

La Estación de Trabajo Lexicográfico (ETL).
La ingeniería lingüística al servicio del profesional de la lexicografía

M. Teresa Cabré, Lluís de Yzaguirre, Mercè Lorente y Anna Matamala
Grupo IULATERM
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra. Barcelona

En el Institut Universitari de Lingüística Aplicada (UPF) se ha puesto en marcha una serie de proyectos de investigación aplicada y desarrollo que tiene como finalidad el diseño y la realización de prototipos precompetitivos (preindustriales) de estaciones de trabajo integradas para profesionales de la lengua. Estas estaciones de trabajo se basan en la idea de la “ergonomización” de actividades profesionales, de manera que integran recursos, herramientas y asistentes automáticos y semiautomáticos para el desarrollo de tareas concretas. Se pretende con ello avanzar hacia la automatización de la cadena de trabajo, no tanto para desvincular al profesional, sino para asistirle de manera pertinente en cada fase de trabajo y en cada toma de decisiones, para ayudarle a controlar la adecuación de procesos y de resultados, y para verificar el trabajo en equipo.

Las estaciones de trabajo diseñadas o en fase de elaboración integran herramientas de procesamiento, recursos lingüísticos, accesos a la comunicación, sistemas de ayuda y sistemas expertos desarrollados *ad hoc* para automatizar la cadena de trabajo de los profesionales. El primer prototipo en fase de desarrollo es la Estación de Trabajo Lexicográfico (ETL), que pretende abarcar desde el diseño de un diccionario hasta su edición.

En esta comunicación reflexionaremos sobre las relaciones que se establecen entre lexicografía e ingeniería lingüística y, a continuación, presentaremos la versión actual de la ETL.

1. La ETL, punto de intersección entre la ingeniería lingüística y la lexicografía

En los últimos años la lexicografía ha vivido cambios importantes, motivados por la incorporación de todo tipo de aplicaciones informáticas que permiten automatizar el acceso a la documentación, el proceso de datos, la gestión, la intercomunicación, la edición, etc. Además, la evolución del hardware y del software hacia planteamientos de consumo mayoritario (menor coste y mejor accesibilidad) ha facilitado que no sólo los grandes proyectos lexicográficos se hayan visto beneficiados de esta tecnologización, sino que las pequeñas editoriales e incluso los profesionales autónomos hayan podido acceder, en mayor o en menor medida, a algunos de estos recursos: programas de tratamiento de textos, gestores de bases de datos, acceso a Internet, correo electrónico.

Pero ha sido la ingeniería lingüística, con el desarrollo de recursos pensados por y para el lenguaje y las lenguas, la que está aportando cambios significativos en la automatización de tareas de los profesionales de la lengua de todos los ámbitos. La adecuación de las aplicaciones diseñadas a estos usuarios profesionales pasa sobre todo por la adaptación de herramientas genéricas para usos particulares (por ejemplo, los gestores de bases de datos terminológicas). Sin embargo, las principales aportaciones de

la ingeniería lingüística son el desarrollo de recursos lingüísticos en entornos digitales (corpus textuales, diccionarios en línea) y el diseño de herramientas de tratamiento de lenguaje mediante el uso de estrategias de base lingüística (sistemas de procesamiento de corpus, sistemas de extracción y recuperación de información). Los diccionarios basados en corpus textuales han sido seguramente los primeros objetivos de esta nueva tecnología y, por lo tanto, sus primeros valedores.

La Estación de Trabajo Lexicográfico (ETL) es un proyecto de investigación que pretende facilitar la tarea del lexicógrafo mediante la integración tecnológica de múltiples productos y recursos. La ETL pretende hacer explícitos todas las decisiones, todas las tareas, todas las necesidades, todos los controles e incluso aquellos implícitos que quedan escondidos en la mecanización del trabajo profesional. La ETL es, en síntesis, un lugar de trabajo integrado desde donde se puede pensar, diseñar, redactar y editar cualquier tipo de diccionario, y desde donde el lexicógrafo puede acceder a la información y a los recursos que precisa.

En este proyecto, realizado por un equipo de investigadores del grupo IULATERM,¹ han colaborado investigadores de las Oficinas Lexicogràfiques del Institut d'Estudis Catalans y del Seminario de Filología e Informática de la Universitat Autònoma de Barcelona, en los contenidos del sistema experto para el diseño de diccionarios y en la versión catalana del prototipo; y se ha incorporado la empresa Editorial SPES SL, para el desarrollo de la versión en lengua castellana del prototipo y para el desarrollo de fases posteriores. La versión catalana del prototipo ha contado con el soporte financiero del Centre de Referència en Enginyeria Lingüística de la Generalitat de Catalunya (1998-2000) y el Ministerio de Ciencia y Tecnología, dentro del programa PROFIT, ha concedido una subvención para el desarrollo de la versión española (FIT-070000-2001-677).

