Convocatorias 2017
Proyectos EXCELENCIA y Proyectos RETOS
AGENCIA ESTATAL DE INVESTIGACIÓN

AVISO IMPORTANTE

En virtud del artículo 16 de la convocatoria **NO SE ACEPTARÁN NI SERÁN SUBSANABLES MEMORIAS CIENTÍFICO-TÉCNICAS** que no se presenten en este formato.

Es obligatorio que la memoria contenga los tres apartados (A, B y C). La parte C de la memoria no podrá exceder de 20 páginas.

Lea detenidamente las instrucciones para rellenar correctamente esta memoria, disponibles en la web de la convocatoria.

Parte A: RESUMEN DE LA PROPUESTA/SUMMARY OF THE PROPOSAL

INVESTIGADOR PRINCIPAL 1 (Nombre y apellidos)**:**

Mercè LORENTE CASAFONT

INVESTIGADOR PRINCIPAL 2 (Nombre y apellidos)**:**

Rosa ESTOPÀ BAGOT

TÍTULO DEL PROYECTO: Evolución del conocimiento científico en medicina: El cambio léxico y semántico

ACRÓNIMO: TERMMED

RESUMEN Máximo 3500 caracteres (incluyendo espacios en blanco):

El conocimiento científico crece permanentemente en medicina, a causa de la identificación de nuevas enfermedades, el desarrollo de nuevos métodos y tecnologías, la aparición de nuevos tratamientos, y sobre todo por la descripción más detallada de los fenómenos. Una de las pistas para detectar este conocimiento es el análisis de las producciones (orales y) escritas de médicos y expertos, dentro de las cuales la terminología es un factor clave. El análisis del léxico de los textos médicos, y en concreto de la variación terminológica detectada a lo largo del tiempo, atendiendo a la presencia o la ausencia de unidades, a las modificaciones que se producen entre paradigmas o a los cambios que se detectan en cuanto al contenido, es determinante para representar esta evolución (oculta) del conocimiento.

La variación terminológica, en sentido general, nos ha de permitir identificar los parámetros de esta evolución. Los estudios de la variación en terminología se han abordado desde la transversalidad temática, el contraste de lenguas o sus causas, e incluyen la variación denominativa, la variación semántica y la neología. Las variables utilizadas habitualmente en los estudios de variación terminológica son el canal (oral o escrito), el registro, los géneros textuales y el nivel de especialización. Este proyecto pivota sobre la variable tiempo, para poder relacionar la variación terminológica con el cambio léxico y el cambio semántico.

Nos centraremos en la terminología de la medicina en español. Los fenómenos de cambio léxico que queremos abordar incluyen la selección de variantes, la polisemia, la fijación de reducciones, la lexicalización de sintagmas, conversiones, nominalizaciones, la neología o la desaparición de formas. La metodología combina la exploración de corpus textuales segmentados por períodos temporales, la extracción automática de terminología, la consulta

de ontologías y otras fuentes léxicas, el análisis estadístico de variantes terminológicas y la experimentación en la detección automática de neología semántica.

La disposición de recursos lingüísticos adecuados para el desarrollo puntero de tecnologías del lenguaje adaptados a la documentación médica en lengua española es uno de los retos del s. XXI. La constitución de estos recursos permite pensar en aplicaciones que cubrirán múltiples necesidades de los profesionales de la salud y de los investigadores en biomedicina (sistemas de recuperación de información, predicción de datos, clasificadores de textos médicos, resumidores, codificadores, asistentes de redacción, etc.), pero también de los pacientes (sistemas de simplificación de textos, reelaboración de textos adaptados a niveles cognitivos, etc.). Ante la proliferación natural de variación terminológica, las tecnologías del lenguaje precisan datos contrastados que les permitan automatizar la toma de decisiones. Con los resultados, las tecnologías del lenguaje podrán inferir modelos de comportamiento de la variación terminológica en medicina. En este proyecto no desarrollaremos tecnología sino que haremos accesibles los recursos léxicos en abierto, derivados de nuestros análisis, para que puedan implementarse en tecnologías lingüísticas adaptadas a dominio.

Para llevar a cabo el proyecto el equipo formado principalmente por lingüistas especialistas en léxico cuenta con el soporte de profesionales de universidades extranjeras (Argentina, Francia y Brasil) de la medicina, la informática y las ciencias de la documentación.

PALABRAS CLAVE: cambio léxico; semántica; terminología; neología; lingüística corpus; documentación médica; recursos lingüísticos; español.

TITLE OF THE PROJECT: Evolution of scientific knowledge in medicine: lexical and semantic change

ACRONYM: TERMMED

SUMMARY Maximum 3500 characters (including spaces):

Scientific knowledge in medicine is constantly growing, due to the identification of new diseases, the development of new methods and technologies, the appearance of new treatments and, especially, due to a more detailed description of different phenomena that occur. One of the clues to detect this knowledge is the analysis of the doctors' and experts' oral and written productions, within which terminology is a key factor. The analysis of medical texts lexicon and, in particular, of the terminological variation detected over time is decisive to represent this (hidden) evolution of knowledge, always paying special attention to the presence or absence of units, the changes that occur between paradigms or the changes that are detected in terms of content.

The terminological variation, in a general sense, must allow us to identify the parameters of this evolution. Terminological variation studies have been approached from different subjects transversely; as well as from languages' contrast; and also from the reasons that produce it. And they include denominative variation, semantic variation and neology. The elements usually considered in terminological variation studies are the channel (oral or written), the register, the textual genres and the level of specialisation. This project revolves on the element "time", in order to be able to relate terminological variation with lexical change and semantic change.

We will focus on medical terminology in Spanish. The phenomena of lexical change that we want to address include selection of variants, polysemy, fixation of reductions; lexicalization of phrases, conversions, nominalizations, neology and the disappearance of forms. The methodology combines exploration of textual corpora; automatic extraction of terminology; query of ontologies and other lexical sources; statistical analysis of term variants; and experimentation on automatic detection of semantic neology.

The preparation of linguistic resources suitable for the development of advanced language technologies adapted to medical documentation in Spanish is one of the challenges for the 21st century. The setting-up of these resources allows us to consider applications that will cover multiple needs not only from health professionals and researchers in biomedicine (information retrieval systems, data prediction, medical texts classifiers, text summarisers, text coders, writing assistants, etc.), but also from patients (text simplification systems, rewriting of texts adapted to different cognitive levels, etc.). Given the natural proliferation of terminological variation, language technologies require verified data that allows them to automate decision making. We propose to provide data that will allow inferring behavioural patterns of terminological variation in medicine. In this project we will not develop technological resources, but we will publish open access lexical resources derived from our analysis, so that they can be implemented in language technologies adapted to different fields.

To carry out this project, the team, consisting mainly of linguists specialized in lexicon, counts on the support of medicine, computer science and documentation sciences professionals from foreign universities (Argentina, France and Brazil).

KEY WORDS: lexical & semantic change; corpus linguistics; terminology; neology; medical documentation ;linguistic resources; Spanish

Parte B: INFORMACIÓN ESPECÍFICA DEL EQUIPO

B.1. FINANCIACIÓN PÚBLICA Y PRIVADA (PROYECTOS Y/O CONTRATOS DE I+D+I) DEL EQUIPO DE INVESTIGACIÓN (repita la secuencia tantas veces como se precise hasta un máximo de 10 proyectos y/o contratos).

1. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos: Rosa Estopà, M. Teresa Cabré, M. Amor Montané, Ona Domènech, Mercè Lorente
Referencia del proyecto: 2015ACUP 00187,
Título: JUNTS. Superant barreres socioeducatives i afavorint l'alfabetització sobre les interferències i dificultats de comprensió d'informació i documentació dirigida a famílies d'infants afectats per malalties minoritàries
Investigador principal (nombre y apellidos): Rosa Estopà (UPF) y Manuel Armayones (UOC)
Entidad financiadora: RecerCaixa 2015. ACUP i Obra Social La Caixa.
Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA):29/02/2016-3/1/2018
Financiación recibida (en euros): 100.000,00 €
Relación con el proyecto que se presenta: Este proyecto ha contribuido a la caracterización de los neologismos léxicos especializados como objeto científico dentro de una teoría del léxico, ampliar esta caracterización con aspectos psicolingüísticos y sociolingüísticos relacionados con la producción, recepción y circulación social de estas unidades, y establecer la especificidad de los neologismos especializados y sus relaciones con los neologismos generales y con el léxico no neológico. Los resultados incluyen un apartado referido a los neologismos del discurso médico.
Estado del proyecto o contrato: concedido, en curso.

2. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): M. Teresa Cabré, Rosa Estopà, M. Amor Montané, Ona Domènech.
Referencia del proyecto: FFI2012-37260
Título: *Neologismos generales y neologismos especializados: análisis y contraste desde el punto de vista de su producción, recepción y circulación social (APLE2)*
Investigador principal (nombre y apellidos): M. Teresa Cabré
Entidad financiadora: Ministerio de Economía y Competitividad
Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA):01/01/2013-31/12/2015
Financiación recibida (en euros): 49.950,00 €
Relación con el proyecto que se presenta: Este proyecto ha contribuido a la caracterización de los neologismos léxicos especializados como objeto científico dentro de una teoría del léxico, ampliar esta caracterización con aspectos psicolingüísticos y sociolingüísticos relacionados con la producción, recepción y circulación social de estas unidades, y establecer la especificidad de los neologismos especializados y sus relaciones con los neologismos generales y con el léxico no neológico. Los resultados incluyen un apartado referido a los neologismos del discurso médico.
Estado del proyecto o contrato: concedido, finalizado.

3. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): Mercè Lorente, Isabel Santamaria, Juan Manuel Torres Moreno.
Referencia del proyecto: FFI2010-21365-C03-01
Título: Procesamiento de corpus textuales especializados para la extracción de estructuras supraléxicas terminológicamente relevantes (RICOTERM 4)
Investigador principal (nombre y apellidos): Mercè Lorente Casafont
Entidad financiadora: Ministerio de Ciencia e Innovación
Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA): 01/01/2011-31/12/2013
Financiación recibida (en euros): 60.500,00 €
Relación con el proyecto que se presenta: En el proyecto RICOTERM4 se han estudiado las características léxicas restrictivas (colocaciones) en el discurso de especialidad y se ha trabajado en las dificultades de acceso de esta información en recursos lexicográficos. En el

proyecto que se presenta ahora, el conocimiento resultante del proyecto RICOTERM4 se puede aplicar en la descripción de las unidades de estructura sintagmática, de difícil segmentación en los corpus textuales especializados y abundantes en medicina.
Estado del proyecto o contrato: concedido, finalizado.

4. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): Rosa Estopà, Ona Domènech
Referencia del proyecto: FCT-11-2501
Título: Jugando a definir ciencia (2): laboratorios de palabras
Investigador principal (nombre y apellidos): Rosa Estopà Bagot
Entidad financiadora: FECYT
Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA): 01/01/2012-31/12/2013
Financiación recibida (en euros): 30.000,00 €
Relación con el proyecto que se presenta: Este proyecto trabajó con las definiciones de términos científicos y desarrolló productos lexicográficos y terminográficos adecuados a las competencias cognitivas y terminológicas de los niños. Se aplicó el Principio de adecuación que pretendemos aplicar también en el proyecto que presentamos y que nos permite adecuar al máximo cualquier aplicación o texto a su destinatario.
Estado del proyecto o contrato: finalizado.

5. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): M. Teresa Cabré, Ona Domènech, Rosa Estopà, M. Amor Montané.
Referencia del proyecto: FFI2009-12188-C05-01
Título: Procesos de actualización del léxico del español a partir de la prensa editada en Cataluña (APLE1)
Investigador principal (nombre y apellidos): M. Teresa Cabré
Entidad financiadora: Ministerio de Ciencia e Innovación
Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA): 01/01/2010-31/12/2012
Financiación recibida (en euros): 98.540,00 €
Relación con el proyecto que se presenta: Este proyecto proporcionó datos fiables sobre las tendencias léxicas observadas en el español en distintas variedades a fin de avanzar en el análisis de la dinámica del léxico en español y contribuir a la adopción de criterios para su actualización en los campos especializados, entre los que se contempló la medicina. El proyecto, de carácter teórico y aplicado, se propuso analizar, a través de los neologismos léxicos procedentes del discurso de la prensa, los recursos y estrategias más productivos que utilizan actualmente algunas variedades del español de España y de Iberoamérica.
Estado del proyecto o contrato: finalizado

6. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): Rosa Estopà, Ona Domènech
Referencia del proyecto: FCT-11-2501
Título: *Jugando a definir la ciencia*
Investigador principal (nombre y apellidos): Rosa Estopà Bagot
Entidad financiadora: FECYT
Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA): 01/01/2011-31/12/2012
Financiación recibida (en euros): 25.000,00 €
Relación con el proyecto que se presenta: El proyecto consistió en reunir un material sistemático y homogéneo que permita estudiar qué saben los niños y las niñas en los estadios iniciales de la escolarización (primer ciclo de Primaria) sobre los conceptos más básicos de ciencia, es decir, conocer las ideas subyacentes y las ideas obstáculo que distorsionan el aprendizaje; además, de proporcionar materiales adecuados y cercanos a sus necesidades. Este proyecto se relaciona con el que presentamos a través de su principios metodológicos de base, basados en la TCT.
Estado del proyecto o contrato: finalizado

7. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): Mercè Lorente, M. Amor Montané.
Referencia del proyecto: HUM2007-65966-C02-01/FILO

Título: Enriquecimiento de ontologías para la extracción de información de discursos de especialidad (RICOTERM 3)
Investigador principal (nombre y apellidos): Mercè Lorente Casafont
Entidad financiadora: Ministerio de Educación y Ciencia
Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA): 01/01/2007-31/12/2010
Financiación recibida (en euros): 71.390,00 €
Relación con el proyecto que se presenta: El objetivo básico de este proyecto fue la adaptación de la herramienta YATE a diversos ámbitos especializados, mediante el enriquecimiento de EuroWordNet (EWN), y para las lenguas castellana, catalana y vasca. Además mejoró la plataforma de acceso a esta herramienta, con la incorporación de otras aplicaciones auxiliares, como la detección automática de expresiones metalingüísticas que reflejen relaciones conceptuales en los textos, y mejores prestaciones de la plataforma de acceso. Yate será una de las herramientas básicas que utilizaremos para el proyecto con el fin de extraer la terminología que contienen los IAH.
Estado del proyecto o contrato: finalizado.

8. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): M. Teresa Cabré, Mercè Lorente, Rosa Estopà, Ona Domènch, M. Amor Montané.
Referencia del proyecto: HUM2006-09458
Título: Fundamentos, estrategias y herramientas para el procesamiento y extracción automáticos de información especializada
Investigador principal (nombre y apellidos): M. Teresa Cabré
Entidad financiadora: Ministerio de Ciencia y Tecnología
Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA): 01/10/2006 - 30/09/2009
Financiación recibida (en euros): 70.180
Relación con el proyecto que se presenta: En este proyecto se desarrollaron herramientas para el procesamiento del lenguaje natural especializado que utilizaremos para la exploración de los corpus del proyecto.
Estado del proyecto o contrato: finalizado

9. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): M. Teresa Cabré, Mercè Lorente, M. Amor Montané
Referencia del proyecto: PT2012-S03-ALEGRET.
Título: ExtracCIT. Extracció terminològica de les revistes de la Hemeroteca Científica Catalana
Investigador principal (nombre y apellidos): Salvador Alegret
Entidad financiadora: Institut d'Estudis Catalans
Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA): 01/01/2012-31/12/2014
Financiación recibida (en euros): 14.400
Relación con el proyecto que se presenta: La extracción de terminología de las revistas científicas especializadas en temas biomédicos y químicos del IEC constituirá unas bases de referencia terminológica para el proyecto que presentamos.
Estado del proyecto o contrato: finalizado

10. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): Mercè Lorente
Referencia del proyecto: FFI2009-08484-E
Título: Multilingual Central Repository 2.0: IULA
Investigador principal (nombre y apellidos): Mercè Lorente Casafont
Entidad financiadora: Ministerio de Educación y Ciencia, Subprograma de Acciones Complementarias para los Proyectos de Investigación Fundamental no orientada
Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA): 1/4/2010-31/3/2011
Financiación recibida (en euros): 25.000
Relación con el proyecto que se presenta: Las herramientas que se desarrollaron en el marco de este proyecto serán de utilidad en las tareas de tratamiento automático del corpus.
Estado del proyecto o contrato: finalizado

B.2. RELACIÓN DE LAS PERSONAS NO DOCTORES QUE COMPONEN EL EQUIPO DE TRABAJO (se recuerda que los datos de los doctores del equipo de trabajo y de los componentes del equipo de investigación no se solicitan aquí). Repita la siguiente secuencia tantas veces como precise.

1. Nombre y apellidos: ANDRÉS TORRES RIVERA
Titulación: Licenciado en Lengua y Literatura Hispánicas, 2013; Máster en Lingüística Teórica y Aplicada, UPF, 2015.
Tipo de contrato: Beca predoctoral CONACYT, México
Duración del contrato: temporal (2016-2019)

2. Nombre y apellidos: OSCAR POZUELO OLLÉ
Titulación: Licenciado en Traducción e Interpretación, UPF, 2013; Máster en Lingüística Teórica y aplicada, 2014
Tipo de contrato: Contrato predoctoral de investigador en formación IULA-UPF
Duración del contrato: temporal (2015-2019)

3. Nombre y apellidos: CORALIE SCHNEIDER
Titulación: Máster de Inglés para propósitos específicos, 2014
Tipo de contrato: Beca predoctoral (Université Diderot, Francia)
Duración del contrato: temporal (2016-2019)

Parte C: DOCUMENTO CIENTÍFICO. Máximo 20 páginas.

C.1. PROPUESTA CIENTÍFICA

C.1.1. Introducción y estado de la cuestión

El conocimiento científico crece permanentemente en medicina, a causa de la identificación de nuevas enfermedades, el desarrollo de nuevos métodos y tecnologías, la aparición de nuevos tratamientos, y sobre todo por la descripción más detallada de los fenómenos. Una de las pistas para detectar este conocimiento es el análisis de las producciones (orales y) escritas de médicos y expertos, dentro de las cuales la terminología es un factor clave. El análisis del léxico de los textos médicos, y en concreto de la variación terminológica detectada a lo largo del tiempo, atendiendo a la presencia o la ausencia de unidades, a las modificaciones que se producen entre paradigmas o a los cambios que se detectan en cuanto al contenido, es determinante para representar esta evolución (oculta) del conocimiento.

La investigación que proponemos en TERMMED responde a una doble motivación:

a) La necesidad de avanzar en el conocimiento sobre la evolución del conocimiento científico a través del análisis de l'evolución del léxico y la semántica del ámbito médico; y

b) la necesidad de disponer de recursos lingüísticos adecuados para el desarrollo puntero de tecnologías del lenguaje adaptados a dominio, en especial para un sector de actividad prioritario, la informática médica.

La variación terminológica, en sentido general, nos ha de permitir identificar los parámetros de esta evolución. Los estudios de la variación en terminología se han abordado desde la transversalidad temática, el contraste de lenguas o sus causas, e incluyen la variación denominativa, la variación semántica y la neología. Las variables utilizadas habitualmente en los estudios de variación terminológica son el canal (oral o escrito), el registro, los géneros textuales y el nivel de especialización. En cambio, este proyecto pivota sobre la variable tiempo, para poder relacionar la variación terminológica con el cambio léxico y el cambio semántico.

Partimos de la idea de que los cambios semánticos y formales que se dan en la terminología de un ámbito pueden ser observados en cortes periódicos más breves que los que habitualmente se dan en los estudios de cambio fonético o de cambio sintáctico. Disponemos de un corpus de textos médicos en español, dentro del Corpus Técnico del IULA-UPF, de más de 4 millones de palabras, que provienen de documentos fechados entre 1990 y 2013. Se propone ampliar este corpus con el procesamiento y la anotación de un subcorpus de informes clínicos del período 2014-2018, debidamente anonimizados, cedidos por el Hospital Italiano de Buenos Aires, que actua de entidad interesada en los resultados del proyecto (EPO). El corpus total integrado será segmentado en 4 cortes temporales de 7 años, con el fín de analizar la variación terminológica en cada corte temporal y la evolución del léxico y la semántica a través de la comparación de los 4 cortes temporales.

La serie de cambios léxicos que proponemos analizar diacrónicamente contempla:

- La confluencia y la selección progresiva de variantes
- La ampliación de la polisemia
- La substitución de unidades con reducciones y su lexicalización
- La lexicalización de estructuras sintagmáticas
- La compleción de paradigmas categoriales (conversiones, nominalizaciones)
- La aparición de neologismos
- La obsolescencia de formas

La disposición de recursos lingüísticos adecuados para el desarrollo puntero de tecnologías del lenguaje adaptados a la documentación médica en lengua española es uno de los retos del s. XXI. La constitución de estos recursos permite pensar en aplicaciones que cubrirán múltiples necesidades de los profesionales de la salud y de los investigadores en biomedicina (sistemas de recuperación de información, predicción de datos, clasificadores de textos médicos, resumidores, codificadores, asistentes de redacción, etc.), pero también de los pacientes (sistemas de simplificación de textos, reelaboración de textos adaptados a niveles cognitivos, etc.). La informática médica es un ámbito priorizado por las políticas públicas, con crecientes necesidades de recursos textuales y léxicos que permitan avanzar en el desarrollo de tecnologías para el procesamiento automático de la ingente documentación médica, la actualización de ontologías, la comunicación entre sistemas informáticos, la historia clínica electrónica, la gestión hospitalaria o la comunicación médico-paciente. Nuestra investigación se propone ofrecer datos que permitan inferir modelos de comportamiento de la variación terminológica en medicina: condiciones para la lexicalización, selección prioritaria de variantes, contextos para la desambiguación de variantes, entre otras. Ante la proliferación natural de variación terminológica, las tecnologías del lenguaje precisan datos contrastados que les permitan automatizar la toma de decisiones. En este proyecto construiremos recursos léxicos, directamente derivados de nuestros análisis sobre la evolución léxica y semántica, que haremos accesibles en abierto, para que puedan implementarse en tecnologías lingüísticas adaptadas a dominio y otros proyectos de informática médica.

Orientación teórica del proyecto. La investigación que proponemos se situa en la intersección entre los estudios de cambio lingüístico y la terminología médica.

