El papel del diccionario en la traducción automática de compuestos

Marta Carulla
UPC

Introducción

La composición es un fenómeno creativo y productivo de la lengua que es especialmente relevante en el tipo de textos susceptibles de ser tratados por un sistema de traducción automática. Muchos compuestos estan lexicalizados, es decir, no tienen un significado composicional y como tales han de ser entrados en los diccionarios monolingües y bilingües. Pero a la vez existen muchos otros compuestos que claramente no están lexicalizados y que se crean de la misma manera que los sintagmas o oraciones nuevas. Especialmente en los textos técnicos y formales se emplean frecuentemente formaciones del momento (nonce-formations) para referirse a entidades particulares y a procesos en un dominio dado. El ejemplo que presentamos a continuación (tomado de una propuesta para una Decisión del Consejo en Investigación y Desarrollo) ilustra estos hechos:

> *European-based integrated circuit manufacturers* supply 30% of their *home market* and represent 13% of *world production*

La traducción de este fragmento del texto inglés a otras lenguas germánicas y a lenguas románicas supone una seria dificultad para la T.A., puesto que no hay una correspondencia de estructura:

> *In Europa gevestigde fabrikanten van geintegreerde schakelingen* voorzien 30% van hun *thuismarkt* en hebben een aandeel van 13% in de *wereldproduktie*

> *Los fabricantes de circuitos integrados de origen europeo* cubren el 30% de *su mercado nacional* y aseguran el 13% de *la producción mundial*

En la traducción neerlandesa observamos que, aparte de la diferencia obvia de que en esta lengua los compuestos forman una sola palabra ortográfica, *European-based* recibe una traducción sintagmática (*in Europa gevestigte*) al igual que *integrated circuit manufacturers (fabrikanten van geintegreerde schakelingen),* en que además el equivalente de *circuit* aparece en plural. En español *integrated circuit manufacturers* también corresponde a un sintagma (*los fabricantes de circuitos integrados)* en el que *circuit* también equivale a una forma plural *(circuitos)* , y en el que además aparece un determinante (***los*** *fabricantes...)* . Por otra parte en la traducción de *world production* se usa un adjetivo relacional (*producción mundial).*

Estos ejemplos, pues, ilustran algunos tipos de desajuste, tanto léxicos como estructurales relacionados con la traducción de compuestos.

Esta comunicación presenta propuestas concretas sobre la información lingüística necesaria para traducir estas estructuras compuestas y analiza en que medida la información ya existente en el léxico y la gramática para tratar estructuras sintácticas equivalentes satisface estas necesidades.

1. Información necesaria para la traducción de compuestos

En el marco de nuestro proyecto[1] y con la finalidad de establecer algunas de las características de la composición en las lenguas románicas y germánicas, llevamos a cabo un estudio empírico sistemático, recogiendo datos y evaluándolos posteriormente. Para ello tomamos un corpus de 23.000 palabras del campo de la tecnología de la información en tres lenguas: español, inglés y neerlandés[2]. El análisis sistemático tanto de los datos monolingües como contrastivo nos permitió establecer una tipología de compuestos productivos de cada lengua y sacar algunas conclusiones acerca de los patrones de traducción. Los apartados siguientes estan dedicados a las áreas de información lingüística que resultaron ser cruciales para la T.A. a partir de este trabajo de investigación.

1.1 Estructura argumental

Uno de los objetivos de nuestro trabajo era integrar el tratamiento de compuestos en el marco lingüístico especificado independientemente para estructuras sintácticas, efectuando el menor número de cambios o ampliaciones posible. El marco lingüístico, naturalmente, es específico de Eurotra (Copeland, 1991), pero creemos que el principio seguido puede ser aplicado también en otros contextos.

Las representaciones de interfície utilizadas se basan en estructuras de dependencia, consistentes en un elemento rector (o núcleo) y sus complementos. Estos últimos se dividen en argumentos y modificadores. En el caso de los compuestos Nombre + Nombre es esencial contar con una teoría de la estructura argumental de los nombres bien fundamentada. Los ejemplos que presentamos a continuación muestran algunas de las etiquetas empleadas[3]:

a. pipe-smoker — fumador de pipa
ARG2 — ARG2

b. staff exploitation of privileges
ARG1 — ARG2

---

[1]Este proyecto de investigación se incluye en el programa comunitario de continuación del proyecto Eurotra 91/92. En este contexto merecen mención especial Fernando Sánchez, Paul Bennett y Kerry Maxwell y también Lieve De Wachter y Dieter Maas. El trabajo que aquí se presenta es tan suyo como mio.

