La integración de semántica léxica en HPSG

Toni Badia
toni.badia@trad.upf.es
Roser Saurí
roser.sauri@trad.upf.es
Institut Universitari de Lingüística Aplicada — Universitat Pompeu Fabra
La Rambla 32 — 08002 Barcelona — Catalunya - Spain

Resumen

En esta comunicación discutimos los límites de los tratamientos habituales de la sintaxis en PLN y proponemos la integración de semántica léxica en los mismos. Para ello, escogemos dos modelos compatibles tanto teóricamente como formalmente: HPSG para la representación sintáctica y GL para la representación de la semántica léxica. Finalmente mostramos la viabilidad de la propuesta mediante algunos ejemplos: los complementos de los nombres, los adjetivos modificadores de nombres y los complementos semánticamente presentes pero no expresados sintácticamente.

1 Tratamientos sintácticos tradicionales

Los sistemas de tratamiento del lenguaje natural suelen situar la sintaxis en un lugar central del procesamiento y suelen considerar la semántica como un complemento al análisis sintáctico para tratar aquellos casos que la sintaxis no puede resolver por sí sola. En esta comunicación nos ocupamos del tratamiento de la semántica en los sistemas de PLN, es decir, de la interacción entre la sintaxis y la semántica. En los últimos años se han producido unos avances que permiten replantear esta interacción a partir de unas nuevas bases: han aparecido sistemas de tratamiento de la sintaxis que permiten representar las relaciones entre la sintaxis y la semántica de una forma mucho más natural, y también maneras nuevas de enfocar la representación semántica, especialmente la llamada semántica léxica.

Los tratamientos estándar de la sintaxis en los sistemas de PLN introducen un nivel abstracto de descripción de las relaciones que se dan en el interior de las oraciones: se trata de lo que se suele denominar la estructura argumental, es decir la relación de los predicados con sus argumentos y con sus modificadores. En las teorías sintácticas más usadas en el procesamiento lingüístico (LFG y HPSG), se establecen en este nivel las relaciones gramaticales generales. Así por ejemplo, en la estructura-f de LFG y en el atributo CONTENT de HPSG se expresan las relaciones entre los núcleos y sus argumentos que difieren de las que mantienen en la forma superficial (y que se expresan en forma de estructura-c o en las listas de subcategorización). Así, las relaciones de control se expresan mediante coindexación de valores argumentales en la estructura-f o en la *psoa* (*parametrised states of affairs*), de manera que un único elemento en la estructura-c o en la lista de subcategorización

provee de contenido a dos posiciones argumentales distintas; y la pasiva se trata como un cambio en la correlación entre elementos de la estructura-c o de la lista de subcategorización y elementos de la estructura-f o de la *psoa*.

A pesar de todo, este nivel en el que se representa la estructura argumental no deja de ser el resultado de una proyección directa de la estructura superficial; es cierto que no se trata de una proyección simple (como acabamos de ver), pero también lo es que no permite eliminar las características básicas de la misma. Así por ejemplo, se dan claras dificultades para incorporar posiciones argumentales que no se corresponden con posiciones explícitas de complementos en la sintaxis superficial. Y esto es así porque tanto la estructura-f como la *psoa* no han sido pensadas como representaciones semánticas completas, sino simplemente como representaciones profundas de la estructura sintáctica. Esto es obvio en LFG, donde la estructura-f se define precisamente como sintáctica y donde se ha propuesto un nuevo nivel de representación independiente para dar cuenta de la semántica de las expresiones lingüísticas (la estructura-s).

En el seno de HPSG la situación no está tan clara. La estructura de *psoa* está incorporada en la parte del signo correspondiente al atributo CONTENT, que está al mismo nivel que los de CATEGORY y CONTEXT; en ella se usa mucha terminología proveniente de la semántica de las situaciones. Parecería por lo tanto que estamos ante un tratamiento plenamente semántico. No obstante, una observación detallada de los fenómenos más estudiados y del tratamiento propuesto muestra claramente que éste se centra especialmente en la interfaz entre sintaxis y semántica, y no en una descripción puramente semántica. Por ejemplo, la tipificación más rica del tipo *psoa* de que se dispone es la que se usa para la resolución del ligamiento (Pollard & Sag, 1994:c.6); y la propuesta más extensa que conocemos sobre la implementación computacional de la estructura argumental (Badia & Colominas (1998)) no pretende representar las implicaciones semánticas de las clases de predicados, sino sólo tipificar de manera consistente las diferentes clases de complementos.