Nuestras referencias en el marco teórico del cambio se restringen a los estudios sobre **el cambio léxico y el cambio semántico**, así como al papel de la frecuencia y a las causas externas e internas que permiten caracterizar los fenómenos de cambio. Para el establecimiento de hipótesis parciales y de protocolos de análisis, partiremos de principios y condiciones consensuados en la bibliografía clásica sobre el cambio léxico y semántico (Ullmann,1962; Lehmann & Malkiel, 1968; Lyons, J., 1977; Labov , 1972, 1994**)** y en autores más recientes (Bybee 1985-2012; Hopper & Traugott 1993; Geeraerts 1997; Heine & Kuteva 2002; & Dasher 2002; Stubs 2002; Traugott Dressler 2003; Briton & Traugott 2005; Phillips 2006).

Destacamos, entre otros, una serie de principios generales sobre el cambio lingüístico, que lo regularían universalmente:

- Los cambios sociales condicionan el cambio lingüístico (Bybee 1994, 2010; Hopper & Traugott 2003).
- La difusión del cambio se produce fundamentalmente en géneros discursivos concretos (Bybee 2000, 2010).
- Los procesos de gramaticalización se dan en condiciones de regularidad, productividad y transparencia y se refuerzan en estructuras lingüísticas menos predecibles (Bybee 1994, 2000). Nos interesan particularmente los procesos de lexicalización y los de gramaticalización menor (Hopper & Traugott 2003), que son los que afectan al cambio léxico.

En este proyecto queremos responder a las siguientes preguntas de investigación:

- ¿Cuáles son los cambios relevantes en la terminología médica: la lexicalización de estructuras sintagmáticas o de formas reducidas (siglas y acrónimos), la gramaticalización menor de afijos, el cambio semántico de formas afijadas o de unidades, la reducción de variantes, la delimitación de unidades sintagmáticas, etc.?
- ¿Qué condiciones externas (de la propia actividad científica, de la tecnificación, de las expectativas sociales de la medicina sobre la salud y la enfermedad) podrían condicionar (y explicar) estos cambios?

- ¿La detección del cambio léxico ofrece diferencias significativas según el género de la actividad médica que se analice (informes clínicos, artículos científicos, textos de divulgación)?
- ¿Las condiciones internas de los procesos de gramaticalización menor y de lexicalización son los mismos en la terminología médica que los expuestos en los estudios sobre el cambio centrados en la lengua general o común?
- Y sobre todo ¿el cambio léxico y el cambio semántico en terminología responden al mismo comportamiento que en el léxico general?

Desde los años noventa se han producido cambios teóricos y metodológicos relevantes en el estudio de la **terminología** y del discurso de especialidad, lo que ha permitido disponer de datos objetivables para la caracterización del discurso médico, entre otros ámbitos de especialidad. La constitución de corpus de textos especializados y la constitución de herramientas para la extracción automática de terminología ha sido fundamental para el análisis de las características del léxico especializado y de su combinatoria, en contexto. Así se ha demostrado que la variación en terminología es un fenómeno natural (e inevitable) en discurso especializado (Cabré 1998, 2009), que ha venido siendo caracterizada a lo largo de estos años (Faulstich 1995; Temermann 2000; Freixa 2002, 2005; Domènech 2006; Cabré & Bach 2010; Fernández Silva 2011, Seghezzi 2011, Faber 2012; Faber & L'Homme 2014). Del mismo modo, la atención mostrada por los investigadores hacia la **neología**, sobre todo a partir de los últimos diez años (Cabré 2009, 2012, 2015; Cabré & Nazar 2012; Estopà 2015; Freixa 2014; García Palacios 2015; Guerrero Ramos & Pérez-Lagos 2009, 2012; Lorente 2015) ha significado añadir, a la descripción sincrónica de estructuras, categorías, afijos y significados de las unidades terminológicas, estudios de orientación diacrónica sobre las tendencias en la innovación léxica y en la selección de neologismos en estos mismos ámbitos especializados.

A pesar de toda esta actividad investigadora, no se ha trabajado todavía en la evolución de la terminología de un ámbito especializado de épocas recientes (otra cosa distinta es el estudio de la historia de los lenguajes de especialidad). Tampoco se ha observado si la variación y los cambios detectados en el uso de la terminología de un ámbito concreto responden a tendencias identificadas en los estudios teóricos sobre el cambio lingüístico del lenguaje común o general (no especializado) o, si por lo contrario, están motivados por nuevos argumentos. Sabemos que en los textos especializados hay variación terminológica que va más allá de los vocabularios controlados o normalizados. Conocemos qué tendencias sigue la neología especializada o neonimia en algunos ámbitos del conocimiento. Pero aún no se ha caracterizado el cambio léxico (formal y semántico) que presenta la terminología de un ámbito de especialidad concreto, como reflejo de la evolución de ese ámbito.

Orientación aplicada del proyecto. El espacio de intersección entre la documentación, la comunicación y la informática médicas está recibiendo muchas muestras de interés desde las políticas públicas, tanto estales como europeos, (con el impulso de las tecnologías del lenguaje adaptadas a dominio, con la adopción y la adaptación de recursos para la codificación de la documentación clínica, con el desarrollo de la historia digital) y también de parte de los agentes del sector público y del sector privado, involucrados en la generación de documentación, recursos digitales y herramientas informáticas asociadas. Conviene destacar el Plan de Impulso de las Tecnologías del Lenguaje, dentro de la Agenda Digital para España (http://www.agendadigital.gob.es/tecnologias-lenguaje/Paginas/plan-impulso-tecnologias-lenguaje.aspx), que marca la hoja de ruta en materia de Tecnologías de la Información y las Comunicaciones (TIC) y de Administración Electrónica para el cumplimiento de los objetivos de la Agenda Digital para Europa en 2015 y en 2020, e incorpora objetivos específicos para el desarrollo de la economía y la sociedad digital en España.

Nuestro grupo de investigación IULATERM tiene experiencia en el desarrollo de tecnologías de procesamiento del lenguaje natural, extracción automática de terminología, infraestructuras lingüísticas, traducción automática, diccionarios electrónicos, bancos de datos de terminología y neología, detección automática de neología formal y semántica; y estamos interesados en participar en el desarrollo de aplicaciones tecnológicas que sean relevantes para esta intersección entre la documentación, la comunicación y la informática médicas. No obstante, esta experiencia també nos hace conocedores de los obstáculos con los que se encuentra en la actualidad el desarrollo de estas tecnologías adaptadas a dominio, el ámbito médico en nuestro caso. Para que estas futuras tecnologías adaptadas a dominio sean eficientes se precisa previamente desarrollar recursos textuales actualizados y recursos léxicos que traten adecuadamente la variación terminológica del ámbito (la controlada y la no controlada) en las lenguas de trabajo.

En esta línea existen diferentes corpus anotados para diferentes tipos de textos en inglés. Los más conocidos son GENIA; 2010 I2b2, MiPACQ, SHARPn NLP Seed Corpus, CLEF eHealth, BioScope, BioInfer, Biocreative. En cambio, para el español, el volumen de datos anotados y accesibles es reducido. Tenemos un equivalente al corpus Quaero francés https://quaerofrenchmed.limsi.fr/) y el corpus gold estándar CLEF-ER anotado por el proyecto EU-MANTRA, disponible junto con otras lenguas en https://files.ifi.uzh.ch/cl/mantra/gsc/GSC-v1.zip. Existen también pequeños corpus específicos: el de reacciones adversas anotado para síntomas y fármacos (Oronoz et al.; 2016) y los corpus anotados para negación Cruz et al. (2017) y Marimon et al. (2017), siendo este último el único accesible. En este proyecto queremos constituir un subcorpus actualizado y anotado con informes clínicos anonimizados, que se integraría en el corpus técnico del IULA-UPF y que también se podría hacer accesible para que, en proyectos tecnológicos posteriores, se puedan adaptar herramientas de procesamiento del lenguaje natural ya existentes a las características y distribución léxica de los textos médicos. Así mismo, como uno de los retos detectados en la adaptación de herramientas a dominio es precisamente la segmentación y tokenización de los corpus textuales, nos proponemos desarrollar también diversos recursos léxicos derivados a partir del corpus, que permitan gestionar la variación de estructuras léxicas (lexicalizadas o en proceso de lexicalización), la polisemia de las unidades terminológicas, la relación entre las formas reducidas (siglas, acrónimos, abreviaturas) y las correspondientes unidades expandidas, entre otros.

Estos recursos léxicos y textuales podrán facilitar también el desarrollo de tecnología específica para la resolución de problemas de comunicación médico-paciente, que es otra de las líneas de investigación cruciales para este espacio de intersección entre salud, informática, documentación y lingüística.. El acceso generalizado a los informes médicos por parte de los pacientes (o de sus familiares), que aumenta con la digitalización de estos sistemas, ha permitido detectar una distancia clara entre los documentos inteligibles por parte de los expertos emisores y la incomprensión parcial o total de esos documentos por parte de los pacientes no expertos (Proyecto RECERCAIXA-JUNTS, 2016-2018). La tecnología para la generación automática de textos puede ser una alternativa para transformar textos opacos expertos en textos más transparentes y legibles para los pacientes. Sería óptimo contar con un reelaborador textual o traductor interlingua de un nivel de especialización A a un nivel de especialización B. Pero para el desarrollo de sistemas de este tipo también es preciso constituir previamente recursos textuales anotados y recursos léxicos específicos de dominio que resuelvan la opacidad cognitiva y faciliten la comprensión de textos a los usuarios-receptores a los que se dirige.

En síntesis, este no es un proyecto para desarrollar nueva tecnología (programación y evaluación de productos informáticos), sino para estudiar la evolución del léxico y la semántica en textos médicos y para poner en abierto recursos léxicos y textuales en español, para que estas tecnologías del lenguaje adaptadas a dominio médico, tan urgentes en este momento, se puedan desarrollar con buenos resultados.

La relación entre ambas orientaciones. Deseamos subrayar que, aunque hayamos presentado las dos motivaciones, como perspectivas diferenciadas, teórica y aplicada, no se

trata de dos planteamientos independientes que debieran vehicularse a través de proyectos separados. En general, porque los proyectos en lingüística aplicada deben integrar las dos perspectivas: no hay avance teórico sin el análisis de datos reales actualizados y no es posible desarrollar aplicaciones sin la guía que nos proporciona el marco teórico. Y, más concretamente, en nuestro caso, porque la ingente productividad de bibliografía médica nos obliga a actualizar con nuevos textos el apartado de medicina del Corpus Técnico del IULA (Instituto de Lingüística Aplicada de la Universidad Pompeu Fabra), que actualmente cubre el período de 1990 a 2013. Además, las necesidades de descripción y de aplicabilidad demandan añadir a este corpus otros géneros textuales como la documentación técnica (prospectos médicos) y sobre todo informes clínicos, debidamente anonimizados, cedidos por hospitales colaboradores del proyecto. No podemos describir la variación terminológica ni la neología en textos médicos sin corpus actualizados y complejos desde el punto de vista discursivo. Por otro lado, el estudio del cambio léxico y la detección de neología hacen necesaria la segmentación del corpus por periodos marcados para poder aplicar análisis de tipo estadístico (herramienta ESTEN http://esten.iula.upf.edu/). Los recursos léxicos resultantes del proyecto derivan tanto del análisis y la explotación del corpus, como del marco teórico planteado.

