6. Conclusiones

Durante este trabajo he desarrollado una herramienta flexible y fácil de usar a partir de Ecode, logré optimizar el sistema original y lo expandí para abarcar un poco más del análisis de los contextos definitorios. Esta herramienta servirá como base para futuros sistemas similares.

6.1 Recapitulación del trabajo

Durante la realización de esta tesis tuve que aprender teoría lingüística y de PLN y aplicar mis habilidades como desarrollador y programador en la práctica.

Estudié Ecode en su totalidad y pude encontrar las necesidades inmediatas del sistema y resolverlas. Busqué parámetros adecuados para su medición, además de permitir que los usuarios extrajeran contextos definitorios durante todo el desarrollo del presente sistema. Se lograron los objetivos de la tesis. Las conclusiones se exponen a continuación.

6.2 Conclusiones

Las conclusiones, al igual que el desarrollo del trabajo, las divido en tres partes, a saber, la optimización, la expansión y finalmente la adaptación.

6.2.1 Conclusiones sobre la optimización

La optimización fue significativa en cuanto al núcleo del Ecode: se redujo la lectura y escritura al disco duro, de modo tal que únicamente el acceso al disco es en los módulos de entrada y de salida, lo que repercute de manera directa en el desempeño del sistema.

Se redujo el código a un 80%, lo cual impacta directamente en el tiempo de ejecución del sistema, además de presentar un código más legible, más organizado y más fácil de modificar en futuros desarrollos.

De la reducción de operaciones de entrada y salida, y de la reducción del código se desprende directamente el tiempo de ejecución, que es una medición que muestra el impacto de los puntos anteriores en el sistema. Para el núcleo del Ecode se logró reducir el tiempo de ejecución en aproximadamente 10%, lo cual quizá no parezca significativo, pero sí lo fue para Ecode, pues se tiene que tomar en cuenta que el tiempo de ejecución crece de manera exponencial con el tamaño de los datos.

Se redujo el tiempo que tarda el sistema en etiquetar las palabras con su categoría gramatical en un 95%; además, este proceso no estaba integrado al sistema antes y era necesario llevarlo a cabo.

En la medición de la precisión y la exhaustividad influyen factores propios de la expansión, sin embargo, cabe mencionar que después de la optimización y de agregar los módulos de entrada y salida de la expansión, la precisión se redujo de 41% a 39%, y la cobertura aumentó de 82% a 88%.

6.2.2 Sobre la expansión

La expansión consistió en añadir un módulo de entrada y uno de salida que permitiera un abanico de opciones de entrada o de salida de acuerdo con las distintas necesidades.

En el módulo de entrada se agregó un separador de oraciones, se integró el etiquetador POS TreeTagger y se permitió el paso de información durante el sistema por medio de etiquetas especiales al inicio de las líneas de texto.

El módulo de salida consiste en la presentación de los CDs con sus diferentes conjuntos de etiquetas (POS, de partes del CD y especiales), la agrupación por términos semejantes y la búsqueda de un solo término en la salida.

6.2.3 Sobre la adaptación

Adaptar el sistema consistió en condicionarlo para que además de funcionar como una herramienta individual, funcione como una librería que pueda ser incluida y utilizada en futuros proyectos, sin necesidad de reescribir o proveer una futura adaptación; esta librería es el núcleo del sistema y está escrita en Perl.

Otro punto de la adaptación fue permitir a Ecode interactuar con la base de datos de Describe –y en general puede interactuar con cualquier base de datos con una ligera adaptación–. Para esto ya está programada la conexión con las bases de datos, tanto para la obtención de los datos como para insertar los contextos de salida con los datos pertinentes a estos, el origen, término buscado, línea en la que aparece, etc.

Estas adaptaciones a Ecode ya se encuentran plenamente funcionales y listas para su aplicación a futuros proyectos.

6.3 Trabajo futuro

El Ecode, si bien es una herramienta completa e integral, requiere de mucho desarrollo para abarcar las necesidades que no fueron planeadas pero fueron surgiendo durante la realización del sistema.

El trabajo futuro consiste en varias acciones tanto de carácter lingüístico como ingenieril y las describo a continuación:

6.3.1 Trabajo futuro sobre los módulos de entrada y salida

Los módulos de entrada y salida requieren de mayor desarrollo para tener un sistema que permita obtener y entregar los resultados aún con más flexibilidad que la que actualmente se tiene.

El módulo de entrada requiere una librería o programa que permita extraer el texto de los PDFs con mayor precisión y para diferentes versiones de este formato de archivo. Requiere además extracción de texto en formato de Open Office (odt) y también de otros formatos de archivo relevantes (DjVu, PowerPoint, Ghost Script, Latex, etc.).

Además se requiere explorar, con la ayuda de un lingüista, las separaciones de las oraciones dentro de los archivos, ya que no siempre son separadas de manera adecuada, ya sea porque vienen mal separadas de origen o por el separador de oraciones implementado.

El etiquetador podría mejorar en gran medida los resultados de Ecode si fuera entrenado con un texto especializado que contenga contextos definitorios y si se contara con un experto que pueda indicar las palabras que son etiquetadas de manera correcta y las que no. Además, algunas palabras funcionales, que pueden tener diferente categoría gramatical según su contexto, frecuentemente son confundidas por el etiquetador y debería agregarse la opción de corregir dichas palabras dentro del módulo de entrada.

El problema más complicado que requiere de solución es el manejo de las codificaciones de caracteres, debido a que las múltiples entradas de datos por lo regular manejan codificaciones diferentes, son difíciles de controlar y las herramientas disponibles son ineficientes. Este sistema tiene un módulo para el manejo de las codificaciones, pero se requiere de una investigación sobre los tipos de codificaciones posibles en las entradas y cómo distinguir cada una de ellas, ya que algunas veces es imposible determinarlo.

En el módulo de salida es necesario mejorar el agrupador de términos semejantes con la ayuda de un lingüista, ya que esta tarea es complicada y requiere de un análisis para determinar en cuáles circunstancias es conveniente agrupar los términos. De manera similar, es necesario profundizar en el filtrado de términos erróneos o mal formados que suelen aparecer en la salida.

Se necesita implementar que la salida sea en XML, crear una hoja de estilo para que la presentación sea más agradable. Además se pretende que haya una estandarización en las etiquetas de los CDs, tanto en el CORCODE como en los otros proyectos existentes dentro del GIL y en los proyectos futuros.

Otro requerimiento es permitir la configuración de los accesos a bases de datos en un archivo o en los argumentos que recibe al ejecutarse. Por ahora es estático y la base de datos de Describe cambia su estructura regularmente, por ello la única forma de configurarlo para los cambios es en el código.

6.3.2 Trabajo futuro sobre el núcleo del sistema

El núcleo del sistema requiere de varios cambios para hacerlo configurable en sus parámetros y que por lo tanto pueda ser acondicionado a los requerimientos de futuras investigaciones.

Se debe implementar una interfaz de configuración con las gramáticas ya existentes. Buscar como agregar a las gramáticas parámetros que aun no son configurables, como la escritura de etiquetas POS, agregar nuevos filtros, etc.

Si bien el sistema cubre una gran cantidad de casos en la identificación de los contextos definitorios, aún queda mucho por investigar. Además deberían agregarse al Ecode los recientes avances en la teoría lingüística sobre contextos definitorios.