En lo que sigue nos concentramos en HPSG (y no en LFG), puesto que parece mejor equipada para permitir la integración de la estructura superficial (representada por los rasgos de valencia en el tipo CAT), la estructura argumental (representada por los valores *psoa* del rasgo CONTENT), y la semántica léxica del signo (representada en un rasgo adicional). Por supuesto, esta integración puede aplicarse explícitamente gracias al recurso del *structure sharing*, de manera que se pueden expresar en un único sistema de tipos las varias dimensiones de los signos lingüísticos (la formal, la estructural y la semántica). De hecho nuestra propuesta en la sección 3 va a basarse en la integración de los tres elementos.

2 Datos lingüísticos

Por supuesto, teorías lingüísticas de base sintáctica (como LFG o HPSG) permiten codificar y tratar satisfactoriamente un gran número de construcciones lingüísticas: en general todas aquellas en que los argumentos se expresan mediante relaciones estrictas de subcategorización o los modificadores pueden ser interpretados intersectivamente. No obstante hay numerosos ejemplos de construcciones que no encajan de manera natural en estos planteamientos. En esta sección repasamos brevemente algunas de estas construcciones: complementos opcionales de verbos, complementos (casi siempre opcionales) de

los nombres, y modificadores no intersectivos.

Naturalmente la distinción habitual entre complementos regidos (obligatorios) y adjuntos (opcionales) no es suficientemente rica para dar cuenta de todas las clases de complementos. Y fuerza a los planteamientos lingüísticos de base sintáctica a incluir en su diccionario múltiples entradas para cada verbo que tiene variación en su estructura de subcategorización.[1]

A veces los complementos verbales no son simplemente opcionales: algunos están implícitos en la semántica del verbo y otros pueden estar presentes sólo en condiciones muy especiales; son los que Pustejovsky (1995:63s) trata como *default* y *shadow arguments* respectivamente:

(1) a. **D-Arg**: John built the house **out of bricks**

b. **S-Arg**: Mary buttered her toast **with an expensive butter**

Aunque como vemos, la opcionalidad de los complementos de los verbos es alta, la de los complementos nominales es aún mayor: de hecho prácticamente todos los complementos de los nombres pueden no aparecer en la superficie. Veamos unos pocos ejemplos:

(2) a. Hoy he visto al padre **de Juan**

b. Hay dos padres que no han venido a buscar a su hijo

(3) a. Compraré dos hojas **de cartulina**

b. Escríbelo en una hoja

En el caso de los complementos de los nombres, la opción de listar en entradas léxicas diferentes todas las opciones resulta muy poco satisfactoria, ya que no hay (casi) ningún aspecto gramatical que pueda restringir la aplicación de una entrada léxica o de otra. Incluso la simple distinción entre el valor objetivo o subjetivo de un complemento verbal resulta problemática desde un planteamiento puramente sintáctico:

(4) a. el estudio de las plantas

b. la evaluación de los inspectores

En los ejemplos de (4) se muestra claramente que la decisión de si un complemento de un nombre deverbal transitivo es objetivo o subjetivo depende estrictamente de su valor semántico, puesto que su estructura sintáctica puede ser exactamente la misma. Esto, claro está, apunta a la necesidad de incorporar información plenamente semántica en el tratamiento de estos complementos.

