http://dx.doi.org/10.12795/PH.1997.v11.i02.01

TRADUCCIÓN AUTOMÁTICA DE TÉRMINOS CIENTIFICO-TÉCNICOS UNIDOS POR GUIÓN DE INGLÉS A ESPAÑOL[1]

Gabriel Amores
Rubén Chacón
Marina Martín
José Mª Montero
Patricia Román
Pablo Salas

This paper describes a strategy to translate hyphenated compounds from English into Spanish using JULIETTA, a Lexical-Functional Grammar-based machine translation prototype. A corpus of more than 500 hyphenated compounds has been used in order to establish a classification based on syntactic and functional criteria. Hyphenated compounds made out of more than two elements and those formed by nouns only have not been taken into account in our investigation. Our aim is to discover a general-purpose strategy that can be applied to all technical texts, regardless the semantic domain. Our results show that LFG offers a valid framework for the analysis and machine translation of hyphenated compounds from English into Spanish.

1. Introducción.

Una de las aplicaciones más comunes de los programas de traducción automática (TA, en adelante), consiste en la traducción de publicaciones de carácter cientifico-técnico. Se asume que estos textos tienen una finalidad primordialmente informativa, y, por tanto, evitan un uso figurado del lenguaje. Los sistemas de TA son, en principio, adecuados para traducir este tipo de textos ya que el lector especializado puede compensar con su conocimiento sobre la materia, los posibles errores gramaticales que el sistema haya

[1] La investigación contenida en este artículo se realizó como trabajo de grupo entre los autores, dirigidos por el Dr. Gabriel Amores como parte práctica del curso de doctorado "Introducción a la Traducción Automática", del Programa de doctorado en Lengua y Lingüística Inglesas del Departamento de Filología Inglesa (Lengua Inglesa) de la Universidad de Sevilla durante el curso académico 1995-96.

cometido. En contrapartida, el científico puede obtener de forma barata y rápida la información que precisa.

Entre los recursos más frecuentes del lenguaje técnico se encuentra la utilización de sintagmas nominales complejos, encaminados a conseguir un texto lo más informativo posible empleando los mínimos recursos lingüísticos. Esto es especialmente cierto en inglés, como nota Levi (1978:58):

> By the formation of complex nominals, a head noun and its underlying proposition can be reduced to just two words... complex nominal formation is a device of tremendous utility in English, if only for expanding the amount of information we can store in short term memory.

En este artículo nos centraremos en los sintagmas nominales complejos unidos por guión que **no** están formados por secuencias de N + N en inglés. Asimismo, descartaremos aquellos compuestos que han completado un proceso de lexicalización, y, por tanto, aparecerían como entradas en un diccionario técnico. Tal es el caso de compuestos como *full-track* (oruga), *far-sighted* (previsor), *single-engine* (monomotor), etc. Es decir, nuestro objetivo consiste en sistematizar los sintagmas complejos unidos por guión que se crean por la aplicación de reglas productivas del lenguaje, y, por tanto, no pueden predecirse. Dado que estos compuestos aparecen con frecuencia en los textos cientifico-técnicos, cualquier sistema de TA debe contar con una rutina específica para tratarlos y ofrecer una traducción lo más aceptable posible

Para tal fin hemos analizado un corpus aproximado de 500 términos entre las áreas científicas de la medicina, la genética, la aeronáutica y el lenguaje militar técnico. Pretendemos conseguir reglas generales aplicables al mayor número posible de áreas de conocimiento, aunque somos conscientes de que este es sólo un primer intento, que deja sin solucionar muchos casos particulares.

El análisis incluye una clasificación formal y funcional de los elementos que componen los compuestos, así como un análisis semántico que permita prever una traducción apropiada del compuesto independientemente del contexto en el que se encuentre.

2. Descripción del corpus.

Para la elaboración de este corpus se han utilizado libros y artículos de revistas especializadas en las materias mencionadas anteriormente. Para el área científica de genética y medicina se recopilaron unos 300 compuestos; y para el área técnica de aeronáutica y lenguaje militar técnico, unos 200 compuestos.

