El Léxico PAROLE del Español

Marta Villegas tona@gilcub.es
Isabel Brosa isabel@gilcub.es
Nuria Bel nuria@gilcub.es
GILCUB (Grup Investigació Lingüística Computacional Universitat Barcelona)

Introducción

PAROLE (LE-4017) es un proyecto financiado por la UE bajo el IV Programa Marco para la creación y desarrollo de recursos léxicos (corpus y diccionarios electrónicos) a gran escala para 14 lenguas europeas con una estructura de codificación común en SGML (Standard Generalized Markup Language, International Standard 8879).

Los léxicos PAROLE siguen las recomendaciones del grupo EAGLES[1]. El modelo PAROLE es, por lo tanto, un modelo descriptivo, flexivo (permite acomodar diferentes niveles de granularidad en sus descripciones) y neutral en cuanto a teorías lingüísticas se refiere.

El léxico Parole incluye cuatro niveles de descripción: (i) nivel morfológico, (ii) nivel sintáctico, (iii) nivel semántico, y (iv) nivel relacional que permite relacionar la información codificada.

Información morfológica

El léxico español Parole contiene algo más de 20.000 entradas. Cada entrada contiene información morfológica y información sintáctica. La información morfológica sigue el enfoque "Word and Paradigm", con lo cual cada entrada especifica el conjunto de raíces posibles, sus correspondientes terminaciones y la información morfosintáctica de éstas. Así, un sustantivo como '*patrón*' se define como sigue:

raíz1: *patrón*	sufijos: ∅
raíz2: *patron*	sufijos: *es / a / as*

Información sintáctica

Cada Unidad Morfológica[2] tiene asignada como mínimo una Unidad Sintáctica. Toda Unidad Sintáctica describe el comportamiento sintáctico de la entrada léxica y se define por tener una Descripción básica obligatoria y una lista opcional de Descripciones no-básicas.

Las Descripciones llevan información sobre la estructura sintáctica (o Construcción) en la cual se inserta la unidad léxica y la información morfosintáctica relevante de dicha unidad en esa Construcción[3].

Toda Construcción consta de una lista (abierta o cerrada) de Posiciones. Cada Posición tiene asignada una función sintáctica y un papel temático. Las Posiciones se definen por la función y el papel temático asignados, así como por la lista de Sintagmas (objetos del tipo Construcción) que pueden ocupar esa Posición.

Este modelo permite estructurar la información sintáctica de diferentes maneras. Así, por ejemplo, la alternancia

[1] Expert Advisory Group on Laguage Engineering Standars es una iniciativa de la UE coordinada por expertos, que se ocupa de evaluar los modelos existentes para después elaborar recomendaciones de estandarización encaminadas a armonizar los trabajos que se realicen en el ámbito de la ingeniaria lingüística.

[2] Para evitar problemas de termonología, en lo sucesivo marcaremos los términos Parole en mayúsculas.

[3] Así, por ejemplo la descripción pasiva de un verbo transitivo consta de una construcción formada por un SN sujeto y un SP oblícuo opcional y de un elemento Self encargado de codificar el hecho de que en esa construcción el verbo toma la forma de participio.

SN/SV/OCOMP para la posición objeto de un verbo como *querer* puede recogerse en el nivel de Posición, en este caso obtendríamos una única Unidad Sintáctica con una Descripción cuya segunda Posición recogería la alternancia:

Ejemplo 1:
Querer1: descripción[SN, SN/SV/OComp]

Otra posibilidad, quizás la más distante a la anterior, consiste en estructurar la alternancia de patrones como la simple enumeración de Descripciones, en este caso tendríamos tres Unidades Sintácticas con una Descripción diferente cada una:

Ejemplo 2:
Querer1: descripción[SN,SN]
Querer2: descripción[SN,SV]
Querer3: descripción[SN,Ocomp]

A continuación mostramos con detalle y para cada categoría

- qué información se codifica
- cómo se ha estructurado y
- qué criterios se han aplicado en cada caso.