2. *Las versiones disponibles de la ETL*

Las versiones disponibles de la ETL en la actualidad son versiones **parciales** del prototipo porque solamente incluyen las fases de documentación y diseño de un diccionario. El proyecto continuará ahora con el desarrollo de la fase de gestión y redacción del diccionario.

Y, por otro lado, las versiones existentes son **limitadas**, en el sentido que únicamente se han desarrollado para el diseño de diccionarios generales, monolingües y sincrónicos, en castellano y en catalán. Más adelante se prevé continuar con la ampliación del prototipo para diccionarios bilingües o diccionarios escolares, por ejemplo.

A continuación presentaremos la versión 1.1. de la ETL. Se ha concebido como una aplicación modular que consta de dos elementos fundamentales: (i) una colección de recursos de ayuda al lexicógrafo disponibles en Internet y (ii) un sistema experto de diseño lexicográfico. Tal como se puede observar en la figura 1, estos módulos son accesibles desde la página inicial, que también da acceso a la presentación y a la ayuda.

¹ M. Teresa Cabré (dir.), Lluís de Yzaguirre, Cristina Gelpí, Mercè Lorente, Anna Matamala, Carlos Rodríguez, Oscar Talamino y Jesús Carrasco.

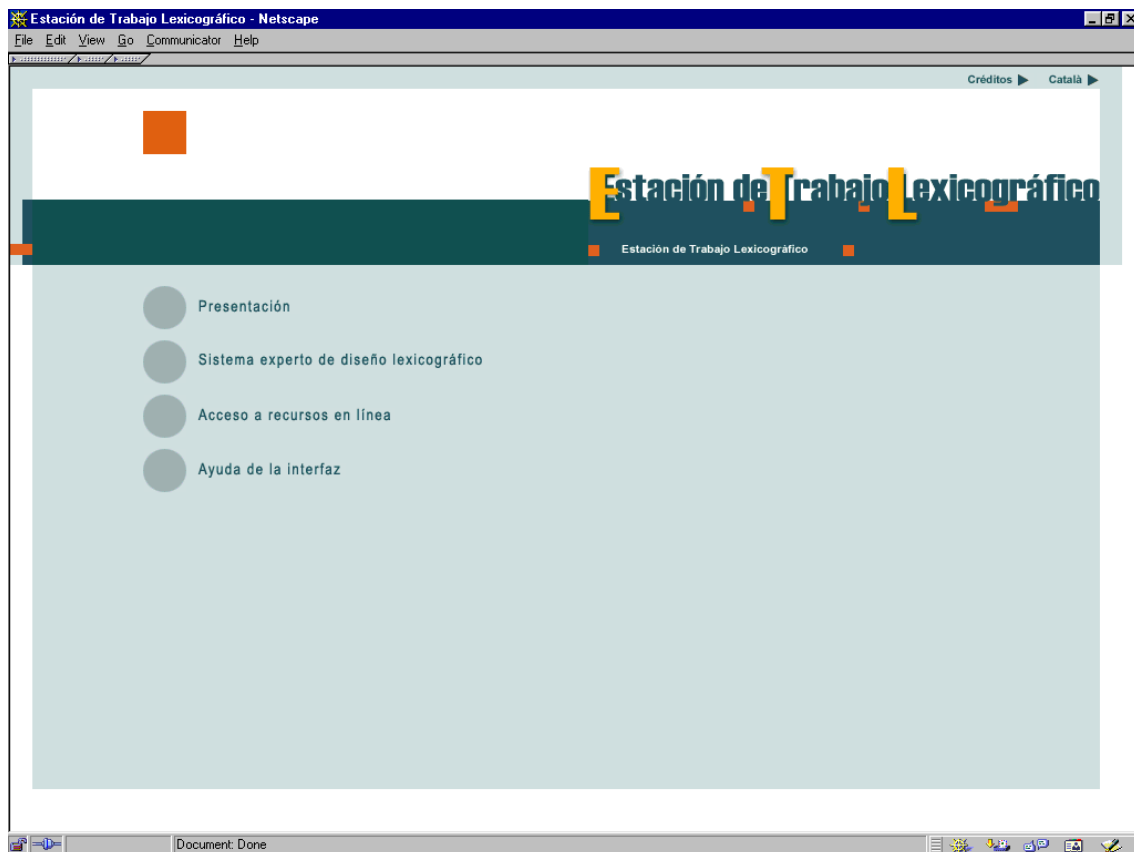


Figura 1. Pantalla inicial de la ETL

En los dos subapartados siguientes expondremos más detalladamente los contenidos de la colección de recursos y del sistema experto de diseño lexicográfico.