C.1.2. Antecedentes

Del equipo de investigación

Todos los doctores vinculados a la Universidad Pompeu Fabra forman parte del grupo IULATERM, grupo consolidado con financiación dentro del Mapa de Investigación de Catalunya (última referencia 2014SGR00946). Este grupo de investigación, fundado en 1994 por M. Teresa Cabré, se ha centrado desde sus inicios en dos grandes líneas: la terminología, la neología y el discurso especializado; y el desarrollo de herramientas para el tratamiento automático de documentos y la extracción de la información. Se trata de un grupo compuesto actualmente por 15 doctores y 18 investigadores en formación. Participan en este proyecto 6 doctores del grupo y 3 investigadores en formación.

En los proyectos del grupo IULATERM siempre se abordan de manera integral las perspectivas teórica, descriptiva y aplicada. También desde sus inicios una parte considerable de estos trabajos se han aplicado a textos médicos. Una de sus principales aportaciones en esta área del conocimiento ha sido el desarrollo de una teoría de los términos, la Teoría Comunicativa de la Terminología (Cabré, 1999; 2003). El equipo ha desarrollado también un trabajo relevante en el campo del procesamiento automático de las lenguas, centrado básicamente en la española y la catalana, primero bajo la denominación de LATRAL (ref. 1995SGR00342; 1996SGR00045; 1998SGR00085; 2000SGR00085), y, a partir del año 2001, como IULATERM (ref. 2001SGR00273; 2005SGR00232, 2009SGR1306, 2014SGR00946 y pendiente de renovación y referencia para el período 2017-2020)

Los trabajos del grupo han seguido una línea progresiva y sistemática: durante el período **1994-1996,** el grupo se concentró en la creación del corpus multilingüe de textos especializados en SGML y en el diseño y desarrollo de herramientas de tratamiento de este corpus, proyecto que recibió varias ayudas externas, la más importante corresponde al proyecto Corpus textual plurilingüe especializado (CIRIT CS-93-4009). **De 1996 a 1999**, período para el que se obtuvo el proyecto La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica (DGES PB96-0293), el grupo se concentró en el análisis y automatización de operaciones de detección de información a partir del corpus constituido y de la creación de un corpus de contraste de lengua general que permitiera comparar los datos entre los llamados textos especializados y los no especializados. Los resultados del proyecto fueron publicados en Cabré y Feliu (eds.) (2001) La terminología científico-técnica. Barcelona, IULA. ISBN: 84-477-0744-X, cuya publicación contó con financiación de la Acción Especial del Plan Nacional de I+D+I PGC-2000-3038 del MEC. **De 2000 hasta 2009**, período para el que se obtuvieron los proyectos Textos especializados y terminología: recuperación automática de la información (TEXTERM) (BFF-2000-0841), Fundamentos, estrategias y herramientas para el procesamiento y extracción

automáticos de información especializada (TEXTERM2) (BFF2003-02111), y Fundamentos, estrategias y herramientas para el procesamiento, extracción y representación de información especializada (TEXTERM3) (HUM2006-09458), la investigación del grupo se centró en: a) la formulación de un marco teórico para la terminología, de base lingüística, sin prescindir de los aspectos cognitivos y discursivos; b) el establecimiento de las condiciones de especialidad de un texto y los criterios formales de su clasificación en niveles de especialidad; c) un primer análisis de las unidades de conocimiento especializado; d) una primera clasificación de las relaciones conceptuales que se dan entre las unidades de conocimiento especializado y el análisis de las unidades verbales que expresan relaciones; e) un informe evaluativo sobre los distintos sistemas de extracción de terminología existentes y disponibles y el establecimiento de algunas estrategias de análisis y reconocimiento automático de candidatos a términos, con la incorporación de información semántica procedente de EuroWordNet. En este mismo período se desarrollaron también los proyectos RICOTERM. Sistema de recuperación de información con control terminológico y discursivo (TIC2000-1191), RICOTERM2. Control terminológico y discursivo para la recuperación de información en ámbitos comunicativos especializados, mediante recursos lingüísticos específicos y un reelaborador de consultas (HUM2004-056658-00) y RICOTERM3. Enriquecimiento de ontologías para la extracción de información de discursos de especialidad (HUM2007-65966-C02-00), de los cuales se obtuvieron resultados aplicados en sistemas de recuperación de información en textos especializados, en reelaboración de consultas especializadas multilingües en Internet y en adaptación temática y de lengua para la extracción automática de terminología, respectivamente. **Desde 2010 a 2015,** el grupo se ha centrado en la línea de investigación de la neología, en el marco de los proyectos APLE. Procesos de actualización del léxico del español a partir de la prensa (FFI2009-12188-C05-01), sobre la neología general en español y en catalán, y APLE2. Neologismos generales y neologismos especializados: análisis y contraste desde su producción, recepción y circulación social (FFI2012-37260), en el contraste entre neologismos generales y neologismos especializados. En otros proyectos de este período se ha trabajado en infraestructuras lingüísticas, como METANET4U. Enhancing the European Linguistic Infrastructure (UE-270893), o en el análisis y la extracción de colocaciones en el discurso especializado, con RICOTERM4. Procesamiento de corpus textuales especializados para la extracción de estructuras supraléxicas terminológicamente relevantes (FFI2020-21365-C093).

De los miembros del equipo investigador o de trabajo de otras universidades. Antecedentes de trabajo conjunto.

El Dr. **Juan Manuel Torres-Moreno**, doctor en ingeniería informática del Laboratorio Traitement Automatique de la Langue Naturelle e Ecrite (TALNE) de la **Université d'Avignon**, es especialista en el tratamiento del lenguaje con herramientas estadísticas. Con anterioridad miembros de nuestro equipo han colaborado con el grupo del Dr. Torres-Moreno (en el marco del proyecto RICOTERM4). Diversos investigadores del grupo IULATERM han realizado estancias postdoctorales en el laboratorio TALNE: Iria da Cunha realizó un postdoctorado de un año en 2014 para trabajar con medidas estadísticas aplicadas a textos médicos con la finalidad de extraer información semántica relevante; Jorge Vivaldi realizó una estancia en 2009 para trabajar la integración de extracción de terminología y resumen automático. El Dr. Torres-Moreno ha trabajado en más de una ocasión con el grupo IULATERM en la mejora de herramientas de PNL que han dado múltiples lugar a múltiples publicaciones que se recogen en su CV (Cabré et al., 2013). Las universidades respectivas firmaron hace 5 años un convenio de cooperación en la investigación, que actualmente se está renovando. En el último año se ha inicado una cotutela de doctorado por la cual la tesis de Andrés Torres Rivera està codirigida por Rosa Estopà y Juan Manuel Torres-Moreno.

La Dra. **Isabel Santamaría**, doctora en Filología Hispánica de la **Universidad de Alicante**, es especialista en lexicografía y recursos léxicos. Imparte la asignatura El español de la medicina y las ciencias de la salud en el Máster universitario en inglés y español para fines específicos. La investigadora ha colaborado anteriormente con el grupo en el proyecto RICOTERM4, dirigido por Mercè Lorente. En el último año ambas han participado como

coautoras en una publicación de los resultados de este proyecto (Lorente et al. 2017).

La Dra. **Fabiana França**, doctora en Ciencias de la Información, es investigadora de la Universdad Federal de Paraíba (Brasil). Durante el curso 2014-2015 realizó una estancia predoctoral en el grupo IULATERM con una beca CAPES-Brasil. La Dra. Rosa Estopà ha sido su codirectora de tesis doctoral, dedicada a las siglas médicas en portugués.

El Dr. **Daniel Luna** es médico y especialista en informática médica. Dirige el Departamento de Informática Médica del Hospital Italiano de Buenos Aires (Argentina). Se trata de un gran departamento (100 técnicos) ubicado en un hospital universitario de referencia, lo que los ha hecho situarse en un lugar preeminente en los desarrollos de informática médica. Su interés por incorporar técnicas del procesamiento del lenguaje natural ha propiciado el inico de una colaboración en investigación, que ha conducido a la firma de un conveno regulador (en preparación) y a su incorporación en este proyecto como asesor médico y como responsable de la entidad EPO.

Proyectos vigentes y antecedentes que confluyen en el espacio de documentación, comunicación e informática médicas

Desde 2015 el grupo ha iniciado una línea de investigación en la **educación en salud** con el proyecto RecerCaixa 2015 JUNTOS: superar barreras socioeducativas y favorecer la alfabetización sobre las interferencias y dificultades de comprensión de la información y documentación dirigida a families de niños afectados por enfermedades raras. El proyecto JUNTOS, dirigido por la Dra. Rosa Estopà (UPF) con el Dr. Armayones de la Universitat Oberta de Catalunya, parte de un problema concreto: el hecho de que las interferencias lingüísticas y cognitivas de los textos generados por profesionales y asociaciones destinados a pacientes y familias impiden su asimilación. En el contexto de una enfermedad rara pediátrica, el proyecto se plantea el reto de acercar la información médico-sanitaria a la familia y aporta soluciones (un protocolo de recomendaciones para los profesionales y las asociaciones, y una app con información adecuada para las familias afectadas) que permitan una alfabetización más inclusiva de este colectivo.
https://www.upf.edu/es/web/medicina_comunicacio]
Este proyecto finaliza en enero de 2018 y solo coincide con el nuevo proyecto planteado por ubicarse en la intersección entre comunicación y medicina. No obstante, como indicábamos anteriormente, los resultados sobre el cambio léxico y semántico en el discurso médico y la elaboración de los nuevos recursos léxicos que elaboraremos en este proyecto, podrían ser retomados para futuros proyectos en la línea de educación en salud.

Dentro del grupo IULATERM, la Dra. Núria Bel dirige, conjuntamente con el Dr. Horacio Saggion, el subproyecto de la UPF del proyecto coordinado TUNER (TIN2015-65308-C5-1-R) para el período 2016-2018, con la finalidad de investigar y desarrollar **técnicas de adaptación a dominio** para el procesamiento de big data en diversas lenguas. Los resultados del proyecto que presentamos aquí son de interés para los grupos de lingüística computacional del TUNER porqué pueden ser reutilizados para la adaptación al domino médico. Una muestra de la necesaria colaboración entre datos terminológicos y discursivos y tecnología adaptada a dominio es una primera publicación (Marimon, Vivaldi & Bel 2017).

La experiencia del grupo en la terminología y el discurso médico arranca en ya en su fundación con los trabajos de constitución del corpus médico en español y en catalán, pero también por la presencia del tema médco en diversos proyectos de terminología, neología y aplicaciones tecnológicas. Además, diversos miembros del equipo han realizado su tesis doctoral en el ámbito de la terminología médica: sobre estrategias para la creación de un extractor de terminología adecuado a las necesidades profesionales de diferentes colectivos (Estopà, 1999), en el diseño e implementación de un extractor de terminología médica que combina estrategias estadísticas, formales y semánticas (Vivaldi, 2001); en un resumidor automático de textos médicos (da Cunha, 2008); o estableciendo índices e instrumentos que permiten distinguir niveles de especialización de los textos (Domènech, 2006).

