

El uso de rasgos en el tratamiento de las dependencias no acotadas en un sistema de traducción automática

Sergio Ballari y Elizabet Gilboy
Eurotra- Barcelona

Resumen

En este artículo hemos querido demostrar que, dada la naturaleza del formalismo de Eurotra, de los distintos tratamientos de las DnA que nos proporcionan las diversas teorías lingüísticas actuales, parece más natural adoptar un enfoque próximo a modelos como GPSG o HPSG basados en una teoría elaborada de los rasgos y las categorías sintácticas. Asimismo, hemos querido mostrar que toda la aportación de las teorías lingüísticas a un sistema de TA acaba allí donde empieza la teoría de la traducción. En nuestro caso, hemos señalado la necesidad de complementar un análisis de las DnA basado en la GPSG con un mecanismo de indexación que permita un proceso de transferencia adecuado. Es imprescindible, pues, distinguir dentro de un sistema de TA aquellos mecanismos y supuestos teóricos que pertenecen a la teoría lingüística del sistema de aquellos que pertenecen a la teoría de la traducción.

DEPENDENCIAS NO ACOTADAS Y TRADUCCION AUTOMATICA

Uno de los aspectos más complejos del análisis de las lenguas naturales, tanto para el lingüista teórico como para el lingüista computacional, es el tratamiento de las dependencias no acotadas (en adelante, DnA). Generalmente, una DnA se define como una construcción en la que a) se establece algún tipo de relación sintáctica entre dos subestructuras, y b) la distancia estructural entre ambas subestructuras no está restringida a un dominio finito, por ejemplo, cuando ambas estructuras no se encuentran en la misma cláusula (Gazdar et al., 1985). Así, las topicalizaciones, las oraciones de relativo, las oraciones interrogativas, etc. son construcciones de este tipo.

Los problemas que plantean este tipo de construcciones son básicamente dos: que la distancia que separa a las dos subestructuras es arbitrariamente larga, y que la relación sintáctica se establece, por lo general, entre un elemento léxicamente realizado y una categoría vacía. La existencia de estas categorías vacías se justifica principalmente por criterios de subcategorización.

En el ámbito de la traducción automática (TA), el análisis de las DnA plantea algunos problemas adicionales. Por un lado, discrepancias estructurales entre lenguas, como las de (1), donde un elemento vacío en una lengua puede traducirse por un pronombre léxico en la otra:

- (1)
a. Who_i did he give ____j the book?
b. ¿A quién_i le_j dio el libro?

Por otro lado, variaciones en cuanto a los rasgos de concordancia que comparten los elementos relacionados:

- (2)
a. Die Part_{Fem}ie, die me_{Fem}ine Frau ___ wäh_{Fem}lte.
Sing Sing Sing
b. El part_{Masc}ido al que votaba mi mu_{Masc}jer ___
Sing Sing Sing

- (3)
a. The house in which I live ___
Sing Sing Sing
b. La casa en la que vivo ___
Fem Fem Fem
Sing Sing Sing

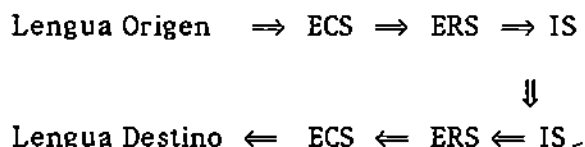
En todo sistema de TA se pueden distinguir tres etapas: el análisis, la transferencia y la generación. En la fase de análisis es preciso determinar qué subestructuras están relacionadas, mientras que en las fases de transferencia y generación es preciso preservar esta relación.

Lo que sigue es la descripción del enfoque que hemos adoptado para resolver el problema de las DnA en Eurotra.

EL SISTEMA EUROTRA

El sistema Eurotra es un sistema de traducción multilingüe basado en la estrategia de transferencia. La idea principal es la de reducir la distancia entre lenguas llegando a una representación canónica básica (Interface Structure, IS) que reduzca el componente de transferencia a la traducción de elementos léxicos, entendidos como colecciones de rasgos.

