Flexión nominal: un problema de contexto

Antonio Moreno
Dpto de Lingüística
Facultad de Filosofía y Letras
Universidad Autónoma de Madrid

Resúmen

Este trabajo trata sobre un problema concreto de morfología computacional, la flexión nominal, enfocado desde el marco teórico del proyecto EUROTRA, sistema de traducción automática basado en el transfer. Este sistema descompone las relaciones de traducción, tanto para el análisis como para la generación, en distintos lenguajes de representación o niveles que describen composicionalmente las estructuras morfológicas, sintácticas y semánticas de los objetos que van a ser traducidos.

Introducción

Este trabajo tratará sobre un problema concreto de morfología computacional, enfocado desde el marco teórico del proyecto EUROTRA, sistema de traducción mecánica basado en el transfer. Este sistema descompone las relaciones de traducción, tanto para el análisis como para la generación, en distintos lenguajes de representación o niveles que describen composicionalmente las estructuras morfológicas, sintácticas y semánticas de los objetos que van a ser traducidos.

El propósito de esta comunicación es hacer una observación a las conclusiones del **Seminario sobre Traducción Automática y Lingüística Computacional**, organizado por la **U.I.M.P.** en abril de 1988. Allí se dijo que el problema de la traducción por ordenador no es tanto una cuestión computacional como una falta de un mejor conocimiento del lenguaje natural por parte de los lingüistas. Creo que la flexión nominal aporta un contraejemplo de cómo fenómenos bien explicados por la morfología teórica no son resueltos de una manera completamente satisfactoria, como cabría esperar, por la lingüística computacional.

Flexión nominal: un problema de contexto

La flexión nominal en español presenta dos tipos de fenómenos relacionados estrechamente entre sí, pero con un comportamiento morfológico distinto. Se trata del género y del número. Me referiré en primer lugar al número.

Es bien conocido que el morfema de plural es único, /s/, y que se realiza por medio de tres alomorfos: [-es], [-s] y [0].

1. [-es]: camión, -es; gris, -es; revés, -es; jabalí, -es (1)
2. [-s]: niño, -s; estudiante, -s; jabalí, -s; casa, -s
3. [0]: crisis, virus, lunes.

Es posible dar cuenta de esta cuestión con una regla de contexto como la que sigue:

/s/ --> [es] / (V) C ___#

/s/ --> [s] / V ___#

/s/ --> [0] / V C V s ___# (2)

donde
V es vocal
C es consonante
mayúsculas y /s/ son elementos no terminales
minúsculas son elementos terminales

Vemos, por tanto, que la alomorfía es una cuestión explicable dentro de una gramática contextual. Pero una de las limitaciones -por motivos prácticos(3)- que tenemos no solo en EUROTRA sino en la mayoría

de los proyectos de MT (**Machine Translation**) y de procesamiento de lenguaje natural (**Natural Language Processing** **NLP**) es que se utiliz como modelo teórico una gramática independiente del contexto (**Context Free Grammar**). Se ha propuesto en EUROTRA utilizar para la morfología otro tipo de gramáticas como son las de estados finitos, concretamente, la morfología de Koskenniemi. Pero la complicación que acarrearía tener que trasladar la información de un modelo a otro, tanto en el análisis como en la generación, así como la incoherencia con el resto del sistema han desaconsejado su utilización (4).

Todo esto obliga a dividir la morfología en dos niveles: el más bajo (ENT) se encarga de los aspectos morfo-grafémicos, por así decirlo, y el superior (EMS) trata las cuestiones puramente morfosintácticas. Es decir, la gramática de ENT se encarga de la normalización de caracteres y, sobre todo, de la alomorfía. Por otro lado, el generador gramatical de EMS realiza la tarea de asignar información morfológica a las cadenas basicas (**strings**) y de combinar estas cadenas en las palabras complejas.

Veamos a continuación cómo funcionan estos niveles. La información que llega a ENT del nivel inferior se refiere a las delimitaciones de palabras, oraciones y otros componentes que forman el texto (párrafo, sección, capítulo). Los elementos básicos son la letra, el dígito, el blanco y el signo de puntuación. En este primer nivel morfológico se establece qué combinaciones de letras forman las posibles cadenas básicas de una lengua determinada. Es importante señalar que en muchos ocasiones alguno de los "alomorfos" elegidos para establecer estas cadenas no se corresponden con los utilizados en la morfología teórica (de ahí que se evite el término "morfema" en EUROTRA). Los motivos que justifican este tratamiento son puramente pragmáticos: la tarea que se pide a los analizadores morfológicos es que envíen a ECS la información básica que requiere el análisis sintáctico superficial i.e., la categoría gramatical, el género, el número, el tiempo, el aspecto, la persona, etc. Un ejemplo de regla de formación de una cadena básica en el nivel ENT podría ser este: (5)

```
regla-camión = {cat=raíz, clase=cadena_básica, unid_lex=camión,
                cadena=camión}

                [{unid_lex=c},{unid_lex=a},{unid_lex=m},{unid_lex=i},
                 {unid_lex=o},{unid_lex=n}]
```