Verbos. La información sintáctica asignada a los verbos corresponde a las relaciones de subcategorización. Cada verbo lleva asignado un conjunto de patrones de subcategorización (o Descripciones) que especifican la categoría, el régimen preposicional, la función y aspectos morfosintácticos relevantes tanto de sus complementos (el modo de una oración completiva, las relaciones de control, etc.) como de la propia entrada verbal (voz, forma verbal etc.).

La organización de las diferentes Descripciones asignadas a una entrada verbal está estructurada en torno a lo que generalmente se entiende por diátesis. Esto es, cuando la alternancia de patrones de subcategorización (o Descripciones) para una unidad léxica no es particular de esta unidad sino que es común a un grupo bien definido de entradas, los patrones (o Descripciones) se agrupan en una misma Unidad Sintáctica relacionados mediante FrameSets[4] (o conjunto de Descripciones relacionadas).

Las alternancias de patrones de subcategorización (o diátesis) recogidas en el léxico mediante FrameSets son: activa/pasiva, causativa/decausativa (distinguimos aquellos verbos que pronominalizan al decausativizar de los que no lo hacen), alternancia entre SV/OCOMP (se incluye información sobre las relaciones de control y obviación y el modo de la cláusula completiva), y los predicados simétricos. Así, por ejemplo, un verbo como *querer* tiene asignadas dos Unidades Sintácticas diferentes. En una, la Descripción base corresponde a la estructura transitiva y la Descripción no-base corresponde a la estructura pasiva, ambas Descripciones están relacionadas mediante el correspondiente FrameSet. En la otra, la Descripción base corresponde a la estructura de control, mientras que la Descripción no-base corresponde a la estructura obviativa relacionadas vía FrameSet:

Ejemplo 3:
Querer1:
desc1[SN/SN] *desc2*[SN/(PPpor)]

Querer2:
desc1[SN/SV]
desc2[SN/OCOMP]

Este enfoque nos permite racionalizar el conjunto de Descripciones verbales en el sentido de que el léxico no modela la información como idiosincrasia léxica si no que la estructura. Una estrategia como la del ejemplo (1) prácticamente nos obligaría a definir una Descripción diferente para cada entrada léxica. Por el contrario, un enfoque 'estructurado' no sólo permite reducir el número de objetos del modelo sino que permite establecer generalizaciones.

[4] En el léxico español todas las descripciones de una misma unidad sintáctica deben estar relacionadas mediante un FrameSet.

El léxico español tiene 162 Descripciones verbales diferentes, 9 objetos Self y 164 Construcciones. Estos objetos nos han permitido codificar el comportamiento sintáctico de 3060 entradas verbales con un total de 7500 realizaciones posibles estructuradas en 4300 Unidades Sintácticas diferentes.

Adjetivos. La información sintáctica asignada a los adjetivos incluye:

(i) su naturaleza +/- predicativa (los adjetivos predicativos tienen una Posición sujeto correspondiente al argumento externo y llevan información sobre el tipo de verbo copulativo con el que puede aparecer).

(ii) su posición (pre/post-nominal) con respecto al nombre que modifican (ej: *el mero hecho...* vs *un chico listo)*

(iii) para el caso de los adjetivos predicativos, el tipo de verbo copulativo que requieren, *ser* o *estar*[5] (ej: *estar divorciado* vs *ser capaz*).

(iv) si admiten especificador de grado (*muy interesante* vs. **muy divorciado*). Los especificadores de 'grado' ocupan una Posición opcional.

(v) posibles complementos y preposición de régimen (*libre de impuestos*)

(vi) relaciones de control para aquellos adjetivos que subcategorizan SSVV (*fácil de hacer, harto de escuchar lo mismo*).

En el léxico español hemos codificado 5.000 entradas adjetivales.

Nombres comunes. Se han clasificado en argumentales, no-argumentales y aposiciones.

(i) Los nombres argumentales llevan información sobre sus posibles complementos (categoría, preposición de régimen...). Dado el estatus particular que los complementos nominales *de*+SN tienen en las lenguas románicas, hemos distinguido los complementos nominales introducidos por la preposición *de* del resto de complementos nominales. Así, los primeros se consideran SSNN 'marcados' y se les asigna la función NCOMPLEMENTO independientemente de la función sintáctica profunda que puedan desempeñar, ya sean sujetos u objetos de nombres de-verbales (*la aparición de Juan, la destrucción de la ciudad*), sujetos de nombres de-adjetivales (*la belleza de Juan)* o complementos de nombres 'colectivos' (*un grupo de estudiantes, tres kilos de peras*). El resto de complementos nominales se consideran SPP y se les asigna la función PCOMPLEMENTO.

