ELUA
ISSN 2171-6692
Núm. 38, 2022, págs. 281-300
https://doi.org/10.14198/ELUA.22357

Extracción automática de unidades terminológicas guiada por Framenet: una aplicación al corpus electrónico CORD-19

Automatic framenet-guided extraction of terminology: an application to the cord-19 electronic corpus

Mario Crespo Miguel
Universidad de Cádiz, España
mario.crespo@uca.es
https://orcid.org/0000-0002-8570-8678

Resumen
La evolución de la terminología va unida a las nuevas tecnologías y al desarrollo de plataformas de trabajo o interfaces, que permitan crear un glosario técnico de forma semiautomática o incluso automática. Las unidades terminológicas y sus propiedades están conectadas a la expresión del conocimiento de los campos especializados en los que aparecen, por lo que estos enfoques automáticos no solo se enfrentan a la tarea de determinar cuáles son las unidades terminológicas de un campo determinado, sino a expresar cómo se estructura dicha información en esa disciplina. Muchas de las unidades terminológicas que se pueden encontrar en un ámbito científico determinado aparecen en otros campos e, incluso, en el uso general. Los términos suelen estar presentes en el acervo léxico de las lenguas y comparten con las unidades léxicas un complejo conjunto de relaciones. La semántica de marcos es un modelo particularmente

Abstract
The evolution of Terminology is joined to new technologies and the development of work platforms or interfaces that allow creating a technical glossary semi-automatically or even automatically. Terms and their properties are connected formally to the expression of knowledge of specialized fields in which they occur, so automatic approaches are not only faced with the task of determining which are the terminological units of a given field, but to express how such information is structured in their technical field. Most of the terms occurring in a scientific domain are also found in other disciplines and even in everyday language. Terms are often present on the lexical stock of languages and share with lexical units a complex set of relationships. Frame Semantics is a particularly attractive model for the terminological work, interested in accounting for the connection between the conceptual structure of a specialized field and the

Para citar este artículo: Crespo Miguel, M. (2022). Extracción automática de unidades terminológicas guiada por Framenet: una aplicación al corpus electrónico CORD-19. *ELUA*, (38), 281-300. https://doi.org/10.14198/ELUA.22357

Recibido: 29/03/2022 Aceptado: 14/05/2022

atractivo para el trabajo terminológico, interesado en dar cuenta de la conexión entre la estructura conceptual de un campo de conocimiento especializado y los elementos utilizados para transmitir este conocimiento. Esto ha llevado a muchos investigadores a utilizar FrameNet como forma de representar la terminología. FrameNet es un recurso en línea para el inglés basado en la semántica de marcos y respaldado por pruebas de corpus. Un marco se fundamenta en el hecho de que ciertas palabras evocan determinadas situaciones en las que tienen lugar determinados participantes. Estas situaciones o marcos son estructuras estereotipadas que representan áreas de experiencia y conocimiento sociocultural. Presentamos un enfoque estadístico basado en corpus que es capaz de seleccionar los marcos de FrameNet que mejor representan un conjunto de textos electrónicos sobre COVID-19 e indicar cuáles de sus unidades léxicas funcionan como unidades terminológicas de ese corpus. Los resultados muestran cómo esta metodología puede ser un buen apoyo al trabajo terminográfico, ya que no solo permite la extracción de unidades terminológicas, sino el uso del esqueleto formal de FrameNet como medio para estructurar este conocimiento.

PALABRAS CLAVE: Terminología, FrameNet, Lenguaje especializado, Extracción de términos, Análisis de corpus, Frecuencia de aparición del término - Frecuencia inversa del documento (*tf-idf)*

elements used to transmit this knowledge. This has led to many researchers to use FrameNet as a way of representing terminology. FrameNet is an online resource for English based on Frame Semantics and supported by corpus evidence. A frame is founded on the basis that certain words evoke certain situations in which particular participants take place. These situations or frames are stereotyped structures representing areas of sociocultural experience and knowledge. We present a statistical approach based on corpus able to select most representative FrameNet frames that best represent a set of electronic texts on COVID-19 and show which of their lexical units work as terminological units. Results confirm that this methodology can be a good support for terminographic work, since it not only allows the extraction of terminological units, but also the use of the FrameNet framework to structure this knowledge.

KEYWORDS: Terminology, FrameNet, Specialised language, Term extraction, Corpus analysis, Term frequency – Inverse document frequency (*tf-idf)*.

1. INTRODUCCIÓN

La terminología es una rama multidisciplinar de la lingüística aplicada que analiza cómo se estructura el conocimiento en los diferentes dominios especializados y cuáles son las características de las unidades usadas para expresar ese conocimiento (L'Homme 2018). Existe un esfuerzo continuo por aumentar el número de recursos terminológicos que recojan las unidades terminológicas del ámbito científico, analicen su interpretación y permitan su inclusión en glosarios o léxicos especializados para la comunidad científica (Ramírez Salado 2019). Sin embargo, la construcción de este tipo de recursos es bastante difícil de realizar. En primer lugar, la continua investigación hace que la lista de nuevos términos esté en constante crecimiento. Además, las unidades terminológicas se relacionan entre sí según diferentes relaciones semánticas y conceptuales que se deben determinar (Casas Gómez 2014). Es por ello que esta disciplina se está ligando a las nuevas tecnologías y al desarrollo de plataformas de trabajo o interfaces que permitan crear un glosario técnico

de forma semiautomática o incluso automática como parte de la tendencia general en Lingüística (Crespo 2020b). Los términos y sus propiedades están conectados formalmente a la expresión del conocimiento de los campos especializados en los que aparecen, por lo que los enfoques automáticos no solo se enfrentan a la tarea de determinar cuáles son las unidades terminológicas de un campo determinado, sino indicar cómo se estructura dicha información en el conocimiento general de esa disciplina.

La semántica de marcos (Fillmore 1977) es un modelo particularmente atractivo para el trabajo terminológico, interesado en dar cuenta de la conexión que existe entre la estructura conceptual de los campos de conocimiento especializados y las unidades lingüísticas utilizadas para transmitir este conocimiento (L'Homme, Subirats Rüggeberg y Robichaud 2016; Azoulay 2017). Esta corriente se basa en la suposición de que los significados de las unidades léxicas se construyen sobre los conocimientos previos de los individuos (experiencia, creencias, convenciones, etc.). Esta conexión se establece mediante marcos semánticos: estructuras estereotipadas que representan áreas de experiencia sociocultural que describen un tipo de evento o situación, y los participantes que aparecen en él. Se trata de un modelo que hace explícita la relación entre las unidades léxicas y este conocimiento. Los hablantes entienden los significados de las palabras en virtud de los marcos que evocan (Potęga 2017), sin importar que pertenezca a dominios generales o especializados. Dentro de esta forma de entender las relaciones léxicas, FrameNet (Baker *et al*. 1998; Ruppenhofer *et al*. 2006) es un recurso online basado en la semántica de marcos que pretende agrupar el vocabulario del inglés en diferentes marcos y establecer sus relaciones conceptuales. Se trata de un proyecto en desarrollo que, hasta ahora, ha definido 1.224 marcos para 13.640 unidades léxicas diferentes del inglés.