2.1 Integración de recursos lexicográficos en Internet

Uno de los módulos de la ETL corresponde a un portal desarrollado en HTML que reúne los recursos de Internet que pueden ser más útiles para la definición del producto lexicográfico que se quiere desarrollar. En este sentido, se ofrecen en una interfaz amigable una colección de enlaces que dan acceso a fuentes bibliográficas, librerías, buscadores y catálogos en línea, obras lexicográficas de Internet, corpus, bases de datos, etc. También se incluyen unas guías didácticas que explican cómo utilizar algunos diccionarios electrónicos disponibles en CD-ROM o en la red, cómo consultar catálogos y buscadores de interés y cómo explotar al máximo los corpus. La intención es que el lexicógrafo pueda aprovechar al máximo los recursos en línea y pueda acceder a la información de manera rápida, ya sea para ver cómo se han resuelto determinadas cuestiones en otras obras lexicográficas, ya sea para extraer datos y explotarlos para su producto.

La pantalla que se presenta al usuario de la estación cuando quiere consultar este módulo es la siguiente:

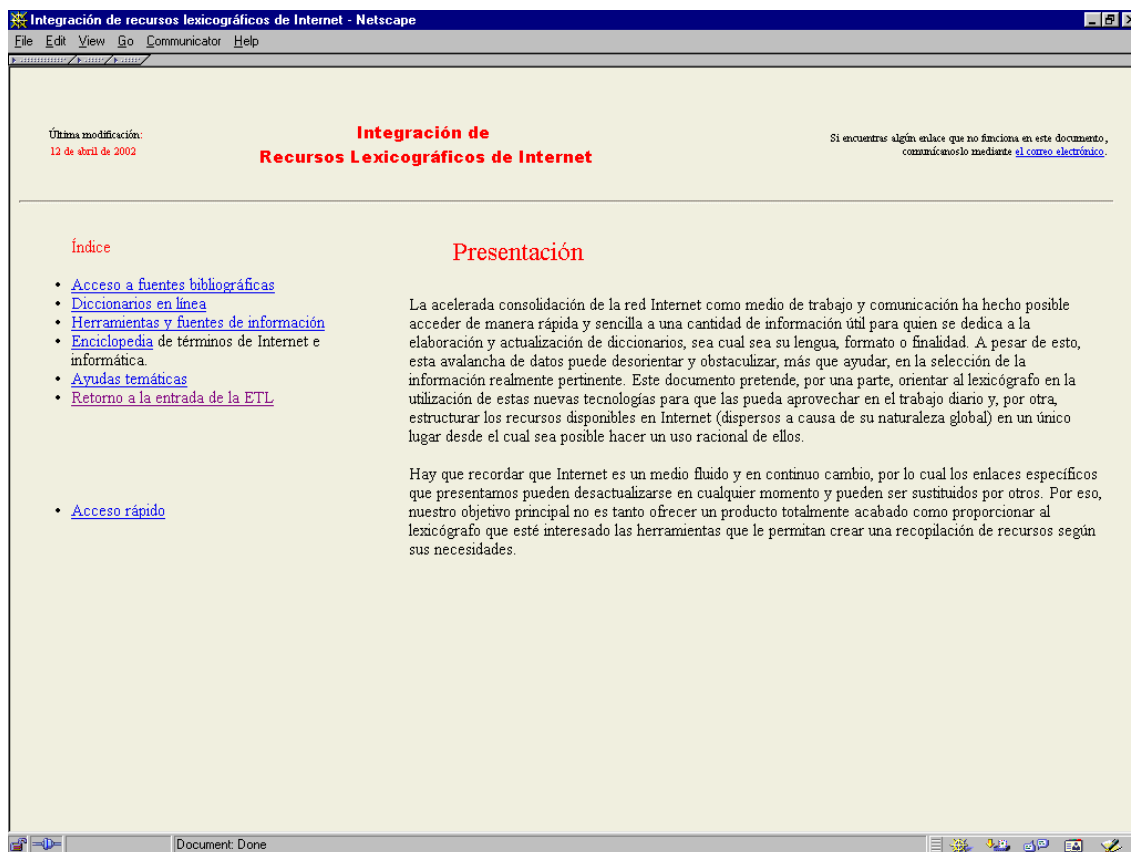


Figura 2. Integración de recursos lexicográficos de Internet

En la pantalla precedente se puede ver que mediante los enlaces de la columna de la izquierda este módulo da acceso a fuentes bibliográficas, diccionarios en línea, herramientas y fuentes de información variadas, una enciclopedia de términos de Internet y distintas ayudas temáticas.

