Terminología y Lexicografía asistidas por ordenador. La experiencia de UZEI

Miriam Urkia y Andoni Sagarna

El objetivo de esta comunicación consiste en la presentación de la experiencia acumulada en UZEI en lo que se refiere a la utilización de medios informáticos para las actividades terminológicas y lexicográficas que esta entidad desarrolla.

1 QUE ES UZEI

UZEI es una asociación cultural sin ánimo de lucro fundada en 1977 con la finalidad de impulsar la modernización del léxico vasco dentro del proceso de normalización de esta lengua. A lo largo de la primera década de existencia, UZEI ha publicado una treintena de diccionarios multilingües de diversas disciplinas como son: Matemática, Física, Quimica, Biología, Medicina, Derecho, Economía, Tecnología mecánica, Estadística, Psicología, Filosofía, Meteorología, etc.

Actualmente sigue desarrollando esta labor y desde 1988 está llevando a cabo además un proyecto denominado EEBS que consiste en el estudio lexicográfico de una muestra del corpus que constituye la producción escrita en lengua vasca desde 1900 hasta 1987. UZEI está reconocido por el Gobierno de la Comunidad Autónoma Vasca como Entidad tutelada de investigación para la planificación lingüística y tiene suscritos convenios de colaboración con el Gobierno de la Comunidad Autónoma de Navarra y con la Universidad del País Vasco. Es así mismo miembro de la red TERMNET.

2. LA INFORMATICA EN LA ACTIVIDAD TERMINOLOGICA DE UZEI

2.1 Desarrollo histórico

En 1977, cuando nació UZEI, no se había producido aún la expansión que ha experimentado la informática algunos años después. No obstante, desde el principio se planteó la posibilidad de utilizar sistemas informáticos que ayudaran en la pesada labor de gestión de ficheros terminológicos y lexicográficos.

Tras un breve período de un par de años en el que se encomendaron algunas tareas muy elementales de grabación y listado de ficheros a un centro de cálculo, se procedió a instalar un ordenador en el propio centro de trabajo. La mayor dificultad para informatizar el trabajo de UZEI procedía de los escasos medios económicos con los que se contaba, ya que si se pensaba en utilizar un sistema de gestión de bases de datos convencional, del tipo de los que en aquel tiempo eran habituales en informática de gestión, hacía falta pensar también en un equipo grande y costoso. Puesto que no era posible contar con los medios necesarios para una solución de esta naturaleza, se decidió adquirir un miniordenador Bull Mini 6 y recurrir al desarrollo propio de un sistema de gestión de bases de datos terminológicos, que permitiera trabajar con registros de longitud variable. Este sistema, que fue desarrollado en Fortran, ha estado prestando servicio en el ordenador antes citado hasta el presente año.

En la actualidad el parque informático de UZEI consta de un ordenador Data General MV2500, de 8 ordenadores personales Apple Macintosh de varios modelos y de un PC-AT compatible IBM. El ordenador DG alberga una base terminológica de más de 150.000 conceptos en varias lenguas y hace de servidor Videotex para ofrecer al público un servicio de consulta de esta base y de un sistema de mensajería para resolver dudas en el uso de la lengua.

Los ordenadores personales se utilizan para todas aquellas tareas en las que es importante una interface de usuario con abundantes recursos gráficos y gran facilidad de manejo, mientras que el miniordenador está destinado a ofrecer su potencia de procesamiento y su capacidad de almacenamiento en memoria de masa. Los ordenadores personales están interconectados en una red TOPS y se dispone de un sistema de transferencia de ficheros entre el ordenador DG y los ordenadores personales.

A partir de la incorporación de los primeros Macintosh al centro se decidió desarrollar para estas máquinas el sistema de gestión de bases de datos que había estado funcionando en el miniordenador, pero esta vez en lenguaje C. El resultado de este desarrollo ha sido un producto denominado LexM que pasamos a describir a continuación.

2.2 LexM

La aplicación LexM permite la creación, la modificación, la consulta y la edición de un diccionario. Las fichas que se crean en esta aplicación tienen una estructura arborescente que se puede representar de diversas formas, pero que en cualquier caso está almacenada en memoria o en disco de modo que la jerarquía viene representada por un sistema de paréntesis.

Un diccionario está constituído en LexM por fichas y por índices que permiten acceder a estas fichas de diversas maneras. Cada dato se almacena en forma de pareja nombre/valor y los diversos datos de la ficha están organizados según la estructura arbórea ya citada. Esta configuración permite almacenar y manipular las fichas sin pies forzados de tamaño de campo o de estructura de la ficha. Aunque la estructura de la ficha es libre, hay que definir un modelo de ficha que se guarda en un fichero de modelos junto con las indicaciones propias de los datos que sirven de índices. Además de esto los nombres de los datos deben figurar en un metadiccionario. Por otra parte el tamaño máximo de una ficha completa es de 32.000 caracteres. Hay una aplicación llamada MetaDico que gestiona el metadiccionario y los ficheros de modelos de fichas.

