De Reglas Léxicas a Marcos de Subcategorización Complejos en la Jerarquía de Tipos

Jordi Porta & Marta Villegas
GILCUB
{jordi, tona}@gilcub.es

Mayo, 1996

Abstract

In HPSG the grammar consist of a type hierarchy and a set of principles. In fact, principles of the grammar (including ID Schemata) are constraints over feature structures and can easily be expressed as so. With this, HPSG manages to model all linguistic knowledge in terms of typed feature structures without resorting to principles or rules. Inheritance and Lexical Rules (LR) allow to eliminate redundancy. Broadly speaking, inheritance avoids 'vertical' redundancy and LR avoid 'horizontal' redundancy. LRs, however, are not part of the formalism, they are not typed and fall outside the linguistic taxonomy.

In this paper we investigate ways of expressing the generalization power of LRs without using LRs. We integrate LRs as feature structures using disjunction and simulating closure operator by means of the inheritance mechanism of the hierarchy itself. We can express LRs generalizations using new types with extended feature structures and modified inheritance mechanisms allowing for disjunctive inheritance. This allow to homogenize linguistic knowledge representation.

HPSG modela el conocimiento lingüístico por medio de estructuras de rasgos tipificadas. Estas estructuras de rasgos están en correspondencia biunívoca con los diferentes tipos de expresiones del lenguaje natural. Se entiende, entonces, que las teorías lingüísticas deben establecer qué estructuras de rasgos son admisibles. Los tipos de entidades lingüísticas correspondientes a estas estructuras de rasgos contituyen, entonces, las predicciones de la teoría.

De algún modo, las estructuras de rasgos son estructuras de datos que especifican valores para los atributos. El poder expresivo de estas estructuras es inmenso gracias a la recursividad (el valor de un atributo puede ser otra estructura) y al *structure-sharing* (una misma estructura puede aparecer como valor de diferentes atributos dentro de una misma super-estructura).

Las estructuras de rasgos deben estar tipificadas. Así, las estructuras de rasgos pertenecen a diferentes tipos de acuerdo con el tipo de objeto lingüístico que describen. Cada tipo tendrá un conjunto diferente de atributos. El tipo de una estructura de rasgos nos dice qué tipo de objeto lingüistico está describiendo[1].

[1] Para una caracterización formal de las estructuras de rasgos ver Carpenter 1992.

Una característica importante de las estructuras de rasgos es que pueden tomarse como parcialmente ordenadas de acuerdo con su grado de información: una estructura es subsumida por otra cuando contiene como mínimo tanta información como la segunda. El conjunto de tipos estará entonces parcialmente ordenado. Así, el tipo *sign* se ordena debajo de los tipos *word* y *phrase* ya que contiene a ambos.

A grandes rasgos la gramática de HPSG consiste en una jerarquía de *tipos* y un conjunto de principios. La jerarquía de tipos se representa como una taxonomía en forma de árbol con la raíz tipificada como *sign*. En realidad, los principios de la gramática son constricciones sobre las estructuras de rasgos. Así, por ejemplo, el Head Feature Principle puede fácilmente expresarse como:

$$\begin{bmatrix} \textit{phrase} \\ \text{synsem} \mid \text{local} \mid \text{cat} \mid \text{head} \ \boxed{1} \\ \text{dtrs} \mid \text{head-dtr} \mid \text{synsem} \mid \text{local} \mid \text{cat} \mid \text{head} \ \boxed{1} \end{bmatrix}$$

Del mismo modo, los principios de Esquemas de Dominancia Imediata (ID Schemata) pueden tambien expresarse como restricciones sobre estructuras de rasgos. HPSG logra con esto modelar todo el conocimiento lingüístico mediante estructuras de rasgos sin tener que recurrir a reglas o principios.

Es bien conocido que la organización de léxico constituye sin duda alguna un aspecto muy imporante en cualquier teoría lexicalista. Las gramáticas de unificación recurren a la jerarquía de tipos para reducir la información requerida en cada entrada léxica. Se definen, pues, jerarquías de tipos ortogonales entrelazadas que permiten eliminar redundancia en el léxico.

Otro mecanismo ampliamente utilizado son las reglas léxicas. Las reglas léxicas permiten establecer generalizaciones en el léxico. Si la jerarquía de tipos permite eliminar gran cantidad de redundancia 'vertical' (vía herencia), las reglas léxicas permiten eliminar redundancia 'horizontal'. En otras palabras, las reglas léxicas proyectan tipos de palabras de otros tipos.