Por su parte, Rosa Estopà ha participado en diversos proyectos externos a la UPF relacionados con la terminología y la medicina: NeuroNEO. Regulación de los procesos

neológicos y los neologismos en las áreas de Neurociencias (FFI2012-34596), dirigido por Joaquín García Palacios de la Universidad de Salamanca; "Proposta de construção de um vocabulário de nomes populares das doenças e sua relação com a terminologia da área da saúde na perspectiva de melhorar a comunicação no contexto do "programa mais médicos" (2014-2016) y "Aplicabilidade da terminologia para a representação de conceitos visando a interoperabilidade semântica no compartilhamento de informação de prontuários do paciente" (2012-2014), dirigidos por la Dra. Virginia Bentes de la Universidade Federal do Ceará (Brasil).

C.1.3. Hipótesis

1. El análisis del cambio léxico y semántico de la terminología médica nos permite representar la evolución del conocimiento de este ámbito. En concreto, el análisis de la variación terminológica detectada a lo largo del tiempo, atendiendo a la presencia o la ausencia de unidades, a las modificaciones que se producen entre paradigmas o a los cambios que se detectan en cuanto al contenido, es determinante para representar esta evolución (oculta) del conocimiento.

2. La evolución permanente en medicina facilita que el análisis del cambio léxico y semántico se aplique a segmentos diacrónicos de corpus de corto alcance (7 años), adecuados para el estudio de los cambios lingüísticos enclavados en las categorías de lexicalización y gramaticalización menor. Actualmente, la disponibilidad de un corpus de textos médicos de 28 años (1990-2018) en español nos permite realizar 4 cortes diacrónicos para el análisis evolutivo.

3. El seguimiento detallado de la evolución de la terminología médica nos permite generar nuevos recursos léxicos específicos, con informaciones necesarias para el desarrollo de tecnologías del lenguaje adaptadas a dominio, accesibles en abierto, relacionadas con el procesamiento, la extracción y la generación automáticas.

4. El cambio léxico y semántico de la terminología médica, si bien comparte principios y condiciones externas e internas, se distingue del cambio léxico de la lengua general por la inmunidad a ciertas gramaticalizaciones, lexicalizaciones y cambios semánticos, a causa de su función representativa del conocimiento experto.

C.1.4. Objetivos

Generales y específicos

1. Analizar el cambio en la variación terminológica en textos médicos a partir del eje temporal, a través de la presencia y ausencia de formas, las modificaciones entre paradigmas y los cambios de contenido, para tener más conocimiento sobre la evolución léxica y semántica (a menudo oculta) que acompaña la evolución del conocimiento científico en medicina.

2. Construir una serie de recursos textuales y léxicos en lengua española, óptimos para el desarrollo de tecnologías del lenguaje adaptadas a dominio y a las necesidades de los diferentes usuarios (médicos generales, médicos especialistas, investigadores, pacientes). Los recursos construidos serán accesibles en abierto para los desarrolladores de tecnologías del lenguaje del ámbito de la documentación, la comunicación y la informática médicas.

Específicos

1. Constituir y anotar un subcorpus de informes clínicos anonimizados y cedidos por la EPO, e integrarlo en la plataforma del Corpus Técnico del IULA-UPF.
2. Realizar segmentaciones de 7 años en el corpus médico en español (de más de 4.000.000 de palabras) para el análisis de la evolución.

3. Analizar la variación estructural de las unidades léxicas en textos médicos de géneros diversos.
4. Analizar el papel de los adjetivos no especializados en la delimitación y establecimiento de las unidades terminológicas
5. Analizar la variación intracategorial (la relación entre unidades léxicas de categoría distinta) en textos médicos de géneros diversos.
6. Analizar la variación semántica (polisemia) de las unidades terminológicas y de la neología en textos médicos de géneros diversos.
7. Analizar las reducciones léxicas (siglas, acrónimos y abreviaturas) de las unidades terminológicas y de la neología en textos médicos de géneros diversos.
8. Detectar neologismos y términos obsoletos.
9. Analizar estadísticamente, con la ayuda de la herramienta ESTEN (IULA-UPF), la variación terminológica en los fragmentos del corpus segmentado y la evolución de esta variación.
10. Construir un servicio web de expansión de siglas médicas.
11. Constituir una Base de datos de los resultados de variantes terminológicas y su evolución.
12. Constituir una Base de datos de acepciones o variación semántica (polisemia).

C.1.5. Colaboraciones y apoyos

Las entidades EPO

El Hospital Italiano de Buenos Aires (Argentina), hospital universitario de referencia, dispone de un gran departamento de Informática Médica, dirigido por el Dr. Daniel Luna, con más de 100 técnicos informáticos, lo que los ha hecho situarse en un lugar preeminente en los desarrollos de informática médica. En esta línea, el HIBA apoya el proyecto para iniciar una vía de cooperación efectiva en la creación de recursos léxicos y textuales y en el desarrollo de tecnologías para el procesamiento de la documentación médica.

El coordinador del Área de Tecnología del Gabinete de la Secretaría de Estado para la Sociedad de la Información y la Agenda Digital (SESIAD), D. David Pérez Fernández, conocedor de la necesidad de disponer de recursos textuales y léxicos específicos para el desarrollo de tecnologías adaptadas a dominio, apoya este proyecto, en el marco de la búsqueda de sinergias y cooperaciones entre grupos para el Plan Impulsa de las Tecnologías del Lenguaje.

Relación con otros centros médicos

Además de la colaboración iniciada con el hospital Italiano de Buenos Aires, el equipo de investigación mantiene relaciones con otros centros hospitalarios, como el Consorcio Sanitario de Terrassa, el Hospital o el Servicio de Documentación del Hospital Clínico de Barcelona, o de formación sanitaria, como el Instituto Bonanova FP Sanitaria, para la cesión de materales o para el asesoramiento de médicos y otros profesionales de la salud.

Relación con grupos de investigación nacionales

En las líneas de educación y comunicación en la salud, cabe destacar la cooperación con el el grupo liderado por el Dr. Armayones de la U. Oberta de Catalunya que trabaja en el campo de la alfabetización médica digital desde una aproximación psicológica (Armayones et al.,2012) y con el equipo de OBSER de la Federación Española de Enfermedades Raras, que lidera el Dr. Solves, que trabaja en la concepción mediática de las enfermedades minoritarias y del colectivo afectado a través de la prensa diaria (Solves, 2013).

En documentación médica, conviene citar las colaboraciones docentes de Rosa Estopà en el máster de la Universidad Autónoma de Barcelona y en la formación FP de grado superior del Institut Bonanova FP Sanitaria.

En las líneas de terminología y neología en intersección con la medicina, el grupo mantiene relaciones con los siguientes grupos:

LingMed de Lingüística Médica de la U. Jaume I, liderado por el Dr. Vicent Salvador que colabora con otros investigadores como la Dra. M. Josep Marín de la U. de Valencia y el Dr. Josep Bernabeu de la U. de Alicante y trabajan desde principios de los 2000 en diversos proyectos sobre discurso médico (Salvador, et al., *Caplletra*, 50, 2011)

NeoUsal de la U. de Salamanca (http://neousal.usal.es/), dirigido por el Dr. J. García Palacios, que trabaja en proyectos de terminología en el ámbito de las neurociencias, como Mecanismos de generación neológica y actualización léxica del español a través del estudio de la prensa castellanoleonesa (NeoUSAL II). Consejería de Educación. Junta de Castilla y León, ref. SA274U1

LEXICON de la U. de Granada (http://lexicon.ugr.es/), dirigido por la Dra. Pamela Faber, que ha trabajado en ontologías del campo de la medicina como en el proyecto ONCOTERM (http://www.ugr.es/~oncoterm/), un proyecto sobre terminología médica realizando conjuntamente con el Hospital Universitario Virgen de las Nieves. Miembros del grupo lexicon, conjuntamente con investigadores de la U. de Valladolid (Dra. Méndez), de U. Pablo de Olavide (Dr. Prieto) y de Rutgers U. (Dr. Jiménez) y Carleton U. (Dr. D'Anguilli) han trabajado en el proyecto VariMed (FFI2011-23120), dirigido por la Dra. Maribel Tercedor, para estudiar la variación denominativa como fenómeno cognitivo y comunicativo entre médico y paciente (http://varimed.ugr.es/). Actualmente la Dra. Tercedor y la Dra. Clara Inés López Rodríguez dirigen el proyecto COMBIMED: Léxico combinatorio en medicina: cognición, texto y contexto FFI2014-51899-R (2015-2017), que plantea el estudio de las combinaciones léxicas de la medicina y de su representación y visualización desde una perspectiva pragmática y cognitiva de cara a la comprensión de los procesos psicolingüísticos, comunicativos y estilísticos de selección léxica que llevan a cabo tanto proveedores y usuarios de los sistemas de salud como redactores y traductores.

En la línea de las tecnologías del lenguaje, y en especial aplicaciones léxicas, el equipo mantiene relaciones con:

Grupo de Procesamiento del Lenguaje Natural (GRPLN), dirigido por Dr. Lluís Padró de la Universidad Politécnica de Catalunya, que trabaja en el desarrollo de herramientas de procesamiento del lenguaje; en concreto, desde hace más de una década hemos colaborado estrechamente con uno de los miembros de GRPLN, el Dr. Horacio Rodríguez (Vivaldi; Rodríguez, 2015), interesado en los últimos años en PNL médico.

Ontology Engineer Group (OEG) http://www.oeg-upm.net/, dirigido por Asunción Gómez Pérez, de la Universidad Politécnica de Madrid. El grupo IULATERM participa de la red de excelencia ReTeLe. Red de recursos para tecnologías de la lengua. TIN2015-68955-REDT para el período 2016-2017, coordinada por la IP del grupo OEG. Además, a través de la Asociación Española de Terminología, su presidenta, la Dra. Guadalupe Aguado de Cea, miembro del grupo OEG, se han abierto colaboraciones directas con los miembtos del equipo de investigación del proyecto, expertos en terminología, para la cooperación en datos enlazados (Open Linked Data).

Grupo IXA de la Universidad del Páis Vasco, dirigida por Arantza Dáz de Ilarraza, grupo de referencia en tecnologías del lenguaje, que integra a su vez algunos investifgadores de la línea de terminología, como Igone Zabala o Mikel Lersundi, que han participado en proyectos de la serie RICOTERM, dirigidos por Mercè Lorente.

Otros centros o grupos de referencia, cuya actividad es de interés para el proyecto: UCA de la Universidad de Cádiz, Dir. Miguel Casas; la red Morfoterem, que integra grupos de investigación que trabajan en morfología léxica, incluidos algunos que lo hacen en morfología diacrónica, como Isabel Puyol de la U. de Girona, etc.

Relación con grupos de investigación internacionales

En equipo también mantiene intercambios científicos con **grupos extranjeros** que trabajan con temas afines como, por ejemplo, el equipo de documentalistas médicos liderado por la Dra. Virginia Bentes (Bentes et al., 2013) de la U. de Fortaleza (Brasil).

En el ámbito de la terminología y la neología el equipo coopera con grupos como los siguientes:

CRTT Lyon 2 (Francia), dirigido por Vincent Renner. Participación de membros del equipo en el proyecto #Néo de la Université de Lyon 2

Proyecto Logoscope de la Université de Strasbourg, dirigido por Christophe Gerard, sobre herramientas de análisis de la neología.