[2]Concretamente, la primera propuesta para la decisión del Consejo sobre el programa Esprit.

[3] Las etiquetas corresponden a la terminología tradicional de Eurotra para los argumentos. Así, p.ej., ARG1 corresponde nocionalmente al argumento externo de los nombres deverbales o al argumento clasificador de otros nombres con estructura argumental; ARG2 denota al argumento interno con función de objeto directo

explotación de privilegios por el personal
ARG2 ARG1

En ambos ejemplos el argumento realizado dentro del compuesto en inglés se realiza sintácticamente en castellano. Pero también en inglés puede existir esta posibilidad, como es el caso del ARG1 de b.:

c. exploitation of privileges by staff
ARG2 ARG1

Vemos, pues, que una misma relación con el núcleo puede estar realizada dentro del compuesto o en una estructura sintáctica. No se precisa ninguna entrada léxica especial para los nombres que aparecen en compuestos, ya que sus marcos de subcategorización son idénticos en ambos casos.

La estructura argumental es directamente relevante para la traducción de compuestos en aquellas situaciones en que el compuesto no corresponde a un compuesto en otra lengua. Este es el caso, p.ej., de los compuestos cuyo núcleo nominal denota una propiedad, partición o conjunto. Estos nombres subcategorizan un ARG1 o argumento clasificador. En neerlandés y alemán este tipo de nombres puede ocupar la posición de núcleo en un compuesto, sin embargo en inglés estos nombres suelen resistirse a la composición y su argumento se expresa mediante un sintagma preposicional con 'of':

bedrijfssnelheid	Ärzteteam
speed of industry	team of doctors
* industry speed	* doctor team

En castellano estos argumentos se expresan mediante la preposición 'de' (la velocidad de la industria, equipo de médicos). En estos casos la información necesaria para generar una estructura correcta tanto en inglés como en español se encuentra en la estructura argumental y en la entrada léxica de 'speed' (velocidad) y 'team' (equipo), que especifica mediante qué preposición se realiza el ARG1 en sintaxis superficial.

Como muestran los ejemplos anteriores, un compuesto de una lengua no corresponde necesariamente a un compuesto en otra lengua. El caso extremo de discrepancia se produce entre las lenguas germánicas y las lenguas románicas; nuestro estudio de corpus demostró que ninguno de los tipos de compuestos productivos del inglés o neerlandés se puede traducir por un compuesto en castellano. Como demuestra el ejemplo anterior estas discrepancias existen también entre las lenguas germánicas, aunque en menor grado. Una manera de tratar este fenómeno en un sistema que tiene como objetivo reducir la complejidad de los componentes de transfer y siguiendo el principio mencionado anteriormente, es adecuar las representaciones de los compuestos a las de las construcciones sintácticas. Concretamente, nuestra propuesta consiste en asignarles representaciones estructuralmente idénticas a las de sus correspondencias sintácticas y etiquetarlas con las mismas relaciones argumentales. Esta estrategia implica, entre otras cosas, que los nombres que forman el compuesto sean representados como sintagmas nominales. De esta manera b. y c. del inglés reciben la misma representación estructural, diferenciándose sólo en un rasgo que indica si la construcción se ha realizado superficialmente como un compuesto o no.

1.2 El tratamiento de los adjetivos relacionales

Los compuestos Nombre+Nombre en las lenguas germánicas equivalen frecuentemente a construcciones de nombre+adjetivo relacional en español:

technology goal — objetivo tecnológico

La relación de estos adjetivos con su nombre puede ser modificadora o argumental. En el caso de núcleos con estructura argumental los adjetivos relacionales pueden ocupar alguna de sus posiciones argumentales:

intervención comunitaria (ARG1) — Community intervention

producción automovilística (ARG2) — car production

En ambos casos, la información necesaria para calcular la relación entre el adjetivo y su núcleo se halla en el nombre que subyace al adjetivo (y que a menudo constituye su base derivativa). Así, en el segundo ejemplo son los rasgos semánticos del sustantivo *automóvil* los que determinan si se cumplen las restricciones seleccionales que *producción* impone a su argumento interno.

