ENEKO AGIRRE,
XABIER ARREGI,
XABIER ARTOLA,
ARANTZA DÍAZ DE ILARRAZA,
KEPA SARASOLA,
AITOR SOROA

Informatika Fakultatea (Computer Science Faculty)
P.K. 649, 20080 DONOSTIA (Basque Country)
e-mail: jiparipx@si.ehu.es

Un diccionario activo vasco-castellano en un entorno de escritura

Tema: Lingüística computacional, diccionarios automatizados

Resumen

La aparición de los diccionarios electrónicos ha supuesto un cambio sustancial en el área de la lexicografía. El uso de los ordenadores ha posibilitado no sólo un cambio de soporte, sino la utilización de técnicas de procesamiento de lenguaje natural que permiten explotar la información contenida en los diccionarios. Entre las distintas aproximaciones al diccionario electrónico moderno, nos parece interesante la visión del diccionario como un instrumento *activo* orientado a la resolución de problemas léxicos.

Presentamos en este artículo un trabajo que desarrolla estas ideas (Arregi, 1995). En primer lugar se da cuenta de un estudio que ha permitido describir las formas de uso de los diccionarios por parte de los escritores humanos. Este análisis revela cuáles son las necesidades reales y, por tanto, cuáles son los aspectos en los que deberían incidir los diccionarios electrónicos.

Como aplicación práctica de dicho estudio, se presenta un proyecto en el que se está desarrollando un plug-in para Microsoft OFFICE. En dicho plug-in se integran un diccionario electrónico bilingüe vasco-castellano (Elhuyar, 1996) y herramientas de análisis y generación morfológica. El diccionario así enriquecido tiene ya características activas.

1. Motivación

La adecuación de los diccionarios al mundo de la informática está sujeta, a nuestro entender, a consideraciones tales como:

- Los diccionarios electrónicos no son (y no deben ser tratados como) una variante de los diccionarios convencionales, sino como un producto distinto. Tanto el diseño como la especificación y análisis de funcionalidad de los diccionarios electrónicos se debe llevar a cabo desde esta premisa.
- Ello no obsta para que las aportaciones de los diccionarios electrónicos se hagan a partir del estudio de las carencias de los diccionarios en papel. Es conveniente, y hasta necesario, abordar cuestiones tales como: ¿Hasta qué punto son útiles los diccionarios en papel para resolver problemas léxicos? ¿Cuándo, en qué situaciones, frente a qué clase de problemas se manifiestan las carencias de los diccionarios? ¿Qué se podría esperar de los diccionarios electrónicos?
- Es necesario replantear las características del interfaz entre el usuario y el diccionario, entendido éste último como un sistema de información, consulta, navegación y búsqueda, dotado, por qué no, de capacidad inferencial. La cuestión que se plantea es dónde finaliza la tarea del usuario y dónde empieza la del sistema diccionarial.

Por tanto, la hipótesis que defendemos es que el diccionario electrónico inmerso en un entorno de escritura debe tener características propias y diferenciadas respecto a los diccionarios clásicos en soporte de papel, y dichas características deben redundar en un uso eficiente del diccionario, así como en una mayor ayuda a la hora de resolver problemas de naturaleza léxica.

En este contexto, el enfoque planteado en (Al, 90:394; Martin, 92:200), según el cual el diccionario debe ser una herramienta activa y dinámica capaz de interactuar con el usuario en la resolución de problemas léxicos, es una buena aproximación de cara a especificar la naturaleza del diccionario electrónico moderno.

2. Análisis del uso de los diccionarios

El manejo de los diccionarios ha sido y es objeto de estudio (Bujas, 75; Tomaszczyk, 79; Baxter, 80; Ard, 82; Hatherall, 84; Hartmann, 85; Atkins & Knobles, 90; Nuccorini, 94). La mayoría de estos estudios se basan en datos empíricos ofrecidos por los usuarios o extraídos a partir de ellos.

En (Marchionini, 1989) y en (Large & Beheshti, 1994) se analizan y comparan versiones en papel y electrónicas de enciclopedias. Las conclusiones vertidas por los autores son fácilmente aplicables al uso de los diccionarios, y revelan, entre otros aspectos, que para sacar provecho a las fuentes electrónicas hace falta una correcta adaptación de nuestras aptitudes cognitivas.