2.2 El sistema experto de diseño lexicográfico

El módulo más novedoso de la ETL es sin duda el sistema experto de diseño lexicográfico, un recurso que mediante preguntas breves va guiando al lexicógrafo y lo ayuda a establecer de modo sistemático las características de la obra lexicográfica que quiere elaborar. Se accede a este bloque clicando en el enlace correspondiente de la página inicial (ver figura 2), y la página que se visualiza es la siguiente:

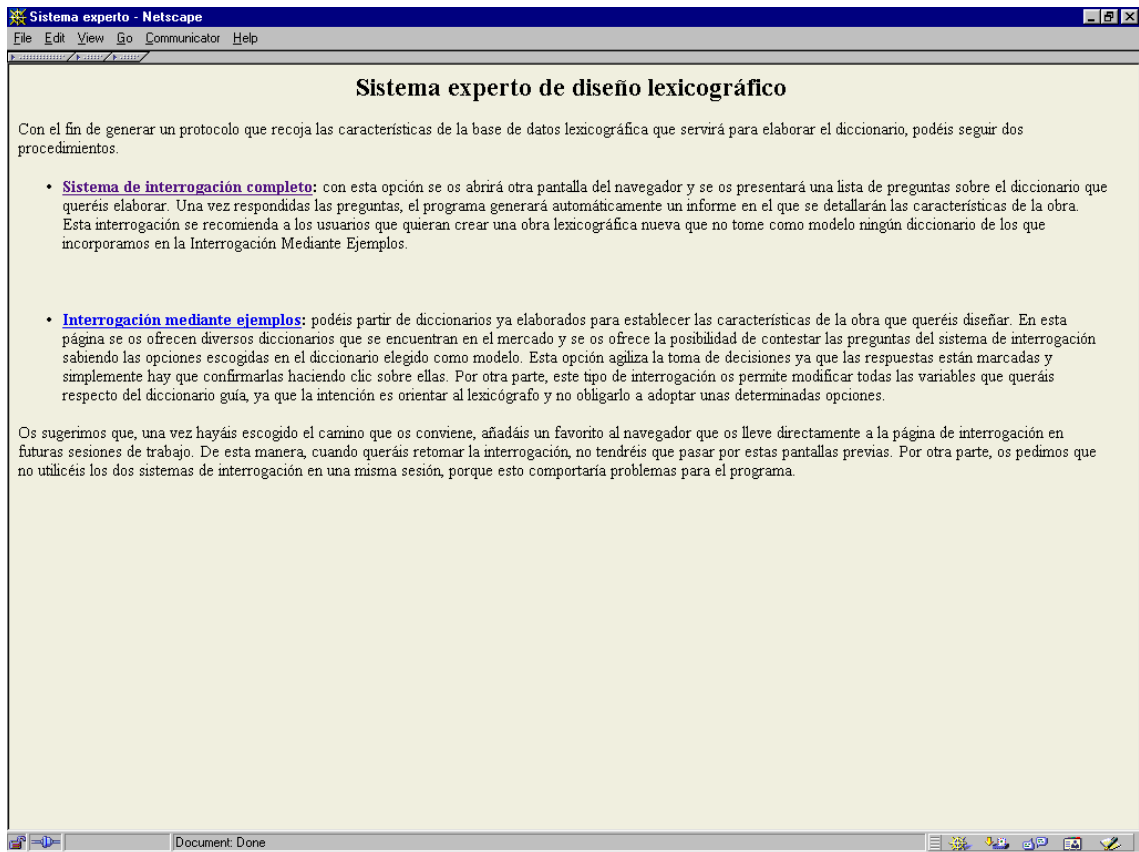


Figura 3. Página inicial del sistema experto

Como se puede observar, el proceso se puede llevar a cabo tomando como referencia algún diccionario existente (Interrogación Mediante Ejemplos) o bien se puede seguir la interrogación completa sin tomar ningún modelo (Sistema de Interrogación Completa). La interfaz en ambos casos es la misma: el usuario debe responder una serie de preguntas que se le presentan automáticamente y que van acompañadas de unas ayudas. Después de responder todas las preguntas, el programa genera automáticamente un informe con las características de la obra en proyecto. A continuación presentamos más detalladamente el funcionamiento de la herramienta, distinguiendo entre los dos sistemas expuestos.

a) Sistema de interrogación completa

Al seleccionar este sistema de interrogación, la aplicación abre automáticamente la página que se puede ver en la figura 4. El usuario debe responder las preguntas que se le presentan en la pantalla blanca y, al final, el programa genera un informe sobre el producto que se pretende desarrollar. Las respuestas posibles son múltiples, por lo que las vías que se irán definiendo también lo serán.

