3) Análisis empírico de las relaciones anafóricas contenidas en un corpus de textos para conocer, con datos objetivos, el comportamiento real de la anáfora en el discurso así como los tipos de información lingüística que intervienen en la localización de antecedentes.

4) Creación de la estrategia de identificación de antecedentes empleando para ello un modelo de inspiración vectorial. La construcción de ésta se apoya, como hemos mencionado, en el empleo de notación vectorial para la codificación de las fuentes de información localizadas en el análisis de corpus.

5) Automatización de la estrategia propuesta mediante la creación de un programa informático que lleve a cabo todas las tareas.

6) Evaluación de la aproximación adoptada con el fin de determinar su validez.

Una propuesta de codificación morfosintáctica para corpus de referencia en lengua española

(Tesis doctoral leída por Aurora Martín de Santa Olalla- Universidad Alfonso X El Sabio)

Esta tesis presenta una propuesta de codificación morfosintáctica para corpus de referencia en lengua española basada en los estándares de la Text Encoding Initiative (TEI), The Network of European Reference Corpora (NERC) y The Expert Advisory Group on Language Engineering Standards (EAGLES).

Nuestra propuesta de codificación morfosintáctica para corpus de referencia en lengua española consiste en la creación de un sistema taxonómico que toma como unidad de análisis la palabra ('conjunto de signos entre dos blancos') y describe todos aquellos rasgos que presentan una marca formal

explícita que supone, además, un comportamiento gramatical específico.

El formalismo se expresa mediante un sistema de pares atributo-valor. Las etiquetas o membretes tienen una estructura atómica o jerarquizada que, junto a la pertenencia de las palabras a clases ('categorías') y subclases ('tipos'), refleja rasgos recurrentes y específicos de las distintas formas.

Como resultado de la aplicación de estos estándares a la descripción morfosintáctica de nuestra lengua, nuestro conjunto de etiquetas consta de 660 etiquetas morfosintácticas o 'entidades de segundo orden' que son el resultado de la combinatoria de 117 pares atributo-valor o 'entidades de primer orden'.

Las principales clases coinciden básicamente con las 'partes de la oración' tradicionales a las que se añaden una clase 'residual' y una clase 'puntuación'. Los atributos codificados en cada clase son sus rasgos morfosintácticos específicos. Los valores son los correspondientes a cada clase más tres valores para subespecificación ('invariante', 'cualquiera' o 'no-aplicable') y dos valores booleanos Y y O.

Nuestra propuesta de codificación contiene además un 'manual del codificador' que caracteriza y describe cada una de las clases ('categorías gramaticales' o 'partes de la oración', en nuestro caso).

Finalmente, incluimos los formalismos necesarios para su utilización con un analizador SGML: definición de entidades de primer y segundo orden, Declaración de un Sistema de rasgos (o fichero FSD) con sus Declaración de Tipo de Documento (DTDs) correspondientes.