Las reglas léxicas pueden verse como procedimientos declarativos o procedurales. En el primer caso, las reglas simplemente capturan las generalizaciones sobre relaciones estáticas entre miembros de dos o más clases. En el segundo caso, las reglas describen procesos por los cuales, de determinado *input*, se produce un *output*.

A pesar de que las reglas léxicas son ampliamene utilizadas, lo cierto es que son externas al formalismo. Las reglas léxicas no están tipificadas y, por lo tanto, no se encuentran en la taxonomía. Tener reglas léxica supone, entonce, tener un tipo de entidad lingüística diferente no contemplada por el formalismo. La reglas léxicas constituyen, por lo tanto, una violación del principio fundamental de la HPSG según el cual todo conocimiento lingüístico puede modelarse mediante estructuras de rasgos. Las reglas léxicas son operaciones sobre estructuras de rasgos, no estructuras de rasgos.

Ya hemos dicho que en el formalismo de la HPSG, la gramática consiste en una tipología organizada jerárquicamente dónde cada tipo especifica un objeto lingüístico. Pollard y Sag definen las reglas lexicas como operadores sobre estructuras de rasgos. La existencia de reglas léxicas obliga, entonces, a definir un formalismo de orden superior. Pollard y Sag 1994 apuntan que el conjunto de reglas léxicas puede abordarse como un operador de clausura que, de un léxico básico, genera un léxico completo. El nuevo formalismo sería algún tipo de álgebra dónde las reglas léxicas serían operaciones

algebraicas aplicadas sobre un léxico básico[2]. En otras plabras, podríamos definir el léxico como una álgebra **L** consistene en un léxico L y un conjunto de reglas léxicas *rl*:

(1) $\mathbf{L} = <\mathrm{L}, \mathrm{rl}_1, \mathrm{rl}_2, \ldots, \mathrm{rl}_n>$

Una forma más sencilla de expresar las generalizaciones las reglas léxicas sin que ello suponga incrementar el formalismo consiste simplemente en expresarlas como estructuras de rasgos a la Krieger y Nerbone (1993)[3]. En su artículo Krieger y Nerbone proponen definir las reglas léxicas de formación de palabras (flexión, derivación y compuestos) en términos de estructuras de rasgos evitando, así, tener objetos lingüísticos diferentes. Siguiendo su propuesta, podemos abordar la flexión en español definiendo paradigmas de flexión en términos de estructuras de rasgos. Estas estructuras de rasgos participan en las relaciones de herencia de la taxonomía. Así, una forma flexionada resultaría de la unificación de una raíz y un paradigma de flexión. En otras palabras, las formas flexionadas serían tipos (terminales) que heredarían la información de algún tipo del conjunto de tipos *raiz* y del conjunto de tipos *paradigma de flexión.* Así, por ejemplo, el paradigma nominal sería:

$$\left\{ \begin{array}{l} \textit{paradigma-nominal} \\ \left[\begin{array}{l} \text{morph} \left[\begin{array}{ll} \text{stem} & \boxed{1} \\ \text{ending} & \boxed{2}o \\ \text{form} & \boxed{1}\&\boxed{2} \end{array} \right] \\ \text{synsem} \mid \text{local} \mid \text{cat} \mid \text{head} \mid \text{agr} \left[\text{gen } masc \right] \end{array} \right] \\ \left[\begin{array}{l} \text{morph} \left[\begin{array}{ll} \text{stem} & \boxed{1} \\ \text{ending} & \boxed{3}a \\ \text{form} & \boxed{1}\&\boxed{3} \end{array} \right] \\ \text{synsem} \mid \text{local} \mid \text{cat} \mid \text{head} \mid \text{agr} \left[\text{gen } fem \right] \end{array} \right] \end{array} \right\}$$

Este paradigma será heredado por todos los nombres cuyo tipo fuera compatible con él (en otras palabras, todos los sustantivos flexionados en *o/a* pertenecerían al tipo que hereda de *raiz-nominal* y *paradigma-nominal.*).

Como podemos observar, no resulta difícil abordar la flexión en términos de jerarquía de tipos. El verdadero problema lo constituyen las reglas léxicas que pueden ser vistas como 'procedurales'. Las reglas léxicas que generalizan la diátesis verbal probablemente no pueden abordarse mediante jerarquías de tipos ortogonales. En este caso la información no es vertical sino horizontal.