A partir de un diccionario se pueden extraer ciertos datos para producir listas en forma de ficheros de texto. Estas listas pueden ser simples documentos de trabajo o bien listas preparadas para ser impresas en papel. El formato de estas listas se puede parametrizar mediante un sistema de scripts que LexM interpreta.

LexM puede importar ficheros elaborados anteriormente y también exportar fichas para poderlas tratar con otras aplicaciones distintas. Las exportaciones e importaciones se realizan en formato texto con paréntesis.

Las operaciones que se pueden ejecutar con las fichas son la creación, la consulta, la modificación, la supresión de la ficha completa. LexM permite el uso de un

palabras, que pueden ser iniciales, finales, combinaciones de iniciales y finales, así como porciones situadas en posiciones intermedias.

2.3 La consulta del banco terminológico EUSKALTERM por medio de videotex

UZEI ofrece a los usuarios de su banco terminológico EUSKALTERM la posibilidad de acceder a una base de datos terminológica de más de 150.000 conceptos. Las consultas facilitan los equivalentes de un término vasco en español, francés e inglés y los equivalentes de términos españoles, franceses e ingleses en vasco. La consulta puede ser hecha a través de terminales de tipo Ibertex o Minitel o también mediante ordenador dotado de un modem V23 y de un programa de emulación de los terminales anteriormente citados.

3. LA LEXICOGRAFIA ASISTIDA POR ORDENADOR EN UZEI

En 1987 el Gobierno de la Comunidad Autónoma del País Vasco encomendó a UZEI la redacción y ejecución de un proyecto de lexicografía que diera cuenta del léxico utilizado en los textos publicados en lengua vasca a lo largo del presente siglo.

La puesta en marcha de este proyecto ha exigido un esfuerzo de puesta a punto de una serie de procedimientos y recursos informáticos que asisten a los lexicógrafos en su trabajo. En diversas tareas no específicas del trabajo lexicográfico se vienen utilizando distintas aplicaciones que corren sobre Macintosh: el procesador de textos Microsoft Word, la hoja de cálculo Microsoft Excel, el sistema de gestión de bases de datos 4th Dimension, el paquete gráfico Freehand, el OCR OmniPage y el entorno de desarrollo MPW. Además de estas aplicaciones convencionales, ha sido preciso desarrollar una aplicación más específica para asistir a los lexicógrafos en su trabajo. Se trata de RTerm.

3.1 RTerm

RTerm es una herramienta para el vaciado o despojo sistemático de textos y la lematización, así como para la edición de listas y obtención de algunas estadísticas. A partir de los textos cuyo vaciado se desea obtener, RTerm produce tres ficheros indexados que son: el vaciado de las frases —entendiendo como tales frases los fragmentos de texto comprendidos de punto a punto—, el vaciado de las palabras del

Se puede utilizar en caso necesario un fichero filtro que permite hacer que determinadas palabras que no tienen interés en el estudio no figuren en el fichero de vaciado de palabras, aunque naturalmente estarán presentes en los contextos de otras palabras. Se puede recurrir también a una lematización asistida en el transcurso del proceso de vaciado utilizando para ello un fichero de correspondencias entre palabras y lemas.

La lematización es una operación por medio de la cual se hace corresponder a una palabra del texto la entrada de diccionario adecuada. Mediante Rterm se puede realizar la lematización ya sea manualmente, después de haber concluído el vaciado del texto, ya sea de forma semiautomática con ayuda de un fichero de correspondencias.

Se pueden fusionar los vaciados de varias obras en uno solo o extraer los vaciados de varias obras de una fusión previamente realizada. A partir del vaciado, lematizado o no, se pueden producir ficheros de texto en los que figuren las listas de frases del texto, las listas de palabras, las listas de lemas, la concordancia —es decir la lista de palabras del texto ordenada alfabéticamente y con los contextos correspondientes— y las estadísticas referentes a las palabras clasificadas de acuerdo con la frecuencia o en orden alfabético. Estos ficheros se pueden tratar a continuación en un procesador de textos o un paginador.

El documento que se desea vaciar debe ser preparado siguiendo unas determinadas reglas, así por ejemplo los principios de página son indicados mediante el número de página entre dos signos de arroba, los únicos puntos que deben figurar en el texto son los de final de frase, de modo que cualesquiera otros puntos deben ser sustituídos por otros signos; se admiten los caracteres ASCII clásicos y los caracteres con tilde.