Las Antenas Neológicas, red en la que trabajan cooperativamente en neología del español, grupos de investigación de España, Argentina, Colombia, Chile, Perú y México , bajo la coordinación del Observatorio de Neología, en el que participan las miembros del equipo de investigación del proyecto.

C.1.6. Metodología

Hemos establecido 5 paquetes de trabajo (4 de contenido y 1 logístico), relacionados con los resultados o *entregables* que pretende conseguir el proyecto. Para cada uno de los paquetes de trabajo, hemos detallado los responsables, el equipo de trabajo, las tareas concretas a realizar, así como la metodología principal a seguir.

PT1. Corpus de medicina en español
Coordinadora del PT1: Mercè Lorente
Equipo de trabajo: Técnico informático contratado a tiempo parcial, como se detalla en el presupuesto, con el apoyo de Óscar Pozuelo y Andrés Torres y el asesoramiento de Daniel Luna y Juan Manuel Torres-Moreno.
Tareas previstas:
1.1. Constitución y anotación del subcorpus de informes clínicos cedidos por la EPO.
1.2. Segmentación diacrónica del corpus de medicina del español.
1.3. Extracción de datos del corpus global y de los subcorpus segmentados: extracción automática de terminología, detección de neología formal, extracción de siglas y otras reducciones, análisis de frecuencias, extracción de unidades por categorías morfosintácticas y por estructuras y combinatorias.

PT2. Análisis del léxico: variación terminológica, neología y otras unidades
Coordinadora del PT2: Rosa Estopà
Equipo de trabajo (incluido el equipo de investigación): M. Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Feliu, Fabiana França, Mercè Lorente, Amor Montané, Isabel Santamaría, Óscar Pozuelo, Coralie Schneider, Andrés Torres.
Tareas previstas (los participantes):
2.1. Análisis de la variación estructural (variantes terminológicas según procedimiento de formación léxica)
2.2. Análisis de la variación intracategorial (paradigmas léxicos con distintas categorías relacionadas formal o semánticamente: substantivos, verbos, adjetivos, nominalizaciones)
2.3. El problema de la segmentación de unidades sintagmáticas. El papel de los adjetivos.
2.4. Detección de siglas, acrónimos y abreviaturas, y sus posibles expansiones.
2.5. Análisis de la polisemia (variación semántica)
2.6. Detección de neologismos y de términos obsoletos.

PT3. Estudio del cambio léxico en el discurso médico
Coordinadora del PT3: Mercè Lorente

Equipo de trabajo (incluido el equipo de investigación): M. Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Feliu, Fabiana França, Mercè Lorente, Amor Montané, Isabel Santamaría, Óscar Pozuelo, Coralie Schneider, Andrés Torres.
Tareas previstas:
3.1. Seguimiento de un conjunto relevante de variantes léxicas por períodos de 7 años (segmentos del corpus). Análisis de la evolución a lo largo de los segmentos. Aplicación de la herramienta ESTEN de análisis estadístico diacrónico del léxico para cada uno de los ítems del PT2.
3.2. Seminario regular sobre el cambio léxico (análisis de la bibliografía y de los resultados parciales, y discusión)
3.3. Establecimiento de una guía de análisis del cambio léxico en discurso médico. Documentación de la metodología de trabajo y de los resultados.

PT4. Creación de recursos léxicos
Responsable del PT4: Rosa Estopà
Equipo de trabajo: Técnico informático contratado a tiempo parcial, como se detalla en el presupuesto, con el apoyo de Fabiana França, Juan Manuel Torres-Moreno, Óscar Pozuelo, Coralie Schneider, Andrés Torres y el asesoramiento de Daniel Luna.
Tareas previstas:
4.1. Base de datos de la variación terminológica (formal) y su evolución
4.2. Servicio web para la expansión de reducciones (siglas, acrónimos y abreviaturas)
4.3. Base de datos de las acepciones detectadas para una misma forma (variación semántica o polisemia)
4.4. Base de datos de neologismos y términos obsoletos detectados
4.5. Prueba de detección automática de la neología semántica

PT5. Logística
Equipo de trabajo del PT5: Mercè Lorente y Rosa Estopà, con el apoyo de técnico informático y de los servicios centrales de la UPF.
Tareas programadas:
5.1. Elaboración de protocolos de trabajo y planificación de actividades
5.2. Web del proyecto
5.3. Seguimiento de los convenios e informes parciales
5.4. Diseminación de resultados parciales y finales a los EPO y en congresos y publicaciones
5.5. Revisión de materiales y emisión de informes de resultados (anexos al informe final)

Infraestructura

Desde 1994 hasta la actualidad el grupo ha creado recursos (corpus) y herramientas informáticas que serán de gran utilidad para la realización del proyecto que proponemos y que pueden consultarse en: http://www.iula.upf.edu/recurs02ca.htm. Algunas de estas herramientas utilizables en el proyecto son: preprocesamiento, etiquetador (tagger) y desambiguador morfológicos, analizador sintáctico (treebank), extractores automáticos de terminología YATE y WIKIYATE, analizador estadístico de variantes terminológicas ESTEN, gestor de corpus y terminología TERMINUS 2.0.

C.1.7. Cronograma

PRIMER AÑO (2018)

PT	Nº tarea	TAREAS	1/18	2/18	3/18	4/18	5/18	6/18	7/18	9/18	10/18	11/18	12/18
1	1.1.	Subcorpus de informes clínicos anotado	X	X	X	X	X	X	X	X	X	X	X
1	1.2.	Segmentación del corpus						X	X	X	X		
1	1.3.	Extracción de datos y mantenimiento herramientas				X	X	X	X	X	X	X	X
2	2.1.	Análisis de la variación estructural								X	X	X	X
2	2.2.	Análisis de la variación categorial								X	X	X	X
2	2.3.	La segmentación de sintagmas. Adjetivos								X	X	X	X
2	2.4.	Detección de siglas y expansión								X	X	X	X
2	2.5.	Análisis de la polisemia								X	X	X	X
3	3.2.	Seminario sobre el cambio léxico	X	X	X	X	X	X	X	X	X	X	X
5	5.1.	Protocolos de trabajo y planificación anual	X	X	X								
5	5.2.	Creación y edición de la web del proyecto				X	X	X					

SEGUNDO AÑO (2019)

PT	Nº tarea	TAREAS	1/19	2/19	3/19	4/19	5/19	6/19	7/19	9/19	10/19	11/19	12/19
1	1.3.	Extracción de datos y mantenimiento	X	X	X	X	X	X	X	X	X	X	X
2	2.1.	Análisis de la variación estructural	X	X	X								
2	2.2.	Análisis de la variación categorial	X	X	X								
2	2.3.	La segmentación de sintagmas. Adjetivos	X	X	X								
2	2.4.	Detección de siglas y expansión	X	X	X								
2	2.5.	Análisis de la polisemia	X	X	X								
2	2.6.	Detección y análisis de neologismos	X	X	X								
3	3.1.	Seguimiento del léxico por períodos: ESTEN	X	X	X	X	X	X	X	X	X	X	X
5	5.3.	Seguimiento de convenios e informes	X	X	X								
5	5.4.	Diseminación de resultados parciales				X	X	X	X	X	X	X	X

TERCER AÑO (2020)

PT	Nº tarea	TAREAS	1/20	2/20	3/20	4/20	5/20	6/20	7/20	9/20	10/20	11/20	12/20
3	3.1.	Seguimiento del léxico por períodos: ESTEN	X	X	X	X	X	X	X				
4	3.1.	BD variación terminológica y evolución	X	X	X	X	X	X	X				
4	3.2.	Servicio web de expansión de siglas médicas	X	X	X	X	X	X	X				
4	3.3.	BD de variación semántica	X	X	X	X	X	X	X				
4	3.4.	BD de neologismos y obsoletos	X	X	X	X	X	X	X				
4	3.5.	Pruebas de detección autom. de neología semántica	X	X	X	X	X	X	X				
5	5.3.	Seguimiento de convenios e informes	X	X	X								
5	5.4.	Diseminación de resultados parciales y finales	X	X	X	X	X	X	X	X	X	X	X
5	5.5.	Revisión de materiales e informes finales								X	X	X	X

C.1.8. Contratación de personal

En el equipo de trabajo hemos incluido aquellos doctorandos cuyas tesis tienen algunos resultados aplicados coincidentes con el proyecto:

- Óscar Pozuelo, tesis de tercer año (2016-17), con previsión de defensa durante el año 2018. Algunos de sus resultados descriptivos están vinculados directamente a tareas del PT2 y puede colaborar también en el resto de PT.
- Coralie Schneider, tesis de segundo curso (2016-17), con previsión de defensa durante el año 2018. Sus resultados están vinculados indirectamente (otras lenguas de trabajo) a tareas del PT2.
- Andrés Torres, tesis de segundo curso (2016-17), con previsión de defensa durante el año 2018. Sus resultados están vinculados directamente al PT2 y PT4, pudiendo colaborar también en el PT1.

La solicitud de contratación de un técnico informático a tiempo parcial se justifica necesariamente por el trabajo previsto en las siguientes tareas:

Constitución y anotación del subcorpus de informes clínicos cedidos por la EPO.
Segmentación diacrónica del corpus de medicina del español.
Explotación del corpus y extracción de datos del corpus global y de los subcorpus segmentados
Mantenimiento de la herramienta ESTEN
Apoyo en la constitución y el mantenimiento de las bases de datos
Desarrollo del Servicio Web de siglas

En los proyectos anteriores del grupo disponíamos de un analista informático asociado al grupo, el Dr. Jorge Vivaldi, que podía ocuparse de estas tareas, pero está prevista su jubilación y la no reposición de su puesto de trabajo para noviembre de 2017. Por lo tanto la contratación de un informático , aunque sea a tiempo parcial, se hace imprescindible para llevar adelante las tareas del proyecto.

Se solicita asimismo la asignación de un **becario FPI** para el proyecto, con el fin de potenciar una nueva investigación doctoral integrada al 100% en el tema del proyecto, y con proyección de futuro en un sector en crecimiento como es el de la informática médica, con competencias transversales en ingeniería lingüística, descripción lingüística y documentación automática.

C.1.9. Bibliografía citada en este documento científico

Armayones, M., et al. (2012). "Guiametabolica.org: empowerment through internet tools in inherited metabolic diseases". Orphanet Journal of Rare Diseases. Núm. N:7, ISSN.1750-1172. DOI: 10.1186/1750-1172-7-53.

Bentes, V.; de Holanda, H. (org.) (2013) *Diálogos paradigmáticos sobre informaçao para a área de saúde*. Fortaleza: UFC. ISBN 978-85-7282-533-7

Benveniste, É., 1968. Mutations of linguistic categories. En W.P. Lehmann y Y. Malkiel (eds.) *Directions for historical linguistics,* Austin y Londres: University of Texas Press, pp. 83-94.

Brinton, L.J. & Traugott, E.C., 2005. *Lexicalization and language change,* Cambridge: Cambridge University Press.

Brown, E.L. & Raymond, W.D., 2012. How discourse context shapes the lexicon: explaining the distribution of Spanish f- / h- words. *Diachronica,* 92(2), pp. 139-161.

Bybee, J., 1985. *Morphology: a study of the relation between meaning and form*, Philadelphia: John Benjamins.

Bybee, J., 2015. *Language Change,* Cambridge: Cambridge University Press.