Hemos de tener en cuenta que la traducción tiene lugar entre dos lenguas de naturaleza considerablemente distinta. El inglés permite sintetizar la información de una cláusula en un compuesto de dos o más elementos unidos por guión, mientras que el español tiende a hacer más uso de cláusulas subordinadas. Esto plantea un problema en TA, ya que el texto destino ha de resultar natural al lector. Esta dificultad se resuelve mediante reglas de traducción que permitan adaptar los compuestos con guión en inglés a sintagmas nominales complejos en español.

3. Método de subcategorización.

Para conseguir un análisis apropiado se ha optado por una clasificación a dos niveles. El primer nivel aporta información sobre la categoría gramatical de cada uno de los elementos del compuesto, y el segundo, información semántica que nos ayude a descubrir las relaciones internas entre los elementos.

En el nivel puramente sintáctico consideramos las siguientes categorías gramaticales: sustantivo (N) ej. *acid-supplemented*; sustantivo en "-ing" (Ning); verbo en forma base (V); verbo terminado en "-ed" (Ved) ej. *food-borne*; verbo terminado en "-ing" (Ving) ej. *oil-degrading*; adjetivo (Adj) ej. *soft-tissue*; adjetivo terminado en "-ing" (Adjing); adverbio (Adv), preposición (P); prefijo (Pref) ej. *intra-aqueductal*; numerales que pueden ser cardinales (Card) ej. *two-step* u ordinales (Ord) y otra categoría a la que hemos denominado acrónimos (Acr) y que comprende todos aquellos términos que, o bien son siglas, o bien incluyen cualquier combinación de letras y dígitos, ej. *AIDS-defining*, *URA3-marked*.

En el nivel semántico hemos optado por el uso de rasgos semánticos del tipo animado, humano, tiempo, lugar, etc., y una generalización de los papeles temáticos más comunes (agente, paciente, instrumento, etc.). Nuestro objetivo ha sido estudiar hasta dónde se podría llegar con este tipo de información general, para evitar tener que introducir demasiada información semántica específica de cada campo en el programa de traducción, que se convertiría en un sistema de TA, para un sublenguaje determinado. Es decir, el sistema de TA al detectar un compuesto con guión, ha de tomar una decisión lo más acertada posible con los datos que posee. Con esto no queremos decir que la semántica y el conocimiento del mundo y del dominio específico no sean de utilidad. Todo lo contrario. Lo ideal sería que el programa de ordenador contara con esa información, pero es del todo imposible (al menos por ahora) formalizar absolutamente todo el conocimiento que sobre cada disciplina científica y técnica tienen todos los expertos del mundo y codificarlo para que sea utilizable en un tiempo razonable por el ordenador.

4. Reglas de traducción.

Tras recopilar un número apropiado de ejemplos se procedió a buscar una traducción lo más literal posible, manteniendo siempre la estructura de un sintagma nominal complejo. Esto se ajusta a los principios de economía del lenguaje y capacidad mnemotécnica propios del estilo científico-técnico.

Una vez estudiados los términos según el método de subcategorización explicado anteriormente y realizadas las correspondientes traducciones establecimos las bases para las reglas de traducción. Estas reglas han sido deducidas a partir de la información semántica y sintáctica de la lengua de origen que para este trabajo hemos considerado relevantes, ya que resultaría difícil codificar dicha información en su totalidad. A continuación pasamos a explicar con detalle las reglas de traducción obtenidas.

Compuestos N-Ving/Acr-Ving: El verbo en "ing" se traducirá por la forma adjetival del verbo en español seguida de la preposición que subcategorice el verbo. En el caso de que no subcategorice ninguna, se usará la preposición "de" por defecto. Si el sustantivo (N) es

contable, se traducirá en español por un plural, y si es incontable, en singular. En todo caso, la traducción pasará a ser la inversión del orden inglés ("N-Ving" => "Adj + Prep + Sust"). Ejemplos: *cholesterol-lowering* se traducirá como "reductor del colesterol".

Compuestos N-Ved/Acr-Ved: El verbo en "ed" se traducirá por la forma adjetival del verbo en español. Si el sustantivo (N) es un locativo y el verbo es de movimiento, se usará la preposición "a". Si es otro tipo de verbo, se utilizará "en". Si el sustantivo (N) es una sustancia, se usará la preposición "con". Finalmente, si el sustantivo no es ni locativo ni sustancia se utilizará "por". En todo caso, la traducción pasará a ser la inversión del orden inglés ("N-Ved" => "Adj + Prep + Sust"). Ejemplos: *site-directed* se traducirá como "dirigido al sitio"; *armor-protected* como "protegido por coraza" y *lipoprotein-associated* será "asociado con lipoproteinas".