Una razón más en este sentido proviene de ejemplos como los de (5), en los que se muestra que factores de tipo discursivo pueden intervenir también:

(5) a. La decoración del puente nos ha llevado mucho tiempo, !pero ha quedado muy bien terminada!

b. Traducir este folleto me ha costado mucho, pero al final creo que me ha quedado muy natural

c. Ha venido una madre esta mañana. Venía a decir que su hijo no podrá venir a la excursión

d. Hemos aliñado la ensalada y la hemos tenido que tirar porque el aceite estaba rancio

En estos ejemplos se muestra que complementos que no están presentes explícitamente en el SV o SN pueden ser objeto de una referencia anafórica o de una implicación en el discurso: el sujeto de *terminada* (5a) y *natural* (5b) sólo se puede recuperar como el resultado del acto de decorar o de traducir respectivamente; y el uso

[1] Es el llamado *Sense Enumerative Lexicon* por Pustejovsky (1995).

de los determinantes definidos *su* y *el* en (5c) y (5d) está permitido por el complemento "oculto" de *madre* y *aliñar*.

Finalmente podemos examinar algunos casos problemáticos de modificadores adjetivales. Muchos adjetivos denotan de manera diferente según el contexto en el que aparecen. El adjetivo *rápido*, por ejemplo, que modifica habitualmente eventos, puede aparecer en expressiones como (6), donde predica sobre individuos.

(6) a. un mecanógrafo rápido

b. un conductor rápido

c. un coche rápido

Además algunos adjetivos pueden expresar una propiedad diferente en el mismo contexto local (permitiendo así tanto una interpretación intersectiva como una de no intersectiva):

(7) a. un lápiz rojo

b. un brazo roto

El ejemplo en (7a) puede referirse tanto a un lápiz pintado de color rojo como a un lápiz que colorea de rojo. Asimismo *roto* en (7b) puede aplicarse tanto a toda la entidad referida por *brazo* como a una única parte de la misma (que es claramente la lectura preferida).

3 Propuesta de tratamiento

A continuación desarrollamos una propuesta de tratamiento para casos como los que hemos presentado en la sección anterior. Asumiendo que una aproximación de orientación sintáctica a estos datos es claramente insuficiente, el requisito básico para el nuevo tratamiento es disponer de un nivel de codificación de la información semántica que sea independiente de la información sintáctica. Este nivel debe partir de una concepción rica y robusta de la semántica de las unidades léxicas. Rica, en el sentido que permita dar cuenta de las implicaciones de distintos participantes y eventos en la denotación de cada unidad (véase por ejemplo (5)), y que sea suficientemente expresiva para dar cuenta de las restricciones de selección que los predicados imponen a sus argumentos (como en (4)). Por otro lado, una concepción robusta en el sentido de que proporcione una representación semántica con información de distintos niveles interrelacionables entre ellos, para poder tratar de una forma natural la interpretación no intersectiva de ciertas expresiones (como en (7)).

Estos requisitos nos llevan a incorporar la propuesta de tratamiento de la información semántica de GL (*Generative Lexicon*) (Pustejovsky, 1991, 1995) al tratamiento formal del léxico que proporciona HPSG.[2] La estructura básica resultante para la representación de las unidades léxicas se muestra a continuación:[3]

$$\begin{bmatrix} \text{LOC} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \textit{head} \\ \text{VALENCE} & \textit{lists-of-complements} \end{bmatrix} \\ \text{SEM} & \begin{bmatrix} \text{INDEX} & \textit{index} \\ \text{EVSTR} & \textit{event-structure} \\ \text{ARGSTR} & \textit{argument-structure} \\ \text{QUALIA} & \begin{bmatrix} \text{FORMAL} & \textit{formal-relation} \\ \text{CONSTIT} & \textit{constit-relation} \\ \text{AGENT} & \textit{agent-relation} \\ \text{TELIC} & \textit{telic-relation} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figura 1

El nuevo nivel de representación semántica

[2] En la linea de otros trabajos previos, como Copestake (1992), Johnston (1996) y Badia & Saurí (1998).