(ii) Los nombres no-argumentales se han clasificado en contables, incontables y masa. Aunque esta clasificación no es estrictamente sintáctica hemos considerado oportuno incluirla dado que la naturaleza del nombre determina su comportamiento sintáctico. Los criterios formales utilizados para esta clasificación son los siguientes:

Nombres contables:
- Admiten la distinción singular/plural (*silla/sillas* vs. *paz/*paces víveres*).
- Requiren un determinante cuando están en singular (*se sentó en la silla* vs. *voy a por pan*).
- Pueden ser enumerados (*dos, tres, diez libros*).
- Con cuantificadores como *poco, bastante* o *nada de*, exigen la forma plural (*bastantes sillas* vs *bastante pan*).

Nombres incontables:
- No pueden ser enumerados: **dos paces*
- No admiten cuantificadores 'distribucionales' del tipo *cada, qualquier, ambos, cierto...*
- No admiten partitivos (**un buen trozo de paz/agricultura*).

Nombres de Masa:

[5] Cuando un adjetivo admite las dos construcciones, con *ser* y *estar*, se le asigna a la entrada dos descripciones diferentes.

- No admite enumeración —a menos que se recategorice— (**dos leches*).
- Con cuantificadores como *poco, bastante* o *nada de*, exigen la forma singular (*mucha leche, poco pan*).
- Tienden a evitar las formas plurales.
- Pueden aparecer en construcciones del tipo 'esto es N': *esto es petróleo.*
- Admiten partitivos: *dos litros de leche.*

Ciertos nombres tienen una lectura contable y una de masa: (*dos plátanos* o *pastel de plátano*). En este caso no hemos definido un tipo que podría llamarse 'variable' sino que asignamos dos Descripciones, una 'contable' y otra 'masa'. Este enfoque facilita la transportabilidad entre lenguas (ej. *peix* del catalán vs. *pez/pescado* en español).

Todos los nombres no-argumentales arriba mencionados llevan información sobre el tipo de determinante que admiten. La posición Determinante de las construcciones nominales únicamente recoge determinantes simples y de manera prototípica. Básicamente la información estriba en establecer el tipo de quantificador que admiten.

El léxico español contempla 39 Descripciones sintácticas diferentes para los nombres. Se han codificado un total de 11.500 nombres comunes.

Adverbios. La información asociada a los adverbios incluye:

(i) el elemento que modifican. Así, tenemos modificadores de nombre: *calle arriba*, de adjetivo: *muy alto*, de adverbio: *muy cerca*, de SV *llegó tarde*, de cláusula *evidentemente* y de participio: *totalmente cansado.*
(ii) su capacidad para ser modificados por adverbios de gradación (ej: *muy tarde vs *muy ayer*)
(iii) posibles complementos (únicamente para adverbios deadjetivales terminados en 'mente')
(iv) si admiten aposición (*hoy martes, ayer tarde*)

De ello se deriva que, contrariamente al caso de los verbos y nombres, las Construcciones utilizadas para definir el comportamiento de los adverbios van más allá de su proyección máxima. En toda Descripción adverbial, el adverbio es el elemento modificador de un SN, SV, SADV u O dependiendo de su función.

El léxico PAROLE español contiene 600 entradas adverviales con algo más de 750 comportamientos diferentes.

Determinantes. En el modelo Parole tenemos como categorías básicas: Artículo, Numeral y Determinante; y como subcategorías: Definido/Indefinido, Ordinal /Cardinal y Demostrativo/Posesivo/Indefinido respectivamente.

Así, en el léxico español tenemos: Art-Definidos, Art-Indefinidos, Num-Ordinales, Num-Cardinales, Det-Demostrativos, Det-Possesivos y Det-Indefinidos.

