Projecte TERMMED. Evolución del conocimiento científico en medicina: El cambio léxico y semántico (FFFI2017-88100-P)

Acta de la reunión 2 (Aprobada el 15/5/2020)

Fecha: 27 de febrero de 2019
Lugar: Barcelona, Universitat Pompeu Fabra, Campus Poblenou, sala 52.701
Hora: 11:30

Asistentes: Rosa Estopà, Alejandra López, Jorge Porras, Òscar Pozuelo, Isabel Santamaria, Ana Rosa Terroba, Andrés Torres, Laia Vidal, Jordi Vivaldi, Yingfeng Xu, Mercè Lorente i Zebing Lu.

Se han excusado: Judit Feliu y M. Teresa Cabré

Orden del día:

1. Aprobación acta de la reunión anterior
2. Incorporación de nuevos miembros al proyecto
3. Tema dentro de la medicina a tratar en el proyecto
4. Constitución y tratamiento del corpus textual
5. Paquetes de trabajo, tareas y calendario
6. Ruegos y preguntas

Desarrollo de la reunión y acuerdos

M. Lorente empieza a hablar sobre el acta de la primera reunión (ver acta nº 1), pregunta a los asistentes qué opinan y si están de acuerdo y todos los asistentes dicen que sí, que aprueban sin cambios el acta de la reunión anterior.

M. Lorente continua la reunión presentando dos nuevas investigadoras del proyecto: Ana Rosa Terroba, de la Universidad de la Rioja, y Zebing Lu, doctoranda del grupo IULATERM que realiza su tesis doctoral bajo la dirección de R. Estopà. Por lo tanto es necesario darlas de alta en el Ministerio.
M. Lorente aprovecha este momento de presentación de las nuevas incorporaciones para hablar de la web (https://www.upf.edu/web/termmed/inici), que ya está organizada con todos los miembros y también con la documentación general del proyecto. Propone también poner el enlace a las herramientas y a los corpus que se utilicen o que resulten del proyecto. Se han comenzado también a subir las publicaciones que hay hasta ahora y se continuará haciendo así. Es decir, subir a la web toda la información y la documentación relacionada con el proyecto.

R. Estopà comienza a hablar sobre el tema de medicina que se tratará en el proyecto. Comenta que se reunieron ella y M. Lorente para hablar del corpus textual, teniendo en cuenta los objetivos planteados inicialmente, para determinar cuál sería el tema que permitiera hacer un mejor análisis. Se tuvieron en cuenta diversas posibilidades:

reproducción asistida (corpus de Alicante), una de las principales ventajas es que, al ser un corpus relativamente joven, el estudio diacrónico sería más fácil. Otro posible tema discutido es el del cáncer, del cual se puede encontrar mucha información, lo cual es bueno, pero también puede ser contraproducente. También se pensó en el tema de las enfermedades raras, al ser un tema conocido ya en el grupo.

Finalmente, se decidió que la mejor opción era la de la reproducción asistida, ya que este tema, desde el punto de vista del conocimiento, seguramente tendrá cambios por su rápida evolución y, por lo tanto, es interesante observar cómo estos cambios se trasladan al ámbito lingüístico. Además, es un tema que tiene una fecha de nacimiento concreta, 1973 en el ámbito internacional y en el ámbito español se comenzó a hablar al respecto unos 11 años más tarde, en 1984. Teniendo esto en cuenta, M. Lorente y R. Estopà comenzaron a hacer una subdivisión para constituir los corpus de trabajo, teniendo en cuenta seis franjas temporales específicas de 7 años: 1973 a 1984[1], 1985 a 1991, 1992 a 1998, 1999 a 2005, 2006 a 2012 y 2013 a 2019. Comenzaron a hablar también de cantidad de palabras por corte y del tamaño final del corpus. Comenta R. Estopà que muchos de los textos ya están en el corpus de Alicante, solo es necesario buscar las fechas de los artículos.

A partir de aquí es que surge la primera tarea para el año 2019, la cual es recoger los textos para constituir el corpus. Por esta razón R. Estopà y M. Lorente consideran que es necesario distribuir los investigadores del proyecto en grupos de tres personas en los que cada investigador buscará 5 artículos de la franja temporal que les corresponda. Es importante que los textos que se seleccionen sean escritos por especialistas y destinados a especialistas, es decir, se buscarán solo artículos publicados en revistas especializadas en español y, si en algún caso es necesario buscar textos diferentes, también se podría tener en cuenta alguna tesis doctoral.

Para la segunda parte, después de haber escogido los textos, R. Estopà comenta que es necesario hacer el tratamiento del corpus. Explica que, después de haberlo hablado con M. Lorente, consideran que no es necesario perder tiempo con el etiquetaje del corpus, sino mejor usar directamente una herramienta como Terminus para hacer la extracción de los candidatos a término. Ò. Pozuello, J. Vivaldi y M. Lorente comentan sobre la posibilidad de hacer un etiquetaje del corpus, ya que, como lo explica J. Vivaldi, hoy en día es relativamente fácil hacer un etiquetaje de un corpus. M. Lorente dice que hacer esto sería una opción adicional para dejar un corpus construido al terminar el proyecto. Sin embargo, R. Estopà insiste en que Terminus permite hacer una buena extracción y rápida de los candidatos a término, la prioridad del proyecto. A. Montané comenta al respecto que, si se tiene un buen diccionario para entrenar Terminus, los resultados de la extracción mejoran significativamente. J. Vivaldi comenta que también se podría usar una herramienta como el WikiYate, pero convendría antes revisar qué tan actualizado

[1] Esta primera franja es más extensa para contrarrestar la posible falta de publicaciones durante los primeros años de evolución del ámbito.

está el tema de la reproducción asistida en la Wikipedia, ya que de esto dependerá en gran medida los resultados de una extracción con esta herramienta.