[3] Únicamente por razones de espacio, en algunas de las representaciones léxicas que siguen separamos el atributo QUALIA del resto de la información semántica.

mantiene el nivel de información referencial estipulado en HPSG mediante el rasgo INDEX (IND). Sin embargo, sigue GL en proporcionar un tratamiento más rico para la información de la estructura argumental, expresado en HPSG con el atributo RESTRICTION (RESTR) para los signos nominales y NUCLEUS para los verbales,[4] y adaptado aquí como ARGUMENT STRUCTURE (ARGSTR). De este modo es posible representar no sólo los argumentos sintácticamente obligatorios que participan en la semántica de la entidad, sino también los opcionales (los D-ARG y S-ARG). Igualmente, se introducen dos niveles adicionales de información semántica: la EVENT-STRUCTURE (EVSTR) y la QUALIA-STRUCTURE (QUALIA). El primero presenta la información de la eventualidad que expresan las entidades: determina su estructura eventiva y la caracteriza según la clasificación de eventualidades, entre estados, procesos y transiciones. El segundo describe la semántica léxica de cada entidad y la relación que mantiene con sus argumentos – sean sintácticamente expresables o no. Se concreta en cuatro niveles de información: la qualia formal (FORMAL) especifica las propiedades de la entidad que la distinguen del resto de entidades dentro del dominio a que pertenece; la qualia constitutiva (CONSTIT) indica los elementos o partes de que está constituida la entidad; la qualia agentiva (AGENT), de qué forma se ha originado, y finalmente la qualia télica (TELIC) expresa su finalidad. Además de los niveles de representación semántica que acabamos de ver, GL utiliza la jerarquía de herencia para controlar formalmente (mediante la utilización de tipos) la información semántica de las unidades léxicas, y dispone de mecanismos generativos que permiten dar cuenta de la creatividad evidente en los procesos de interpretación de las expresiones lingüísticas (*co-composition, type coercion* y *selective binding*).

[4]Para una mayor simplicidad, en este artículo no entraremos en el tratamiento de la cuantificación

4 Aplicación de la propuesta

Empezaremos por el tratamiento de los complementos opcionales. Precisamente por su opcionalidad, el tratamiento estándard de los complementos obligatorios en HPSG no es pertinente. Sin embargo, una propuesta reciente de Sanfilippo (1998), que trata como adjuntos sintácticos a ciertos complementos temáticamente ligados al verbo, permite una aproximación adecuada a estos casos. Concretamente, representa estos complementos como miembros del conjunto del nivel de información no local (NONLOC). En nuestra propuesta adoptamos su mecanismo para los argumentos D-ARG y S-ARG, que pasan pues a concebirse como adjuntos temáticamente ligados.

Consideremos un nombre ambiguo entre una interpretación de proceso y una de resultado como *construcción*, la nominalización del verbo *construir*. Este verbo presenta dos complementos obligatorios: el agente y el resultado del proceso de *construir*. Tiene además un tercer argumento, de tipo D-ARG, el cual expresa el material (Pustejovsky, 1995). Se trata de un complemento sintácticamente opcional pero que participa en la expresion lógica del evento. En la representación de este verbo (fig. 2), los primeros dos complementos se expresan en la lista de complementos obligatorios, mientras que el último se trata, siguiendo a Sanfilippo (1998), como miembro en el conjunto del nivel de información no local.