El trabajo que aquí presentamos introduce un enfoque automático para reutilizar el modelo e información codificada en FrameNet con la finalidad de estructurar el conocimiento y las unidades terminológicas que aparecen en textos de especialidad. Este trabajo parte del corpus en línea COVID-19 Open Research Dataset (CORD-19), de 100 millones de palabras para uso científico. En primer lugar, se determinan cuáles de los 1.224 marcos codificados actualmente en FrameNet son los más representativos de este ámbito médico. Para ello, seguimos un método estadístico que selecciona los marcos de FrameNet más relevantes a partir de la comparación de la frecuencia de las unidades codificadas en este proyecto tanto en el corpus online COVID-19 Open Research Dataset como en uno de temática general. De la lista de marcos resultante, elegimos las unidades léxicas más relevantes de acuerdo con un análisis *tf-idf* (del inglés *Term frequency – Inverse document frequency*) usando el mismo corpus especializado. Los resultados muestran cómo FrameNet puede utilizarse para estructurar el conocimiento de un dominio especializado, determinar los marcos más representativos del mismo, establecer relaciones entre ellos y analizar cuáles de las unidades léxicas de estos marcos funcionan como unidades terminológicas de este corpus técnico. Por último, observamos que los resultados se ven afectados por la falta de cobertura de unidades léxicas de los marcos de FrameNet, ya que este se encuentra en proceso de creación. Para resolver este problema, ampliamos el número de palabras mediante su conexión de esta base de datos con WordNet (Miller et al. 1993), una base de datos léxica que organiza el vocabulario inglés en grupos según conceptos y relaciones semánticas. Tras la conexión de ambos recursos, los resultados generales mejoran.

2. TERMINOLOGÍA BASADA EN LA SEMÁNTICA DE MARCOS

La semántica de marcos (Fillmore 1985; Fillmore y Baker 2010) es un desarrollo dentro de la lingüística cognitiva que ha llamado la atención de muchos investigadores interesados en dar cuenta de las asociaciones entre el léxico y el conocimiento que se supone que comparten los hablantes de una lengua (L'Homme, Subirats Rüggeberg y Robichaud 2016). Se parte de la base de que los significados de las unidades léxicas se construyen en relación con el conocimiento representado en los marcos semánticos de los que estas forman parte. Los marcos semánticos son estructuras que representan situaciones específicas (por ejemplo, una situación médica) (Bernier-Colborne y L'Homme 2015) y participantes implicados en ellas. Los modelos cognitivos (Lakoff 1987, cit. en Durán-Muñoz 2016) dan la posibilidad de definir las unidades léxicas de una forma sistemática y flexible según la situación comunicativa en la que se producen.

Como indicábamos en el apartado anterior, FrameNet es un recurso en línea para el inglés basado en la semántica de marcos con el objetivo de identificar y definir todas las posibles situaciones evocadas por las unidades léxicas de una lengua, identificar los participantes de tales situaciones (lo que en la práctica supone la determinación de los roles semánticos evocados) y dar cuenta de las relaciones sintáctico-semánticas de todos ellos mediante pruebas de corpus (Ruppenhofer *et al*. 2006). Para ejemplificar cómo debe concebirse cada una de estas estructuras, podemos partir del marco 'cure'. En él se describe una situación estereotipada en la que un *Healer* ('sanador') trata y cura una *Affliction* ('afección') de un *Patient* ('paciente'). Otros participantes o roles semánticos centrales de este marco serían: *Body_part* ('parte del cuerpo'), *Medication* ('medicación') y *Treatment* ('tratamiento'). De esta manera, unidades léxicas como *alleviation.n, curable.a*, *curative.a* o *cure.v* evocarían este conocimiento al aparecer en el discurso (Ruppenhofer et al (2016)[1]. A partir de aquí el objetivo de FrameNet es analizar y anotar diferentes enunciados de un corpus lingüístico para mostrar las estructuras sintácticas en las que aparecen estas unidades y cómo se estructuran tales roles semánticos sobre ellas (Verdaguer 2020). De esta manera, las anotaciones de FrameNet son formalmente conjuntos de triples que representan las realizaciones de los elementos del marco para cada frase anotada, cada una de las cuales consiste en rol semántico (por ejemplo, *Affliction*), una función gramatical (e.g., objeto) y un tipo de frase (e.g., sintagma nominal (NP)):

1 https://framenet.icsi.berkeley.edu/fndrupal/

Roles semánticos	Estructuras en las que aparece y número	Ejemplo en el corpus
Affliction	2nd.-- (2) DNI.-- (4) INI.-- (1) NP.Ext (6) NP.Obj (12) PP[of].Dep (6) PPing[of].Dep (1)	On the way Jesus CURED the woman with the haemorrhage . with the haemorrhage If the procedure succeeds , Carly will be able to lead a normal life , offering hope to thousands of sufferers with diseases like cystic fibrosis and muscular dystrophy who could be CURED by similar techniques . I could never CURE him .DNI Rosemary apparently has to cope with her husband 's anxiety that she could only be CURED by becoming more ` assertive " .DNI I 'm very positive that Natalie will be CURED .CNIDNI Isis then CURED Re by reciting a spell and with her new power became one of the mightiest goddesses .DNI You can not CURE yourself .INI I do n't want to suggest that social problems can be CURED by the application of wealth .INI [It was almost as if she was suffering from some dreadful disease that could only be CURED by his physical removal .DNI By a visit to a doctor , a psychiatrist , or even through a programme of counselling and guidance , mental illnesses can be CURED , and almost always are .INI It occurred mainly in communities with poor diets and was eventually identified as a deficiency disease which could be promptly and completely CURED by supplying adequate food .INI All forms of early syphilis can be completely CURED , and some of the later manifestations can be markedly improved , or , at least , the progression of the disease can be arrested .CNIINI It could point to plenty of ailments that the Spanish economic rejuvenation so far has failed to CURE .INI The others are dozing while I chatter , attempting to CURE their somnambulism with my words .
Degree	AVP.Dep (2)	
Healer	NP.Ext (13) CNI.-- (4)	
Manner	AVP.Dep (3)	
Patient	NP.Obj (9) NP.Ext (4) PP[re].Dep (1) INI.-- (12) DNI.-- (1) 2nd.-- (3) PP[in].Dep (2)	
Treatment	PP[by].Dep (4) PPing[by].Dep (4) NP.Ext (9) PP[with].Dep (2) PP[without].Dep (1)	

Figura 1. Anotación en FrameNet para la unidad 'Cure.v'. Fuente: Ruppenhofer *et al.* (2016). https://framenet.icsi.berkeley.edu/fndrupal/.