Cabré Castellví, M. Teresa (2015). "La neologia: un nou camp a la cerca de la seva consolidació científica". En *Caplletra* 59. València: Institut de Filologia Valenciana. 125-136.

Cabré, M. Teresa (2015). "Bases para una teoría de los neologismos léxicos: primeras reflexiones". En Alves, Ieda Maria; Simões Pereira, Eliane (eds.) *Neologia das Línguas Românicas.* São Paulo: CAPES; Humanitas. 79-107

Cabré, M. Teresa; Domènech, Ona; Estopà, Rosa (ed.) (2014). *Mots nous en català: una panoràmica geolectal = New words in Catalan: A diatopic view*. Amsterdam: John Benjamins. ISBN 9789027240132.

Cabré Castellví, M. Teresa; Montané March, M. Amor (2015). "ESTEN, nueva herramienta de medida de la implantación". En Alves, Ieda Maria; Simões Pereira, Eliane (eds.) *Neologia das Línguas Românicas.* São Paulo: CAPES; Humanitas. 1095-1106.

Cabré, M. Teresa; Nazar, Rogelio (2012). "Towards a new approach to the study of neology". En *Neologica: revue internationale de néologie* 6. Paris: Laboratoire de linguistique informatique (LLI, Université Paris XIII - CNRS); Éditions Garnier. 63-80.

Cabré y Bach 2010 Coneixement, llenguatge i discurs especalitzat. Barcelona: IULA-UPF; Documenta Universitaria.

Cabré, M. T; da Cunha, I.; SanJuan, E.; Torres-Moreno, J-M.; Vivaldi, J. (2013) Automatic Specialized vs. Non-specialized Text Differentiation: The Usability of Grammatical Features in a Latin Multilingual Context. *Languages for Specific Purposes in the Digital Era*, Springer, pp 223-241, 2013

Cruz, N. P., R. Morante, M. J. Mana, J. Mata, and C. L. Parra. 2017. Annotating negation in Spanish clinical texts. In Proceedings of the Workshop SemBEaR 2017 Computational Semantics Beyond Events and Roles. The Association for Computational Linguistics, pages 53- 58.

Da Cunha, I.; Vivaldi, J.; Torres-Moreno, J-M.; Sierra, G. (2014) SIMTEX: An Approach for Detecting and Measuring Textual Similarity based on Discourse and Semantics. *Computacion y Sistemas*, 18(3):505--516, 2014.

Domènech Bagaria, Ona. Textos especialitzats i variació vertical: la diversitat terminològica com a factor discriminant del nivell d'especialització d'un text. Tesis doctoral. Universitat Pompeu Fabra; 2006.

Dressler, W.U., 2003. Naturalness and morphological change. En B.D. Joseph y R.D. Janda (eds.) *The handbook of historical linguistics,* Oxford: Basil Blackwell, pp. 461-471.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A. and Biemann, C. (2016): A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In Proceedings of the LT4DH workshop at COLING 2016, Osaka, Japan.

Estopà, R. (2012) "Medicina y llenguatge: Les paraules de la salut". LSC, 2012, 10.

Estopà Bagot, Rosa (2015). "Sobre neologismos y neologicidad: reflexiones teóricas con repercusiones metodológicas". En Alves, Ieda Maria; Simões Pereira, Eliane (eds.) *Neologia das Línguas Românicas.* São Paulo: CAPES; Humanitas. 111-150.

Estopà, Rosa (2012). "El léxico de la salud: necesidades terminológicas y acceso al conocimiento especializado". En Bente, V.; de Holanda, H. (eds.) *Diálogos paradigmaticos: informacaçao para a àrea de saúde.* Fortaleza: Universidad de Ceará.

Faber, Pamela (2012). A Cognitive Linguistics View of Terminology and Specialized Language. Berlin: Walter de Gruyter

Faber, P. and M.C. L'Homme (eds.) 2014. *Lexical Semantic Approaches to Terminology.* Special issue of *Terminology* 20(2).

Feliu, Judit. Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica. Universitat Pompeu Fabra; 2004.

Fernández-Silva, Sabela. Variación terminológica y cognición: factores cognitivos en la denominación del concepto especializado. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada; 2011.

Freixa, Judit. "Variación terminológica: ¿Por qué y para qué?."Meta 504 (2005): None–None. DOI : 10.7202/019917ar

Freixa, Judit (2014). "La variación denominativa en terminología: tipos y causas". En Negri Isquerdo, Aparecida; Mantovani Dal Corno, Giselle Olivia (org.) *As ciências do léxico: lexicologia, lexicografia, terminologia*, vol. VII. Campo Grande: UFMS. 311-329. ISBN 978-85-7613-163-2 (UFMS).

Freixa, Judit; Bernal, Elisenda; Cabré, M. Teresa (ed.) (2015). *La neologia lèxica catalana.* Barcelona: Institut d'Estudis Catalans. Secció Filològica. (Biblioteca Filològica; 76).

Geeraerts, D., 1997. *Diachronic prototype semantics: a contribution to historical* lexicology, Oxford: Clarendon Press.

Goldberg, A., 1995. *Constructions: a construction grammar approach to argument structure,* Chicago: University of Chicago Press.

Hamilton, W. L., Leskovec, J. & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *CoRR*, abs/1605.09096.

Hay, J., 2001. Lexical frequency in morphology: is everything relative? *Linguistics,* 39(6), pp. 1041-1070.

Heine, B. & Kuteva, T., 2002. *World lexicon of grammaticalization,* Cambridge: Cambridge University Press.

Hopper, P.J. & Traugott, E., 1993. *Grammaticalization,* 1st edition, Cambridge: Cambridge University Press.

Keller, R., 1994. *On language change: the invisible hand in language,* Londres y Nueva York: Routledge.

Labov, W., 1972. *Sociolinguistic patterns,* Philadelphia: University of Pennsylvania Press.

Labov, W., 1994. *Principles of linguistic change,* Vol. 1 *Internal factors,* Oxford: Basil Blackwell.

Labov, W., 2001. *Principles of linguistic change,* Vol. 2 *Social factors,* Oxford: Blackwell Publishers.

Labov, W., 2010. *Principles of linguistic change: cognitive and cultural factors,* Chichester: Wiley-Blackwell.

Liu, Hongfang, Kavishwar Wagholikar, and Stephen Tze-Inn Wu. "Using SNOMED-CT to Encode Summary Level Data – a Corpus Analysis." AMIA Summits on Translational Science Proceedings 2012 (2012): 30–37. Liu et al. 2012

Lorente Casafont, Mercè (2014). "Hacia la terminología 3.0: evolución del uso de las tecnologías en la terminología". En Vargas Sierra, Chelo (ed.) *TIC, trabajo colaborativo e interacción en terminología y traducción*. Granada: Comares. 9-26. ISBN 978-84-9045-046-8.

Lorente, Mercè (2015). "Neología y gramática". En Alves, Ieda Maria; Simões Pereira, Eliane (eds.) *Neologia das Línguas Românicas.* São Paulo: CAPES; Humanitas. 219-248.

Lyons, John . 1977. *Semantics*. Cambridge: Cambridge Unversity Press.

Lüdtke, H., 1998. *El cambio lingüístico,* Barcelona: Universitat Autònoma de Barcelona.

Marimon, Montserrat; Bel, Nuria 2015. Dependency structure annotation in the IULA Spanish LSP Treebank, Language Resources and Evaluation, 49, 2, 433-454, 2015, Springer Netherlands

Marimon, Montserrat; Vivaldi, Jorge; Bel, Núria. 2017. Annotation of negation in the IULA Spanish Clinical Record Corpus, SemBEaR 2017.

Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. Fourth

Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing - BioTxtM2014. 2014:24-30

Oronoz, M., K. Gojenola A. Pérez A. D. de Ilarraza A. Casillas "On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions", Journal of Biomedical Informatics. vol. 56 no. 0 pp. 318-332 2015.

Phillips, B.S., 2006. *Word frequency and lexical diffusion,* New York: Palgrave.

Poplack, S., 2011. Grammaticalization and linguistic variation. En H. Narrog y B. Heine (eds.) *The Oxford handbook of grammaticalization,* Oxford: Oxford University Press, pp. 209-224.

Prieto Velasco, J.A. & Tercedor Sánchez, M. (2014) The embodied nature of medical concepts: image schemas and language for pain. *Cognitive processing*, 15(3):283-296. Springer

Quiros, F.; López, A. et al. (2007) Medical Informatics Department, Hospital Italiano of Buenos Aires, Argentina. Studies in health technology and informatics 02/2007; 129(Pt 1):621-5.

Robert, S., 2008. Words and their meanings. En M. Vanhove (ed.) *From polysemy to semantic change,* Studies in Language Companion Series, Amsterdam y Philadelphia: John Benjamins, pp. 55-92,

Safeer, R. S., & Keenan, J. (2005). Health literacy: the gap between physicians and patients. American Family Physician, 72(3), 463–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16100861).

Seghezzi, Natalia. Variación terminológica y canal de comunicación. Estudio contrastivo de textos especializados escritos y orales sobre lingüística. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada; 2011

Sorense, K. et al. (2012) *Health literacy and public health: A systematic review and integration of definitions and models.* BMC Public Health. 2012; 12:80 doi:10.1186/1471-2458-12-80

Suárez de la Torre, Mercedes. Análisis contrastivo de la variación denominativa en textos especializados: del texto original al texto meta. Universitat Pompeu Fabra; 2004.

Stubbs, M., 2002. *Words and phrases: corpus studies of lexical semantics,* Oxford: Blackwell Publishers.

Temermann, Rita (2000). Towards New Ways of Terminology Description. The sociocognitive approach. Amsterdam: John Benjamins

Traugott, E.C. & Dasher, R.B., 2002. *Regularity in semantic change,* Cambridge: Cambridge University Press.

Ullmann, S. 1962. *Semantics: an Introduction to the Science of Meaning*. Oxford: Blackwell.

Uzuner et al. 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):519-23.

Vivaldi, Jorge; Rodríguez, Horacio (2015). "Medical Entities Tagging Using Distant Learning". Gelbukh, Alexander (ed.) *Computational Linguistics and Intelligent Text Processing : 16th International Conference, CICLing 2015 Cairo, Egypt, April 14-20, 2015 : Proceedings* 2. Heidelberg: Springer. 631-642. ISBN 978-3-319-18110-3

Vivaldi, Jorge; Rodríguez, Horacio (2012). "Using Wikipedia for Domain Terms Extraction". En Gornostay, Tatiana (ed.) *Proceedings of CHAT 2012: The 2nd Workshop on the Creation, Harmonization and Application of Terminology Resources: co-located with TKE 2012; 2012 June 22; Universidad Politécnica de Madrid, Madrid, Spain.* Linköping, Sweden: Linköping University Electronic Press, Linköpings universitet. 3-10. (Linköping Electronic Conference Proceedings; 72). http://www.ep.liu.se/ecp/072/001/ecp12072001.pdf

Walsh SH. The clinician's perspective on electronic health records and how they can affect patient care. BMJ. 2004; 328 : 1184-7

Y. Yamamoto, A. Yamaguchi, H. Bono and T. Takagi, "Allie: a database and a search service of abbreviations and long forms.", Database, 2011:bar03.