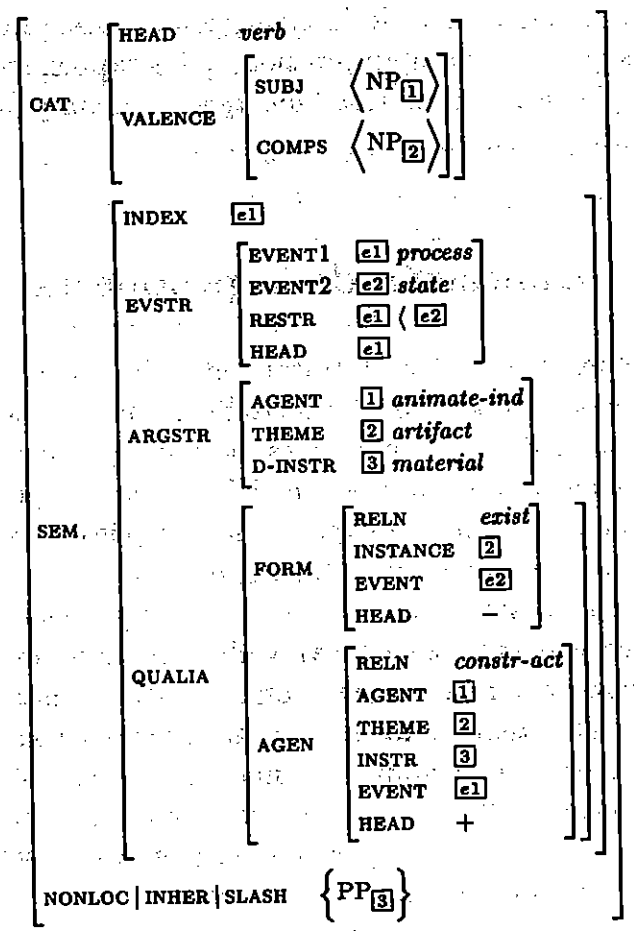

Figura 2: *construir*

La entrada léxica para la nominalización de proceso de este predicado se muestra en la fig. 3. Nótese que todos los argumentos que intervienen en la predicación están tratados como argumentos D-ARG.[5]

[5]En la representación de la lectura de resultado, el tipo semántico introducido por el rasgo INDEX sería un individuo, en lugar de un evento (concretamente, se trataría del valor de INSTANCE de la qualia FORMAL), y el núcleo (HEAD) de la EVSTR sería el estado de existir de este individuo (el EVENT de FORMAL), en lugar del proceso de creación que se expresa en el nivel de qualia AGENTIVE. Para mayor detalle sobre los mecanismos formales para el tratamiento adecuado de estos casos de polisemia en el sistema de tipos, véase Badia & Saurí (1998).

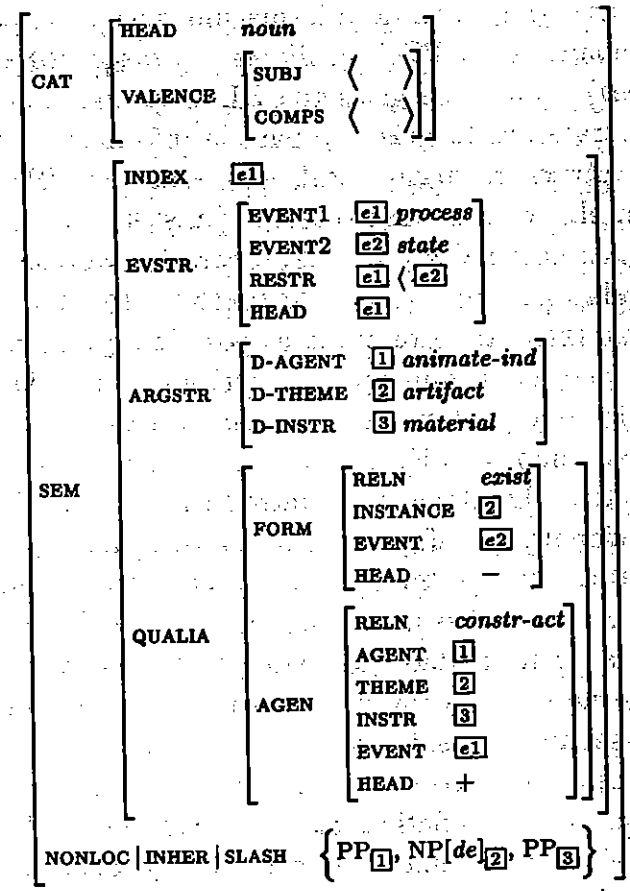

Figura 3: *construcción* (lectura de proceso)

Vemos pues que los complementos verbales y nominales que son opcionales pero que estan implicados lógicamente por la semántica del predicado (D-ARGS y S-ARGS), se distinguen de los complementos sintácticamente obligatorios a partir del nivel sintáctico en el que se declaran.