La anotación se realiza para cada uno de los sentidos que poseen las palabras analizadas. Así, una palabra polisémica debe pertenecer a marcos semánticos diferentes ya que cada uno de sus sentidos evoca un marco o estructura conceptual particular. La Figura 1 ilustra los diferentes significados asignados a la unidad léxica *'accept.v'*, cada uno de los cuales lleva parejo unos roles y una estructura sintáctica diferente:

Marcos asociados a la unidad léxica *'accept.v'*	
Respond_to_proposal	*Receiving*
A Speaker addresses a Proposal made by an Interlocutor, either agreeing to it or rejecting it.	A Recipient comes into possession of the Theme as a result of the joint action of the Donor and the Recipient.
accept.v, acceptance.n, rebuff.n, rebuff.v, refuse.v, reject.v, rejection.n, turn down.v	accept.v, receipt.n, receive.v

Figura 2. Unidad léxicas 'accept.n' en dos marcos diferentes de FrameNet.

Actualmente FrameNet está en proceso de creación y hasta ahora se han completado 202.232 anotaciones. La versión utilizada en este trabajo es la R1.7. El estado actual de FrameNet puede verse en la Tabla 1. Esta versión contiene 1.224 marcos que cubren 13.640 unidades léxicas diferentes y 10.542 roles semánticos:

	FrameNet versión: R1.7
Marcos	1.224
Unidades léxicas (‘Triggers’)	13.640
Promedio de unidades léxicas por marco	12,5
Roles semánticos	10.542
Promedio de roles semánticos por marco	9,7

Tabla 1. Estado de FrameNet en su version R1.7.

Como señalan Johnson y Fillmore (2000), las situaciones representadas en FrameNet se pueden agrupar en torno a dominios del conocimiento y la experiencia humana, tales como “Body, Chance, Cognition, Communication, Emotion, Health, Life Stages, Motion, Perception, Society, Space, Time, Transaction, or a General domain”. Además, los marcos se conectan entre sí según diferentes relaciones conceptuales e interrelacionan el léxico de múltiples maneras. La Figura 3 (Gildea y Jurafsky 2002) muestra los marcos vinculados al dominio de la cognición (‘cognition’):

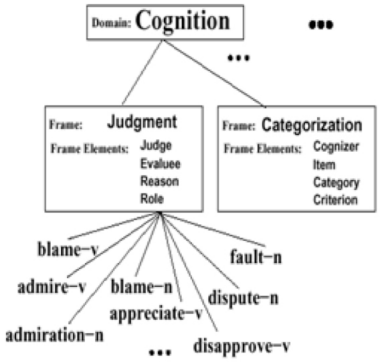

Figura 3. Dominio de la cognición. Fuente: FrameNet.

Como ya indicábamos en el apartado anterior, la ligazón que existe en FrameNet entre estructura conceptual y unidades léxicas ha llevado a muchos investigadores a utilizar FrameNet como forma de representar la terminología. A este respecto, Malm *et al.* (2018) indican que la mayoría de los términos que se encuentran en un dominio científico también se encuentran en otros campos (por ejemplo, tense.n en Lingüística). Aparte de su terminología específica, un dominio científico hace uso de las mismas unidades léxicas y patrones sintácticos que la lengua general. Según Verdaguer (2020), el lenguaje del dominio de la biomedicina no es muy diferente del lenguaje ordinario. Asimismo, Guerrero Ramos y Pérez Lagos (2003), y Cabré (2005) señalan que muchas unidades terminológicas que se

encuentran en el léxico del hablante general pueden llegar a adquirir valores terminológicos en determinadas situaciones. Muchas unidades terminológicas están presentes en el acervo léxico de las lenguas y comparten con las unidades léxicas un complejo conjunto de relaciones (L'Homme, Robichaud y Subirats Rüggeberg 2020). Entre los estudios que han aplicado FrameNet a registros especializados se encuentran Dolbey et al. (2006) para textos biomédicos, Venturi (2013) para el lenguaje jurídico, Malm et al. (2018) para la lingüística o L'Homme et al. (2020) para la ecología.

Estos proyectos suelen utilizar una metodología *bottom-up* que parte del análisis de las unidades terminológicas que aparecen en los corpus científicos para asignarles el marco que mejor representa su significado en ese contexto. La gran desventaja de esta forma de trabajar es que analiza secuencialmente las unidades que van apareciendo, por lo que no se discrimina de ninguna manera entre marcos que realmente sean específicos y representen ese campo del conocimiento. Por el contrario, una metodología *top-down* partiría del proceso contrario: define las unidades de conocimiento (o marcos) y sus participantes (roles semánticos), y a partir de ahí enumera qué elementos son los que se asocian a esos marcos conceptuales. Sin embargo, esta última metodología se enfrenta al reto de identificar por adelantado la gama de marcos que mejor representa un determinado conjunto de textos científicos. Entre las posibles soluciones a este problema se encuentra el recurrir a la opinión del experto (que deberá familiarizarse previamente con la semántica de marcos), o recurrir a un análisis macroscópico de las unidades del corpus, que permita identificar patrones y los contenidos que se desarrollan en un corpus técnico de manera general.

Este trabajo se sitúa en esta segunda opción y presenta un método estadístico para hacer esa selección de una manera fiable a partir del análisis automático del corpus. La siguiente sección presenta la metodología necesaria para determinar los marcos más significativos de FrameNet para un corpus determinado. Para ello se compara cómo se comportan las unidades léxicas de cada marco desde el punto de vista de su frecuencia en un corpus específico respecto a uno general. En este estudio se ha trabajado con COVID-19 Open Research Dataset (CORD-19) de 100 millones de palabras y el Corpus of Contemporary American English (COCA) de 560 millones de palabras. Una vez realizada la selección de marcos, vinculamos sus unidades léxicas con los términos *tf-idf* más representativos del corpus CORD-19. Los resultados se presentan al final de este artículo.

3. METODOLOGÍA

1.1. Selección automática de marcos de FrameNet

Yarowsky (1995) afirma que el tema global de un determinado documento o corpus se refleja en el vocabulario que se puede encontrar en el mismo (por ejemplo, en un documento médico no es raro encontrar el término 'cardiología' o 'auscultación' frente a otros tipos de texto donde tales unidades no aparecerán). Esto no solo tiene implicación respecto a la aparición o no de determinadas unidades en el texto en función de la temática de los mismos, sino que existen diferencias significativas respecto a las frecuencias observadas del léxico que comparten todo estos textos según el corpus donde aparezcan. De esta manera, habrá elementos que tengan una mayor frecuencia en los textos científicos que en los generales (por ejemplo, es el caso de términos como "pulmón" o "cura" en un ámbito médico). En la

Tabla 2 se comparan la frecuencias por millón de palabras de algunas unidades de los marcos FrameNet *Medical_conditions* y *Medical_professional* en un corpus médico y otro general según los datos proporcionados por el corpus COCA, del que hablaremos más adelante.

