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

Acta de la reunión 3 (Aprovada 16/12/2020)

Fecha: 15 de julio de 2020
Hora: 11:30
Via telemàtica:
Join Zoom Meeting
https://zoom.us/j/98975997791?pwd=dHlydEp2WW9HMXcwb2RZRUtZOXdHdz09
Meeting ID: 989 7599 7791
Password: 531752

Asistentes: Mariona Arnau, M. Teresa Cabré, Ona Domènech, Rosa Estopà, Fabiana da Silva França, Alejandra López Fuentes, Mercè Lorente, Zebing Lu, Amor Montané, Jorge Porras, Òscar Pozuelo, Mariona Reverté, Isabel Santamaria, Ana Rosa Terroba, Laia Vidal, Jordi Vivaldi, Yingfeng Xu.
Se ha incorporado al final de la reunión: Andrés Torres,
Se ha excusado: Judit Feliu.
Damos por excusados, por posibles incompatibilidades horarias: Daniel Luna, José Castaño, Juan Manuel Torres-Moreno, Coralie Schneider.

Orden del día:

1. Aprobación del acta de la reunión anterior (27/1/19)
2. Cuestiones de gestión:
    a. Nuevos participantes y colaboradores en el proyecto
    b. Respuesta del Ministerio sobre el informe de seguimiento a medio proyecto.
    c. Petición de prórroga de un año: hasta diciembre de 2021.
3. Corpus de reproducción asistida del proyecto: Estado de la cuestión
4. Vocabularios de reproducción asistida: Creación del vocabulario de entrenamiento de TERMINUS 2.0
5. Herramientas de procesamiento:
    a. Entrenamiento de Terminus con el vocabulario creado.
    b. Otras informaciones: cambio de servidores, utilidades
6. Distribución de los análisis y calendario
7. Turno abierto

Desarrollo de la reunión y acuerdos

Sin modificaciones, se aprueba por asentimiento el acta de la reunión anterior, enviada con la convocatoria.

M. Lorente informa que, desde la reunión del año pasado, mediante el informe de seguimiento preceptivo, se ha notificado la incorporación de los nuevos miembros del

equipo de trabajo, que no se habían incluido en la solicitud inicial. Las novedades de los participantes desde entonces son las siguientes:

- La incorporación reciente en el equipo de trabajo de Mariona Arnau i Mariona Reverté, estudiantes de máster.
- El paso a doctores de los siguientes miembros: Andrés Torres, Coralie Schneider, Òscar Pozuelo y Yingfeng Xu, a quienes felicitamos.

Estas modificaciones en el equipo se notificarán este mismo mes de julio al Ministerio. Entre el equipo investigador y el equipo de trabajo somos 23 participantes en el proyecto. La web se ha actualizado con la inclusión de todos los miembros.

M. Lorente informa también de la respuesta del Ministerio al informe de seguimiento, que reproducimos parcialmente a continuación:

> El plan de trabajo es adecuado y se han completado los aspectos del proyecto previstos para la primera mitad. Los reajustes en el plan de trabajo y los cambios de personal están adecuadamente justificados. Si bien el corpus que se está creando y las herramientas analíticas son en sí un resultado excelente y de gran impacto sería deseable publicar en revistas indexadas en JCR o similar; se recomienda por tanto a los investigadores cuidar un poco más este aspecto en el futuro y tratar de posicionar su producción en medios de mayor impacto internacional.

Como se observa, nos emplazan a publicar más y en revistas de impacto.

M. Lorente expone la necesidad de solicitar una prórroga de un año, hasta diciembre de 2021, para poder alcanzar los objetivos previstos en el proyecto. Se ha consultado al Servei de Recerca de la UPF y nos han confirmado que los miembros de este proyecto pueden solicitar su inclusión en un nuevo proyecto de investigación en la próxima convocatoria, aunque nos concedan la prórroga para presentar resultados. Los asistentes aprueban la solicitud de la prórroga, que se tramitará al Ministerio este mismo mes de julio.