Compuestos N-Adj/Acr-Adj: El adjetivo se traducirá por su equivalente español seguido de la preposición que subcategorice. En el caso de que no subcategorice ninguna, se usará la preposición "de". En todo caso, la traducción pasará a ser la inversión del orden inglés ("N-Adj" => "Adj + Prep + Sust"). Ejemplos: *glycine-rich* se traduce como "rico en glicina"; *species-specific* como "específica de especies".

Compuestos Adj-N/Adj-Ning/Adjing-N: Tanto el adjetivo como el sustantivo se traducirán por sus correspondientes en español y se realizará la inversión de los términos ("Adj(ing)-N(ing)" => " Sust + Adj"). Ejemplos: *short-wave* pasará a ser "onda corta"; *rolling-circle* pasará a ser "círculo rodante".

Compuestos Card-N: Se traducirán en el mismo orden. Si el cardinal es "uno", el sustantivo irá en singular, y si el cardinal es mayor que uno, irá en plural. Ejemplos: *one-third* pasará a ser "un tercio"; *two-stage* pasará a ser "dos fases".

Compuestos Adv-Ved: El verbo en forma "ed" se traducirá por la forma adjetival de verbo español seguido del adverbio terminado en "-mente". Ejemplo: *chromosomally-integrated* pasará a ser "integrado cromosómicamente".

Compuestos Adj-Ving/Adj-Ved: El verbo, ya sea en forma "ed" o "ing", se traducirá por la forma sustantiva de éste en español. Se realizará la inversión de los términos y se introducirá la preposición "de" precediendo a ambos. Ejemplo: *slow-migrating* se traduce como "de migración lenta".

Compuestos Adv-Adj: Se traducirán en el mismo orden. Ejemplo: *nearly-precise* pasará a ser "casi preciso".

5. Casos especiales.

A lo largo de nuestro análisis hemos encontrado numerosos ejemplos que necesitaban un tratamiento distinto de las reglas generales. Este tratamiento especial consiste en una traducción dependiente de la naturaleza del segundo elemento del compuesto unido por guión. Hemos incluido los siguientes casos: aquellos compuestos que contienen *ill* y *non* como primer elemento del compuesto, y aquellos que tienen *free* como segundo elemento del compuesto.

Para que nuestro sistema no aplique las reglas de traducción propuestas en los compuestos que contengan estas palabras, (lo que daría lugar a una traducción errónea) ha sido necesario implementarlo con un filtro. El diccionario general tendrá que incluir la posibilidad de una traducción diferente de estos tres términos cuando aparezcan en un compuesto con guión. A continuación examinaremos con detalle cada uno de los casos:

***Compuestos con* ill**: En un ejemplo como "*ill-defined*", la traducción resultante acorde a nuestras reglas sería "de definición enferma". Para evitar este error, cada vez que una entrada presente el elemento en cuestión, el filtro evitará que se aplique la regla de traducción general (en este caso la regla para "Adj-Ved"). En segundo lugar, cuando el programa busque la palabra "*ill*" en su diccionario, tomará la traducción especifica para compuestos con guión. La traducción resultante de "*ill*" será "mal", y la traducción de "defined" será la forma participial del verbo por defecto, resultando en la traducción "mal definido".

***Compuestos con* non**: Después de aplicarse el filtro, el programa encontrará dos posibilidades de traducción en su diccionario. Cuando el segundo elemento sea un adjetivo, *non* se traducirá por "no". Así, por ejemplo, *non-necrotic* resultaría en "no necrótico". Sin embargo, si el segundo elemento es una sustancia o una enfermedad, la traducción resultante sería "sin"; por ejemplo "*non-AIDS*" se traducirá por "sin SIDA".

***Compuestos con* free**: Una vez aplicado el filtro, la traducción que deberá tomar para compuestos con guión será "sin". Por ejemplo, *fat-free* pasará a ser "sin grasa".

