

Creación de un corpus electrónico bilingüe en Ciencias Ambientales (CEBCA)

CLARA HELENA BELTRÁN¹

ANDRÉS TORO²

DAVID ABELLA³

KAREN BELTRÁN⁴

Resumen

Este proyecto consiste en la creación de un corpus electrónico bilingüe en Ciencias Ambientales como herramienta de investigación, estudio y traducción. El corpus textual se delimita, según Suárez (2004), ya que consta de artículos de especialidad media, media-alta y alta, de los cuales se extraen los contextos que contienen los términos que conforman el corpus electrónico bilingüe. Esta herramienta se alojará en el micro-sitio del Instituto de Lenguas de la Universidad Santo Tomás. En este recurso se encontrarán los términos de la búsqueda en sus respectivos contextos y los hipervínculos a la fuente original de donde fueron extraídos. Se extrajeron aproximadamente 10.000

-
- 1 Clara Helena Beltrán: Docente- investigadora Instituto de Lenguas Fray Bernardo de Lugo O.P. Grupo de Investigación USTA LEARNING. Universidad Santo Tomás Bogotá. clarabeltran@usantotomas.edu.co
 - 2 Andrés Toro: Docente- investigador Instituto de Lenguas Fray Bernardo de Lugo O.P. Grupo de Investigación USTA LEARNING. Universidad Santo Tomás Bogotá josetorog@usantotomas.edu.co
 - 3 Wilson David Abella: Docente- investigador Instituto de Lenguas Fray Bernardo de Lugo O.P. Grupo de Investigación USTA LEARNING. Universidad Santo Tomás Bogotá wilsonabella@usantotomas.edu.co
 - 4 Karen Beltrán: Auxiliar de Investigación. Egresada Ingeniería Ambiental. Universidad Santo Tomás Bogotá. karenn.beltran@usantotomas.edu.co

contextos cuyos candidatos a término fueron validados por herramientas de gestión terminológica. Así, se brinda una herramienta que cubre, en parte, las necesidades de quienes buscan correspondencias de términos en textos paralelos –inglés-español, español-inglés– en textos de Ciencias Ambientales.

Palabras clave: corpus electrónico bilingüe, ciencias ambientales, bases de datos, herramientas de gestión terminológica, contextos.

Creation of a Bilingual Electronic Corpus of Environmental Sciences (CEBCA)

Abstract

This project is about the creation of a bilingual electronic corpus on environmental sciences as a source of study, research and translation. The textual corpus is delimited according to Suárez (2004), because it takes into account medium, medium-high and high specialized leveled articles from which contexts containing the terms, make up the bilingual electronic corpus. This tool will be hosted on the micro-site of the Instituto de Lenguas at Universidad Santo Tomás. On the site, the terms from the search will appear within their contexts together with the original source links from where they were extracted. About 10000 contexts were taken from where the term candidates were validated by terminology management systems. Thus, we are offering a tool that, in part, helps to cover the needs of those who want to look for term correspondences in parallel texts –english-spanish, spanish-english– in environmental science texts.

Keywords: Bilingual electronic corpus, environmental sciences, data base, terminology management tool, contexts.

Justificación

Los problemas estudiados por las Ciencias Ambientales, son consecuencia del rumbo de evolución que la sociedad ha tomado y nace como resultado del comportamiento de las personas, de las formas en las que la sociedad se ha relacionado con la naturaleza y el medio ambiente. Dicha relación es inminente y necesaria, de tal manera que se hace imperativo el cuidado que le debemos dar al medio ambiente para prolongar nuestra calidad de vida en armonía con la naturaleza. De esta manera, se ha hecho necesario que el estudio científico del área de las Ciencias Ambientales se profundice y expanda para garantizar el cuidado y mantenimiento del mismo.

La complejidad de estas relaciones se presenta como una oportunidad para que la academia facilite y articule el trabajo de los profesionales formados en las diferentes especialidades de las Ciencias Ambientales, mediante el desarrollo de herramientas que les permitan interactuar en el mundo del análisis de la problemática ambiental con visión amplia y holística, para que desarrollen en mejores condiciones su actuar y puedan proponer cambios en el manejo sustentable de nuestro planeta. De esta manera, y debido a la gran cantidad de textos científicos en inglés producidos en el área, y como medio para facilitar la tarea de estudiantes, investigadores y traductores, decidimos llevar a cabo la creación de un corpus electrónico bilingüe inglés-español en Ciencias Ambientales para que sirva como herramienta de consulta y referencia. Dicha herramienta brindará la posibilidad de encontrar términos relacionados con las Ciencias Ambientales en contextos extraídos de artículos en inglés-español. El usuario no solo encontrará la equivalencia del término en la otra lengua, sino que también podrá acceder al texto original del cual es tomado dicho término.

Objetivos de investigación

Objetivo general

Crear una herramienta electrónica bilingüe en Ciencias Ambientales que comprenda corpus paralelos (inglés-español, español-inglés junto con el manejo y la presentación del vocabulario especializado.

Objetivos específicos

1. Proveer una herramienta que ayude en la elaboración de futuros trabajos de investigación y traducción en temas referentes a las Ciencias Ambientales.
2. Establecer la equivalencia de los términos en contextos reales bilingües en Ciencias Ambientales.
3. Ofrecer vínculos a las fuentes de textos o artículos originales desde la ventana de resultados de la herramienta CEBCA.