Esta representación es el último de los tres niveles sintácticos en los que se organiza la fase de análisis y el primero de la fase de síntesis, según el siguiente esquema:



El nivel ECS (Eurotra Configurational Structure) utiliza información puramente estructural. En el nivel ERS (Eurotra Relational Structure) la información estructural deja de tener relevancia en favor de la información funcional¹

EL USO DE RASGOS EN EL TRATAMIENTO DE LAS DNA

La mayoría de las teorías lingüísticas, transformacionales o no, comparten una estrategia común en el momento de abordar las DnA: reducir una DnA a una serie de relaciones acotadas susceptibles de ser tratadas localmente². En el modelo transformacional, el "problema de la localidad" se resuelve aplicando la regla transformacional Move α de forma cíclica sucesiva, insertando huellas intermedias coindexadas en todos los nudos COMP que cruza el movimiento (o, en versiones más recientes, en todas las posiciones de especificador de SC). Es decir, la estructura resultante muestra una cadena de elementos vacíos coindexados, cuya cabeza es el antecedente y cuya cola es la categoría vacía que ocupa la posición desde donde se ha iniciado el movimiento. Los índices son los encargados de asegurar la identidad de rasgos entre el elemento vacío y su antecedente.

¹ Por problemas de espacio no nos extenderemos en la descripción del sistema; para una descripción más exhaustiva del mismo véase Hutchins, W.J. (1986), *Machine Translation: Past Present and Future*, Ellis Horwood, Chichester, pp. 264-271.

² El concepto de "localidad" varía en función del marco teórico; por ejemplo, la localidad en el modelo GB es menos restrictiva que la localidad en el modelo GPSG. Véase p.ej. Chomsky (1981b) y Gazdar et al. (1985).

Efectivamente, la coindexación fue introducida por la Gramática Transformacional para cumplir dos funciones distintas: 1) para establecer la identidad de rasgos entre un sintagma desplazado y su huella; 2) para relacionar dos sintagmas correferenciales. Tal distinción se establece explícitamente en Chomsky (1980), donde se asumen dos tipos de índices: los referenciales y los anafóricos. Los primeros se asignaban de abajo a arriba en el árbol por la aplicación cíclica y sucesiva de la regla Move α , mientras que los segundos se asignaban de arriba a abajo por una regla especial de coindexación³. Nótese que dos reglas distintas en dos niveles de representación distintos se encargaban de asignar los índices: Move α , una regla transformacional, y coindexación, una regla interpretativa. Así pues los índices tenían una interpretación distinta en la estructura-S y en la FL. La coindexación de sintagmas en la estructura-S era necesaria para asegurar la identidad de rasgos. La coindexación en la FL era necesaria para la aplicación de las reglas interpretativas y las relaciones de ligamiento. Posteriores desarrollos del modelo han permitido adoptar un enfoque unificado de la coindexación, la cual se interpreta siempre en términos de ligamiento, es decir, como una relación entre un antecedente y una anáfora o un antecedente y una variable. Sin embargo, sigue asumiéndose una cierta diferencia entre los índices de la estructura-S y los de la FL (Chomsky, 1981a y 1981b).

Sin embargo, la Gramática Transformacional carece de una teoría de rasgos desarrollada (véase Muysken & van Riemsdijk, 1986, donde se aborda por primera vez el problema). Además, no existe en este modelo nada parecido a una definición de unificación ni a un conjunto de principios de instanciación de rasgos. Así, el modelo transformacional sólo dispone del mecanismo de coindexación para asegurar la identidad de rasgos entre dos sintagmas relacionados.

Esto no ocurre sin embargo en las teorías y formalismos basados en la unificación como GPSG, LFG, HPSG, FUG, PATR-II, etc. En general, estos modelos conciben las categorías sintácticas como conjuntos o estructuras de rasgos, y utilizan operaciones de combinación de información tales como la unificación. Los índices en estas teorías generalmente no tienen estatus teórico, ya que son innecesarios, al menos por lo que a la identidad de rasgos en las DnA respecta. Ninguna de estas teorías utiliza los índices en el tratamiento de las DnA, aunque el estatus teórico de éstos varía. En GPSG no se usan en ningún caso los índices, mientras que en HPSG y en LFG sí, pero únicamente para representar la correferencialidad semántica (Gazdar, 1981; Gazdar et al., 1985; Pollard & Sag, 1987; Kaplan & Bresnan, 1982). A veces, los índices se utilizan para representar gráficamente la unificación, pero, obviamente, esto es otra cuestión.

³ El proceso es un poco más complejo, ya que algunos índices referenciales podían asignarse de arriba a abajo mediante la regla de coindexación a todos aquellos SSNN que todavía carecieran de índice.