[2]Smolka & Ait-Kaci (1989) Feature Constraint Logics for Unification Grammar. IWS Report 93, Nov 1989, IBM Deutchland, Stuttgart, W. Germany.

[3]Krieger y Nerbone 1993. Feature-Based Inheritance Networks for Computational Lexicons. En *Inheritance Defaults and the Lexicon.* Briscoe, de Pavia y Copestake (eds). Cambridge University Press.

Así, por ejemplo, la regla léxica de pasiva establece que si tenemos el tipo ... tenemos también el tipo *pasivo*:

$$\left[\begin{array}{l} activo \\ \text{subcat} \left\langle \text{NP}\boxed{1}, \text{NP}\boxed{2} \right\rangle \\ \text{content } \boxed{3} \end{array}\right] \rightarrow \left[\begin{array}{l} pasivo \\ \text{subcat} \left\langle \text{NP}\boxed{2}, \left(\text{PP}\left[\text{por}\right]\boxed{1} \right) \right\rangle \\ \text{content } \boxed{3} \end{array}\right]$$

Observamos, sin embargo, que no podemos obtener el tipo *pasivo* del resultado unificar el tipo *activo* con algún otro tipo. La regla de algún modo modifica el ... para generar el *output*. El problema fundamental estriba en que las coindexaciones entre los elementos de la lista SUBCAT y los argumentos el CONTENT deben obedecer la unificación establecida en la regla mediante los sub-indices: es decir, que elemento agentivo de la activa coindexa con el sujeto mientras que el en la pasiva hace con el objeto. Este requerimento no puede expresarse vía taxonomía ortogonal. Es decir, vía una taxonomía que incluya jerarquías disjuntas (una que recoja el marco de subcategorización sintáctico y otra que recoja la estructura argumental):

(2) a. *biargumental* [CONTENT X]

b. *bivalente1* [SUBCAT <SN,SN>]

c. *bivalente2* [SUBCAT <SN,(SPpor)>]

d. *pasivo* = *biargumental* ⊔ *bivalente2*

El tipo *pasivo* que obtenemos en (2-d) como consecuencia de la unificación de los supertipos *biargumental* (para verbos con dos argumentos) con el tipo *bivalente2* (para verbos con un SN y un SP[por]) no garantiza la necesaria coindexación que establece la pasiva.

Nuestra propuesta consiste en integrar las reglas léxicas como estructuras de rasgos utilizando la disyunción y simulando el operador de clausura sobre las reglas por medio del propio mecanismo de herencia de la taxonomía. Así, podemos conseguir las mismas generalizaciones que expresaban las reglas léxicas utilizando nuevos tipos con estructuras de rasos extendidas y mecanismos de herencia modificados que permitan la herencia disjunta. Con ello evitamos tener reglas léxicas y logramos homogeneizar la representación del conocimiento lingüístico.

En el caso de la alternancia activa/pasiva definimos un nuevo tipo *activo+pasivo* que resulta de la herencia disjunta de los supertipos *bivalente1* y *bivalente2* definidos en (2):

$$\left[\begin{array}{l} activo+pasivo \\ \text{subcat} \left\{ \left\langle \text{NP}\boxed{1}, \text{NP}\boxed{2} \right\rangle, \left\langle \text{NP}\boxed{2}, \left(\text{PP}\left[por\right]\boxed{1} \right) \right\rangle \right\} \\ \text{content } \boxed{3} \end{array}\right]$$

Dado que el subtipo *activo+pasivo* exige que el sujeto y el objeto del esquema activo coindexen con el SP y el sujeto del esquema pasivo respectivamente, las proyecciones sintactico-semánticas exigidas por la pasivización se cumplen en ambos casos. Así, si tomamos un verbo transitivo como *comer* obtendremos la siguiente estructura:

$$\left[\begin{array}{l} \textit{activo+pasivo} \\ \text{pho } \textit{comer} \\ \text{subcat } \left\{\left\langle \text{NP}_{\boxed{1}}, \text{NP}_{\boxed{2}}\right\rangle, \left\langle \text{NP}_{\boxed{2}}, \left(\text{PP}\left[\textit{por}\right]_{\boxed{1}}\right)\right\rangle\right\} \\ \text{content } \left[\begin{array}{ll} \text{rel} & \textit{comer} \\ \text{arg1} & \boxed{1} \\ \text{arg2} & \boxed{2} \end{array}\right] \end{array}\right]$$