6. Implementación.

El sistema de TA que hemos utilizado para implementar y verificar nuestras hipótesis ha sido JULIETTA, desarrollado en el Departamento de Filología Inglesa (Lengua Inglesa) de la Universidad de Sevilla (Amores 1992). El sistema se ha implementado en lenguaje de programación Prolog, y sigue los postulados lingüísticos de la Gramática Léxico-Funcional (LFG) (Bresnan, ed. 1982). En LFG, a cada oración se le asigna una doble representación sintáctica. De una parte, la estructura de constituyentes (estructura-c) muestra la configuración sintagmática de la oración, en forma de árbol sintáctico. Mediante unas anotaciones a las reglas de derivación (ecuaciones funcionales), la estructura de constituyentes genera un nivel de representación más abstracto, que contiene información funcional (sujeto, objeto, etc) en forma de una matriz de pares atributo-valor. A este nivel se le denomina estructura –f.

Desde el punto de vista de la TA, JULIETTA sigue un enfoque de transferencia (*transfer*). En un enfoque de transferencia, el sistema primero obtiene una representación de la oración en la lengua fuente. A continuación, transfiere esa representación a una equivalente en la lengua destino, y, finalmente, genera la oración en lengua destino a partir de dicha representación. Este enfoque encaja perfectamente con los dos niveles de representación propuestos en LFG (Netter 1986 y contra lo que opina Zajac 1990). Mientras que la estructura-c incluye información específica al idioma, y es, por tanto, descartada durante la transferencia, las relaciones gramaticales que encontramos en la estructura-f proporcionan una información más abstracta, ideal para ser manipulada durante la transferencia.

El sistema genera estructuras-c y -f para cada oración en lengua fuente. La transferencia se efectúa desde la estructura-f de la lengua fuente y produce una estructura-f en lengua destino, desde la cual se genera su correspondiente estructura-c. Este enfoque sigue las líneas propuestas por Kudo y Nomura (1986), en contra de lo que proponen Kaplan et al. (1989). Como ejemplo ilustrativo, presentamos a continuación el resultado de los tres módulos del sistema para una frase simple.

Entrada en inglés `The boy ate a cake.`

Estructura-c inglés

```
s(clh(cl(np(detp(det2(det(the))),
            n2(n1(n(boy)))),
         vph(vp(vg(v(ate)),
                np(detp(det2(det(a))),
                   n2(n1(n(cake))))))))
```

Estructura-f inglés

```
pred:eat([subj,obj])
num:sing
tense:past
obj:pred:cake
    spec:a
    count:yes
    role:theme
    agr:num:sing
subj:pred:boy
     spec:the
     count:yes
     role:ag
     agr:num:sing
```

Estructura-f español

```
pred:comer([subj,obj])
tense:past
num:sing
pronominal:yes
subj:pred:niño
     spec:el
     count:yes
     agr:gen:masc
         num:sing
obj:pred:pastel
     spec:un
     count:yes
     agr:gen:masc
             num:sing
```

Estructura-c español

```
o(prop(snh(sn(sres(res(el)),
              stbar(st(niño)))),
       sv(vbg(clt(se))
              vb(comió)),
          snh(sn(sres(res(un)),
                 stbar(st(pastel))))))
```

Salida en español `El niño se comió un pastel.`

7. Fase de análisis.

El primer paso en el proceso de traducción consiste en la identificación del compuesto, para lo que aprovecharemos el guión como identificador. Una vez que sabemos que se trata de un compuesto con guión, trataremos de identificar sus componentes sintácticos. Así, por ejemplo, la siguiente regla nos dice que un compuesto con guión (`hyph`) puede estar formado por dos palabras (`Palabra1 y Palabra2`), de las cuales, una debe ser un sustantivo (`lex(n,Palabra1,F1)`) y la siguiente un verbo en participio presente (`lex(v,Palabra2,F2,partpres)`). Si es así, añade el rasgo (`hyph:nving`) para identificar el tipo de compuesto con guión. Esta información será utilizada por el módulo de transferencia. Nótese además, que la relación entre el sustantivo y el participio presente es la de ser su objeto, tal como indica la última línea del código de esta rutina: (`unify(F3,[hyph:nving,obj:F1],Fh)`.

```
% hyphenated compounds
hyph(hyph(Palabra1,Palabra2),Fh) --> [Palabra1,-,Palabra2],
{
      /*--          N          Ving          hyphens          ---*/
lex(n,Palabra1,F1),lex(v,Palabra2,F2,partpres),
      delfeat(num:_,F2,F3),
      unify(F3,[hyph:nving,obj:F1],Fh)}
```