El propósito de nuestro estudio es profundizar en aspectos cualitativos ligados a la interacción entre el usuario y el diccionario. De esta forma, lo que pretendemos es detectar las lagunas de los diccionarios y tratar de explorar los medios mediante los que los diccionarios electrónicos podrían subsanar dichas lagunas.

Los datos recogidos de los estudios previamente citados han sido clarificadores en muchos aspectos, pero hemos considerado necesario elaborar nuestro propio estudio, mediante una metodología ajustada a nuestras necesidades.

2.1. Metodología seguida en nuestro análisis

La técnica que se ha seguido en la recogida de datos se ha basado en la observación directa y en protocolos verbales y escritos. Los siete escritores (en su mayoría traductores) que han participado en las pruebas han ido narrando todas las dificultades encontradas, así como sus impresiones y resoluciones. Esta información se ha registrado en cintas magnetofónicas para, posteriormente, rellenar registros con la siguiente estructura: diccionario seleccionado, término que se ha querido consultar, entrada diccionarial seleccionada, tipo de consulta, resultado de la consulta y otras anotaciones. Estos registros se han ordenado secuencialmente.

En el desarrollo de la tarea el observador ha jugado un rol activo: no se ha limitado a grabar los comentarios del escritor sino que le ha requerido sobre sus impresiones, objetivos, etc.

La formación y experiencia de los escritores ha sido heterogénea. Si bien todos ellos están habituados a traducir o escribir textos con la ayuda del diccionario, tres de ellos son profesionales y el resto aficionados. Los textos usados han sido diversos (textos técnicos, literarios y comunes). A veces se han traducido párrafos enteros, y en otras ocasiones se ha limitado la traducción a frases, oraciones, locuciones o palabras.

2.2. Resultado del análisis

Vamos a resaltar a continuación algunas consideraciones generales que hemos extraído de los ejercicios realizados.

Tal y como se revela en distintos estudios efectuados sobre el manejo de diccionarios (Neubach & Cohen, 88), hemos podido constatar que el uso de los diccionarios es una tarea difícil y compleja, y que el porcentaje de fracaso es elevado.

Veamos cuáles son las causas que conducen al fracaso:

〈 Elección inadecuada del tipo de diccionario. Dependiendo del tipo de trabajo y del nivel de conocimiento de las lenguas involucradas, es conveniente

compaginar el uso de diccionarios monolingües y multilingües. Estrategias inadecuadas en el uso de estos recursos conducen a decisiones erróneas.

- Síntesis de formas léxicas. La mayoría de los diccionarios están concebidos para la comprensión o traducción de palabras. Una de las actividades en la que los escritores manifestaron mayor descontento, en lo que se refiere a la utilidad del diccionario, fue la síntesis o producción de formas léxicas. La búsqueda de palabras a partir de ideas o conceptos es una tarea difícilmente abordable por los diccionarios convencionales, pero que debería tenerse muy en cuenta en el diseño de los diccionarios electrónicos.
- Manejo de locuciones. La comprensión y uso de locuciones, de formas idiomáticas y de términos multi-palabra es, en general, problemática. Los diccionarios convencionales presentan deficiencias, tanto en la forma de acceso de esos términos, como en la manera de presentarlos y de dar información acerca de ellos.
- Dicotomía texto-diccionario. La transición entre las unidades textuales y las entradas diccionariales da lugar a serias dificultades. Ciertamente, hay una notable diferencia entre las unidades léxicas del texto, que están cargadas de información contextual, y las entradas diccionariales, que son independientes del contexto. Estas diferencias se manifiestan en aspectos gramaticales (morfosintácticos) y semántico-pragmáticos. Los diccionarios no facilitan la transición entre ambos niveles.

Dado que los diccionarios en soporte de papel tienen severas limitaciones para abordar con éxito los problemas expuestos, los sistemas diccionariales electrónicos deben tratar de subsanar esas carencias.

