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

Acta de la reunión 5 (Corregida y aprobada el 22/2/2021)

Fecha: 1 de febrero de 2021
Hora: 16:30
Via telemática: Google Meet meet.google.com/trz-deho-jxh

Asistentes: Mariona Arnau, Fabiana da Silva França, Rosa Estopà, Alejandra López Fuentes, Mercè Lorente, Amor Montané, Jorge Porras, Òscar Pozuelo, Isabel Santamaría, Ana Rosa Terroba, Juan-Manuel Torres-Moreno, Laia Vidal, Jorge Vivaldi, Yingfeng Xu.

Se disculpa Andrés Torres, que hoy mismo está haciendo el traslado de domicilio a Zaragoza.

Orden del día:

1. Aprobación del acta de la reunión anterior (16/12/2020).
2. Nueva distribución de paquetes de trabajo (resultado de la reunión anterior).
3. Resultados de los primeros análisis de cada paquete de trabajo.
4. Previsiones de participación en congresos y publicaciones.
5. Debate abierto.

Desarrollo de la reunión y acuerdos

1. Aprobación del acta de la reunión anterior (16/12/2020).

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

M. Lorente revisa el cumplimiento de los acuerdos:

Andrés Torres nos pasó la visualización realizada en HTML y Jorge Porras la ha compartido en una carpeta del Drive.
Jorge Porras hizo un contraste de las extracciones y está en el Drive (carpeta Anàlisi comparativa).
Se ha mandado a los investigadores dos artículos (carpeta Bibliografía_referencia) de Teresa Cabré y de Rogelio Nazar. Dentro de la misma carpeta tendremos un documento con la bibliografia utilizada, para que todo el mundo pueda compartirla.

Se mandó el recordatorio a Andrés Torres y a Juan Manuel Torres-Moreno sobre si era posible aplicar un detector de neología al corpus del proyecto. Están trabajando en ello.
Mariona Reverté y Alenjandra López han acabado la descripción del corpus para todos los paquetes de trabajo. Ahora se podrá completar la descripción del corpus.
Teníamos una reunión programada para el 20 de enero, la estamos haciendo el 1 de febrero. A mediados de febrero podríamos hacer la próxima con más resultados.
Además, todos los paquetes han empezado a trabajar.

Nos queda pendiente:

Realizar un informe sobre las distintas estrategias de extracción automàtica y los resultados discordantes.
La presentación de los resultados de la aplicación del detector de neología semántica.
Enviar a todos los participantes la descripción completa del corpus.

2. <u>Nueva distribución de paquetes de trabajo (resultado de la reunión anterior).</u>

Se repasa la nueva distribución resultado de la reunión anterior, y sin cambios los asistentes la aprueban por asentimiento.

Incidencias:
Coralie Schneider no había recibido siempre las convocatorias, y nos manifiesta que no puede trabajar en el proyecto.
Judit Feliu no ha podido trabajar en el paquete y nos comunica por correo electrónico que tampooco podrá participar.
Mercè Lorente comenta que, como ya estamos en el tramo final, en el primer caso no se trata de una baja ya que constaba como doctoranda y, por lo tanto, su tesis doctoral es un resultado del proyecto, aunque sea indirecto. En el caso de J. Feliu veremos cómo justificar su baja al Ministerio en el informe final.

Comentarios:
En el paquete 3, teniendo en cuenta la baja de Judit Feliu, trabajarán Mercè Lorente y Mariona Arnau.
Rosa Estopà explica que en el paquete de lexicografía seguirán trabajando Isabel Santamaría, Rosa Estopà y Mariona Reverté.

3. Resultados de los primeros análisis de cada paquete de trabajo.

PT1. Análisis contrastivo de las nuevas unidades terminológicas (posibles neóminos) en un período concreto en relación con el período anterior y en correlación con la evolución científica. **Rosa Estopà**, Teresa Cabré, Amor Montané, Jorge Porras.

R. Estopà (presenta un documento): Primero se reunieron para hacer una primera aproximación a los datos.
Partieron de una hipótesis de trabajo que tienen que ver con el paso de la generalidad a la especificad del conocimiento. La especialización del conocimiento tendrá un correlato terminológico. Decidieron analizar solo unidades sintagmáticas con un mismo núcleo en cada uno de los periodos. Escogieron las doce unidades más representativas, las más frecuentes en los últimos periodos. Las unidades que tenían más nodos representativos (siguiendo el trabajo de Andrés). También teniendo en cuenta la representatividad temática del corpus.
A partir de las doce unidades han hecho un vaciado selectivo del corpus.
En un Excel vieron como se distribuyen estas unidades en cada uno de los periodos.
Han utilizado la suma de extracciones YATE y TERMINUS, pero la han depurado porque había mucho ruido.
Se han subdivido en dos grupos (Jorge y Teresa; Amor y Rosa) para seguir trabajando y analizar cada uno de los términos. Ya tienen el objeto de trabajo, las hipótesis y el material vaciado, ahora deben empezar a analizar los datos.