Formalmente la regla presenta dos partes: la cabeza, entre llaves, y sus argumentos (uno o más) que van debajo y separados entre sí por comas. Tanto la cabeza como los argumentos son un conjunto de rasgos, de por lo menos un par atributo=valor:

```
e.g.          cat=raíz       es  un  par  atributo=valor

        {clase=cadena_básica, cadena=...} es  un conjunto de rasgos,
                                          en este caso es la cabeza

        {unid_lex=c}, {unid_lex=a}... son varios conjuntos de rasgos,
                                 de un solo par atributo=valor, y
                                 a su vez argumentos de la cabeza
```

Por último, falta especificar que las cadenas en este nivel sólo dan una información elemental acerca de las posibilidades combinatorias

de cada una, es decir, de si se trata de una raíz, un prefijo, un elemento flexivo o una invariante. En EUROTRA se calcula que con unas 10.000 entradas de este tipo se cubriría cada lengua. El profesor De Kock, especialista en lingüística computacional y ajeno al proyecto, estima que para el español es necesario al menos distinguir unos 18.000 morfemas productivos (comunicación personal).

Los objetos producidos por las reglas de la gramática están en condiciones de ser "transferidos" al nivel siguiente. Una vez allí se pasan por el diccionario específico de ese nivel, donde reciben los rasgos pertinentes que necesitarán para combinarse con otros objetos. En EMS, concretamente, tomarán rasgos como "gen=fem" o "clas_pal=n". Veamos como ejemplo la regla de formación de plural:

Diccionario:

```
camión = (cat=raíz, clas_pal=n, unid_lex=camión, cadena=camión,
            tipo_pl=2)

s1 = (cat=flexión, unid_lex=s, cadena=s, num=plural, tipo_pl=1)

s2 = (cat=flexión, unid_lex=s, cadena=es, num=plural, tipo_pl=2)
```

Gramática:

```
regla-plural = (cat=palabra, clas_pal=X, unid_lex=Y, num=plural)

               [(cat=raíz, clas_pal=X, unid_lex=Y, tipo_pl=T),
                (cat=flexión, unid_lex=s, num=plural, tipo_pl=T)]
```

El proceso sería como sigue: el analizador morfológico tomaría las dos cadenas "camion" y "es", que ya habrían sido reconocidas en ENT, y comprobaría que sus partes casan perfectamente con todos los rasgos que especifica la regla, pues de no ser así desecharía la combinación. Los valores en mayúsculas son variables y en minúscula, constantes. Para que la unificación se produzca es necesario que los rasgos de los distintos argumentos sean compatibles entre sí. Por ejemplo, el rasgo "tipo_pl" se utiliza para distinguir los alomorfos del plural

```
[s]  --> tipo_pl=1
[es] --> tipo_pl=2
[0]  --> tipo_pl=3
```

Tanto el argumento raíz como el argumento flexivo tienen que compartir el mismo valor T para que se aplique la regla. Así se evitan malas formaciones como "camions" o "casaes".

Una vez realizada la operación el resultado sería un nuevo objeto con la misma categoría y unidad léxica que los de su argumento con paradigma radical, y con el número plural recibido del argumento co paradigma flexivo.

Pero si queremos que el objeto lleve ya desde EMS toda la información que necesitará en el nivel sintáctico, hay que añadir el rasgo de género. Es bien conocido que el género es inherente a los sustantivos y que, de hecho, no existe un morfema flexivo para el género. "Mano" y "poeta" son femenino y masculino respectivamente, aunque acaben en "o" y "a". El uso de estas dos terminaciones como marcas

morfológicas de género no es tan productivo como pudiera parecer, y además tendríamos muchos problemas con oposiciones del tipo "caso" / "casa", "cero" / "cera", que pertenecen a entradas léxicas diferentes. Por estas y otras razones, decidimos incluir el género directamente en el diccionario. Aquellas palabras que utilicen la marca **-a** para distinguir el femenino necesitamos dos entradas, aunque con la misma unidad léxica:

niño1 = {cat=raíz, unid_lex=niño, cadena=niño, gen=masc}

niño2 = {cat=raíz, unid_lex=niño, cadena=niña, gen=fem}

Nuestra regla para la flexión nominal no contiene, por tanto, ningún argumento especial para el género, ya que se filtra a la cabeza directamente del argumento de la raíz. La regla unificada de género y número es la siguiente:

flexnom = {cat=palabra, clas_pal=X, unid_lex=Y, num=N, gen=G}

[{cat=raíz, clas_pal=X, unid_lex=Y, gen=G, tipo_pl=T),
^(cat=flexión, unid_lex=s, num=plural, tipo_pl=T)]

El símbolo ^ significa que el argumento que sigue es opcional. Para las palabras en singular, que no tienen ninguna marca, necesitamos una regla especial que añada el rasgo "num=sg" a la cabeza una vez que se haya aplicado la regla anterior y consolidado el objeto.

singular = (num=sg/cat=palabra) [(cat=raíz)]

Este tipo de reglas, llamadas **features rules**, permiten el uso de contexto pero solo para elementos terminales. Se interpreta así: añádase el rasgo "num=sg" a aquellas cabezas que sean "cat=palabra" y contengan un argumento único con el rasgo "cat=raíz".