Esta propuesta es igualmente adecuada para otros tipos de nominales. Consideremos ahora un ejemplo de nominalización de redescripción. Este tipo de nominalizaciones se distinguen de las nominalizaciones anteriores en que la interpretación de proceso no puede expresar como complemento sintáctico el resultado de la acción. Tomemos el nombre *decoración*, derivado de *decorar*, que es un predicado con tres argumentos: el agente del evento, el tema (o sea, el ob-

jeto que se decora) y un argumento D-ARG que expresa el material. Los argumentos agente y el tema son introducidos como obligatorios por parte del verbo, pero se realizan como adjuntos temáticamente regidos en la entrada léxica de la nominalizacíon (fig. 4).

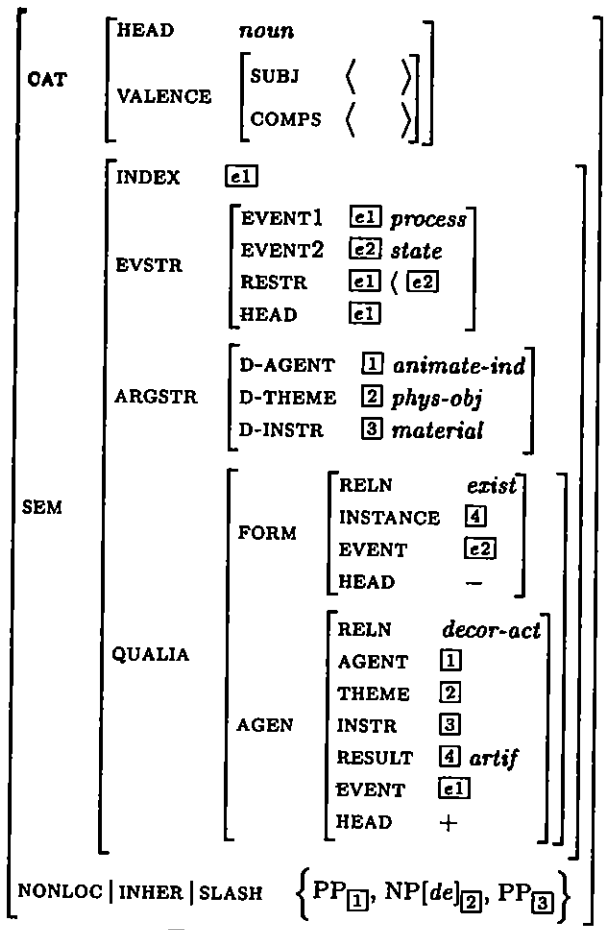

Figura 4: *decoración*

Nótese que esta representación es adecuada para tratar casos de implicaciones como los de (5a-b), en el cual el predicado se interpreta en la primera cláusula como el proceso de decorar, mientras que en la segunda se interpreta como el objeto resultado de este proceso. La recuperación adecuada de la anáfora es posible mediante el cuarto argumento en la qualia AGENTIVE, el cual no esta ligado con ningún otro argumento en la ARGSTR ya que nunca se puede realizar sintácticamente. Por consiguiente, el nivel de información de la estructura argumental funciona como interfaz entre la representación de la información semántica (básicamente QUALIA y EVSTR), y los mecanismos de superficie que controlan la realización de los complementos de los predicados –VALENCE y NONLOCAL. Así, el nivel ARGSTR recoge los argumentos que pueden realizarse en la superficie, obligatoria u opcionalmente, mientras que el nivel de representación de QUALIA expresa los argumentos semánticamente implicados aunque no sean realizables sintácticamente.

Esta concepción rica y estructurada de la semántica permite el tratamiento adecuado de otros fenómenos para los cuales una orientación sintáctica es insuficiente; por ejemplo, la distinción adecuada entre complementos subjetivos u objetivos de los nombres deverbales (como en (4)). La correcta interpretación de estas expresiones se obtiene mediante las restricciones de selección que el nombre impone en relación con la información de QUALIA de su complemento. Así, la interpretación como complemento objetivo de *de las plantas* de (4a) se obtiene de la restricción de *estudio* sobre el tipo semántico de su complemento agente –que sea *humano*.