En las pruebas realizadas se ha puesto de manifiesto la necesidad de dotar al diccionario de una funcionalidad más completa. Entre las funciones que se han detectado como necesarias cabe destacar:

- La búsqueda tesaúrica. Es una utilidad muy relevante en la generación léxica. Facilita la búsqueda de formas léxicas a partir de ideas o conceptos abstractos.
- La presentación de ejemplos de uso. Esta función muestra el manejo contextualizado de las unidades léxicas.
- La búsqueda de coocurrencias léxicas. Se trata de mostrar al usuario combinaciones de unidades léxicas que aparecen estrechamente ligadas, no tanto por criterios sintácticos o semánticos, sino por la frecuencia de uso.
- La incorporación de funciones gramaticales. La idea de integrar analizadores y generadores morfosintácticos y otras herramientas para el tratamiento automático del lenguaje natural parece abrir un amplio abanico de posibilidades.

3. Primera aproximación: incorporación de herramientas morfológicas

Entre las distintas opciones para tratar de dotar al diccionario electrónico de un comportamiento activo, hemos optado en primer lugar por diseñar un proyecto que tiene como objetivo la integración de un diccionario bilingüe vasco-castellano y herramientas de análisis y generación morfosintáctica.

Como ya hemos apuntado en la sección anterior, no hay una correspondencia directa entre las formas léxicas y las entradas diccionariales, ya que éstas últimas carecen de flexiones (casos declinativos, conjugaciones verbales, etc.)

Estas diferencias son aún mayores cuando las lenguas utilizadas son aglutinantes, como es el caso del vasco. Por ejemplo, la forma *"lagunekin"* es una flexión de la entrada *"lagun"*. En este contexto, la incorporación de herramientas gramaticales es necesaria para aquellos escritores que sin tener un conocimiento profundo de la lengua pretendan usar el diccionario.

3.1. Plug-in diccionarial para Microsoft Office

La idea de facilitar y enriquecer el uso del diccionario mediante herramientas gramaticales está en vías de materializarse como un plug-in del paquete Microsoft Office.

Ya anteriormente el grupo IXA de procesamiento de lenguaje natural ha desarrollado herramientas que se han incorporado a aplicaciones Microsoft (Aduriz et al., 97).

La propuesta que se está desarrollando en este proyecto toma como base un contexto de trabajo bilingüe castellano-vasco en el que interactúan:

- Un diccionario bilingüe. Se trata del diccionario *Elhuyar*, del que ya existe una primera versión electrónica.
- Analizadores y generadores morfosintácticos para el vasco y el castellano. El analizador/generador morfológico que se va a usar para el vasco será una variante de XUXEN (Alegria *et al.,* 1996).
- Un interfaz amigable adaptado al entorno Windows.

El sistema diccionarial así construído va a permitir consultar el diccionario directamente a partir de las formas léxicas. Por ejemplo, un usuario que quiera traducir la forma vasca *"lagunekin"* no va a necesitar saber que la entrada diccionarial que se debe consultar es *lagun*, ni que el sufijo declinativo *"–ekin"* expresa el plural del caso asociativo; por otro lado, tampoco va a tener necesidad de generar el texto equivalente en castellano, ya que el sistema diccionarial será capaz de generar la traducción *"con los amigos"*.

4. Trabajos futuros

Actualmente estamos desarrollando trabajos que permitan integrar en un futuro próximo distintas fuentes de información léxica en los entornos de escritura. Creemos que la presentación generalizada de ejemplos y de coocurrencias léxicas será una ayuda importante para los usuarios de diccionarios.

Se está trabajando asimismo en la línea presentada por (Artola, 93), que desarrolla métodos de inferencia sobre bases de conocimiento diccionarial.

Los aspectos de desambiguación léxico-semántica constituyen otra de nuestras líneas actuales de trabajo (Agirre & Rigau, 96; Rigau *et al.*, 97). Sin duda, este campo tiene una repercusión directa en el desarrollo y utilización de diccionarios electrónicos.

El análisis del contexto de las formas léxicas es otro de los aspectos que va a permitir mejorar la precisión de las consultas diccionariales.

5. Conclusiones

Se ha descrito un trabajo que aborda la definición y análisis de funcionalidad de un diccionario electrónico, a partir del estudio de las carencias de los diccionarios convencionales en soporte de papel.

