**Tesis “Recopilación y Almacenamiento de Corpus de Idiomas en Peligro de Extinción"**

"El *Corpus* existente de nuestras lenguas nacionales es reducido, por ello incrementar el *Corpus* es una tarea indispensable. Sin embargo, colectar el número de voces y textos necesarios es una tarea extenuante para un pequeño número de personas, por ello todo método disponible de recolección basada en la participación masiva de voluntarios debe ser empleado sin descuidar la calidad del audio.

Es conocido que la recolección de Corpus, actualmente sigue alguno de los siguientes procesos:

1. la búsqueda de textos escritos y la grabación de las lecturas de estos por un grupo suficiente de personas. En este caso los textos deben estar debidamente revisados por seres humanos (hasta que un *spelling checker* esté debidamente desarrollado). Además, los textos deben recoger diversos estilos literarios y al mismo tiempo ser ricos fonéticamente hablando.
2. la grabación de voces y la posterior transcripción de las grabaciones siguiendo un guion apropiado. Consideramos que esta última modalidad es lo más pertinente, pero requiere de un tiempo mayor.

El repositorio de Corpus es la herramienta necesaria para uniformizar la calidad, para concientizar sobre la cantidad de voces debidamente transcritas y/o textos debidamente leídos requeridos e incluso, en el extremo más favorable, para dirigir y/o indicar aquello que debe ser registrado. Hay una necesidad urgente de garantizar que los investigadores puedan tener acceso a las colecciones de datos, más allá de la vida útil de sus respectivos proyectos. Las cuestiones de sostenibilidad y preservación son cada vez más importantes para la comunidad científica; ya existen esfuerzos notables de estandarización tales como OLAC[1], E-MELD[2], y los agregadores de metadatos tales como la Infraestructura de Repositorios Digitales Europeo[3].

Dos aspectos importantes de la sostenibilidad son:

1. perpetuar el acceso al Corpus independientemente de la duración de la tesis.
2. la apertura de los sistemas operativos, herramientas y aplicaciones usados para construir los repositorios. Existe el gran peligro de que los repositorios de recurso idiomáticos se conviertan en carísimos cementerios de datos si las herramientas de acceso, visualización y búsqueda de recursos se vuelven obsoletas.

Existen dos grandes retos. En primer lugar, debido a la heterogeneidad de información que necesita ser almacenada y revisada, la interfaz debe ser lo suficientemente específica para permitir a los usuarios encontrar la información que ellos están buscando. En segundo lugar, debido a la diversidad de potenciales usuarios, la interfaz de consulta tiene que ser diseñado para favorecer la simplicidad.

Una manera sencilla de resolver este problema es adherirse a un estándar de anotación y codificación particular de modo que sólo una interfaz común necesite ser mantenida para acceder a toda la gama de recursos. Esa es la estrategia que se seguirá en esta tesis, en primer lugar, se hará una revisión de las iniciativas en curso (por ejemplo [5]), se seleccionará la más adecuada, se adaptará para el uso específico planteado, se difundirá y se participará del desarrollo internacional de la herramienta.

[1] [http://www.language-archives.org]( http://www.language-archives.org)

[2] <http://emeld.org>

[3][http://www.driver-repository.eu](http://www.driver-repository.eu/" \t "_blank)