Dado que esta información ya existe en la entrada léxica del nombre correspondiente y que estos adjetivos se traducen frecuentemente por nombres parece natural tratarlos como nombres en el nivel de interfície. De esta manera las estructuras de nombre+adjetivo relacional reciben la misma representación estructural que otros compuestos nominales (véase sección anterior) y no es necesario tratar este caso de no correspondencia entre categorías sintácticas en el transfer.

El tipo de relación existente entre el núcleo y sus complementos (argumentos y modificadores) es básica para la traducción de estructuras compuestas. Para las lenguas estudiadas la manera en que se realizan superficialmente los argumentos nominales y la información que ha de contener el léxico están bastante bien entendidas.

1.3 Las relaciones semánticas modificadoras

Nuestro estudio reveló que el tipo de compuesto más corriente es el de nombre+nombre y a su vez de estos los más frecuentes son los compuestos en que el no-núcleo guarda una relación modificadora (es decir, no argumental) con su núcleo. En estos casos se pueden establecer muchos tipos de relaciones semánticas entre el no-núcleo y el núcleo, siguiendo propuestas teóricas que consisten en determinar un número finito de relaciones semánticas, como la de Levi (1978)[4]. Sin embargo el uso de teorías como esta para el tratamiento de compuestos es un tema controvertido. Los límites entre las relaciones no están siempre claros y a menudo las estructuras encajan en más de una relación. En su estudio experimental sobre la creación e interpretación de compuestos nuevos Downing (1977) demuestra que potencialmente cualquier tipo de relación semántica es

---

[4] En su propuesta Levi utiliza etiquetas semánticas com FOR para compuestos del tipo *cooking utensils*, USE para *steam iron*, ABOUT para *adventure story* ...

posible para estos compuestos en inglés. Su significado viene determinado por el contexto y el conocimiento del mundo, más que por la información lingüística, que es absolutamente indeterminada a este respecto. Por estas razones y teniendo en cuenta que se sabe muy poco sobre la relevancia de esta información para la traducción decidimos no representar la relación entre modificador y núcleo en estos términos.

1.4 La determinación del no-núcleo: genericidad y tipo de nombre

El no-núcleo de un compuesto no contiene ninguna marca morfosintáctica (artículo, número) que permita determinar su carácter referencial o genérico; aparece sin artículo y normalmente en singular. Aunque, de manera general, podamos asumir que los compuestos contienen relaciones genéricas, no siempre es así:

> The project began on the 1st November. The *project management* stipulated the deadlines for deliverables.

En este contexto está claro que el no-núcleo *project* del compuesto *project management* hace referencia específica al proyecto introducido en la oración anterior. Sin embargo y aunque la referencia del no-núcleo sea genérica en la mayoría de los casos, cada lengua la expresa de manera diferente. Compárese los siguientes compuestos ingleses con sus traducciones castellanas:

c. air circulation	circulación del aire
d. circuit complexity	complejidad de los circuitos

Los no-núcleos en posición de ARG1 siempre llevan artículo en español, tanto si se trata de un argumento externo de un nombre deverbal (c.) o de un argumento clasificador (d.)[5] El número depende también del tipo de nombre, así en el ejemplo c. *aire* es un nombre de referente único que obligatoriamente se expresa en singular, mientras que en el ejemplo d. se trata de un nombre contable que expresa la genericidad mediante el plural.

Cuando el no-núcleo es un nombre contable en posición de ARG2 (objeto profundo) está en plural y no lleva determinante:

factory automation	automatización de fábricas

Si el no-núcleo es no-contable tampoco lleva determinante y está en singular:

software production	producción de software

Sin embargo cuando se trata de referentes únicos, vuelve a aparecer del determinante para expresar genericidad, p.ej:

home automation	automatización del hogar

---

[5] Por argumentos clasificadores entendemos los argumentos de nombres simples y deadjetivales que llevan argumentos. Son nombres abstractos que denotan conjuntos, particiones o propiedades y tienen, por tanto, algún tipo de función clasificadora.