Después de hacer esto y tener una lista de términos validados se conformarán tres grupos de análisis, con los miembros del grupo que sean especialistas en lingüística, para estudiar el comportamiento del léxico en cada segmento temporal con relación al segmento posterior:

- analizar las novedades (nuevas unidades léxicas);
- analizar las ausencias (unidades léxicas que ya no se usan);
- analizar las coincidencias (lo que se mantiene)

Del tercer bloque se desprende un cuarto análisis en el que se aprovechará el trabajo que ha hecho A. Torres en su tesis para detectar neología semántica con métodos estadísticos sobre los focos de conocimiento, es decir, un análisis para detectar si las unidades léxicas que se mantienen conservan o no su entorno léxico intacto.
Cambiando de tema, R. Estopà comenta que, con respecto al calendario, había hablado con M. Lorente para que la primera parte de constitución del corpus estuviera lista para septiembre.

Continua M. Lorente comentando que, por ejemplo, Anna Veiga es profesora del centro regulador de genómica y estaría bien incorporar al proyecto especialistas asesores como ella o alguien de su equipo para que sean asesores científicos y para que también, en caso de ser necesario, nos ayuden a acceder a las revistas especializadas de pago de donde se extraerían los artículos. Así mismo, la ayuda de especialistas será importante para saber dónde buscar los artículos que constituirán el corpus y para saber cuál información es importante y cuál no. Al respecto, R. Estopà comenta que sería ideal seleccionar tres temas específicos de la reproducción asistida (por ejemplo: aspectos genéticos; técnicas y tratamientos; y fertilidad y esterilidad) y luego preguntarle a un especialista para que valide la selección.

M. Lorente habla de que habían estimado compilar un corpus de 90 artículos (15 por cada franja) tendrían aproximadamente 1 millón de palabras, más las 300 mil del corpus de Alicante. Expresa también que lo ideal sería tener lo antes posible el corpus de Alicante con las fechas de los artículos para así poder tener el corpus segmentado con las franjas equilibradas, desde el punto de vista de la cantidad de palabras de cada segmento. Es importante, entonces, tener definidos los temas, con ayuda de los especialistas, para buscar los artículos y luego segmentar el corpus equitativamente.

Aprovecha M. Lorente para hacer un resumen de lo que se ha hablado hasta ahora y las tareas que se han definido:

- Organizar el corpus de Alicante para saber qué artículos contiene, de qué fechas son, y así poder segmentarlos de acuerdo a las franjas temporales.
- Búsqueda de fuentes bibliográficas de las que se puedan tomar los artículos (listas de revistas especializadas en español).
- Buscar asesores científicos para el proyecto: por ej. Anna Veiga, para confirmar los tres temas a tratar.
- Buscar los artículos sobre cada tema, 5 artículos por tema por franja temporal, es decir, 15 artículos en total por cada una de las seis franjas.
- WikiYate. Ver cómo está el tema de reproducción asistida en la Wikipedia para contemplar la posibilidad de usar este extractor.
- Terminus: revisar los diccionarios de medicina disponibles en el IULA (también los de que tiene Vivaldi) para hacer entrenamiento y pruebas de extracción de candidatos a término con el corpus de Alicante.
- Detección de neología semántica con métodos estadísticos sobre los focos de conocimiento.

En cada paquete debería haber un responsable. Estos responsables no tendrán que buscar artículos:

1. Paquete WikiYate: J. Vivaldi será el responsable de este paquete y tendrá como punto de referencia a A. López. En este grupo estarán también Y. Xu, Ò. Pozuello y Z. Lu.
2. Paquete de Terminus: A. Montané liderará el paquete de Terminus y trabajará junto con J. Porras. En este grupo también estará L. Vidal.
3. Paquete de neología semántica: A. Torres junto con J. M. Torres se encargarán del paquete de detección de neología semántica.
4. Búsqueda de fuentes bibliográficas relevantes (los investigadores senior).

Teniendo esto en cuenta lo que hay que hacer, R. Estopà y M. Lorente diseñarán un protocolo de trabajo y harán un a propuesta de división para los grupos que buscarán los artículos. Igualmente, organizarán los investigadores que se encargarán del análisis lingüístico.

El calendario propuesto para las tareas de este año es el siguiente:

1. Constitución y exploración del corpus textual (marzo de 2019 a agosto de 2019).
2. Análisis de las unidades léxicas encontradas (septiembre, octubre y noviembre de 2019).
3. Informe de cada análisis (diciembre de 2019).

I. Santamaria aprovecha el turno abierto para preguntar las frecuencias de las reuniones. M. Lorente le contesta que estaría bien que fueran cada 3 meses. Y que, en

caso de que las personas que vienen de fuera no puedan hacerlo, se podrían hacer las reuniones por Skype en la sala de videoconferencias que tiene disponible la UPF.

M. Lorente aprovecha para preguntar si los del HIBA podrían colaborar en alguna cosa relacionada con la reproducción asistida. R. Estopà pensará en qué podrían ayudar, tal vez, compartiendo informes médicos sobre reproducción asistida para un análisis posterior.

Se cierra la reunión a las 13h30.