Zanotti, Cristian; Vivaldi, Jorge; Lorente, Mercè (2012). "Upgrading WordNet: a Terminological Point of View". En Fellbaum, Christiane; Vossen, Pick (eds.) *GWC 2012: 6th International Global Wordnet Conference: January 9-13, 2012, Matsue, Japan.* Brno: Tribun EU. 390-399.

C.2. IMPACTO ESPERADO DE LOS RESULTADOS

Los resultados del proyecto TERMMED permitirán adquirir más conocimiento sobre el cambio léxico y semántico, completar bases de conocimiento médico a través de información semántica nueva y plantear nuevas estrategias para la detección de la neología semántica. TERMMED pretende generar interés, resultados e impacto científico en diversas áreas del conocimiento: salud, medicina, lingüística, terminología, documentación, comunicación, tecnologías del lenguaje, procesamiento del lenguaje natural.

Contribuciones al avance del conocimiento en lingüística:

- El salto cualitativo desde los estudios de terminología y neología de corte descriptivo hacia una mayor interacción con presupuestos de la lingüística teórica sobre el cambio léxico y semántico.
- Una primera propuesta para la variación terminológica y para la neología especializada que permita evaluar qué impacto tienen los llamados lenguajes de especialidad en el cambio lingüístico.
- La profundización en algunos temas apuntados en investigaciones anteriores sobre la delimitación de unidades terminológicas (reducciones, sintagmas con adjetivos graduales, verbos, locuciones verbales, la dinámica de la variación terminológica).

Beneficios transferibles (recursos lingüísticos construidos y puestos a disposición de la comunidad científica):

- Corpus textual de medicina en español actualizado y anotado, de casi 5.000.000 de palabras, integrado en el Corpus técnico del IULA.
- Un servicio web sobre siglas y otras reducciones en medicina y sus correspondientes expansiones con toda su variación.
- Una base de datos de unidades terminológicas con datos enlazados y variación terminológica extraída de textos médicos, analizada estadísticamente y cualitativamente.

Cooperación internacional: Refuerzo de las colaboraciones en informática clínica con el Hospital Italiano de Buenos Aires (centro universitario y de investigación, puntero en este ámbito). Consolidación de la cooperación en investigación y formación de investigadores con el Laboratorio de Informática de la Universidad de Avignon. Abertura de una línea de investigación común con los servicios de documentación de la Universidad Federal de Paraíba (Brasil), resultado de la formación doctoral de la Dra. França en Barcelona.

Plan de difusión previsto:

- Contribuciones en congresos internacionales. Aspectos lingüísticos de la terminología y la neología: RITERM2018, CINEO2018. Comunicación médica: IV Congreso Internacional de Comunicación en Salud (previsto 2019). Sobre recursos lingüísticos desarrollados, corpus y extracción de terminología y neología: LREC, COLING o ACL.
- Contribuciones en revistas de terminología y neología (*Terminology,Terminàlia, Neologica*), de lingüística (*Natural Language Semantics*, *Language variation and change*, *Semantic web, International Journal of Corpus Linguistics, Lexicology, Ibérica, Caplletra*), de lingüística computacional (*Computational Linguistics, Natural Language Engeenering, Linguamática*, *Procesamiento del Lenguaje Natural)* y de comunicación médica (*Revista Española de Comunicación en Salud*).
- Una jornada internacional de investigación sobre recursos lingüísticos para la medicina
- La publicación de los resultados teóricos del proyecto (publicación electrónica de los informes)

C.3. CAPACIDAD FORMATIVA DEL EQUIPO SOLICITANTE

El equipo solicitante forma parte del grupo de investigación IULATERM (Léxico y Tecnología), considerado grupo consolidado con financiación del mapa de investigación de Catalunya, con referencia 2014SGR00946 (última referencia, pendientes de la resolución de la última convocatoria). IULATERM es un grupo con una alta capacidad formativa, como muestran los datos siguientes:

DOCTORADO

Total de tesis defendidas dirigidas por doctores del grupo (1999 – juny 2017): 51
Tesis defendidas entre enero de 2014 y junio de 2017: 13
Depositadas a 30 de junio de 2017: 2
Tesis en curso a 30 de junio de 2017: 15

Tesis defendidas en los últimos años

Formas de redução léxica na redação de resumos de alta em prontuário do paciente visando a indexação e recuperação da informação. Doctoranda: Fabiana França (Brasil). Tesi Universidad Federal de Paríba, 20 de enero de 2017. Directores: Virginia Bentes; Rosa Estopà. Qualificació: excel·lent, cum Laude.

Llopart-Saumell, Elisabet. La funció dels neologismes: revisió de la dicotomia neologisme denominatiu i neologisme estilístic. Barcelona, Institut de Lingüística Aplicada, Universitat Pompeu Fabra. Data de defensa: 04-11-2016. Menció Europea. Direcció: Judit Freixa Aymerich

Villena Araya, Belén. Innovación léxica en mapudungún: genuinidad, productividad y planificación. Barcelona, Institut de Lingüística Aplicada, Universitat Pompeu Fabra. Data de defensa: 24-10-2016. Menció Europea. Direcció: M. Teresa Cabré Castellví

Cañete González, Paola Isabel. Innovación léxica y género en textos periodísticos del español actual. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, febrer de 2016. Menció Europea. Direcció: Judit Freixa Aymerich

Necsulescu, Silvia. Automatic acquisition of lexical-semantic relations. Gathering information in a dense representation. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, febrer de 2016. Menció Europea. Direcció: Núria Bel Rafecas

Vázquez Suárez, Silvia. Pattern-Based Automatic Induction of Domain Adapted Resources for Social Media Analysis. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, gener de 2016. Direcció: Núria Bel Rafecas

Romeo, Lauren Michele. The structure of the lexicon in the task of the automatic acquisition of lexical information. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, octubre de 2015. Menció Europea. Direcció: Núria Bel Rafecas

Lázaro Hernández, Jorge Adrián. El ejemplo en terminología. Caracterización y extracción automática. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, octubre de 2015. Menció Europea. Direcció: M. Teresa Cabré Castellví, Gerardo Sierra Martínez (UNAM)

Costa, Lucimara Alves da Conceiçao. Reflexões sobre a variação terminológica na lexicografia corrente no brasil e a construção das bases teórico-metodológicas para o dicionário de lexicografia brasileira. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, setembre de 2015. Direcció: M. Teresa Cabré Castellví, Claudia Zavaglia

Morales Moreno, Albert. Estudi lexicomètric del vocabulari del procés d'aprovació de l'Estatut d'autonomia de Catalunya (2006). Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, setembre de 2015. Direcció: M. Teresa Cabré Castellví, Carmen Pineira

Benmakhlouf, Hajar. La comprensión lectora en los manuales de ELE en el sistema educativo marroquí: análisis y propuestas. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, juliol de 2015. Direcció: M. Teresa Cabré Castellví, Hassan Boutakka

Zarnikhi, Abolfazl. Towards a Model for Terminology Planning. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, juliol de 2014. Direcció: M. Teresa Cabré Castellví. Menció Europea.

Burgos Herrera, Diego Alberto. Towards an Image-Term Co-occurrence Model for Multilingual Terminology Alignment and Cross-Language Image Indexing. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, maig de 2014. Direcció: Leo Wanner

Tesis depositadas

Fomicheva, Marina.The role of human reference translation in machine translation evaluation. Direcció: Dra. Núria Bel, Universitat Pompeu Fabra; Dra. Iria da Cunha, UNED. Defensa prevista: 18/7/17

Fathi, Besharat. Terminology planning evaluation: the case of persian languageDirecció: M. Tertesa . Cabré, Universitat Pompeu Fabra. Defensa prevista: setembre 2017

Tesis en curso de los participantes en el proyecto

Les col·locacions en llengua catalana. Doctoranda: M. Àngels Martínez Salom (Illes Balears). Direcció: Mercè Lorente. Pla de tesi defensat. 5è any (a temps parcial)

L'enventivitat en medicina. Doctorand: Òscar Pozuelo (Barcelona). Directores: Rosa Estopà; Mercè Lorente. IULA-UPF. 3r any

La enseñanza del chino con fines específicos basada en el anàlisis de necesidades. Doctoranda: Yingfeng Xu (Xina). Direcció: Mercè Lorente. 3r any

La metàfora en el discurso del genoma humano. Doctoranda: Míriam Suarez (Cuba). Direcció: Mercè Lorente. 2n any

Dinámica de la lengua en las constituciones españolas. Doctoranda: Ségolène Demol. Cotutela: UPF-Université d'Artois. Direcció: Mercè Lorente; Carmen Pineira-Tresmontant. 2n any

La extracción automática de neología semántica. Doctorand: Andrés Torres (Mèxic). Directores: Rosa Estopà; Juan Manuel Torres. Cotutela. UPF-Universitat d'Avignó. 2n any

Étude des phénomènes de création néologiques dans le domaine médical des maladies rares entre 2007 et 2015. Doctoranda: Coralie Schneider (França). Cotutela: UPF-Université Paris Diderot. Directores: Rosa Estopà; Natalie Kubler. 2n any

La inteligibilitat del discurs sanitari: el cas d'una comunitat virtual de pacients de càncer de mama. Doctoranda: Laia Vidal (Lleida). Tesi UPF-ICO. Directores: Rosa Estopà (UPF); Cristian Ochoa (UB). 2n any

La transmisión de conocimiento especializado en biomedicina y genómica: herramientas para la adecuación a usuarios no expertos. Doctoranda: Alejandra López Fuentes (Mèxic). Codirecció: Rosa Estopà (UPF); Julio Collado (UNAM). Pla de tesi defensat. 1r any

Enseñanza superior de la terminología. Doctorand: Jorge Mario Porras (Colombia). IULA-UPF. Direcció: Rosa Estopà. Pla de tesi defensat. 1r any

Les verbes en Tamazight: structure argumentale et sémantique. Doctoranda: Nadia Belkacem (Algèria). Direcció: Mercè Lorente. Pla de tesi defensat. 1r any

La terminología de Geografía en zapoteco: una nueva perspectiva para las escuelas bilingües de Oaxaca, México La terminología zapoteca del siglo XXI. Doctorand: Òscar Méndez. 1r any

MÁSTER

Los miembros del equipo de investigación vinculados a la UPF participan en 8 másters (de la UPF y 6 externos).

M. Teresa Cabré dirige y Rosa Estopà coordina el Màster online de Terminologia, título propio de la UPF que tiene una antigüedad de 10 años. Dentro de este máster de 2014 a 2017 se han defendido un total de 30 TFM, dirigidos por miembros del equipo de investigación.

En el Máster Oficial de Lingüística Teórica y Aplicada de la UPF los miembros del equipo de investigación han dirigido un total de 9 TFM de 2014 a 2017.

C.4. IMPLICACIONES ÉTICAS Y/O DE BIOSEGURIDAD

Los informes clínicos utilizados en la constitución del subcorpus son cedidos por el Hospital Italiano de Buenos Aires ya completamente anonimizados. Para más seguridad, y teniendo en cuenta de que lo que nos interesa es realizar un análisis del léxico en el corpus, no se introducen metadatos de identificación de los documentos de origen y se realiza una mezcla de las oraciones contenidas en todos los documentos, de manera que no se pueda reconstruir el discurso continuado.

Estas operaciones están reguladas en el convenio de cooperación entre el Hospital Italiano de Buenos Aires y la Universidad Pompeu Fabra, y controladas por sus respectivos comités éticos.