En conclusión, todo sistema que trate compuestos requiere un tratamiento de la determinación nominal, que permita identificar la interpretación genérica tanto de *home automation* como de *automatización del hogar*, aunque el determinante esté ausente en inglés y presente en castellano. No vamos a profundizar aquí en la determinación nominal del español, sólo remarcar que es absolutamente básico disponer de información sobre el tipo de nombre para generar la determinación correspondiente en castellano.

2 Requisitos para el léxico y la gramática

De nuestro trabajo resulta que el léxico y la gramática deben cumplir los requisitos siguientes para asegurar un tratamiento adecuado de las estructuras compuestas:

La estructura argumental: Puesto que los compuestos nominales reciben una representación estructural análoga a la estructura de los sintagmas nominales, se pueden analizar y traducir mediante los mismos mecanismos que las construcciones sintácticas correspondientes. De esta manera se eliminan los cambios estructurales en el transfer y se simplifican los diccionarios monolingües y bilingües. En ninguno de estos diccionarios hay que especificar separadamente los requisitos de subcategorización de una unidad léxica según aparezca en un contexto morfológico o sintáctico. Los diccionarios de transfer pueden reducirse, en el caso ideal, a simples proyecciones léxicas entre una lengua y otra, sin tener en cuenta tampoco dicho contexto.

Los adjetivos relacionales: Los adjetivos relacionales aparecen frecuentemente en la traducción de compuestos nombre+nombre. Su tratamiento como nombres aumenta la interlingualidad de la representación y, a la vez, simplifica el transfer y minimiza la redundancia en el léxico, ya que se usa la información presente en las entradas de sus nombres correspondientes. De esta manera se consigue también una mayor consistencia interna.

Las relaciones semánticas modificadoras: El grupo de compuestos nombre+nombre en que el no-núcleo modifica a su núcleo es el más numeroso. Seguramente la información semántica contenida en el léxico puede ser utilizada hasta cierto punto, pero el problema no parece ser predominantemente lingüístico.

Teoría de la determinación: El sistema debe incluir el cálculo de la determinación a partir de estructuras morfológicas. A diferencia de las estructuras sintácticas en los compuestos no existen marcas morfosintácticas (presencia o ausencia de determinantes y número) en el no-núcleo y el cálculo ha de basarse en las características semánticas del nombre, y en todos los demás aspectos que configuran la determinación en general (contexto sintagmático, oracional, discursivo ...).

3 Conclusiones

Una parte del proyecto presentado aquí se ha dedicado a la implementación experimental de compuestos para el alemán, neerlandés, inglés y español. Consecuentemente con nuestra estrategia se ha implementado en las gramáticas ya existentes para el análisis, transfer y generación, cuyo diseño respondía a las necesidades del tratamiento de estructuras sintácticas. La

experiencia ha demostrado que, a pesar de los problemas mencionados, el enfoque de extender los principios de las representaciones sintagmáticas es viable. Inevitablemente se han tenido que añadir reglas para el tratamiento de compuestos en las componentes de análisis y generación; pero no han supuesto cambios sustanciales en los módulos de transfer.

En cuanto al léxico, la estrategia seguida no ha supuesto ningún cambio importante, puesto que la información (sobre todo la semántica) ya existente en el léxico para el procesamiento de sintagmas ha demostrado ser igualmente válida para el procesamiento de compuestos. Principalmente la información sobre estructura argumental se ha explotado en las reglas que generan compuestos. Dada la prevalencia de compuestos en textos técnicos o semi-técnicos este resultado es especialmente satisfactorio.

Bibliografía

Carulla, M. (en prensa). Relational Adjectives: their characteristics and correspondences. In P.Alberto & P.Bennett eds: Lexical Issues in Machine Translation. CEC.

Copeland, C. &. J. D., S.Krauwer, B.Maegaard (Ed.). (1991). The Eurotra Linguistic Specifications. 1. Luxembourg: Commission of the European Communities.

Downing, P. (1977). On the Creation and use of English compound nouns. Language, 53, 810-42.

Braasch et al. (1993). Eurotra- Final Report on Compounding. Documento interno. CEC.

Levi, J. (1978). The Syntax and Semantics of Complex Nominals. . New York: Academic Press.