La lematización asistida se puede realizar de dos maneras: con gestión de las ambigüedades, en cuyo caso las homonimias del fichero lematizador son detectadas al vaciar el texto y son advertidas al lexicógrafo en ese mismo momento para que las resuelva, a la vista de los lemas que se han hecho corresponder a la palabra en cuestión en el fichero lematizador y del contexto actual, o bien sin gestión de las ambigüedades, en cuyo caso si hay ambigüedades en el fichero lematizador, las palabras a las que correspondan varios lemas posibles no serán lematizadas. A cada vaciado le puede

caso de que se haya utilizado alguno, el nombre del lematizador, si se ha utilizado alguno, la naturaleza de los errores, en caso de que se hayan producido y la hora de terminación de la operación si ésta ha concluído con éxito.

Una vez terminada la operación se puede consultar una lista en la que figuran:

- Un código de referencia, cuyos nueve primeros dígitos corresponden a la obra de la que se trate y los nueve siguientes dan cuenta de la página, la frase dentro de la página y del número que corresponde a la posición del primer carácter dentro de la frase.
- El contexto de la palabra, que puede visualizarse de dos maneras: de forma que se vea el contexto que cabe en una sola linea o de forma que se vea la frase completa.
- El o los lemas que correspondan a la palabra.

Las palabras, sus referencias y sus contextos pueden aparecer listados de acuerdo con el orden alfanumérico de las referencias, es decir de acuerdo con el orden de aparición de las mismas en la obra, o de acuerdo con el orden alfabético de las palabras, o de acuerdo con el orden alfabético de los lemas. En este modo consulta se pueden realizar búsquedas de acuerdo con una palabra, con una referencia o con un lema

Se pueden también eliminar palabras, añadir, cambiar o suprimir lemas y hacer corresponder el mismo lema a varias palabras simultáneamente.

3.2 La optimización del fichero lematizador por aproximaciones sucesivas en el curso de una lematización en dos etapas.

El proceso que se va a describir a continuación da unos resultados bastante satisfactorios cuando se vacían y lematizan textos en lengua vasca. Es necesario explicar este extremo ya que la lematización de palabras de textos vascos tiene una serie de dificultades que no se suelen dar en la lexicografía de las lenguas de origen indoeuropeo. Por otra parte, las dificultades que surgen en este proceso se derivan del sistema morfológico de la lengua vasca y por lo tanto aparecen continuamente en cualquier trabajo de procesamiento de lenguaje natural en esta lengua.

En lingüística se han solido diferenciar las lenguas aglutinantes de las flexivas,

aquellas en las que las palabras están provistas de morfemas gramaticales que indican la función de las unidades, no siendo posible la segmentación de los elementos que constituyen cada morfema.

Del vasco se puede decir que tiene un carácter más aglutinante que la mayoría de las lenguas indoeuropeas sin que ello quiera decir que sea una lengua aglutinante pura. Por poner un ejemplo, lo que en español se diría *los de para con las hijas* en vasco se podría decir *alabenganakoak*. Esta palabra se podría analizar de la siguiente manera:

alaba : radical de la palabra *alaba* = hija
-en : sufijo que indica genitivo posesivo, determinado y plural
-gana:: marca del caso adlativo
-ko: sufijo que se puede considerar semiflexivo y semiderivativo y que realiza determinadas funciones de enlace desde el punto de vista sintáctico que no merece la pena describir aquí
-ak: determinante plural

Esta morfología junto con la importancia de la composición y la dificultad de distinguir entre derivación y flexión hacen que a la hora de lematizar las palabras procedentes de un texto vasco nos encontremos, por ejemplo, con un número relativamente alto de formas asociadas a radicales de nombre en comparación con el que aparece al vaciar textos en otros muchos idiomas y también con muchos casos de ambigüedad.

A la vista de estas circunstancias, en UZEI hemos elegido una estrategia de vaciado/lematización aprovechando las distintas formas de trabajar que nos ofrece RTerm. Inicialmente se lematizaron una a una y manualmente una serie de textos y se obtuvo una primera lista de correspondencias entre palabras y lemas. Esta lista iba a constituir el primer fichero lematizador. Naturalmente, en los casos en los que una misma palabra había presentado más de una ocurrencia en los textos que se habían lematizado, la correspondencia estaba repetida. Un sencillo programa permitió eliminar estas redundancias.

Por otra parte, después de algunos tanteos, se dedicidió utilizar la opción de RTerm que permite vaciar y lematizar el texto automáticamente sin gestión de las ambigüedades. El motivo de la elección fue el siguiente: la lematización con gestión de las

gestión de las ambigüedades, a continuación se fusionan varios textos y luego se procede a completar la lematización manualmente, se pueden visualizar simultáneamente varias ocurrencias de la misma palabra en diferentes contextos e incluso lematizar de una sola vez todas aquellas palabras a las que corresponda el mismo lema y sean contiguas en la ordenación alfabética.