El grupo de Det-Indefinidos (o 'quantificadores') es muy heterogéneo, así que hemos creído oportuno subclasificarlos atendiendo al tipo (contable, incontable o masa) y número del nombre que especifican. Así, hemos obtenido la siguiente clasificación:

Det-Indef-Contable/Singular: (o 'quantificadores distribucionales') aquellos que requiren denotaciones nominales contables/singular: *ningún libro, cada chico.*

Det-Indef-Contable/Plural: aquellos que requiren denotaciones nominales contables/plurales: *ambos chicos, diversos libros*

Det-Indef/Contable: aquellos que requiren denotaciones nominales contables *cierto(s) / algún(s) / determinado(s) libro(s)*

Det-Indef/Masa/Singular: aquellos que requiren denotaciones nominales de masa en singular: *poco pan, mucho viento*

Todos los Artículos, Numerales y Determinantes son el elemento 'Self' de una Construcción de SN con almenos una Posición/núcleo ocupada por un N. El tipo de N que ocupa esta posición dependerá del tipo de Artículo, Numeral o Determinante de que se trate.

Los determinantes forman un grupo complejo e idiosincrático que admite pocas generalizaciones y difícilmente son 'transportables' a otras lenguas. Por consiguiente, la estrategia seguida para la organización de la información sintáctica en el caso de los determinantes es la de asignar tantas unidades sintácticas como patrones sintácticos tengamos reflejando así su carácter idiosincrático. Por ejemplo el Determinante-Indefinido *mucho* tiene asignadas tres unidades sintácticas diferentes de acuerdo con el tipo de sustativo que especifican y las posibilidades de co-aparecer con otros determinantes:

Ejemplo:
Mucho1: Contable/Pl: muchos niños
Mucho2: Contable/Pl: muchos (otros) niños
Mucho3: Masa/Sing: mucho pan

Este enfoque facilita la transmisión de información entre lenguas entre las cuales pude no haber una correspondencia directa:

Muchos niños	many boys
Muchos otros niños	many other boys
Mucho pan	much bread

Los determinantes pueden combinar entre si formando Sintagmas Determinantes complejos. En el léxico español los SSDD complejos se analizan como estructuras planas. Así, las Construcciones de determinantes complejos constan de tantas posiciones como determinantes siguen a la entrada en cuestión más el sustantivo al que determinan.

Así, por ejemplo quantificador universal TODO tiene asignada 7 Unidades Sintacticas diferentes dependiendo de la naturaleza del sustantivo que determina y del tipo de determinante con el que puede co-aparecer:

Todo1: Cont/Sing: todo hombre
Todo2: Cont/Plu: todos los (otros) niños
Todo3: Cont/Plu: todos esos (otros) niños
Todo4: Sing: todo el día, toda la leche
Todo5: Sing: todo ese día, toda esa leche
Todo6: Sing: toda una mujer

Anexo:

Unidades Morfológicas:

Verbos:	3060
Nombres comunes:	11500
Adjetivos:	5000
Adverbios:	600
Modos flexivos:	137

Unidades Sintácticas:

Verbos:	4277
Nombres comunes:	18854
Adjetivos:	5305
Adverbios:	756
Descripciones:	324
Construcciones:	265
Self:	54
FrameSets:	106

Bibliografía:

1994, Emilio Alarcos Llorach. *Gramática de la Lengua Española*. Espasa Calpe, Madrid.

1989, Ignacio Bosque. *Las Categorías Gramaticales*. Editorial Síntesis, Madrid.

1993, Ted Briscoe, Valeria de Pavia y Ann Copestake Editores. *Inheritance Defaults and the Lexicon*. Cambridge University Press.

1987, M. Luisa Hernanz y José M. Brucart. *La Sintaxis.*Editorial Crítica, S.A., Barcelona.
1986, *Esbozo de una Gramática de la Lengua Española.* (Real Academia Española). Espasa Calpe, Madrid.

1994, Monica Monachesi y otros. *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages.* EAGLES Report.

1995, Antonio Sanfilipo y otros. *Subcategorization Standars.Report of the EAGLES Lexicon/Syntax Group.*

1989, Manuel Seco. *Gramática Esencial del Español.* Espasa Calpe, Madrid.