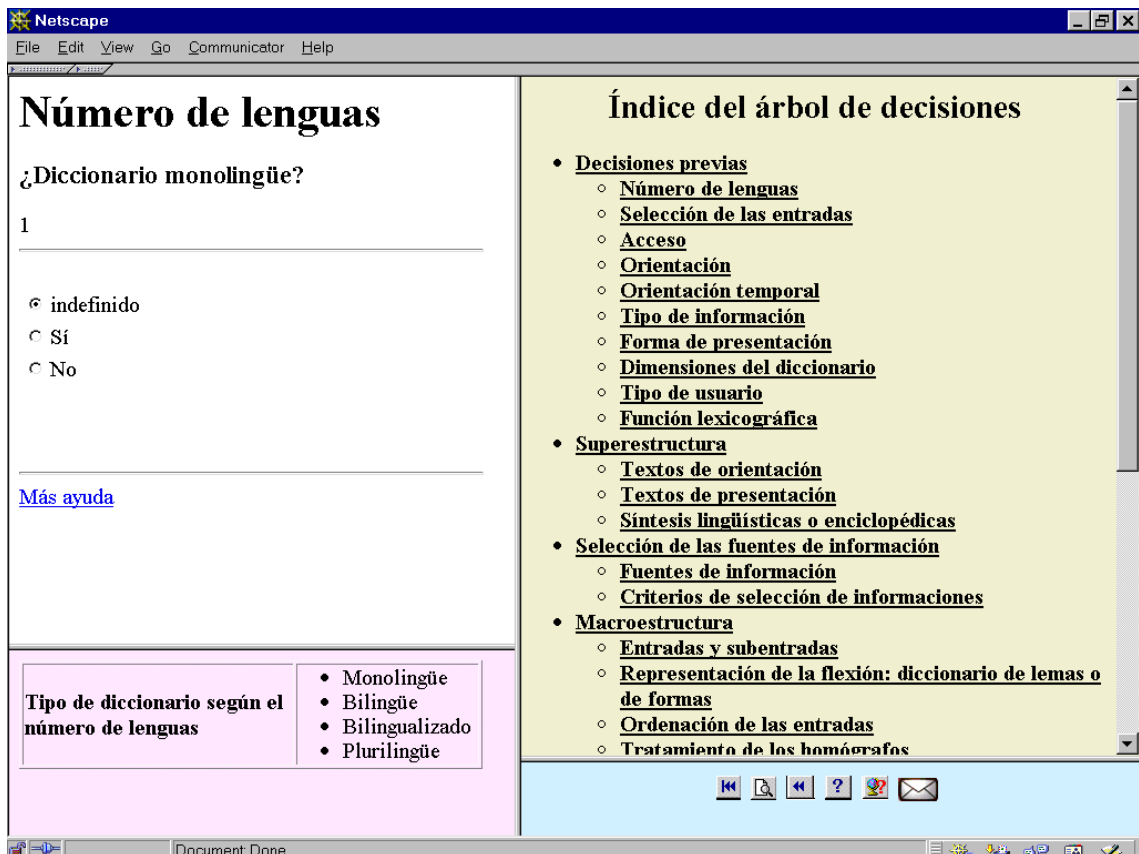


Figura 4. Página de interrogación

La interfaz de consulta consta de los elementos siguientes:

- en la *ventana blanca*, en la parte superior izquierda, se presenta el título —que indica la sección a la que pertenece la pregunta—, la pregunta que se debe responder, seguida de las respuestas posibles. En la parte inferior hay un enlace que indica "Más ayuda" y que sirve para abrir una pantalla de ayuda en la ventana amarilla;
- en la *ventana malva*, en la parte inferior izquierda, se encuentra la ayuda sintética. Esta ayuda tiene como objetivo ofrecer de modo resumido todas las opciones previstas por el programa para una determinada sección. Esta ventana es útil a la hora de prever posibles respuestas y concebir cada bloque globalmente;
- en la *ventana amarilla*, en la parte superior derecha, se recogen distintas informaciones. Al inicializar el sistema experto se visualiza un índice de todos los bloques de preguntas, el llamado "Índice del árbol de decisiones", gracias al cual el usuario tiene una perspectiva global de todas las decisiones. Si se desea más información sobre algún aspecto concreto, con sólo clicar en el enlace correspondiente se abre una ayuda detallada —ayuda que también se abre clicando sobre el enlace "Más ayuda" de la ventana blanca;
- finalmente, en la *ventana azul*, de la parte inferior derecha, hay distintos botones con las funciones siguientes:
 - *inicialización*: sirve para reinicializar la interrogación, es decir, para borrar todas las respuestas dadas y volver a empezar el proceso;

