalmente, el modelo es simple y poco costoso en recursos computacionales como lo muestra nuestra propia implementación en LKB, y de amplia cobertura, como lo atestigua su validación con el Corpus DIME.

Referencias

- COPESTAKE, A. (2002) «The LKB system», Stanford University: <http://www-csli.stanford.edu/symbol/~aac/lkb.html>.
- GILI GAYA, S. (1991) *Curso superior de sintaxis española*, Barcelona: Biblograf: 103-153.
- MILLER P. H. And SAG, I. A. (1995) «French clitic movement without clitics or movement», Stanford: CSLI Technical Report.
- MONACHESI, P. (1999) *A lexical approach to Italian cliticization*. Lecture Notes Series No. 84. Stanford: CSLI-Cambridge University Press.

- NEVIS, J. A., JOSEPH, B. D., WANNER, D. and ZWICKY, A. M. (1994) *Clitics, A comprehensive bibliography 1892-1991*. Library and information sources in linguistics, 22. Amsterdam/ Philadelphia: John Benjamin's Pub.
- PINEDA, L. A., MASSÉ, A., MEZA, I., SALAS, M., SCHWARZ, E., URAGA, E and VILLASEÑOR, L. (2002) «The Dime project». In *Proceedings of MICAI2002*. Lectures notes in artificial intelligence 2313: 166-175.
- PINEDA, L. A. & MEZA, I. (2003) «The Spanish pronominal clitic system». Depto. de Ciencias de la Computación, IIMAS, UNAM
- POLLARD, C. & SAG, I. (1994) *Head-driven phrase structure grammar*. Chicago & London: Stanford CSLI-University of Chicago Press.
- SAG I. & WASOW T. (1999) *Syntactic theory: a formal introduction*. Stanford: CSLI Publications.
- QUIRINO, A. P. (2001) *Spanish clitics: a computational model*, Master's Thesis. Bielefeld: Universität Bielefeld: <<http://coral.lili.uni-bielefeld.de/~aquirino/thesis.html>>
- VILLASEÑOR, L., MASSÉ, A. & PINEDA, L. A. (2001). «The DIME Corpus». En C. Zozaya, M. Mejía, P. Noriega y A. Sánchez (eds.). *Memorias ENC01*, Tomo II. SMCC, Aguascalientes, Ags. México.

Agradecimientos

Se agradece el apoyo de CONACyT Proyecto C092A.