Informe sobre los objetivos y la metodología del proyecto

Objetivos

Este proyecto se proponía dos objetivos generales: a) la caracterización de los neologismos léxicos en general y el análisis contrastivo entre los neologismos generales y los especializados en distintos ámbitos del saber, y b) el desarrollo de herramientas para el trabajo neológico.

Estos objetivos se desplegaban en objetivos específicos que, para que resulten más claros los resultados, resumimos en los siguientes:

- Objetivo 1. Elaborar un corpus general de prensa.
- Objetivo 2. Elaborar un corpus de especialidad en cinco ámbitos de tallas aproximadas.
- Objetivo 3. Anotar dichos corpus automáticamente.
- Objetivo 4. Utilizar y adaptar en su caso las herramientas de procesamiento del IULA.
- Objetivo 5. Crear una base de datos textuales accesible.
- Objetivo 6. Crear o adaptar una herramienta de exploración de corpus con finalidades neológicas.
- Objetivo 7. Crear un corpus lexicográfico de exclusión específico y adecuado para detectar neologismos especializados. Creación de un diccionario máquina acumulativo a partir de las unidades terminológicas del corpus IULA, de los datos recogidos en el período anterior al analizado por el Observatorio de Neología.
- Objetivo 8. Crear o adaptar una herramienta de análisis automático de la estructura de los neologismos para poder clasificarlos morfosintácticamente.
- Objetivo 9. Caracterizar los neologismos generales y los especializados por separado y por ámbitos de especialidad.
- Objetivo 10. Extraer los contrastes relevantes.
- Objetivo 11. A la vista de los datos, formular cuestiones. Se ha tratado de dar respuesta a tres cuestiones específicas:
    a. ¿Presentan diferencias los ámbitos de especialidad y el ámbito general de la lengua en el uso de unidades neológicas tomadas como préstamos de otras lenguas?
    b. ¿Presentan diferencias estructurales los neologismos de los ámbitos de especialidad y los del ámbito general de la lengua?
    c. ¿Serían neologismos las unidades detectadas en caso de utilizar Internet como corpus? ¿En qué grado están estabilizados en el uso en la red?

Metodología

A continuación se detallan las diferentes etapas incluidas en la metodología del proyecto:

1. **Constituir un corpus textual especializado de donde extraer neologismos terminológicos**. El corpus debía estar formado por textos publicados entre el 2008 y el 2013. Los ámbitos temáticos de los textos seleccionaron debían ser Derecho, Economía, Informática, Medicina y Medioambiente, que son los ámbitos que incluye el Corpus Técnico (CT) del IULA.

1.1. *Seleccionar los textos*. Debían ser textos en castellano peninsular de nivel de especialización alto, concretamente 2 tesis doctorales y 10 artículos de investigación por año y ámbito. Para realizar la selección de las subáreas de cada ámbito se empleó principalmente el árbol de campo correspondiente al CT. Se intentó cubrir la mayor parte de subáreas de cada ámbito para obtener un corpus representativo y equilibrado.

1.2. *Procesar los textos con las herramientas semiautomáticas del IULA*. Para realizar este procesamiento, se elaboró un protocolo en el que se incluyeron todos los pasos necesarios (ver Anexo 1). En síntesis, las etapas de procesamiento fueron:

- Convertir los documentos de PDF a texto plano.
- Insertar la información de los textos en una base de datos bibliográfica Access (título, autores, año de publicación, editorial, etc.).
- Marcar el texto estructuralmente y preprocesarlo (análisis morfológico y desambiguación).
- Crear la cabecera y otros ficheros auxiliares.

2. **Extraer automáticamente los términos del corpus especializado mediante WikiYATE**, un sistema de extracción de candidatos a término que se basa en la Wikipedia. Para ello, se elaboró otro protocolo con los pasos que debían seguirse (ver Anexo 2). Este sistema ofrece diversas informaciones sobre cada unidad extraída, como por ejemplo un coeficiente de dominio. Cuanto más alto es el coeficiente, más probabilidades hay de que la unidad sea un término del ámbito. En función de este coeficiente, las listas de unidades obtenidas para cada uno de los cinco ámbitos se clasificaron en cuatro bloques:

- 1: coeficiente 0,99-0.01.
- 2: coeficiente 1.
- 3: coeficiente 0.
- 4: coeficiente -1.