PT2. Análisis contrastivo de las unidades terminológicas que desaparecen (posibles términos obsoletos) en un período concreto en relación con los períodos posteriores y en correlación con la evolución científica. Mercè Lorente, **Amor Montané,** Yingfeng Xu, Jorge Porras.

A. Montané comenta que en algunas cosas coinciden con lo que ha explicado Rosa.
Primero establecieron el ámbito de estudio: términos obsoletos. Segundo, recuperaron la suma de extracciones YATE y TERMINUS (con el fichero de Jordi Vivaldi). En ese fichero hay mucho ruido, pero están en proceso de limpiarlo.
Luego detectaron cuáles son las unidades obsoletas, las aparecen en un periodo, pero no vuelven a aparecer en periodos posteriores.
Decidieron dejar fuera el primer periodo porque los dos textos están muy sesgados, y también los dos últimos, porque son demasiado recientes para hablar en términos de obsolescencia de las unidades terminológicas. Por tanto, trabajan con los periodos del 1973 al 2005.
Ahora hacen un análisis de las unidades partiendo de 3 escenarios (o hipótesis):

- Fenómeno de desaparición de un concepto: medicamentos que dejan de utilizarse, técnicas que dejan de utilizarse o que no se utilizan con la misma frecuencia.
- El concepto permanece, pero hay cambios en las denominaciones. hay que tener en cuenta de que en los primeros períodos hay muchas cuestiones morales que condicionan algunas denominaciones, que van cambiando con el tiempo. En esta línia, se detectan también posibles cambios de denominaciones por internacionalización de la investigación, ya se empiezan a usar préstamos para nombrar las técnicas.
- Se mantiene el concepto, pero el cambio está en la categoría gramatical. Algunos términos aparecen como sustantivos en los primeros períodos y después pasan a utilizarse como verbos o adjetivos.

M. Lorente lanza dos temas para compartir con los demás grupos.

- Inicialmente los textos estaban clasificados en 4 subáreas. Esto iba muy bien para manejar el corpus, pero para los análisis tal vez sea necesario fragmentar estas 4 subáreas, ya que son demasiado amplias.
- En el PT2 han observado la necesidad de disponer de una lista de variables externas que puedan explicar cambios significativos en los diferentes periodos. Se han visionado algunos vídeos cortos de Ana Veiga (entrevistas, charlas) que detallan muy bien los primeros pasos, pero queda muy inconcreta la época desde 1990 hasta la actualidad. ¿Hay algún documento de historia de la RA en España, por ejemplo, que nos pueda ayudar?

A. R. Terroba explica que le suena que había unos artículos de los años 2002-2003 que explicaban las diferentes técnicas. Lo consultará.

PT3. Análisis de la morfología y de los patrones sintagmáticos de las unidades terminológicas más frecuentes en cada período. **Mercè Lorente**, Mariona Arnau.

M. Lorente expone que, de momento, se describieron más de 4.000 ejemplos con los patrones categoriales o sintácticos. Se observa claramente que los patrones se repiten en todos los periodos. Del primer análisis vemos que, sin limpiar, entre las más de 20.000 extracciones tenemos de todo, pero que los patrones más relevantes y frecuentes son NA y NPN y extensiones de estas dos.
Otro elemento destacable es el uso de sintagmas con paratérminos (*proceso, acción, factor*, etc.) como núcleo y con complementos (A o SP) que se corresponden con otros términos.
A continuación, los 4000 descritos, los hemos ordenado según patrones para ver su productividad y ver si se mantenían estables en todos los periodos. En algunas

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Unión Europea
Fondo Europeo de Desarrollo Regional "Una manera de hacer Europa"
AGENCIA ESTATAL DE INVESTIGACIÓN

extensiones como NAA, vemos que cuando el segundo adjetivo tiene valor terminológico, en períodos anteriores aparece el NA correspondiente, lo que sería una marca de mayor especialización en períodos posteriores.
Los patrones más complejos (con 5, 6, 8 ítems) introducen elementos sin valor terminológico (normalmente adjetivos, muchos en posición prenominal) o bien son el resultado de la combinación de dos términos predocumentados.
Hemos eliminado los errores y hemos ordenado alfabéticamente cada patrón. Para ver qué patrones se mantienen inalterables y cuáles se expanden.
Ahora queremos profundizar la descripción de los patrones morfológicos y sintácticos de los 4000 descritos inicialmente y comprobar si las tendencias indicadas se confirman con datos una revisión general del resto de datos extraídos (20.000 aprox.)