- *informe sintético*: permite visualizar todas las respuestas dadas hasta el momento;
- *retroceder un bloque*: se utiliza para retroceder un bloque de preguntas y borrar todas las respuestas dadas en dicho bloque;
- *ayuda de la interfaz*: da acceso a una ayuda sobre el programa;
- *ayuda global*: contiene el módulo de recursos en línea;
- *correo y FAQ*: botón que da acceso a una FAQ y al correo, recursos que facilitan la interacción con los usuarios.

Las preguntas de la interrogación corresponden a distintos ámbitos del diseño lexicográfico que se pueden agrupar en los bloques siguientes:

- *Decisiones previas*: sirven para definir los rasgos generales del producto, como el número de lenguas, la orientación temporal, el grado de especialización, etc.;
- *Superestructura*: las preguntas de este bloque definen los elementos que tradicionalmente se han agrupado bajo la etiqueta de superestructura o hiperestructura.
- *Selección de informaciones*: en esta fase de la interrogación se deciden qué fuentes de información se utilizarán y los criterios que guiarán la selección de las informaciones;
- *Macroestructura*: las preguntas de esta sección definen los rasgos macroestructurales, a pesar de que en la versión actual sólo se prevén las opciones correspondientes a un diccionario monolingüe general y sincrónico;
- *Microestructura*: en este último bloque, el lexicógrafo establece la estructura del artículo lexicográfico.

La tarea del lexicógrafo consiste en responder las preguntas de los distintos bloques y, una vez terminado el proceso, el programa genera automáticamente un informe como el que presentamos en la figura 5.