	Médico	General
cancer.n	407	111
asthma.n	165	8
doctor.n	162	19
surgeon.n	198	24

Tabla 2. Frecuencias por millón de palabras en un corpus médico y otro general.

Como se puede observar, estas unidades léxicas 'cancer.n', 'asthma.n', 'doctor.n' y 'surgeon.n' tienen una mayor aparición en el dominio médico que en el general. Esto indica que la distribución de frecuencias de las unidades del corpus podría ser un indicador de la pertenencia a un determinado campo temático. Siguiendo esta premisa, es posible analizar la frecuencia de las unidades léxicas de cada marco de FrameNet en un corpus científico y compararlas respecto a cómo aparecen en un corpus general. La idea es determinar si existen diferencias estadísticas entre ambas. Las unidades de los marcos más relacionados con la temática del corpus estarán sobrerrepresentadas en su frecuencia respecto al otro corpus.

Como ya se ha indicado en la Tabla 1, el estado actual de FrameNet clasifica 13.640 unidades léxicas en 1.224 marcos diferentes. Evidentemente las unidades del inglés exceden las dimensiones actuales de FrameNet. El desarrollo de recursos léxicos requiere de un gran esfuerzo, lo que justifica que sea un proceso lento. Es por ello por lo que poder reutilizar o transferir la información de otros repositorios léxicos a FrameNet se considera como una de las posibles soluciones al problema (Cristea y Pistol 2012). Entre las diferentes posibilidades se ha planteado su alineación con WordNet, de tal manera que la información contenida en este recurso pueda integrarse en otras bases de datos léxicas (Crespo 2021).

WordNet (Miller et al. 1993) es una base de datos léxica que organiza el vocabulario del inglés a partir de relaciones conceptuales y semánticas. Estos grupos, llamados synsets, reflejan la disponibilidad del léxico inglés para reflejar una idea o concepto. Los synsets son conjuntos de "sinónimos cognitivos". WordNet ha sido desarrollado en el Cognitive Laboratory de la Universidad de Princeton bajo la dirección de George A. Miller y contiene unas 150.000 unidades distribuidas entre las diferentes categorías verbales de sustantivos, verbos, adjetivos y adverbios. En la siguiente figura se aprecia los synsets asociados a la unidad 'pain.n' en inglés y las unidades asociadas a cada uno de sus sentidos:

Noun: "pain.n"
S: (n) **pain**, **hurting** (a symptom of some physical hurt or disorder) "the patient developed severe pain and distension"
S: (n) **pain**, **painfulness** (emotional distress; a fundamental feeling that people try to avoid) "the pain of loneliness"
S: (n) **pain**, **pain sensation**, **painful sensation** (a somatic sensation of acute discomfort) "as the intensity increased the sensation changed from tickle to pain"
S: (n) **pain**, **pain in the neck**, **nuisance** (a bothersome annoying person) "that kid is a terrible pain"
S: (n) **annoyance**, **bother**, **botheration**, **pain**, **infliction**, **pain in the neck**, **pain in the ass** (something or someone that causes trouble; a source of unhappiness) "washing dishes was a nuisance before we got a dish washer"; "a bit of a bother"; "he's not a friend, he's an infliction"

Figura 4. Synsets asociados a la unidad léxica "pain.n".[2]

A partir del repositorio de WordNet, Crespo (2021) amplía la cobertura de FrameNet conectando los marcos de FrameNet con los synsets que los representan en este recurso. Este enfoque alcanza una precisión del 88% y amplía la cobertura de FrameNet de 13.631 unidades a 25.305. Sobre este segundo número de unidades aplicamos un procedimiento estadístico basado en un test de hipótesis para la determinación de los marcos más representativos.

Como es ampliamente sabido, una prueba de hipótesis se utiliza para comparar dos conjuntos de datos y poder afirmar si una propiedad que es constante para una determinada población bajo análisis, es compatible con otra muestra de la misma población. Si no es compatible, se rechaza la hipótesis nula y se considera que las dos muestras proceden de poblaciones diferentes. Este trabajo sigue el procedimiento metodológico propuesto por el Crespo (2020a), que utiliza la prueba de rangos con signo de Wilcoxon para comparar si existen diferencias estadísticamente significativas entre las frecuencias observadas en uno u otro corpus. La prueba de rangos con signo de Wilcoxon se trata de una prueba de hipótesis estadística no paramétrica que compara dos muestras relacionadas, muestras emparejadas o mediciones repetidas sobre una única muestra para evaluar si la población de diferencias tiene una mediana de cero, por lo que las hipótesis nula y alternativa son las siguientes (Triola, 2007):

H0: Los pares emparejados tienen diferencias que provienen de una población con una mediana igual a cero.

H1: Los pares emparejados tienen diferencias que provienen de una población con una mediana distinta de cero.

Si $n \leq 30$, el estadístico de prueba es T.

Si $n > 30$, el estadístico de prueba es

$$Si\ n > 30, el\ test\ estadístico\ es\ z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

2 Ejemplo extraído de https://wordnet.princeton.edu.

Esta es la fórmula de la prueba de rangos con signo de Wilcoxon, donde T es la menor de las dos sumas siguientes:

1. El valor absoluto de la suma de los rangos negativos de las diferencias (d) que no sean cero. U
2. La suma de los rangos positivos de las diferencias (d) que no sean cero.

Sobre los valores de T o z se comprueban sus valores críticos. En nuestros experimentos, utilizamos las frecuencias relativas de las unidades y adoptamos una hipótesis de prueba de cola a la derecha con una significación del 95% y del 99%. Esto significa que la región crítica a considerar estará en la región extrema derecha (cola) bajo la curva. Esto es debido a que solo tomamos aquellas diferencias en las que las frecuencias observadas en el corpus CORD-19 sea estadísticamente superior a las observadas en el ámbito general. Una prueba de cola derecha se realiza cuando la hipótesis alternativa (H1) contiene la condición H1 > x (mayor que una cantidad determinada).