Del mismo modo, para un verbo como *abrir* con una alternancia sintáctica más amplia ejemplificada en (3) definimos el subtipo *intransitivo+transitivo+ pasivo+decausativo* en (4):

(3) a. *intransitivo*: Al final no abrieron

b. *transitivo*: Juan abrió la puerta

c. *decausativo*: La puerta abre mal

d. *pasiva*: La puerta fue abierta

(4) Marco de subcategorización *int+tr+pas+dec*:

$$\left[\begin{array}{l} \textit{int+tr+pas+dec} \\ \text{subcat } \left\{\begin{array}{l} \left\langle \text{NP}_{\boxed{1}}\right\rangle \\ \left\langle \text{NP}_{\boxed{1}}, \text{NP}_{\boxed{2}}\right\rangle \\ \left\langle \text{NP}_{\boxed{2}}\right\rangle \\ \left\langle \text{NP}_{\boxed{2}}, \left(\text{PP}\left[\textit{por}\right]_{\boxed{1}}\right)\right\rangle \end{array}\right\} \\ \text{content } \boxed{3} \end{array}\right]$$

El tipo *int+tr+pas+dec* resulta de la herencia disjunta de los supertipos *intransitivo*, *bivalente1* y *bivalente2*. De nuevo las coindexaciones establecidas por el tipo garantizan que en el caso de *abrir* obtengamos la siguiente descripción:

$$\left[\begin{array}{ll} \textit{int+tr+pas+dec} & \\ \text{pho } \textit{abrir} & \\ \text{subcat} & \left\{\begin{array}{l} \left\langle \text{NP}_{\boxed{1}} \right\rangle \\ \left\langle \text{NP}_{\boxed{1}}, \text{NP}_{\boxed{2}} \right\rangle \\ \left\langle \text{NP}_{\boxed{2}} \right\rangle \\ \left\langle \text{NP}_{\boxed{2}}, \left(\text{PP}\left[\textit{por}\right]_{\boxed{1}} \right) \right\rangle \end{array}\right\} \\ \text{content} & \left[\begin{array}{ll} \text{rel} & \textit{abrir} \\ \text{arg1} & \boxed{1} \\ \text{arg2} & \boxed{2} \end{array}\right] \end{array}\right]$$

Observamos que las coindexaciones entre la información sintáctica (esto es, l.. patrones de la lists SUBCAT) y la información semántica (esto es el CONTENT) ... las correctas. Así, el ARG1 está (correctamente) coindizado el sujeto de las form... intransitivas y transitivas y con el sintagma preposicional de la pasiva, mientras qu... el ARG2 lo está con el objeto de la forma transitiva, el sujeto de la decausativa y ... sujeto de la pasiva.

A modo de conclusión, nuestra propuesta permite definir tipos de marcos de sub... categorización complejos dentro de una jerarquía que admita la herencia disjunta. C... ello logramos expresar las generalizaciones de las reglas léxicas sin tener, en realidad, ... glas léxicas al tiempo que logramos dar una representación del conocimiento lingüísti... modelada exclusivamente en términos de estructuras de rasgos. Finalmente queremo... destacar que esta manera de 'entender' las generalizaciones expresadas mediante regl... léxica nos ha permitido abordar no sólo la diátesis verbal sino también el fenómeno d... la cliticización de forma eficaz y elegante.

Bibliografía

Carpenter, Bob. 1992. *The Logic of Typed Feature Structures.* Cambridge Tract... in Theoretical Computer Science no. 32. New York: Cambridge University Press.

Flikinger & Nerbone. 1992. 'Inheritance an Complementation: a case study of *easy* adjectives and related nouns'. *Computational Linguistics*, vol. 18.3, 269-310.

Krieger & Nerbone 1993. Feature-Based Inheritance Networks for Computational Lexicons. En *Inheritance Defaults and the Lexicon.* Briscoe, de Pavia y Copestake (eds). Cambridge University Press.

Pustejovsky, J. 1991. 'The Generative Lexicon'. *Computational Linguistics*, vol 17.4, 409-441.

Smolka, Gert. 1989. *Feature contraint logics for unification grammars.* IWBS Report 93, IBM - Deutschland GmbH, Stuttgart, Germany. To appear in *Journal of Logic Programming.*