El uso editorial de herramientas lingüísticas: un ejemplo con los descriptores humanos

Adán Cassan, Sergi Cervell, Mireia Colom, Mireia Farrús,
Ignacio González, Rafael Marín y David Trotzig
{acassan, scervell, mcolom, mfarrus, igonzalez, rmarin, dtrotzig}@planeta-actimedia.es
Departamento de Lingüística Computacional
Banco de Contenidos, Planeta Actimedia S.A.

Resumen

Esta presentación muestra cómo funciona la *Interfaz Descriptores*, una aplicación que permite la sistematización de los descriptores humanos en los artículos enciclopédicos del medio editorial.

1. Presentación

El desarrollo de herramientas lingüísticas en el marco de un gran grupo editorial da lugar a la colaboración mutua para satisfacer las necesidades de todos. La sistematización eficiente del material enciclopédico es a un tiempo requisito y objetivo de buena parte del trabajo de nuestro Banco de Contenidos y su Base de Conocimiento (BDCon).

Una muestra de que es posible imbricar aplicaciones de lingüística computacional en la regularización de los contenidos editoriales está en la herramienta que presentamos: una interfaz de descriptores humanos que, combinada con una normativa estructurada, simplifica el tratamiento textual de un repositorio de miles de artículos enciclopédicos, procedentes de fuentes diversas, con lo que los datos de la BDCon están cada vez más afinados.

2. Normativas y descriptores humanos

Entendemos por *descriptores humanos* el conjunto de los términos susceptibles de aparecer en una biografía especificando la actividad principal del biografiado. Un manejo apropiado de los datos ofrece la posibilidad de realizar consultas sobre los descriptores y sobre los antropónimos relacionados con ellos. Otra función de los descriptores es actuar como *trigger-words* para detectar nombres propios.

Para posibilitar la sistematización es necesaria una normativa que defina el conjunto de descriptores aceptados en los artículos biográficos. Dependiendo de la elasticidad de la normativa elegida, puede haber varios niveles: en concreto, parece interesante la distinción entre los descriptores normalizados y los que no lo están, sin ser descartables para otras posibles normativas vinculadas, tal vez, a eventuales obras editoriales.

En concreto, en la normativa propuesta se descartan aquellos descriptores genéricos que se pueden cambiar por otros más específicos. Se ha limitado la información dada por el descriptor, de modo que estilos artísticos o cargos públicos no se consideran parte de la actividad principal del biografiado, sino detalles propios del desarrollo enciclopédico. Por lo demás, en las elecciones entre un descriptor y otro ha influido la facilidad de su tratamiento regular y automático. Así, desde varios puntos de vista es preferible *novelista* a *autor de novelas*, por lo que el segundo será sustituido siempre por el primero.

3. ATIP Descriptores: un subconjunto del ATIP (Árbol tipológico)

Siguiendo la pauta de optimizar los recursos siempre que sea posible, se ha integrado en la interfaz de descriptores humanos un subconjunto del Árbol tipológico (ATIP) de la BDCon, que es una estructura léxico-conceptual con más de 110.000 nodos en su versión completa. Para la herramienta se ha preparado una versión reducida de la rama *Persona* del árbol, que se ha podado hasta dejar únicamente los descriptores normalizados, los no normalizados por ahora (pero aceptables tal vez) y el mínimo número de nodos estructurales necesarios para mantener la

posición todos los nodos. Tras eliminar más de 9.000 nodos han quedado unos 1.000, que han acabado constituyendo el ATIP Descriptores.

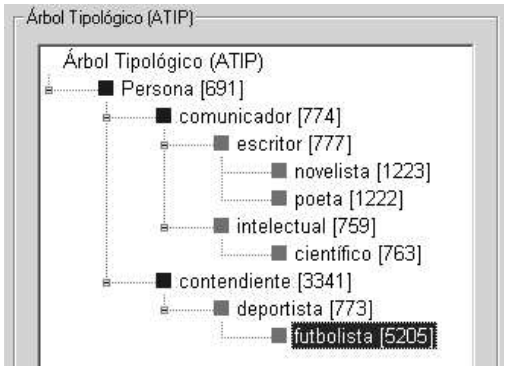

Fig. 4. ATIP Descriptores

Se puede recorrer el árbol accediendo a los hipónimos (hijos) de cada nodo o por medio de búsquedas por nombre del nodo y por sinónimos. La mayoría de los nodos del ATIP están relacionados con acepciones léxicas, a las que se puede acceder mediante su información asociada. Una vez más, el reaprovechamiento de datos de la BDCon se aplica como metodología.

4. El Analizador de texto

La principal utilidad de la herramienta está en su capacidad para analizar cualquier texto y buscar en él todas las ocurrencias de descriptores que el usuario desee. Así, tras seleccionar qué categorías de descriptores se quiere incluir en el análisis, se procede a escribir o insertar un texto. El analizador busca a lo largo de todo el texto y marca con colores los descriptores según la categoría a la que pertenecen. A la vez, elabora una lista ordenada por categorías u orden alfabético sobre la cual se efectúan búsquedas directas sobre el ATIP Descriptores.

Fig. 5. Analizador de texto resultado

Además de los tres tipos de elementos que se dibujan en el árbol, el análisis del texto contempla uno nuevo: los no descriptores, es decir, todo lo que ha sido descartado para aparecer en el árbol. Esto puede tener gran utilidad, ya que permite detectar en un artículo formas no aceptadas por la normativa, probable señal de que habría que sustituirlas por descriptores normalizados.

5. Funcionalidades futuras

Como integrante del conjunto de aplicaciones generadas por la BDCon, la herramienta se beneficiará del continuo enriquecimiento de contenido y estructuras del sistema. Así, la interfaz podría tener en breve un acceso directo a las categorías de la normativa, para que ésta pueda evolucionar o sustituirse por otra sin alterar en absoluto el funcionamiento de la aplicación. Un paso extremadamente útil sería un proceso semiautomático que no sólo reconociera los no descriptores, sino que al momento ofreciera el descriptor normalizado más cercano semánticamente. Automatizarlo del todo exigiría asumir cierto riesgo, ya que un fenómeno tan habitual como la polisemia, mayor aún tratando con acepciones, podría generar errores sin una validación humana. Otra mejora será la vinculación de la normativa con el tipo de texto que se quiera normalizar, más allá de esta primera versión.

Naturalmente, el proceso desarrollado para sistematizar los descriptores en los artículos biográficos es distinto al que se puede realizar con otros dominios léxicos. Actualmente estamos llevando a cabo el tratamiento de los descriptores geográficos relacionados con los más de 100.000 topónimos que la BDCon ha obtenido de diferentes fuentes documentales y bases de datos (los descriptores geográficos conllevan como problema añadido el hecho de que a veces forman parte del nombre del topónimo y a veces no, con lo que la sistematización tiene que contemplar múltiples factores). El método consistente en trabajar a partir de secciones específicas del ATIP tiene un potencial lingüístico y editorial considerable, ya que, en principio, puede hacerse extensible a cualquier material sujeto a una sistemática, posibilitando, así, una gestión avanzada para obras multimedia y online.