3.2. El corpus CORD-19

El COVID-19 Open Research Dataset (CORD-19)[3] es un repositorio en constante actualización con artículos científicos y preprints sobre el COVID-19, y toda la investigación relacionada con otros coronavirus como el SARS y el MERS (Wang et al. 2020). Este corpus ha sido diseñado para facilitar todo tipo de estudios entre los que se encuentra el desarrollo de sistemas capaces de extraer información relevante de los mismos como la minería de textos y la recuperación de información. Los artículos de CORD-19 proceden de PubMed Central (PMC), PubMed, la base de datos COVID-19 de la Organización Mundial de la Salud y los servidores de bioRxiv, medRxiv y arXiv.

De este corpus se descargaron aleatoriamente 40.000 artículos que contenían 102.716.238 palabras. Este corpus fue tokenizado, etiquetado con partes de la oración y lematizado con Treetagger. El resultado de este etiquetador contiene palabras asignadas a partes de la oración (por ejemplo, verbos, nombres, adjetivos) y lematizadas a sus formas canónicas (por ejemplo, formas adjetivas singulares y masculinas, formas verbales de infinitivo). Las etiquetas de TreeTagger se simplificaron a las observadas en FrameNet, de modo que compartiesen el mismo tipo de anotaciones. La frecuencia de los lemas se obtuvo a partir de este corpus.

Para disponer de una lista fiable de ocurrencias de palabras que representaran el ámbito de general, se optó por utilizar el corpus COCA (Davies, 2019). El Corpus of Contemporary American English (COCA) es uno de los mayores corpus de inglés de libre acceso. Contiene más de 560 millones de palabras (20 millones de palabras por cada año del periodo 1990-2017) y está dividido por igual entre textos hablados, de ficción, revistas populares, periódicos y textos académicos. Las unidades han sido lematizadas y etiquetadas con su clase de palabra (POS-tagged). El corpus COCA contiene unos 60.000 lemas. La Tabla 3 muestra el aspecto de este corpus:

3 https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge.

rank	lemma/word	POS	disp	totFreq
25083	piglet	n	0,88	239
25088	woodsman	n	0,70	300
25090	candied	j	0,87	242
25093	metacognitive	j	0,69	306
25107	industry-wide	j	0,89	236
25108	health-food	j	0,85	246
25110	posterior	n	0,88	240

Tabla 3. Ejemplo de unidades en el corpus COCA. Fuente: COCA corpus.

3.3. Selección de las unidades léxicas de FrameNet como posibles unidades terminológicas

Como afirma Witschel (2005), los enfoques automáticos para la extracción de terminología se basan en las frecuencias de las unidades del corpus. Las palabras que se repiten con más frecuencia en un documento son buenos descriptores de su contenido, siempre y cuando estos elementos no aparezcan en muchos otros documentos (como "el", "sobre" o "cree"). La medida *tf-idf* se basa en este supuesto. Se trata de uno de los enfoques más comunes para la extracción automática de unidades terminológicas o palabras clave a partir del análisis de corpus (Lossio-Ventura, Jonquet, Roche y Teisseire, 2014). En lugar de representar a las diferentes unidades del corpus por su frecuencia absoluta o relativa, con *tf-idf* cada elemento se pondera dividiendo su frecuencia de aparición por el número de documentos del corpus que contienen tal unidad. Las unidades léxicas o lemas con una alta clasificación *tf-idf* se calculan de la siguiente manera:

$$W_{x,y} = tf_{x,y} * log \frac{N}{df_x}$$

$tf_{x,y} = frecuencia\ de\ x\ en\ y$

$df_x = número\ de\ documentos\ o\ partes\ que\ contienen\ \ x$

$N = número\ total\ de\ documentos\ o\ partes\ del\ corpus$

El resultado de esta operación es un valor que indica el grado de relevancia de la unidad analizada en ese corpus. A partir del CORD-19 lematizado y etiquetado con POS, se extrajo la lista de términos *tf-idf* ordenados según su puntuación:

Unidad	Valor
cell.n	0,037
virus.n	0,027
use.v	0,025
protein.n	0,021
infection.n	0,019
study.n	0,017

Unidad	Valor
figure.n	0,015
show.v	0,013
patient.n	0,012
gene.n	0,011

Tabla 4. Primeras 10 unidades en el ranking de *tf-idf*.

Una vez que se obtuvo esta lista se cruzó con las unidades léxicas presentes en FrameNet. Partimos de la hipótesis de que la selección de unidades terminológicas que realiza este último algoritmo debería coincidir con los principales elementos que encontramos en los marcos más representativos del CORD-19. La sección de resultados presenta los resultados de todo el proceso metodológico: selección automática de marcos, extracción de las unidades léxicas de cada uno de ellos y cruce de tales unidades con las procedentes de la selección de la herramienta *tf-idf*. Los resultados muestran hasta qué punto la selección de marcos y la lista *tf-idf* coinciden, así como el nivel de precisión al asignar estos términos a un determinado marco.

4. RESULTADOS

4.1. Selección de marcos del CORD-19

Para la selección de los marcos más representativos, el sistema itera a través de todos los marcos de FrameNet tomados uno por uno, y calcula la prueba de rango con signo de Wilcoxon utilizando las frecuencias relativas de sus unidades léxicas en el corpus CORD-19 y COCA. Sin embargo, en este análisis no se incluyeron todas las unidades léxicas de cada marco. Si buscamos en FrameNet, hay unidades que pertenecen a verbos frasales como ‘break up’, ‘look for’, etc., expresiones idiomáticas ‘hold (one’s) tongue’, ‘give thought’, etc. o compuestos ‘family room’, ‘screw-up’, etc. Como ya se indicó anteriormente, el análisis de frecuencias se ha realizado a partir del corpus COCA y este tipo de combinaciones no se incluyen; sólo se recogen elementos léxicos simples. Por eso se decidió descartar estas unidades para calcular la significación estadística del grupo, ya que, en caso contrario, se les asignaría la frecuencia “cero”. Este mismo procedimiento se aplicó en el caso de que la unidad léxica de un marco determinado no tuviera representación ni en el corpus general ni en el médico. Igualmente fueron descartados los marcos con un solo elemento (no hay mediana posible). De los 1.222 marcos de FrameNet, 149 no tienen ninguna unidad y 53 sólo tenían un elemento. Por último, realizamos la prueba de hipótesis con signo de Wilcoxon para comprobar si la media de la muestra procedente de CORD-19 es superior a la obtenida con los mismos elementos en el corpus general. Los resultados a un nivel de significación del 95% y del 99% son los siguientes:

95%	99%
• Biological mechanisms • Building subparts • Condition symptom relation • Containers • Control • Dispersal • Distinctiveness • Diversity • Extreme value • Health response • Identicality • Measure mass • Measure volume • Medical conditions • Medical specialties • Ordinal numbers • Origin • Progression • Quantity • Similarity • Undergo transformation	• Condition symptom relation • Dispersal • Diversity • Health response • Medical conditions • Medical specialties • Ordinal numbers • Origin • Similarity

Figura 5. Resultados de la selección de marcos al 95% y 99% de nivel de significatividad.