Con las reglas expuestas hasta el momento damos cuenta de los fenómenos regulares de la flexión pero quedan sin resolver los casos excepcionales, aunque no menos importantes. Dentro del número, son aquellas palabras que no distinguen morfológicamente el singular del plural como "crisis", "hipótesis", "martes", todas ellas muy comunes en textos científicos y administrativos. La ambigüedad morfológica en cuanto al género es aún mucho más frecuente: aparece en aquellas palabras que son formalmente invariables y que expresan el género por medio de otros elementos del SN. Es el caso, sobre todo, de los adjetivos:

e.g. el / la artista
el árbol **verde** / la hoja **verde**
el hombre **pobre** / la mujer **pobre**

O el caso particular de algunos nombres femeninos que presentan la alternacia "el agua / las aguas", "el arte / las artes". Esto se debe a un conocido fenómeno morfofonológico: cuando la palabra empieza por **á**, el nombre recibe el artículo masculino para evitar una pronunciación difícil.

En todos estos casos el sistema es incapaz de asignar un valor de género o número en el presente nivel porque la información que necesita está en otras palabras. Como aquí se trabaja dentro de los límites de la palabra, la solución propuesta es que el objeto

ascienda al nivel sintáctico sin rellenar el valor correspondiente y que sea desambiguado allí por medio de reglas de concordancia.

Conclusión

Hemos visto la complejidad que supone el tratamiento de la flexión nominal (un caso sencillísimo para la morfología teórica) dentro de un **framework** que utilice como modelo una gramática independiente del contexto. En EUROTRA, concretamente, necesitamos un rasgo especial que especifique el tipo de alomorfo seleccionado para el plural y, además, las ambigüedades tienen que ser resueltas en el nivel superior. Podemos hacer una constatación: el contexto es relevante en el nivel morfológico y tenemos que expresarlo por medio de algún mecanismo. Hay formalizaciones lingüísticas plenamente satisfactorias que no son utilizables tal cual por la máquina:se necesita definir de otra forma,es decir, de una forma que sea implementable, los datos del problema. Nos encontramos por tanto, ante una cuestión no de conocimiento lingüístico teórico sino de formalización de ese conocimiento de una manera adecuada para el tratamiento por el ordenador.

Notas

(1) El **Esbozo** de la R.A.E reconoce la posibilidad de que los substantivos acabados en **-í** formen su plural con **-es** o con **-s**.

(2) Las excepciones del tipo "jabalíes" pueden ser tratadas con una variante de la segunda regla /s/ --> / V (e) __#.

(3) Señala Grishman (1986:21): "Context-sensitive grammars, however, have not proven to be a particularly suitable formalism for stating most grammatical constraints; context-free grammars with others types of restrictions have proven more effective".

(4) Otro argumento contra las gramáticas de estados finitos es que no pueden reconocer compuestos coordinados como los del alemán **Lernfähig___ und ___willigheit** o las derivaciones castellanas del tipo **en-riqu-ec-er**. Siempre hay que definir qué elementos pueden seguir a los que pueden aparecer. En el caso de "enriquecer" se afijan al mismo tiempo un prefijo y un sufijo. En el caso del alemán, los dos compuestos coordinados comparten los mismos sufijos y prefijos y por eso lo explicitan solo una vez.

(5) Me he permitido traducir los términos de los pares atributo=valor, originalmente en inglés.

Referencias

EUROTRA: **Reference Manual, version 4.0.**

Grishman, R. (1986): **Computational linguistics: An introduction.** Cambridge, Cambridge University Press.

Koskenniemi, K.(1983): **Two-level Morphology: A General Computational Model for Word-form Recognition and Production.** Helsinki, University of Helsinki.

Marcos Marín, F. y Sánchez Lobato, J. (1988): **Lingüística aplicada**. Madrid, Síntesis.

Moreno Sandoval, A. y Salamanca, P. (1988): "Regular Phenomena of Spanish Flexion" (interno de EUROTRA).

Niremburg, S. (1987), ed.: **Machine translation: Theoretical and methodological issues**. Cambridge, Cambridge University Press.

R.A.E (1973): **Esbozo de una nueva gramática de la lengua española**. Madrid, Espasa-Calpe.