R. Estopà expone el estado del corpus (En anexo al acta se incluye el informe detallado). Recuerda que ha sido una tarea mucho más dificultosa que lo que estaba previsto. Finalmente, tenemos el siguiente material compilado:
Carpeta 1973-1984: 8 textos
Carpeta 1985-1991: 22 textos (abundan también los de cuestiones éticas y jurídicas)
Carpeta 1992-1998: 30 textos (que corresponden al 100% al corpus inicial de la UA)
Carpeta 1999-2005: 37 textos (17 de los cuales también estaban en el corpus de la UA)
Carpeta 2006-2012: 35 textos
Carpeta 2013-2019: 33 textos
Carpeta ENG 1973-1984: 39 textos en inglés
Carpeta extra 1956-1972: 2 textos de cuestiones ético-religiosas

Como podemos ver, los períodos que conviene completar son los más antiguos 1973-1984 y 1985-1991. R. Estopá propone intentar encontrar textos de estos períodos

durante este mes de julio, si es posible. Con lo que tengamos, en septiembre se podrán extraer los listados.

R. Estopà informa que han participado en este paquete de trabajo, además de las dos IPs, I. Santamaría, A. López Fuentes, M. Reverté y A. Torres Rivera, que segmentó el corpus de la UA y extrajo los metadatos de todos estos artículos.

Aquellos archivos pdf que eran solo de imagen se han convertido a formato pdf legible, tarea a cargo de A. López Fuentes, que informa que de los 168 textos 14 eran imágenes y se han convertido a pdf legibles, y solo ha dado error uno de ellos. Ahora conviene ver si TERMINUS 2.0 realmente los admite como pdfs legibles. M. Lorente cargará los documentos a Terminus en los próximos días e informará a todos los miembros de los resultados por correo electrónico.

Se debate sobre la necesidad de poner a disposición de todos los miembros una descripción completa del corpus. M.T. Cabré pregunta por qué hablamos de cantidad de textos y si no sería más adecuado hablar de número de ocurrencias para ver si los subcorpus están equilibrados. R. Estopà comenta que no lo hemos hecho aún, pero que lo podemos hacer immediatamente, que los textos de las primeras etapas son más cortos que los de las etapas recientes, y que los textos de los subcorpus más recientes son muy equilibrados en cuanto a ocurrencias. M. Lorente informa que A. Torres Rivera empezó a controlar el corpus mediante una tabla EXCEL, que A. López Fuentes ha ido completando con textos que ella ha localizado. Ahora sería conveniente que M. Reverté incluyera los datos de los textos que ha localizado para su trabajo de fin de Máster (TFM). Con esto podremos generar una descripción completa del corpus, que pondremos a disposición de todos el mundo en la web y en el GOOGLE DRIVE, y que puede ser el punto de partida para realizar una de estas publicaciones que reclama en Ministerio.

Diversos participantes comentan que no tienen acceso a la carpeta de GOOGLE DRIVE. A. López Fuentes envia mediante el chat del ZOOM el enlace. No obstante, M. Lorente al finalizar la reunión revisará que todos los miembros del proyecto tengan acceso a las carpetas de GOOGLE DRIVE.

M. Lorente informa de la búsqueda que se ha realizado sobre vocabularios y glosarios sobre reproducción asistida. La mayoría son vocabularios reducidos publicados por centros especializados y publicados en sus respectivas webs. Además se incluyen las versiones española y anglesa del vocabulario de la OMS sobre este tema. Todo este material también se encuentra en el GOOGLE DRIVE. Con ello, M. Arnau ha trabajado cruzando los datos y creando un único vocabulario sin repeticiones en español, de 630 entradas, que ha de servir para el entrenamiento del extractor automático de TERMINUS.