El número de marcos seleccionados difiere entre el nivel de significación del 95% y el del 99%. Este último es más restrictivo. Pasa de una lista de 21 elementos a sólo 9 situaciones de FrameNet. Como se puede observar, en general los resultados son coherentes respecto a la temática del CORD-19.

4.2. Integración de elementos de Tf-idf y FrameNet

Este análisis parte exclusivamente de los verbos y sustantivos del corpus CORD-19. A partir de esta selección analizamos cuántas de estas unidades de FrameNet coincidían con el ranking obtenido de *tf-idf*. Como se ha señalado anteriormente, utilizamos la cobertura de FrameNet propuesta por el Crespo (2021) con 25.305 términos. La siguiente figura muestra la proporción de la lista *tf-idf* encontrada en todo FrameNet desde las primeras 100 unidades hasta las 15.000:

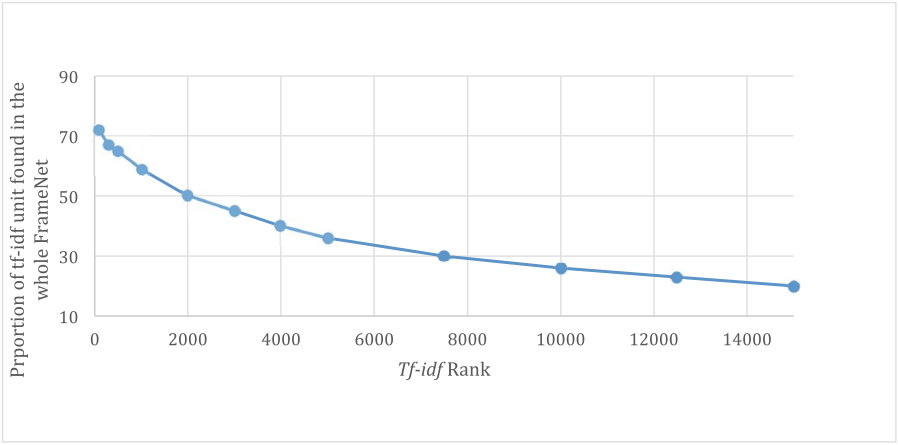

Figura 6. Proporción de coincidencia entre FrameNet y la lista *tf-idf* desde sus 100 primeros elementos hasta 15.000.

A medida que aumentamos la lista, la proporción de coincidencia entre las unidades léxicas de FrameNet y el *tf-idf* disminuye. Debemos recordar que el ámbito de aplicación de FrameNet es general, por lo que las unidades del CORD-19 coincidirán parcialmente con las de FrameNet. Para la experimentación, sólo nos centramos en los primeros 100, 300, 500 y 1000 sustantivos y verbos del ranking *tf-idf*, ya que superan el 50% de coincidencias de acuerdo con la Tabla 5. La proporción de coincidencia para tales valores son las siguientes:

100 tf-idf	300 tf-idf	500 tf-idf	1000 tf-idf
72 de 100: 72%	201 de 300: 67%	326 de 500: 65%	589 de 1000: 59%

Tabla 5. Proporción de coincidencia entre FrameNet y tf-idf para los primeros 100, 300, 500 y 1000.

Para los marcos seleccionados, enumeramos sus unidades léxicas. Si una de estas unidades coincidía con la lista *tf-idf* en alguno de sus grados (100, 300, 500 y 1000), se enumeraba. Las coincidencias fueron revisadas manualmente para determinar si los elementos asociados eran coherentes con los marcos en los que aparecían. Los resultados de la selección de marcos a un nivel de significación del 95% y del 99% son los siguientes:

Marcos al 95%			
100 tf-idf	300 tf-idf	500 tf-idf	1000 tf-idf
Número de unidades de nuestra selección de FrameNet que aparecen en la lista *tf-idf*			
27 de 72: 38%	54 de 201: 27%	68 de 326: 21%	104 de 589: 18%
Número de marcos con alguna de sus unidades en la lista tf-idf			
52% 11 de 21	67% 15 de 21	76% 16 de 21	90% 19 e 21

Marcos al 95%			
Precisión del cruce FrameNet-tf-idf			
85.1% correcto - 23 de 27 posibles-	93% correcto - 50 de 54 posibles-	93% correcto - 63 de 68 posibles-	94% correcto - 98 de 104 posibles-

Tabla 6. Resultados de correcta asociación entre FrameNet y *tf-idf* al 95%.

Marcos al 99%			
100 tf-idf	300 tf-idf	500 tf-idf	1000 tf-idf
Número de unidades de nuestra selección de FrameNet que aparecen en la lista *tf-idf*			
18 de 72 posibles: 25%	38 de 201 posibles: 19%	46 de 326 posibles: 14%	70 de 589 posibles: 12%
Número de marcos con alguna de sus unidades en la lista tf-idf			
44.4% 4 de 9	66.6% 6 de 9	77.7% 7 de 9	89% 8 de 9
Precisión del cruce FrameNet-tf-idf			
100% correcto - 18 de 18 posibles -	100% correcto - 38 de 38 posibles -	96% correcto - 44 de 46 posibles -	94% correcto - 66 de 70 posibles -

Tabla 7. Resultados de correcta asociación entre FrameNet y *tf-idf* al 99%.

Como se observa, la proporción de términos de FrameNet asignados disminuye a medida que aumenta la lista de *tf-idf*. Los mejores resultados proporcionales se encuentran al 95% y al 100 *tf-idf*. En todos los casos, los marcos con términos en *tf-idf* aumentan a medida que se incrementa esta lista. Los mejores resultados de precisión se encuentran en 100 y 300 *tf-idf* y en el 99%. En estos casos el 100% de las unidades terminológicas encontradas son correctas.

5. DISCUSIÓN

Observamos en los resultados algunos problemas señalados por L'Homme, Subirats Rüggeberg y Robichaud (2016):

- Muchos elementos registrados como términos en la lista *tf-idf* no aparecen en FrameNet. A medida que aumentamos esta lista de 100 a 1000, la proporción de términos que aparecen en FrameNet disminuye. Podemos atribuir esto a la falta de unidades léxicas en FrameNet y al hecho de que la lista *tf-idf* también se basa en un corpus que recoge la investigación en curso sobre COVID-19, por lo que contiene mucha terminología nueva.
- Por otra parte, nos encontramos con que un elemento léxico puede estar registrado en FrameNet, pero el significado contabilizado no es el requerido. Este es el caso del marco seleccionado *Building_subparts* y de las unidades léxicas 'room.n', 'study.n', 'cell.n' y 'level.n' que refieren a algo diferente a lo expresado por el marco. Se ha tomado como un error.
- Es necesario crear nuevos marcos para dar cuenta de la información del corpus. Este es el caso de la unidad 'cell.n' que solo se encuentra en el marco *Building_subparts*. Debería incluirse un nuevo marco que la represente.