Los mismos motivos que en estas teorías hacen innecesario el uso de índices, también hacen innecesario postular la existencia de una cadena de huellas intermedias entre el elemento vacío y su antecedente. En GPSG, por ejemplo, donde sólo se acepta un nivel de representación sintáctica, el análisis de las DnA se basa en una serie de principios de instanciación de rasgos y en la operación de unificación. La presencia de un elemento vacío en una construcción se codifica en el rasgo SLASH, cuyo valor es una categoría, que se va copiando hacia arriba en el árbol hasta encontrar el antecedente. Aquí, los rasgos del antecedente se unifican con el valor de SLASH. En este caso, el "problema de la localidad" se resuelve asegurando la presencia del rasgo SLASH en todos los árboles locales intermedios, creando así una "cadena" desde el elemento vacío hasta el nudo en el árbol donde se produce la unificación.

Eurotra puede incluirse dentro de este segundo grupo de teorías y formalismos. Efectivamente, Eurotra es un formalismo donde las categorías sintácticas son conjuntos de rasgos. Dispone además de una serie de reglas (reglas-f) que actúan como principios de instanciación. Así, todas las consideraciones anteriores sobre los índices y las cadenas de huellas intermedias pueden aplicarse en nuestro caso. No hay que olvidar, sin embargo, que Eurotra es un sistema de TA. Cuanto hemos dicho vale, pues, sólo para la fase de análisis.

En un sistema de TA no sólo es relevante la información de que un elemento vacío y su antecedente comparten una serie de rasgos, sino también la relación de dependencia que existe entre ambos. Es decir, en el proceso de traducción es esencial indicar de algún modo que cualquier variación en los rasgos del antecedente repercutirá en los rasgos de la categoría vacía. Véase, por ejemplo (2) y (3). Así pues, la coindexación será necesaria para conservar esta relación en el paso de una lengua a otra. Nótese sin embargo, que en este caso los índices no pertenecerían a la teoría lingüística de Eurotra, sino a su teoría de la traducción.

DESCRIPCION DE LA IMPLEMENTACION: LAS ORACIONES DE RELATIVO

En este apartado describiremos la implementación de un análisis de las oraciones de relativo basado en los principios teóricos expuestos anteriormente. Las DnA se resuelven en los dos primeros niveles de representación (ECS y ERS). En el primero se detecta configuracionalmente la oración de relativo, se codifica toda la información en un conjunto de rasgos que se copian de arriba a abajo en todos los nudos oracionales. En la ERS, a partir de la información proveniente del nivel anterior, se inserta un elemento vacío utilizando la información funcional codificada en los rasgos de subcategorización del verbo.

ECS

La principal función del componente ECS es detectar configuracionalmente la presencia de una oración de relativo. Se trata de obtener una representación que contenga la información de que estamos construyendo una oración de relativo y de que en algún lugar por debajo del nudo O falta algún elemento. Para ello utilizamos las dos reglas siguientes:

```
cSBAR2 = (cat=sbar, type=rel) [
    (cat=sn, lu=L, type=rel),
    (cat=o, form=fin, slash=sn, antlu=L)].
```

```
cSBAR3 = (cat=sbar, type=rel) [
    (cat=sp, lu=L, type=rel),
    (cat=o, form=fin, slash=sp, antlu=L)].
```

Estas reglas construyen SBARs que dominan inmediatamente a un SN o un SP y a una O. El SN y el SP tendrán el rasgo tipo=rel en virtud de la asignación léxica de este rasgo a todos los pronombres relativos. El rasgo tipo=rel está también presente en el nudo superior SBAR. La O recibe un valor para los rasgos slash y antlu, donde se codifica la categoría y el valor para el rasgo de unidad léxica (lu) del sintagma relativo hermano. Como se verá más adelante, el rasgo antlu juega un papel crucial en el momento de determinar la función sintáctica de una categoría vacía.