La desventaja reside en que las correspondencias que dan lugar a ambigüedades no son lematizadas automáticamente. Pero hay que hacerse otra pregunta: ¿son comparables las probabilidades de aparición de las formas ambiguas? Veamos un caso particular: sea un caso de ambigüedad en el que haya dos correspondencias que están en competencia. Podemos distiguir dos casos importantes, aquel en el que una de las formas es mucho más probable que la otra y aquel en el que las dos correspondencias son igualmente probables. Si nos encontramos en el primer caso y deshacemos la ambigüedad eliminando la forma menos probable, la lematización automática nos dará una solución que tendrá una probabilidad muy alta de resultar acertada. Si nos encontramos en el segundo caso, cualquiera de las dos que eliminemos nos dará más o menos una tasa de aciertos de un 50%. Esta elección se basa en la competencia lingüística y la experiencia de una persona especializada en este tipo de trabajo.

Eliminadas de esta manera todas las ambigüedades, se utiliza el fichero resultante como fichero lematizador de un conjunto de textos. A continuación una persona corrige y termina la lematización. Finalmente produce un listado de correspondencias por cada texto que lematiza. Estos listados irán a engrosar el fichero de correspondencias, pero volviendo a introducir redundancias y ambigüedades que deberán ser eliminadas. A base de repetir una y otra vez este ciclo se va logrando un fichero lematizador que lematiza cada vez más palabras.

4 EL ANALIZADOR MORFOLOGICO AUTOMATIZADO

UZEI colabora en un proyecto financiado por la Diputación Foral de Guipúzcoa, junto con la Facultad de Informática de la UPV y la empresa de servicios informáticos APIKA en un analizador morfológico asistido por ordenador, que sería una nueva aportación, entre otras, para facilitar aún más la lematización que se está llevando a cabo.

Este proyecto se ha basado en la morfología de dos niveles propuesta por Koskenniemi para el finlandés, de modo que puede analizar morfológicamente cualquier

La primera aplicación del sistema de Koskenniemi fue implementada en un Burroughs B7800, en el que el programa ocupó unas 2000 líneas de Pascal. En nuestro caso se ha hecho una implementación en C con vistas a una posible mejora para su utilización en PC's y compatibles.

La aplicación de la morfología de dos niveles permite tanto el análisis como la síntesis. Separa el nivel superficial de la palabra a analizar o generar y el nivel léxico o profundo, de modo que se evita el almacenamiento de las distintas formas del mismo morfema debido a cambios morfofonológicos. El analizador lo componen dos elementos básicos: las reglas y el sistema léxico. Las reglas sólo establecen correspondencias entre los dos niveles, sin ejecutar nada, mientras que el sistema léxico define todo el conjunto de morfemas. Este último está compuesto por un conjunto de subléxicos que agrupan los elementos de las mismas características, y las clases de continuación, que regulan las secuencias posibles de raíces y afijos. La estructura es la misma en todos los subléxicos: un identificador y su correspondiente entrada, la cual se compone de la representación léxica, la clase de continuación y la información morfológica.

El programa se sirve de dos módulos auxiliares principales: Fsp y Lex. El primero funciona como autómata y es el que va aceptando los pares de caracteres. El segundo, Lex, es el módulo léxico y realiza la función de acceso al léxico. Existe un fichero dividido en subléxicos, organizados en forma de árbol, donde cada arco es un carácter, de modo que se consigue un acceso incremental. La unión entre subléxicos se realiza por medio de las clases de continuación.

Cualquier palabra puede ser lematizada por medio de esta aplicación y añadir además la información morfológica de cada uno de sus componentes como del total del conjunto analizado.

5 LA FUSION DEL PROYECTO LEXICOGRAFICO Y EL ANALIZADOR MORFOLOGICO

Tanto el proyecto de lexicografía como el analizador morfológico pueden complementarse.

El analizador morfológico necesita un soporte léxico para colaborar a la normalización de la lengua y poner en práctica el corrector ortográfico, que será la [illegible] El proyecto lexicográfico de UZEI será

tiempos y la aplicación de la norma que la normalización de nuestra situación lingüística exige.

Por otro lado, la labor de lematización semiautomática puede ser facilitada por medio del analizador, ya que en éste se recogen sistemáticamente todos los afijos gramaticales de la lengua vasca. Las ambigüedades también pueden ser resueltas hasta cierto punto, si bien siempre tendrá que intervenir una persona en último término.

La información que se obtenga de un diccionario no se limitará a la lista alfabetizada tradicional, sino que ofrecerá datos morfológicos para cada entrada y la posibilidad de su utilización más en profundidad con la ayuda del ordenador.

Así pues, ambos proyectos pueden ser una aportación más a la normalización de la lengua y adaptación a los nuevos tiempos para la elaboración y utilización de diccionarios.