Con este tipo de información, el sistema obtendría el siguiente análisis para el compuesto *'temperature-sensing device'*

```
pred:device
count:yes
hyph:pred:sense
     hyph:nving
     part:pres
     ggf:[subj,obj]
     obj:pred:temperature
          count:no
          role:theme
          agr:num:sing
               per:three
agr:num:sing
    per:three
```

8. Transferencia.

Una vez que hemos identificado y clasificado el compuesto adecuadamente, se pasa a la fase de transferencia. Durante esta fase hemos de aplicar alguna de las reglas de traducción descritas anteriormente. Sabremos cuál hemos de usar dependiendo del tipo de compuesto que hayamos identificado durante el análisis. En el ejemplo que estamos considerando, el rasgo pertinente es `hyph:nving`.

La regla de transferencia se compone de una serie de condiciones que se han de satisfacer y una serie de acciones que se han de llevar a cabo si las condiciones tuvieron éxito. Esta regla de transferencia es de tipo estructural. Esto es, cambia la configuración interna de la estructura-f proveniente del inglés. Así, el rasgo `hyph` se transferirá como `mod`(ificador), y el objeto se transferirá como `pmod` (postmodificador), unidos ambos por la preposición **de** (`pcase:de`). Además, se incluirá el rasgo `vnom:yes`, que se usará durante la fase de generación, como veremos en la siguiente sección. La regla de traducción que estamos aplicando requiere también que el participio se genere como agentivo. De ello se ocupa la función `agent_adj(C1,Agadj)`, que devuelve la forma participial de un verbo cualquiera. Aplicada al verbo **sentir**, dará como resultado **sensor**. Esta operación es posible ya que el módulo de transferencia estructural se aplica tras el de transferencia léxica, y las palabras del inglés han sido sustituidas por sus equivalentes en español.

```
/*-- structural transfer for hyphenated compounds --*/
/* --- Nving hyphens --*/
strtrf([hyph:V|Rest],[mod:[vnom:yes|V2]|Result],Model) :-
       member(hyph:nving,V),
       member(pred:C1,V),
       member(obj:O,V),
       agent_adj(C1,Agadj),
       replace(obj:O,pmod:[pcase:de|O],V,V1),
       replace(pred:C1,pred:Agadj,V1,V2),
       strtrf(Rest,Result,Model).
```

Una vez que se ha aplicado la regla de traducción, obtendremos el siguiente resultado para el ejemplo anterior.

```
pred:dispositivo
count:yes
agr:per:three
    gen:masc
    num:sing
mod:pred:sensor
    vnom:yes
```

```
hyph:nving
asp:imperf
pmod:pred:temperatura
     count:no
     pcase:de
     agr:per:three
         gen:fem
         num:sing
```

9. Generación.

La última fase del proceso de traducción convierte la estructura-f transferida a un árbol sintáctico del español, que contiene las palabras totalmente flexionadas y en el orden establecido por la gramática del español. Para el ejemplo anterior, el resultado será el siguiente, seguido de la cadena de palabras final.

```
o(prop(snh(sn(stbar(st(dispositivo),
                sadj(adje(sensor),
                   sp(pre(de),
                      snh(sn(stbar(st(temperatura))))))))))
Dispositivo sensor de temperatura.
```