Antes de dar la palabra a A. Montané, M. Lorente comenta que los Servicios de Informática de la UPF han substituido algunos servidores. Las herramientas TERMINUS 2.0 y ESTEN, que utilizamos en este proyecto, han sido trasladados de servidores, cosa que ha comportado algunas disfuncionalidades, que afortunadamente han sido resueltas.

Amor Montané confirma que TERMINUS 2.0. funciona sin problemas y que se ha dado acceso a todos los miembros del proyecto, para que lo utilicen autónomamente cuando lo necesiten. De todos modos, ruega que todo el mundo tenga mucho cuidado, que todos compartimos los datos, y que nadie borre ni modifique nada. Pasaremos los accesos a TERMINUS una vez se hayan constituido los corpus y se hayan generado los listados, en septiembre. Respecto al vocabulario de entrenamiento, A. Montané recuerda que la medida ideal de los diccionarios de entrenamiento es de 2500 términos. Ha hecho pruebas con este vocabulario de 630 términos y los resultados son bastante satisfactorios. Propone introducir el vocabulario de reproducción asistida dentro del vocabulario de entrenamiento de medicina para solucionar el posible silencio que provoque un vocabulario de entrenamiento tan reducido. J. Vivaldi comenta que seguramente generaríamos mucho ruído (términos de medicina generales que no nos interesan para nuestros análisis). Se acuerda pues, en principio, entrenar el extractor con el vocabulario reducido. J. Vivaldi solicita el acceso al vocabulario para poder establecer fronteras en los recursos que interactúan con las herramientas de extracción.

Amor Montané continua informando de las herramientas. Plantea que para trabajar con la herramienta ESTEN debemos disponer de los textos en pdf legibles y comprimidos en carpetas zip. Más adelante, A. Montané se ha ofrecido a realizar una sesión sobre el funcionamiento de ESTEN.

Como procedimiento a seguir en las próximas semanas, acordamos que:

(1) Durante el mes de julio se subirán los documentos localizados a TERMINUS 2.0. y se constituiran los corpus por períodos. Así veremos si todos los pdf son realmente legibles o si hay que buscar documentos de sustitución.
(2) Los documentos procesados en TERMINUS se guardaran en formato txt. Los corpus por períodos con los textos en txt se pondrán a disposición de J. Vivaldi para el procesamiento lingüístico y las posteriores extracciones automáticas con YATE y WIKIYATE. También se mandarán a A. Torres-Rivera para que pueda aplicar su herramienta de detección de neología semántica.

R. Estopà introduce el tema de la distribución de tareas y el calendario. Comenta que precisamente el TFM de M. Reverté ha mostrado que la revisión de los datos abre posibilidades diversas de tareas o estudios que no habíamos tenido en cuenta anteriormente. Como ejemplo, comenta el interés de revisar diccionarios generales y especializados para ver qué terminología de la reproducción asistida contienen y cómo la representan; el análisis de webs especializadas dirigidas a pacientes y familiares; o la

realización de encuestas. I. Santamaría también se suma a la idea de que cuando analizamos los datos se abren nuevas perspectivas interesantes que debemos valorar. Expone que ella ha analizado los textos de períodos recientes (1999-2019) con Sketch Engine y para analizar las metáforas del ámbito. M. Lorente recuerda los análisis que habíamos planteado en la solicitud y afinado en las reuniones anteriores: mecanismos de formación, variación, neología, términos obsoletos, etc.