Como se ha visto, la mejor coincidencia entre las unidades léxicas de FrameNet y la lista *tf-idf* es del 38%, pero en la mayoría de los casos la vinculación entre ambas ronda el 20-25%. Esto se explica porque FrameNet incluye términos generales y un corpus científico incluye unidades más específicas.

Podemos intentar paliar esta carencia añadiendo nuevos términos no considerados anteriormente. Como señala Pérez Hernández (2002), los términos científicos se relacionan entre sí mediante distintas relaciones conceptuales. Entre ellas están la inclusión ("hipernimia-hiponimia"), la equivalencia ("sinonimia"), la semiequivalencia ("parasinonimia") o la oposición conceptual ("antonimia") (Casas Gómez 2006). A continuación, podríamos ampliar la cobertura de FrameNet añadiendo la gama de relaciones conceptuales descritas en WordNet. Entre ellas se encuentran (Vossen 1998; 2002): '*hypernym*', '*hyponym*', '*holonym*', '*meronym*', '*near_synonym*', '*near_antonym*' or '*has_derived*'. Sin embargo, WordNet está estructurada jerárquicamente de elementos generales a más específicos, en una relación de tipo is-a, por lo que la hiperonimia y la hiponimia será la relación más numerosa. De estas dos interesan las segundas tal como indica Carrió Pastor (2010) sobre el lenguaje de especialidad, ya que las primeras son elementos más generales.

De esta manera ampliamos los marcos seleccionados con unidades más específicas de acuerdo con el procedimiento que se muestra en la siguiente figura:

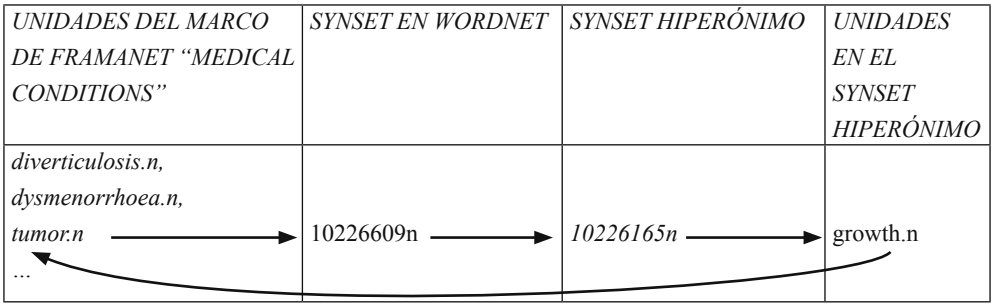

Figura 7. Método de adición de nuevos elementos desde WordNet

Utilizamos el enlace FrameNet-WordNet propuesto por el Crespo (2021) que desarrolla un algoritmo para de alinear las diferentes unidades del marco con los synsets de WordNet que mejor los representan, llegando a una precisión del 88%. Los resultados tras la adición de hipónimos son los siguientes:

Marcos al 95%			
100 tf-idf	300 tf-idf	500 tf-idf	1000 tf-idf
Número de unidades de nuestra selección de FrameNet que aparecen en la lista *tf-idf*			
49 de 72 posibles: 68%	127 de 201 posibles: 63%	185 de 326 posibles: 57%	299 de 589 posibles: 51%
Número de marcos con alguna de sus unidades en la lista tf-idf			
67% 14 de 21	86% 18 de 21	86% 18 de 21	95% 20 de 21

Marcos al 95%			
Precisión del cruce FrameNet-tf-idf			
92% correcto - 58 de 73 -	88% correct - 140 de 159-	88% correcto - 204 de 231-	85% correcto - 255 de 299-

Tabla 8. Resultados de correcta asociación entre FrameNet y *tf-idf* al 95%.

Marcos al 99%			
100 tf-idf	300 tf-idf	500 tf-idf	1000 tf-idf
Número de unidades de nuestra selección de FrameNet que aparecen en la lista *tf-idf*			
23 de 72 posibles: 31.9%	61 de 201 posibles: 30%	83 de 326 posibles: 25.4%	133 de 589 posibles: 23%
Número de marcos con alguna de sus unidades en la lista tf-idf			
67% 6 de 9	89% 8 de 9	89% 8 de 9	89% 8 de 9
Precisión del cruce FrameNet-tf-idf			
96% correcto - 26 de 27 -	93% correcto - 62 de 67-	94% correcto - 87 de 93-	90% correcto - 120 de 133-

Tabla 9. Resultados de correcta asociación entre FrameNet y *tf-idf* al 99%.

Observamos cómo los resultados mejoran al incluir hipónimos. Para la selección de marcos al 95% de significación, el número de términos de FrameNet encontrados en la lista *tf-idf* duplica los resultados anteriores y, en algunos casos, se triplica. El número de marcos con representación en la lista *tf-idf* aumentó en todos los casos. El nivel de precisión disminuye ligeramente, pero sigue siendo muy alto.

A pesar de que el nivel de significación del 99% es más restrictivo, también mejora los resultados. La coincidencia de FrameNet y la lista *tf-idf* duplica los resultados anteriores, pero sólo para los primeros 500 y 1000. En consecuencia, el número de marcos seleccionados con representación en la lista *tf-idf* también aumenta. La precisión de los términos sólo disminuye en torno al 5%. Esta reducción se explica en parte por el hecho de que el enlace FrameNet-WordNet en el Crespo (2021) presenta una precisión del 88%, por lo que deben haberse incluido algunas coincidencias erróneas.

6. CONCLUSIONES

Hemos propuesto una metodología basada en corpus para gestionar la terminología. Este enfoque permite reutilizar la información proporcionada por FrameNet para estructurar la información terminológica de un corpus técnico. Aparte de su terminología específica, un campo científico hace uso de las mismas unidades léxicas y patrones sintácticos que el lenguaje general. Este enfoque permite reutilizar el marco FrameNet para representar la terminología presentada en un corpus y determinar cómo se relacionan los términos conceptualmente. Entre sus ventajas se encuentra el hecho de tratarse de una metodología 'top-down' que define, en primer lugar, las unidades de conocimiento (o marcos), y a partir de ahí enumera qué elementos son los que se asocian a esos marcos conceptuales. Los resultados muestran su utilidad para apoyar el trabajo terminógrafico.

Para la selección de marcos representativos, este estudio parte de la base de que los marcos representativos pueden determinarse observando las frecuencias de las unidades que aparecen en nuestro corpus CORD-19 frente a las mismas palabras en un dominio general. Nuestro enfoque para la selección de marcos hace uso de la prueba de rangos con signo de Wilcoxon, un método no paramétrico, que no requiere una suposición de normalidad para la distribución de los datos. Para las primeras 1000 unidades tf-idf, el sistema es capaz de asignar 299 términos a sus marcos correctos con una precisión del 85%.