Dado que cualquier sistema automático es susceptible de cometer errores, se realizó una **revisión manual de los candidatos a término detectados**, con el objetivo de eliminar tanto los candidatos que no eran términos como aquellas unidades que no eran propias del ámbito. Para realizar esta revisión de una manera más ágil, se desarrolló una interfaz que permite observar información relacionada con cada unidad, como por ejemplo su coeficiente de dominio, el contexto de uso, etc. Asimismo, se elaboró un protocolo para la revisión de los candidatos a términos (ver Anexo 3). Se organizaron las unidades detectadas en función de dos parámetros:

a) *Condición de los candidatos a término*. La clasificación incluyó tres posibilidades en este sentido:
   - 0: la unidad no es un término (porque no es del ámbito o porque es una unidad fraseológica).
   - 1: la unidad es un término del ámbito.

- 2: la unidad es dudosa, tanto en relación con el ámbito como con la fraseología.

b) *Patrón que incluye un núcleo deverbal*. En el caso de las unidades poliléxicas consideradas términos o dudosas (pero no en las descartadas) se especificó si el núcleo era un nombre deverbal. Esta propuesta se basa en el hecho que hay algunos patrones que tienen más posibilidades de ser términos y, en cambio, otros patrones son más prototípicos para la fraseología. Esto ocurre sobre todo en el caso de las estructuras NJ (nombre + adjetivo) con un núcleo nominal deverbal y en menor grado en los patrones NPN (nombre + preposición + nombre) deverbales (ya que acostumbran a ser fraseología). En cambio, estas mismas estructuras sin núcleos deverbales, aunque también pueden ser términos, no son tan prototípicas.

3. **Comparar la lista final de los términos detectados (2008-2013) con un corpus lexicográfico de exclusión**. Este corpus lexicográfico de exclusión incluye:

- Una lista de términos obtenidos automáticamente de cada uno de los ámbitos del CT (que incluye textos de 1994 a 2014, de los cuales se seleccionaron los anteriores a 2008) mediante WikiYATE, revisados manualmente.
- Diccionarios en línea publicados antes del 2008[1]: SNOMED para el área de medicina.

Mediante esta comparación **obtuvimos una lista de neologismos terminológicos** procedentes de nuestro corpus especializado (2008-2013), que llamamos Lista APLE_NEOESP.

4. **Para cada ámbito temático, analizar los fenómenos siguientes**:

- Grado de interferencia de otras lenguas en el ámbito: número de préstamos sobre el total de neologismos terminológicos.
- Tendencias estructurales: número de neologismos terminológicos por estructuras (N, NJ, NPN, etc.).
- Cuantificación de los neologismos del ámbito: ocurrencias de los neologismos terminológicos en Google Académico (restringiendo la lengua al castellano).

Para cada ámbito, se selecciona un 10% del total de candidatos a término detectados inicialmente. De acuerdo con este análisis, se observó, para cada ámbito, el índice de penetración de los préstamos, las estructuras predominantes y la permeabilidad de los neologismos.

5. **Extraer del banco de neologismos OBNEO[2] una lista de neologismos generales de 2008 a 2013**, que llamamos Lista APLE_NEOGEN. El corpus que se emplea en OBNEO para extraer estos neologismos incluye diversos periódicos nacionales en castellano (*El País* y *La Vanguardia*).

6. **Analizar en la Lista APLE_NEOGEN los tres fenómenos indicados en el punto 4**.

7. **Comparar las dos listas de neologismos:** Lista APLE_NEOESP (neologismos terminológicos) vs. Lista APLE_NEOGEN (neologismos generales). El propósito de la comparación fue el

[1] Los demás diccionarios detectados en línea no incluían la fecha de edición o bien eran posteriores a 2007. En el caso de los diccionarios libres del TERMCAT y de los diccionarios que se crearon en el marco del proyecto de investigación TEXTERM III, se trataba de obras con actualizaciones posteriores a esta fecha. Además, los que pasaban estos filtros presentaron dificultades técnicas porque contenían entradas multilingües y otros elementos en el campo *entrada*.

[2] Observatori de Neologia: http://www.iula.upf.edu/obneo/

contraste cuantitativo, la detección de coincidencias y/o divergencias en las unidades, y el contraste entre tipos de estructuras.

ANEXO 1: Protocolo de procesamiento de los textos del corpus (ProcesoCTaple2)
ANEXO 2: Protocolo de extracción de candidatos a término (WikiYATE).
ANEXO 2: Protocolo de revisión de candidatos a término (Guía de revisión).
ANEXO 4: Protocolo de elaboración de los informes por ámbito (Guia de redacción).