Pasemos finalmente a ver la aplicación de esta propuesta al tratamiento de los modificadores que no pueden ser tratados intersectivamente. El tratamiento que Pustejovsky (1995) propone para los casos como (6) consiste en provocar una interpretación de evento al nombre que modifica el adjetivo. Esto es posible aplicando el mecanismo generativo de *Selective Binding*, el cual fuerza al adjetivo a predicar sobre el nivel específico que satisface sus restricciones selecionales (es decir, que denota un evento), en lugar de predicar sobre toda la entidad. Así,

rápido aplicado a *mecanógrafo* predica sobre el evento de mecanografear, expresado en el nivel de qualia TELIC; mientras que cuando modifica a *conductor* hace referencia al proceso de conducir. Este planteamiento pues permite por un lado dar cuenta de la distinta interpretación de un mismo adjetivo según el sustantivo con el que combina como fruto de la información semántica de éste, sin necesidad de tener que establecer entradas léxicas independientes para cada interpretación posible. Y por el otro, permite tratar la interpretación no interesectiva de determinados adjetivos. La entrada léxica para un adjetivo modificador de eventos como *rápido* es la siguiente:

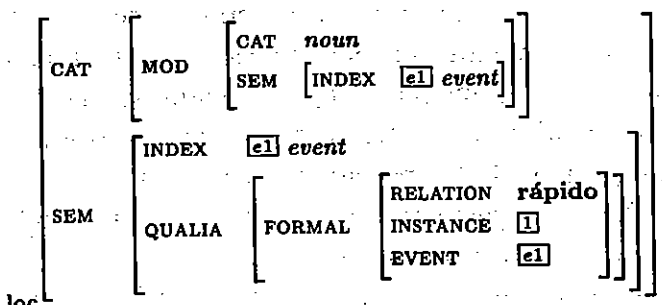

Figura 6: *rápido*

Veámos ahora la aplicación de la propuesta al tratamiento adecuado de la polisemia de algunos adjetivos modificando a un mismo nombre, como ocurre en la expresión *un lápiz rojo*, recuperada de (7a). Esta expresión puede hacer referencia a un lápiz pintado exteriormente de rojo o bien a un lápiz que pinta rojo. Específicamente, el adjetivo *rojo* exige un sustantivo que denote un individuo concreto del cual sea posible predicar sobre color. Así, la interpretación de la expresión como "lápiz que tiene la superficie pintada de rojo" (que corresponde a la intersectiva) se obtiene de la combinación de la semántica del adjetivo y el índice del sustantivo –coindexado éste con el valor de INSTANCE en FORMAL.

Sin embargo, para la interpretación no intersectiva (es decir, la que expresa un lápiz que colorea rojo) se requiere un proceso más complejo, ya que el individuo sobre el que predica el adjetivo corresponde al resultado del proceso expresado en el nivel TELIC de *lápiz*. Podemos considerar que estos casos estan motivados por la información semántica del sustantivo. Concretamente, proponemos refinar el tratamiento, especificando para cada tipo de sustantivo, un nivel de qualia más prominente que los otros. Esta intuición se ve confirmada por los resultados de la tabla que sigue:

	rápido selección subespecificada	*eficaz* selecciona TELIC	*inacabado* selecciona AGENTIVE	QUALIA prominente
construcción	AGENT	info externa	AGENT	Prom: AGENT
análisis	AGENT	info externa	AGENT	Prom: AGENT
pastel	AGENT	info externa	AGENT	Prom: AGENT
estatua	AGENT	info externa	AGENT	Prom: AGENT
cuchillo	TELIC	TELIC	AGENT	Prom: TELIC
teclado	TELIC	TELIC	AGENT	Prom: TELIC
carpintero	TELIC	TELIC	*	Prom: TELIC
mecanógrafo	TELIC	TELIC	*	Prom: TELIC

En esta tabla se puede apreciar que un adjetivo modificador de eventos como *rápido* combinado con sustantivos agentivos o que denotan instrumentos (como *mecanógrafo* y *cuchillo*, respectivamente) predica sobre el nivel de qualia TELIC de éstos; mientras que cuando modifica a nominalizaciones resultativas (como *construcción*) y nombres como *pastel*, predica sobre el nivel AGENTIVE. Dado que *rápido* (a diferencia de *eficaz* o *inacabado*) no especifica el nivel de qualia sobre el cual predica, estos resultados

confirman la prominencia de un nivel de qualia por encima de los otros según el tipo de entidad que denote cada sustantivo.

