Informe sobre el proceso de selección de los neologismos generales y terminológicos

1. HERRAMIENTAS DISPONIBLES

- Recursos del Observatori de Neologia (Plataforma, Buscaneo y Gestor de diccionarios)
- Cadena de procesamiento de textos (preproceso, etiquetador morfológico y lematizador) (ProcesoCTaple2)
- Extractor de términos WikiYATE, creado por Jordi Vivaldi

2. CONSTITUCIÓN DE LOS CORPUS

El **corpus de neologismos generales** se extrajo del banco de neologismos del Observatori de Neologia, que contiene datos desde 1989. Para el proyecto se recuperaron los neologismos registrados en el banco entre 2008 y 2013 por el nodo coordinador del Obervatori, situado en la Universidad Pompeu Fabra de Barcelona.

El **corpus de neologismos terminológicos** del proyecto se constituyó a partir de un corpus de textos de especialidad que cubría las siguientes áreas temáticas, correspondientes a las existentes en el Corpus Técnico del IULA (en adelante, CT-IULA):

- Derecho
- Economía
- Informática
- Medio ambiente
- Medicina

Para la constitución del corpus textual, en un principio se tuvo en cuenta la cantidad de géneros textuales con el objetivo de que el corpus especializado fuera internamente equilibrado. Con este fin, se seleccionaron 12 tesis[1] y aproximadamente 30 artículos de revistas científicas especializadas por área. Los textos se escogieron teniendo en cuenta los siguientes aspectos:

- Período: 2008-2013 (6 años). Se seleccionaron los 6 años anteriores al inicio del proyecto (2013), ya que para el trabajo en neología no interesa estudiar períodos de tiempo muy alejados de la actualidad.
- Lengua: español peninsular.
- Nivel de especialización: alto.
- Longitud: se prefieren los textos breves para favorecer la variedad.
- Temática: se incorporaron en las secciones del árbol de campo del CT-IULA correspondientes.

[1] Finalmente se descartaron las tesis del corpus de especialidad definitivo, por la dificultad de procesamiento que comportaban, al tratarse de textos muy extensos.

El corpus definitivo constituido para el proyecto contiene 1.424.064 palabras. La información para cada uno de los ámbitos es la siguiente:

Ámbito	Formas	Lemas
Derecho	235.036	7.081
Economía	374.504	5.182
Medio ambiente	268.436	6.600
Informática	294.338	5.473
Medicina	251.750	6.182
Total	**1.424.064**	

Tabla 1. Formas y lemas por ámbito

Los miembros del proyecto procesaron los textos siguiendo la cadena de procesamiento de textos establecida en el CT-IULA (preproceso, etiquetado morfológico y lematización). Para la extracción de los neologismos terminológicos se utilizó el extractor que creó Jordi Vivaldi (denominado WikiYATE). Funciona con datos de la Wikipedia, lo cual permite obtener unidades neológicas, ya que es un corpus que evoluciona constantemente. Además, este extractor funciona con cualquier texto de entrada, es fácil de usar y proporciona una lista de candidatos a término bastante afinada, que incluye formas y lemas.

La aplicación del extractor permitió obtener una lista de candidatos a término de cada uno de los corpus constituidos. Para establecer su condición de unidades neológicas, se aplicó un corpus de exclusión, formado por textos especializados de períodos anteriores y por otras obras especializadas. De este modo, se obtuvo una lista de términos que hasta el momento no se habían utilizado en textos de fechas anteriores y, en el caso de medicina, además, no se habían recogido en diccionarios (SNOMED), es decir, obtuvimos una lista de neologismos terminológicos.

Las obras que constituyen el corpus de exclusión utilizado en el proyecto son las siguientes:

- CT-IULA (datos hasta 2004). Con el extractor WikiYATE se extrajeron los términos del CT-IULA (1994-2004) para la creación de un diccionario terminológico de los 5 ámbitos temáticos seleccionados.
- SNOMED (versiones anteriores a 2008), para el ámbito de medicina

El siguiente paso consistió en depurar los resultados, que se presentaban ordenados a partir de un índice de terminologicidad y dominio (porque el extractor puede detectar unidades que no son términos o que no son términos del ámbito especializado de estudio). La depuración se llevó a cabo a partir de la selección manual de las unidades, con la ayuda de una interfaz creada por Jordi Vivaldi a partir de los resultados de WikiYATE que permite observar la información relacionada con el término (índice de terminologicidad, índice de dominio, contextos de uso, etc.) para decidir si se trata de un término del ámbito o no.

La metodología para la selección de las unidades terminológicas siguió unos criterios para la selección de los términos del dominio, teniendo en cuenta, por un lado, los

patrones prototípicos y, por otro, la condición de núcleo deverbal de las unidades poliléxicas (ver el protocolo de revisión de candidatos a término, titulado *Guía de revisión*).

Con el fin de asegurar el equilibrio entre las unidades terminológicas seleccionadas para cada ámbito especializado, se decidió que el número de términos seleccionados debía corresponder aproximadamente al 10 % del total de candidatos inicial. El número final de unidades que constituyen el corpus de términos neológicos del proyecto es el siguiente:

Área temática	Candidatos iniciales	Candidatos seleccionados	
Derecho	6.401	880	13,70 %
Economía	5.126	624	12,70 %
Informática	4.532	320	7,10 %
Medicina	2.941	494	16,79 %
Medio ambiente	4.917	770	15,70 %
General	8.099	-	-

Tabla 2. Términos definitivos