En este contexto, se ha presentado un primer paso en la implementación de un sistema diccionarial electrónico, basado en la utilización de técnicas de procesamiento de lenguaje natural.

Agradecimientos

Este trabajo ha recibido ayudas de la Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV 141.226-TA073/96), del Gobierno Vasco en el marco de las acciones Universidad-Empresa y del CICYT en el marco del proyecto ITEM (TIC96-1234-C03-02).

Referencias

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M.. A spelling corrector for Basque based on morphology, in *Literary & Linguistic Computing*, Vol. 12, No. 1. Oxford University Press. Oxford. 1997.

Agirre E., Rigau G.. Word Sense Disambiguation using Conceptual Density, *Proceedings of COLING'96*, 16-22. Copenhagen (Denmark). 1996.

Al B.. Dictionnaire Bilingue et Transfert Automatique de Donnees Lexicales, in L. Fignoni, C. Peters eds., vol. I, 17-32, *Computational Lexicology and Lexicography (Special Issue dedicated to B. Quemada).* Pisa: Giardani, 1990.

Alegria I., Artola X., Sarasola K., Urkia M.. Automatic morphological analysis of Basque, in *Literary & Linguistic Computing,* Vol. 11, No. 4, 193-203. Oxford University Press. Oxford. 1996.

Ard J.. The use of bilingual dictionaries by ESL student while writing, *ITL Review of Applied Linguistics*, 58, 1-27.1982.

Arregi X.. *ANHITZ: Itzulpenean laguntzeko hiztegi-sistema eleanitza / ANHITZ: Sistema diccionarial multilingüe de ayuda en la traducción,* Tesis doctoral, EHU, 1995.

Artola X.. *HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza / Conception et construction d´un système intelligent d´aide dictionnariale (SIAD),* Tesis doctoral, EHU, 1993.

Atkins B.T., Knowles F.E.. Interim report on the EURALEX/AILA Research Project into Dictionary use, in T. Magay, J. Zigány eds., 381-392, *Proceedings BudaLEX´88*. Budapest: akadémiai kiadó, 1990.

Baxter J.. The Dictionary and Vocabulary Behavior: a Single Word or a Handful?, *TESOL Quarterly*, vol. 14, no. 3, 325-336. 1980.

Bujas Z. Testing the performance of a bilingual dictionary on topical current texts, *Studia Romanica et Anglica Zagrabiensia*, no. 39, 194-204. 1975.

Elhuyar. Elhuyar hiztegia. Usurbil, 1996.

Hartmann R.R.K.. Four perspectives on dictionary use: a critical review of research methods. The Dictionary and the Language Learner, in A.Cowie ed., 11-28, *Papers from the EURALEX Seminar (Leeds), Lexicographica, Series maior*, no. 17, Tübingen; Niemeyer. 1985.

Hatherall G.. Studying Dictionary Use: Some Findings and Proposals, *Proceedings LEXeter'83*, 183-189, *Lexicographica, Series maior*, no.1, Tübingen; Niemeyer, 1984.

Large A. and Beheshti J.. A Comparison of Information Retrieval from Print and CD-ROM versions of an Encyclopedia by Elementary School Students, *Information Processing & Management*, vol. 30, no. 4, 499-513, 1994.

Marchionini G.. Making the transition from print to electronic encyclopaedias: adaptation of mental models, *Int. J. of Man-Machine Studies*, no. 30, 591-618. 1989.

Martin W.. On the organization of semantic data in passive bilingual dictionaries, *Actas del IV congreso Internacional.EURALEX´90* (Benalmádena), 193-201, Biblograph, 1992.

Neubach A. and Cohen A.D.. Processing Strategies and Problems Encountered in the use of Dictionaries, *Dictionaries. Journal of the Dictionary Society of North America,* 10, 1-20, 1988.

Nuccorini S.. On Dictionary Misuse, *Proceedings EURALEX'94* (Amsterdam), 586-597. 1994.

Rigau G., Atserias J., Agirre E.. Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation, *Proceedings of ACL'97/EACL'97*, 48-56. Madrid, Spain. July 1997.

Tomaszczyk J.. Dictionaries: users and uses, *Glottodidactica*, 12, 103-119, 1979.