Partiendo de esta ponderización de los distintos niveles de qualia, vemos que la aparente polisemia del adjetivo *rojo* en el ejemplo (7a) se explica por la configuración semántica del sustantivo. Así, a pesar de que *rojo* requiere un individuo y que esto posibilita de manera natural una interpretación intersectiva de la expresión, la interpretación no intersectiva se explica porque *lápiz* denota un instrumento y que, como tal, presenta la qualia TELIC como prominente (fig. 7):

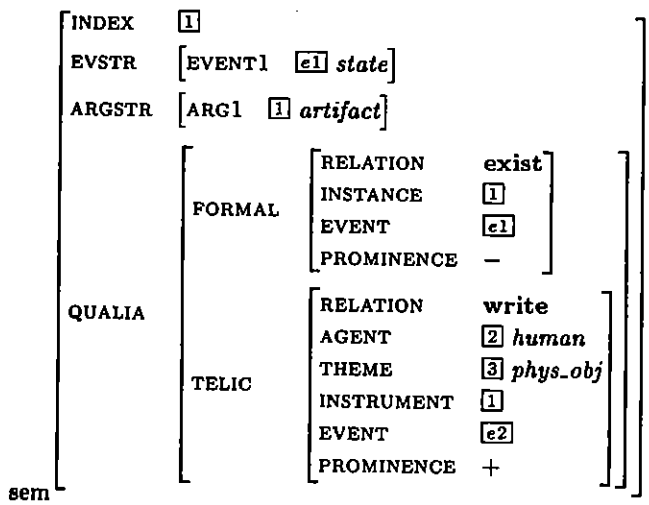

Figure 7: *lápiz*

5 Conclusión

En este artículo hemos discutido la necesidad de incorporar información de semántica léxica en el tratamiento sintáctico-semántico tradicional en procesamiento del lenguaje natural. Hemos propuesto integrar en la estructura del signo lingüístico de HPSG la información de semántica léxica propuesta en el modelo de GL. Esta integración permite formular de una manera satisfactoria las alternacias de significado de varias clases léxicas (nombres deverbales y relacionales, verbos con complementos no expresados, adjetivos modificadores de nombres...), de manera que el léxico resultante resulta altamente simplificado. Al mismo tiempo, esta aproximación ofrece una respuesta a algunos fenómenos lingüísticos que no pueden ser tratados satisfactoriamente en los sistemas basados únicamente en la sintaxis (como ciertas anáforas dependientes de complementos semánticos no expresados sintácticamente, o la relación entre muchos adjetivos y sus núcleos).

Bibliografía

Badia, T. & C. Colominas (1998) "Predicate-Argument Structure", en F. van Eynde & P. Schmidt (eds.) *Linguistic Specifications for Typed Feature Formalisms.* CUE. Luxemburg.

Badia, T. & R. Saurí (1998) "The Representation of Syntactically Unexpressed Complements to Nouns", en *COLING-ACL'98. Workshop on the Computational Treatment of Nominals*, pp. 1-10. Montréal, Canadá.

Copestake, A. (1992) *The Representation of Lexical Semantic Information*, PhD thesis, Sussex University, Cognitive research paper CSRP 280.

Johnston, M. (1996) "Semantic underspecification in lexical types: capturing polysemy without lexical rules". *Acquilex Workshop on Lexical Rules, 1995.* Cambridge.

Pollard, C. & I. Sag (1994) *Head-driven Phrase Structure Grammar.* CSLI, Standford CA.

Pustejovsky, J. (1991) "The Generative lexicon", en *Computational Linguistics*, 17:409-441.

Pustejovsky, J. (1995) *The Generative Lexicon.* The MIT Press. Cambridge MA.

Sanfilippo, A. (1998) "Thematically bound adjuncts". In Balari, S. & L. Dini (eds.) *Romance in HPSG.* CSLI. Stanford CA.