Hemos dejado en un apartado todos los candidatos que provienen del discurso religioso. Curiosamente no aparecen sólo en los primeros períodos, sino que se mantienen durante muchos años, aunque con menor frecuencia. Este vocabulario también va cambiando.

PT4. Estudio lexicográfico de las 10 UT más frecuentes en cada período. Rosa Estopà, Mariona Reverté, **Isabel Santamaría**.

Isabel Santamaría (presenta un documento):
Parten de 5 hipótesis:

- La terminología que aparece en diccionarios generales presenta diferencias respecto a los diccionarios especializados.
- Las definiciones de los diccionarios generales no son suficientes ni adecuadas para las pacientes.
- La variación léxica no está suficientemente representada.
- Los diccionarios generales y especializados no recogen la evolución terminológica del ámbito.
- Los diccionarios no recogen unidades terminológicas muy frecuentes actualmente, como *madre de alquiler*.

Han seleccionado los términos más representativos. A partir del criterio de frecuencia no se podía trabajar con los términos más representativos. Finalmente, han seleccionado en un primer momento 15 términos representativos de los 100 más frecuentes.
Han hecho un primer análisis cuantitativo y cualitativo de los términos en diccionarios generales (DEL, CLAVE, DEA y VOX Uso) y especializados (DTM, IGIF).
Por una parte se han centrado en los términos que aparecen en todos los periodos, pero también han añadido los que van apareciendo a medida que se desarrolla la materia.

Análisis cuantitativo: ver en qué momento aparecen en el diccionario de la RAE los términos relacionados con la reproducción asistida. En contraste, las unidades de los últimos periodos no aparecen aún en los diccionarios.

También han trabajado su presencia en los corpus del español (CREA, CORPES XXI y GOOGLE)
Análisis cualitativo: habría que comparar el DEL, el DTM y el glosario internacional (FEDER). Analizarán las marcas diatécnicas, los descriptores, las definiciones, las marcas ideológicas, etc.

Juan Manuel Torres-Moreno opina que es curioso que algunos términos como *bebe probeta* o *bebe a la carta* no estén ni en diccionarios especializados. ¿Sería interesante hacer búsqueda en otros recursos como el https://es.wiktionary.org/wiki/Wikcionario:Portada

I. Santamaría aclara que algunos están en diccionarios especializados, pero en el general no.
R. Estopà explica que para la población de pacientes de habla española lo importante es que en los diccionarios de referencia no los encontramos.
M. Lorente explica que para los diccionarios tal vez funcionaría como toque de atención.

J. Vivaldi explica que el hecho que en plataformas como Wikipedia estén en una lengua y no en otra puede significar que hay más gente involucrada en el tema. Dentro de medicina separar la reproducción asistida es difícil. Requeriríamos de un especialista.
M. Lorente añade que tiene que ver con si los recursos que utilizamos que son más finos o menos finos. Ahora trabajamos en un ámbito muy específico y encontramos obstáculos en la tecnología que está pensada para ámbitos más generales.

PT5. Estudio de abreviaturas, siglas, acrónimos y símbolos. **Ana Rosa Terroba**, Fabiana França.

A. R. Terroba: Han partido de la hipótesis de que las siglas a lo largo del tiempo varían. Han hecho vaciado de las siglas, pero es muy pobre. Con la extracción de Jorge Vivaldi están consiguiendo más.
Hay una serie de mecanismos que nos ayudan a buscarlas en los textos, como los paréntesis. Pasando los textos a Word, han podido hacer una búsqueda avanzada. Probarán también con corchetes y con una fórmula de dos, tres, cuatro, etc. letras en mayúscula.
Con el vaciado completo, crean una ficha que les permitirá clasificar las siglas. El volumen de siglas es muy amplio, van a ir paso a paso y se centrarán en las siglas.

Tienen que ver como extraer los símbolos que no aparecen en los vaciados que tenemos.
Van a estudiar los criterios de formación de siglas. También el género y número de las siglas y los campos semánticos. Se intentarán centrar en las específicas de la reproducción asistida y comprobarán si están documentadas en los diccionarios especializados y qué criterios de uso tienen. Todo esto estudiando las variaciones entre periodos y su frecuencia dentro de los textos.
Están trabajando en el vaciado, falta el último periodo.
Los criterios para abreviaciones serían totalmente diferentes y podría ser un problema.
Hay que ver que criterios establecer para el vaciado de símbolos. Jorge Vivaldi podría ayudarlas para ver como automatizar el vaciado de siglas y de símbolos.
Han apreciado siglas sinónimas, polisemia, etc.