Introducción

El presente trabajo se basa en la creación de un corpus electrónico bilingüe en Ciencias Ambientales cuyo interés es el de contar con una herramienta virtual de almacenamiento de textos escritos fiables (español-inglés), los cuales sirven como base de investigación y estudio para investigadores, profesores y estudiantes interesados en el ámbito de las Ciencias Ambientales, así como también a docentes y traductores del ILFBL y en general cualquier persona interesada en el análisis lingüístico.

Metodología

Esta investigación (que aún se encuentra en curso) se enmarca dentro del enfoque de corpus que encierra procesos cuantitativos y cualitativos. Como punto de partida, en la primera fase se conforma el corpus textual que consta de textos paralelos inglés-español, español-inglés que se encuentran en forma digital en organizaciones, entidades científicas, revistas científicas, libros y bases de datos relativas a temas de las Ciencias Ambientales.

Después de la creación del corpus textual, se hace la extracción de candidatos a término utilizando algunas herramientas web utilizadas para este fin, como TermoStat Web 3.0 o Term Extraction-Fivefilters. Estas herramientas evalúan y validan los candidatos a término referentes a las Ciencias Ambientales y la frecuencia en la que aparecen

en el corpus textual. Este análisis cuantitativo de las frecuencias nos permite categorizar los términos que se incluyen en el corpus. A continuación, se identifican los contextos donde se encuentran los términos validados y se extrae dicho contexto. Posteriormente, se hace el registro de los contextos en una base de datos de Excel, destacando los términos entre asteriscos para fines técnicos en el diseño de la herramienta. Finalmente, debido a su carácter de corpus electrónico, se crea la base de datos en la nube. Para el desarrollo de este programa, recurrimos a un ingeniero especializado en desarrollo de software y creación de corpus electrónico de la Universidad de Antioquia, quien está encargado de crear la plataforma que finalmente se alojará en el portal web del ILFBL de USTA-Colombia.

Con el fin de conformar el corpus textual se recurrió a fuentes web donde se encuentran textos paralelos de especialidad media y alta, entre las que se encuentran: SciDev.Net, GreenFacts, ScienceDirect, Spinger, Cordis, EcoLexicon, United States Environmental Protection Agency, IDEAM, AGROVOC Tesauro, entre otras. Otras fuentes esenciales para esta parte del proyecto fueron las revistas de difusión científica y tecnológica, journals y publicaciones hechas por entidades involucradas en la investigación del medio ambiente. Por último, se acudió a los repositorios de las universidades donde se encuentran los resúmenes de las tesis y artículos escritos por docentes e investigadores de los grupos de investigación en facultades relacionadas con el medio ambiente.

Para esta conformación del corpus se tienen en cuenta ciertos criterios importantes: en primer lugar la autenticidad, que hace referencia al carácter real que debe tener la compilación, que en ningún caso ha de estar compuesta por textos creados *ad hoc*. Por ser esta una herramienta web, el formato de los textos seleccionados debe ser electrónico para garantizar su procesamiento informático. Además, las dimensiones del corpus han de ser lo suficientemente representativas como para que los datos que se extraigan de él resulten relevantes. Una vez seleccionado el corpus textual y los candidatos a término, se continúa con la extracción automática de términos, una técnica que se lleva a cabo en gestión terminológica y es útil en la creación de recursos como léxicos, diccionarios, glosarios especializados, corpus electrónicos y traducción automática, entre otros. Después de esta extracción, se prosigue con

un escrutinio manual replicado por Roldán y Fernández (2012), en el cual se hace una revisión manual de los datos arrojados por las herramientas, con el fin de identificar términos ya que es necesario tener en cuenta que los programas gestores de corpus que emplean métodos estadísticos producen mucho “ruido”, y en algunas ocasiones no extraen los sintagmas nominales completos (Beltrán et al., 2014).

Ya identificados los términos, se extraen los contextos de donde han sido identificados y se registran en una base de datos en la cual se resaltarán. Finalmente, se procede a cargar la base de datos en el portal web del ILFBL a través del servicio técnico del webmaster de la Universidad Santo Tomás.

Con la implementación de esta herramienta y luego de la etapa de divulgación de la misma, se espera que se inicie la fase de consulta por parte de los usuarios finales y que allí puedan encontrar las equivalencias contextualizadas tanto en inglés como en español de los términos buscados. Por otra parte, se espera también que en sus resultados encuentren dichos términos no solamente en variados contextos, sino que también hagan uso de los vínculos que los llevan a las fuentes de consulta originales y puedan dar amplitud a su búsqueda.

Referencias

- Beltrán, C. et al. (2014, octubre). Propuesta metodológica para la identificación y validación de neónimos. En *Debate Terminológico*, (11), 8-26.
- Suárez, M. M. (2004). *Análisis contrastivo de la variación denominativa en textos especializados del texto original al texto meta*. Tesis doctoral. Universidad Pompeu Fabra.
- Vargas, C. (2006). Diseño de un corpus especializado con fines terminográficos: el corpus de la piedra natural. *Debate Terminológico*, 2(7). París: RITERM (Red Iberoamericana de Terminología).
- Roldán, M., y Fernández, J. (2012). Emergent neologisms and lexical gaps in specialized languages. In *Terminology*, 18(1), 9-26. John Benjamins Publishing Company.