10. Resultados.

Una vez implementadas las reglas de traducción descritas, obtuvimos los siguientes resultados para una batería de ejemplos.

```
| ?- try_all(114,136).Sentence 114  temperature-sensing device.
  Dispositivo sensor de temperatura.
Sentence 115  radiation-suppressing device. Dispositivo supresor de
  radiaciones.
Sentence 116  wing-mounted device. Dispositivo montado en ala.
Sentence 117   absorber-coated device. Dispositivo revestido de
  absorbente.
Sentence 118  voice-actuated device. Dispositivo accionado por la
  voz.
Sentence 119  site-directed device. Dispositivo dirigido al sitio.
Sentence 120  starvation-regulated treatment. Tratamiento regulado
  por hambre.
Sentence 121  rain-resistant device. Dispositivo resistente a la
  lluvia.
Sentence 122  base-specific compound. Compuesto específico de base.
Sentence 123  single-electron atom. Átomo de un sólo electrón.
Sentence 124  steady-state patient. Paciente de estado estable.
Sentence 125  sonic-speed aircraft. Avión de velocidad sónica.
Sentence 126  two-seat aircraft. Avión de dos plazas.
Sentence 127  one-step division. División de un paso.
Sentence 128  chromosomically-encoded feature. Rasgo
  cromosómicamente codificado.
```

```
Sentence 129     perfectly-tightened nut. Tuerca perfectamente
   apretada.
Sentence 130  slow-migrating cell. Celula de migración lenta.
Sentence 131   quick-retracting device. Dispositivo de retracción
   rápida.
Sentence 132  nearly-precise survey. Estudio casi preciso.
Sentence 133  cholesterol-free diet. Dieta sin colesterol.
Sentence 134  ill-planned development. Desarrollo mal planeado.
Sentence 135  non-cholesterol diet. Dieta sin colesterol.
Sentence 136  non-necrotic illness. Enfermedad no necrótica.
yes
```

11. Conclusión.

En este artículo hemos ofrecido un primer acercamiento a la traducción automática de compuestos con guión entre el inglés y el español. Hemos usado el sistema JULIETTA para verificar las reglas de traducción que se han diseñado. Los resultados parecen indicar que un sistema de TA basado en LFG ofrece un marco adecuado para representar y transferir compuestos con guión entre inglés y español. Sin embargo, somos conscientes de que nuestra propuesta no es definitiva, ni abarca todos y cada uno de los casos en todas sus posibilidades. Así, por ejemplo, solamente se tratan compuestos formados por dos elementos, cuando, en realidad, pueden aparecer compuestos con tres y hasta cuatro elementos. No tratamos la coordinación de compuestos o coordinación en la premodificación al núcleo nominal, lo que dificulta enormemente la traducción. La integración de compuestos con guión con otros compuestos con guión tampoco ha sido tenida en cuenta.

Por otro lado, no está claro cuándo un participio presente se debe traducir por su equivalente en español o por una oración de relativo, etc.

BIBLIOGRAFÍA

G. Amores, *A Lexical-Functional Grammar-based Machine Translation System for Medical Abstracts.* Tesis Doctoral (Universidad de Sevilla, 1992).

J. Bresnan ed, *The Mental Representation of Grammatical Relations.* (Cambridge, MA 1982).

A. de Urquia Gómez, *Diccionario Técnico Militar Inglés-Español, Español-Inglés* (Madrid, 1990).

R. Kaplan, K. Netter, J. Wedekind, y A. Zaenen, "Translation by Structural Correspondences". *4th Conference of the European Chapter of the ACL* (Manchester, UK 1989) 272-281.

I. Kudo y H. Nomura, "Lexical-Functional Transfer: A Transfer Framework in a Machine Translation System based on LFG", *Proceedings of COLING 1986* (Bonn, 1986) 112-115.

J. N. Levi, *The Syntax and Semantics of Complex Nominals* (Londres, 1978).

C. Morales Pérez, "Claves para la decodificación de grupos nominales complejos en el inglés científico-técnico". *Actas del IX Congreso Nacional de AESLA*. F. Etxeberria y J. Arzamendi (eds.) (Bilbao,1992) 445-455.

K. Netter, "Getting Things out of Order: An LFG-Proposal for the Treatment of German Word Order", *Proceedings of COLING 1986* (Bonn, 1986) 494-496.

F. Salager, "Compound Nominal Phrases in Scientific Technical Literature: proportion and Rationale", A.K. Pugh y U.M. Ulijn (eds.), *Readings for Professional Purposes: Studies in Native and Foreign Languages*. Heinemann, (Londres, 1983) 136-145.

F. Salager, "Syntax and Semantics of Compound Nominal Phrases in Medical English Literature: A Comparative Study with Spanish", Manuscrito no publicado. Departamento de Lenguas Modernas. Universidad de los Andes. (Mérida,Venezuela 1980)

R. Zajac, "A Relational Approach to Translation", *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language* (Austin, TX 1990) 235-255.