J. Vivaldi puede colaborar. Se puede extraer el mecanismo que habían utilizado en el trabajo de Fabiana. Le parece que será poco frecuente que aparezcan las siglas y el desarrollado al lado. Hará una prueba y les mandará los resultados.
J. M. Torres-Moreno se ofrece a ayudar, tiene un extractor de siglas basado en las propiedades regulares que ha explicado Ana Rosa. Funciona para francés y para inglés, pero cree que podría adaptarse fácilmente al español.

M. Lorente cree que es bueno que al hacer las fichas miren en que documentos aparecen. Podría ser que estuviésemos delante de un texto más divulgativo. En estas revistas o en estos artículos no estamos ante un ámbito de médicos con médicos. Interesante ver cuales son los objetivos de estas revistas.

O. Pozuelo dice que posiblemente les sería útil disponer de los textos verticalizados, los textos con el análisis morfosintáctico. Los sistemas que utilizamos en IULATERM para la desambiguación morfosintáctica de los corpus detectan, por ejemplo, varias palabras consecutivas en mayúscula, y es posible que lo relacionen con la sigla correspondiente si figura en el texto del documento.
A. R. Terroba lo había pensado, si habría un sistema de búsqueda en este sentido ara ver si las siglas coinciden con otras palabras en el mismo texto.
J. Vivaldi añade que en los listados que hay en Drive ahora mismo ya hay secuencias de este tipo pero sin contexto. Puede comprimir todos los ficheros verticalizados y mandarlos.
J. M. Torres-Moreno añade que muchas veces las siglas son ambiguas y Wikipedia puede ayudar a hacer la desambiguación. No es la solución definitiva, pero es una herramienta que puede ayudar. Hay muchos artículos precisos sobre un bien en particular.

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Unión Europea
Fondo Europeo de Desarrollo Regional "Una manera de hacer Europa"
AGENCIA ESTATAL DE INVESTIGACIÓN

J. Vivaldi añade que los artículos más buenos de Wikipedia son los más específicos. Si se va al artículo de la reproducción asistida en Wikipedia lo más probable es que se encuentren los hechos cronológicos.

Correo de A. R. Terroba para mandarle los documentos: arterroba@larioja.org

PT6. Estudio contrastivo entre los textos y la terminología de las webs de clínicas de reproducción asistida y los textos especializados y la terminología del corpus. Análisis de algunas variables de la repercusión social de las webs de clínicas. **Ona Domènech**, Laia Vidal, Jorge Porras, Zebing Lu.

J. Porras explica que aún no han quedado.
R. Estopà explica que Laia está a un mes de entregar la tesis.
M. Lorente propone que Ona revise el material y haga una primera prospección para trabajar con el grupo.

Durante el proceso de revisión de este acta hemos sabido que hubo problemas en la difusión de la convocatoria de la reunión de febrero y que Ona Domènech no la había recibido. Subsanada la recepción de información, se pone en marcha el PT6.

PT-Tecnológicos

J. M. Torres-Moreno explica que A. Torres y él trabajarán en **visibilización** y detección de **neologismos semánticos** con el corpus.

Resultados distintos de los extractores
J. Vivaldi explica que uno de los motivos de las diferencias a candidatos a términos es la diferencia substancial que hay en los patrones. YATE es un muy tradicional (parten de patrones predefinidos), TERMINUS se basa en patrones inducidos, en tener un diccionario de referencia y a partir de aquí inducir cuales son los patrones que son de interés.
A. Montané añade que la extracción de TERMINUS reduce patrones. Eso se debe al hecho que el diccionario de referencia es un diccionario real y normalmente suelen tener unidades más cortas. Tendríamos que hechar un vistazo al diccionario de entrenamiento.
M. Lorente aclara que es la vez que tenemos un diccionario de entrenamiento más pequeño. Además, si realmente los textos son textos menos especializados, este tema de los patrones debe quedar muy afectado.

4. Previsiones de participación en congresos y publicaciones.

A R. Terroba dice que *Cuadernos de Investigación Filológica* sacará un nuevo número, hay que mandar los artículos a finales de marzo como muy tarde.

5. Debate abierto.

M. Lorente propone hacer la próxima reunión después de haber aclarado algunas cuestiones, todos los paquetes tenemos asignado por donde debemos ir para acabar de afinar los resultados. Conviene que no lo retrasemos mucho. Podemos mantener la reunión a mediados de febrero para hablar de como está cada paquete y ver si surgen otras dudas o problemas. A mediados de marzo haríamos otra con los informes redactados.
R. Estopà propone la tercera semana de febrero.

Se acuerda que la próxima reunión sea el 22 de febrero a las 17 h, con un único punto del día: estado de la cuestión de los paquetes.

A Montané pregunta si los informes tienen que ser en forma de artículo.
M. Lorente dice que no es necesario, porque desde el Ministerio solo nos piden informes para redactar la memoria. Pero nos pueden servir de borrador para hacer las primeras publicaciones.

Se cierra la reunión a las 18h20.