Seguidamente, es preciso determinar los rasgos de concordancia que deberá recibir el elemento vacío. Como hemos visto, éstos dependen del N modificado por la oración de relativo. El primer paso es el de asegurar que el N y el sintagma relativo concuerden en género y número; después, codificamos estos rasgos como antgn y antnum (léase género del antecedente y número del antecedente, respectivamente) en el nudo O hermano del sintagma relativo. El siguiente paso es el de asegurar que este conjunto de rasgos aparezca en todos los dominios locales pertinentes. En nuestro caso hemos adoptado una noción de localidad equivalente a la de la GB, es decir, los rasgos se copian en todos los nudos O por debajo de SBAR mediante las reglas-f siguientes:

```
fREL1 = (slash=S, antlu=L, antnum=N, antgn=G, antpform=P/cat=o) [
    *(),
    (slash=S, antlu=L, antnum=N, antgn=G,
    antpform=P/cat=o),
    *() ].
```

```
fREL2 = (slash=S, antlu=L, antnum=N, antgn=G, antpform=P/cat=o) [
    *(),
    (cat=?) [
        *(),
        (slash=S, antlu=L, antnum=N,
        antgn=G, antpform=P/cat=o),
        *() ],
    *() ].
```

En resumen, en la ECS simplemente codificamos la información sobre el tipo de construcción que estamos tratando y sobre el tipo de elemento vacío que debemos encontrar.

ERS

En este nivel todas las estructuras que poseen el rasgo slash se construyen mediante un conjunto especial de reglas. En estas reglas se utiliza el operador ! para indicar los nudos que deberán insertarse. Para insertar correctamente un elemento vacío se recurre a información de subcategorización y a los rasgos cat y antlu codificados en los nudos O:

```
cInt2rell = {cat=o, slash=sn, antlu=que, antnum=N, antgn=G} [
  {fs=gov, cat=v, vtype=vi2},
  ! {fs=subj, cat=sn, num=N, gn=G},
  ^ {fs=iobj, cat=sp, pform=a};
  {fs=iobj, cat=cli} ),
  ^{fs=pcomp, cat=sp},
  *(fs=mod) ].
```

```
cTransrellb = {cat=o, slash=sn, antlu=que, antnum=N, antgn=G} [
  {fs=gov, cat=v, vtype=vt},
  {fs=subj, cat=sn},
  ! {fs=obj, cat=sn, num=N, gn=G},
  ^{fs=iobj, cat=sp},
  ^{fs=pcomp, cat=sp},
  *(fs=mod) ].
```

La idea principal es que, por ejemplo, si tenemos slash=SN y antlu=que, y el núcleo de la construcción es un verbo intransitivo, el elemento vacío sólo puede ser el sujeto. Es decir, la información sobre categoría y lu, y el tipo de verbo (transitivo, intransitivo) se combinan para identificar la función sintáctica del elemento vacío. El resultado final es una estructura en la que se ha insertado un elemento vacío cuyos rasgos de concordancia son los mismos que los de su antecedente. Nótese que, por el momento, no utilizamos índices en nuestras representaciones. Ello no se debe a una incongruencia en nuestros planteamientos, sino a que la función \$index encargada de asignarlos automáticamente se halla todavía en fase de experimentación.

Bibliografía

Balari, S. & E. Gilboy (1988), "A feature-based approach to coindexation in Eurotra" en Allegranza, V. & E. Steiner, *Papers on Interlevel Phenomena*, Comisión de las Comunidades Europeas, Proyecto Eurotra, Luxemburgo.

Chomsky, N. (1980), "On binding", *Linguistic Inquiry*, 11, 1-46.

Chomsky, N. (1981a), "Markedness and core grammar", en Belletti, A., L. Brandi & L. Rizzi, *Theory of Markedness in Generative Grammar*, Actas del 1979 GLOW Colloquium, Scuola Normale Superiore, Pisa.

Chomsky, N. (1981b), *Lectures on Government and Binding*, Dordrecht, Foris.

Gazdar, G. (1981), "Unbounded dependency and coordinate structure", *Linguistic Inquiry*, 12, 155-184.

Gazdar, G. et al. (1985), *Generalized Phrase Structure Grammar*, Basil Blackwell, Oxford.

Kaplan, R. & J. Bresnan (1982), "Lexical Functional Grammar: A formal system for grammatical representation", en Bresnan, J. ed., *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, Mass.

Kaplan, R., J. Maxwell & A. Zaenen (1987), "Functional uncertainty", Technical Note, Xerox PARC, Palo Alto, California.

Muysken, P. & H. van Riemsdijk (1986), *Features and Projections*, Foris, Dordrecht.

Pollard, C. & I. Sag (1987), *Information-based Syntax and Semantics*, CSLI Lecture Notes, 12, Stanford University, Stanford, California.