M. T. Cabré abre el debate planteando que todas estas iniciativas son interesantes, pero que por su experiencia recomienda no apartarnos del objetivo central del proyecto, el análisis del cambio léxico vinculado a la evolución del conocimiento científico. Expone sus temores a quedarnos en muchos pequeños estudios descriptivos, que son interesantes que pero que no aportan novedades realmente. Recomienda no quedarnos en la pura descripción sino intentar, aunque sea más difícil, generar resultados que sean realmente novedades de impacto. R. Estopà está de acuerdo pero piensa que debemos tener una estrategia para generar publicaciones lo antes posible, y que en nuestro ámbito deben ser publicaciones de tres autores como máxima, para que sean valoradas. M. Lorente recuerda que el tratamiento de los datos y la generación de listados debe hacerse antes del 15 de octubre y que la idea es empezar a analizar los datos en octubre, para continuar hablando y concretar los paquetes de análisis y las estrategias de publicaciones. Si no tenemos datos lo antes posible, no podremos pasar de la descripción a los intentos de explicación como deseamos.

A partir del debate se acuerda:

(1) Avanzar el calendario previsto para los próximos meses, de manera que en julio intentemos cerrar el corpus, en septiembre se apliquen las herramientas opara generar los listados y a inicios de octubre nos volvamos a reunir, habiendo revisado los datos, para concretar paquetes y equipos de trabajo, y posibles publicaciones relacionadas, si fuera el caso.
(2) Los listados se generarán por períodos temporales y con tres herramientas de extracción: TERMINUS, YATE y WIKIYATE.
(3) Los listados generados serán validados por especialistas. Además del contacto de Quirón-Dexeus pendiente de concretar, I. Santamaría comenta que puede implicar a los especialistas de la UA y del Centro de Reproducción Humana; A.R. Terroba también puede pedir la colaboración de algún especialista de La Rioja. M. T. Cabré informa que tiene un contacto directo con el I. Dexeus, por si falla el contacto formal realizado la semana pasada.

Antes del cierre de la reunión, y dentro ya del turno abierto, A. R. Terroba y Isabel Santamaría se ofrece también a buscar textos de los primeros períodos en español en la biblioteca del Hospital de La Rioja o de la Universitat d'Alacant, y consultando directamente a los especialistas de sus respectivos centros.

Resumimos aquí las tareas que nos hemos comprometido a hacer hasta el 31 de julio (aprovechamos para introducir aquí la asignación de tareas concretas a personas concretas):

1. A.R. Terroba y I. Santamaría intentarán encontrar textos de los períodos 1973-1984 i 1985-1991.
2. A. López Fuentes, M. Reverté, Yingfeng Xu, Zebing Lu y Òscar Pozuelo completarán el Excel de control con los datos de todos los textos, el número de ocurrencias de cada texto y la clasificación temática de cada texto.
3. M. Lorente, con la ayuda de Laia Vidal y Jorge Porras, colgará los textos en Terminus 2.0. i generará las copias en txt para enviar a J. Vivaldi y a A. Torres Rivera.
4. Se dará acceso al Google Drive a todos los miembros del proyecto.
5. M. Lorente enviará los vocabularios de entrenamiento a J. Vivaldi.
6. Las IPs solicitarán al Ministerio la inclusión de nuevos miembros y la prórroga de un año.
7. Se reclamará la respuesta del grupo Quirón-Dexeus.
8. R. Estopà preparará una tabla con la propuesta de paquetes de análisis o tareas.

Durante el mes de septiembre:

9. Amor Montané, con el apoyo de Laia Vidal y Jorge Porras, realizará las extracciones de términos con TERMINUS 2.0.
10. Jordi Vivaldi, con el apoyo de Yingfeng Xu, Zebing Lu y Òscar Pozuelo, procesará los corpus para la posterior aplicación de los extractores YATE y WIKIYATE.
11. R. Estopà, A.R. Terroba y I. Santamaría contactarán con los expertos para la validación de las listas de términos.
12. A. Torres Rivera i J. M. Torres Moreno aplicarán el detector de neología semántica a los distintos subcorpus.

Se acuerda, para terminar, que con la convocatoria de la próxima reunión, prevista para la primera quincena de octubre, se mandarán el acceso a los listados generados y la propuesta de paquetes de trabajo, para el debate.

Se cierra la reunión a las 12h15.