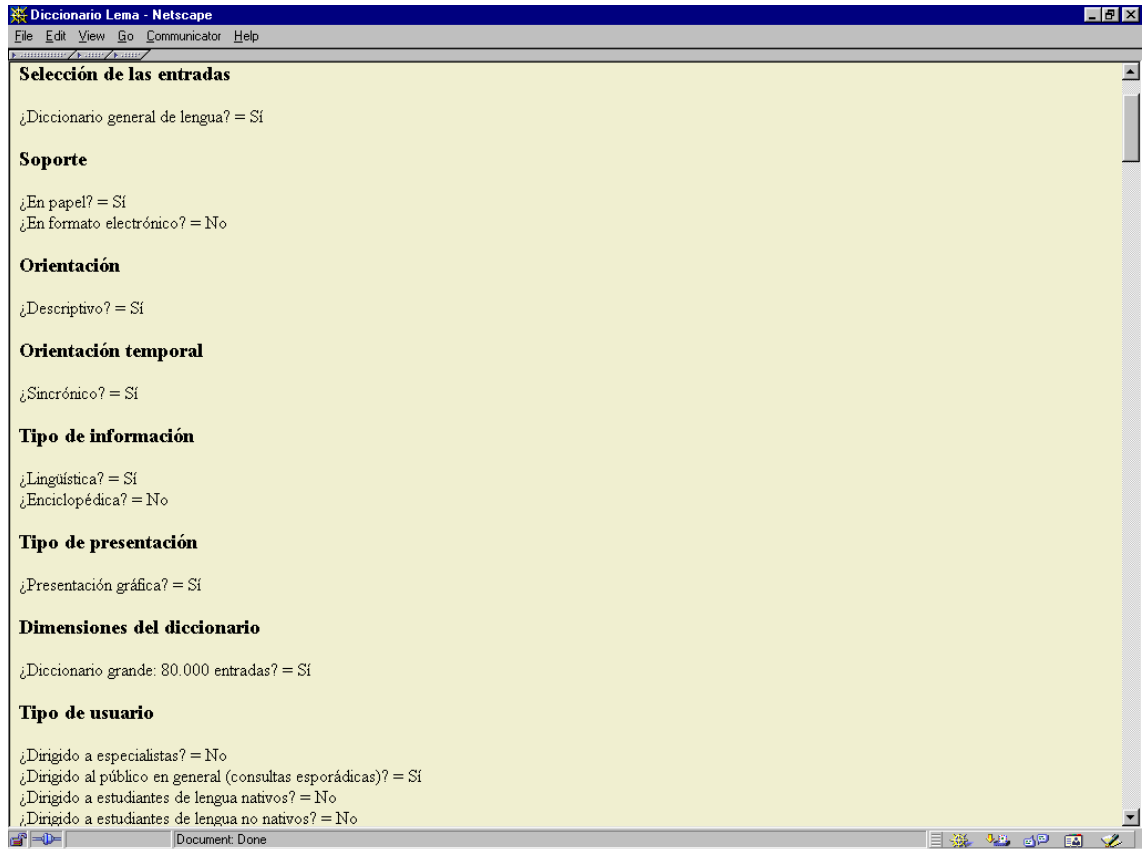


Figura 5. Ejemplo de informe

En este informe se definen de modo sistemático las características que tendrá la obra lexicográfica en proyecto y sirve de punto de partida para la elaboración de la base de datos lexicográfica.

En este apartado hemos expuesto el sistema de interrogación completo, en el que el lexicógrafo define un proyecto desde el principio y sin tomar ninguna obra existente como referente inmediato. Sin embargo, muchas veces los profesionales se basan en obras de referencia para definir su proyecto, y es esto precisamente lo que quiere facilitar la Interrogación Mediante Ejemplos que presentamos a continuación.

b) Sistema de Interrogación Mediante Ejemplos

La voluntad de este módulo es que el usuario seleccione una obra como modelo y edifique su proyecto lexicográfico a partir de ella. A este módulo se accede desde la página que se presenta en la figura 3, y la primera pantalla que se visualiza es la siguiente:

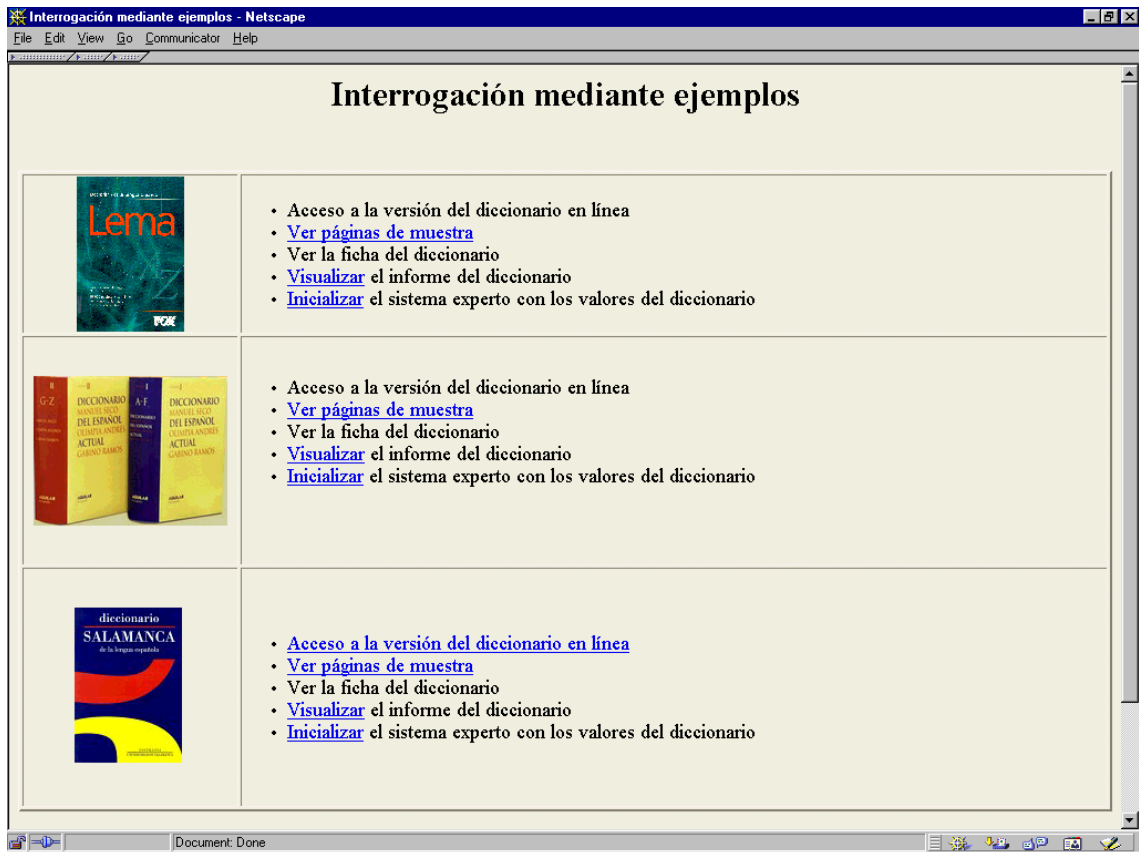


Figura 6. Interrogación Mediante Ejemplos

En la versión disponible actualmente, se prevé la posibilidad de tomar como diccionarios de referencia el *Diccionario Lema*, de Vox, el *Diccionario Salamanca*, de Santillana, y el *Diccionario del Español Actual*, de Aguilar, para el castellano, y el *Diccionari de la llengua catalana* de l'Institut d'Estudis Catalans, el *Gran Diccionari de la Llengua Catalana* d'Enciclopèdia Catalana y el *Gran Diccionari 62 de la Llengua Catalana* d'Edicions 62 para el catalán.

Desde esta página se puede acceder a distintos elementos relacionados con los diccionarios:

- la versión en línea (si está disponible),
- unas páginas de muestra,
- el informe de la obra, y
- el sistema experto inicializado con los valores de dicho diccionario.

Este último enlace abre una interfaz igual que la reproducida en la figura 4, con la particularidad que las respuestas están marcadas según el diccionario seleccionado. El usuario sólo tiene que confirmar las respuestas clicando encima. Sin embargo, esta interrogación es muy flexible y permite cambiar todas las variables que se desee respecto de la obra guía, ya que la intención es orientar al usuario y no obligarlo a adoptar unas soluciones determinadas.

Como en el sistema de interrogación anterior, después de responder todas las preguntas, el programa genera automáticamente un informe en el que se detallan las características de la obra en proyecto.

En resumen, se trata de dos sistemas de interrogación con pequeñas diferencias que conducen a un mismo resultado: un informe sistemático en el que se indican de modo preciso las características de la obra lexicográfica que se está diseñando.

3. Conclusiones y perspectivas de futuro

En este artículo hemos presentado las características de la ETL, una herramienta modular y amigable que ejemplifica claramente la relación estrecha que se ha establecido entre la ingeniería lingüística y la lexicografía. La aplicación principal de esta estación es el diseño de productos lexicográficos en un entorno profesional, pero no se puede ignorar su posible aplicación en el ámbito de la docencia, ya que puede resultar un instrumento útil para llevar a cabo análisis sistemáticos de obras existentes o para definir proyectos lexicográficos ficticios con finalidades docentes.

La versión que hemos presentado corresponde a las fases iniciales del trabajo lexicográfico, para las que la ETL actúa como un asistente para la documentación inicial y la elaboración de un diseño lexicográfico. Este asistente es flexible, ya que permite tomar la iniciativa al profesional, pero le ayuda a sistematizar y controlar su toma de decisiones. Otra ventaja evidente es la generación automática de un informe que refleja esa suma de decisiones en el planteamiento del diseño del diccionario, de manera abierta o a partir de ejemplos ya preexistentes.

Durante el año 2002 tenemos previsto trabajar en la generación automática de una estructura de base de datos lexicográfica a partir del diseño elaborado. Con ello iniciamos la fase de gestión lexicográfica, que conducirá a la redacción del diccionario, con asistentes que permitan nuevas tomas de decisiones no previstas en el diseño y sobre todo diversos controles sobre la selección de la nomenclatura, las remisiones, las definiciones u otro tipo de informaciones de la microestructura del diccionario. Siguiendo con la idea de integración modular no se descarta la posibilidad de incorporar nuevos recursos o herramientas de gestión ya existentes, ni la actualización de recursos ya integrados. Esta nueva fase del proyecto estará también financiada por el MCYT y se está elaborando en cooperación con la empresa SPES. A partir de estos resultados, se prevé solicitar soporte financiero para concluir versiones completas (que incluyan las fases de edición y postedición de diccionarios, en papel y en soporte electrónico) para el próximo bienio 2003-04.

Más adelante, como ya indicábamos al inicio de esta comunicación, está previsto el desarrollo de nuevas versiones del prototipo para la elaboración de otros diccionarios, bilingües en primera instancia.

La filosofía de las estaciones de trabajo del IULA —modularidad, integración de recursos y herramientas existentes, formatos estándares, tecnología multiplataforma y adecuación ergonómica a las necesidades y a las tareas del profesional— nos permite augurar un buen futuro a la serie iniciada con la ETL, en proyectos paralelos para otro tipo de profesionales de la lengua, como ya está sucediendo con el inicio de un nuevo proyecto cooperado para la realización de una estación de trabajo para el asesor lingüístico.