Observamos que este método ofrece buenos resultados en general, por lo que puede aplicarse para la determinación de los marcos que más representan un dominio científico determinado. Sin embargo, FrameNet está orientado a dominios generales, por lo que no cubre todos los usos científicos. Además, al estar aún está en proceso de creación, la lista de activadores de marcos aún no está completa. Intentamos resolver la escasez de cobertura léxica de FrameNet enriqueciéndola con nuevos términos de WordNet. Esta base de datos léxica organiza el vocabulario de diferentes idiomas según conceptos y relaciones semánticas. Estos grupos, llamados synsets, tratan de reflejar la disponibilidad léxica para un determinado concepto. Para la adición de términos se han utilizado las relaciones de hipónimos. Los resultados mejoran notablemente con la inclusión de estos nuevos términos, lo que indica que los lenguajes específicos hacen uso de unidades léxicas más específicas.

Hemos observado igualmente que muchos términos no se incluyeron o el sistema los asignó a un marco equivocado. Esto significa que es necesario construir nuevos marcos e identificar las unidades léxicas a las que se asocian. El trabajo futuro en la terminología basada en FrameNet incluye la ampliación del número de marcos para dominios específicos y la mejora de las conexiones entre WordNet y FrameNet. Pensamos que es interesante pasar a otros campos especializados diferentes y observar cómo trabaja este modelo. Esto nos permitiría establecer equivalencias y diferencias entre diferentes dominios desde el punto de vista de la semántica de los marcos.

Por último, la conexión entre FrameNet y los términos de un corpus nos permite el desarrollo de diferentes aplicaciones semánticas que incluyen la recuperación de información multilingüe, el resumen automático, la extracción de información, la traducción automática, la clasificación de documentos multilingües, los sistemas pregunta-respuesta, el procesamiento de diálogo o la organización del conocimiento.

AGRADECIMIENTOS

Esta investigación se enmarca dentro del proyecto *Lingüística y Humanidades Digitales: base de datos relacional de documentación lingüística* (PY18-FR-2511) Entidad financiadora: Convocatoria 2018 de Ayudas a proyectos I+D+i (Modalidad «Frontera Consolidado») del Plan Andaluz de Investigación, Desarrollo e Innovación Duración del proyecto: 01/01/2020 -31/03/2023. Cuantía de la subvención: 71.800 €. Investigador responsable: Miguel Casas Gómez.

REFERENCIAS BIBLIOGRÁFICAS

Azoulay, D. (2017). Frame-based knowledge representation using large specialized corpora. En L. Steels y J. Feldman (Eds.), *2017 AAAI Spring Symposium on computational construction grammar and natural language understanding* (pp. 119-126). Palo Alto, California: AAAI Press. https://www.aaai.org/ocs/index.php/SSS/SSS17/paper/view/15324

Baker, C., C. J. Fillmore y J.B. Lowe. (1998). The Berkeley FrameNet project. En C. Boitet y P. Whitelock (Eds.), *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational* (pp. 86–90). San Francisco, California: Morgan Kaufmann Publishers.

Cabré, M. T. (2005). La Terminología, una disciplina en evolución: pasado, presente y algunos elementos de futuro. *Debate Terminológico*, 1. http://riterm.net/revista/ojs/index.php/debateterminologico/article/view/23/45

Carrió Pastor, M. L. (2010). La variación del lenguaje de especialidad en artículos científicos. *Pragmalingüística*, 15-16, 71-83. https://doi.org/10.25267/Pragmalinguistica.2017.i25

Casas Gómez, M. (2006). Modelos representativos de documentación terminográfica y su aplicación a la terminología lingüística. *Revista de Lingüística y Lenguas Aplicadas*, 1(1), 25-36.

Casas Gómez, M. (2014). A Typology of Relationships in Semantics. *Quaderni di semantica: Rivista Internazionale di Semantica Teorica e Applicata*, 35 (2), 45-74.

Crespo, M. (2020a). *Automatic Corpus-based translation of a Spanish FrameNet medical Glossary*. Colección Lingüística. Sevilla: Universidad de Sevilla.

Crespo, M. (2020b). Lingüística digital: revisión de su estado actual y retos en el Instituto Universitario de Investigación en Lingüística Aplicada de la Universidad de Cádiz. *Pragmalingüística*, 28, 148-165. https://doi.org/10.25267/Pragmalinguistica.2020.i28.08

Crespo, M. (2021). Aproximación al trasvase automático de predicados de Framenet al español mediante Wordnet. *Revista de Lingüística y Lenguas Aplicadas*, 16, 49-62. https://doi.org/10.4995/rlyla.2021.14408

Cristea, D. y I. C. Pistol. (2012). Multilingual linguistic workflows. Multilingual Processing in Eastern and Southern EU Languages. *Low-resourced Technologies and Translation* (pp. 228-246). Cambridge, Reino Unido: Cambridge Scholars Publishing.

Davies, M. (2019). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Disponible online en https://www.english-corpora.org/coca/

Dolbey, A., M. Ellsworth y J. Scheffczyk. (2006). BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies. En O. Bodenreider (ed.), *Proceedings of the "Biomedical Ontology in Action" Workshop at KR-MED* (pp. 87-94). Baltimore, Maryland: National Library of Medicine.

Durán-Muñoz, I. (2016). Producing frame-based definitions: A case study. *Terminology*. *International Journal of Theoretical and Applied Issues in Specialized Communication*, 22(2), 23-249. https://doi.org/10.1075/term.22.2.04mun

Fillmore, C. J. (1977). Scenes and Frames Semantics. En A. Zampolli (Ed.), *Linguistic Structures Processing* (pp. 55-82). Amsterdam: North Holland.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2): 222–254.

Fillmore, C. J. y C. F. Baker. (2010). A Frames Approach to Semantic Analysis. En B. Heine y H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 313–339). Oxford: Oxford University Press.

Gildea, D y D. Jurafsky. (2002). Automatic Labelling of Semantic Roles. *Computational Linguistics*, 28, 245-288. https://doi.org/10.1162/089120102760275983

Guerrero Ramos, G. y M. F. Pérez Lagos. (2003). Lexicografía, terminología y diccionario. En E. Ortega Arjonilla, A. B. Martínez López y E. Echeverría Pereda (Eds.), *Panorama actual de la investigación en traducción e interpretación* (pp. 541-563). Granada: Atrio.

Johnson, Christopher y C. J. Fillmore. (2000). The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. En J. Wiebe (Ed.), *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics* (ANLP-NAACL 2000) (pp. 56–62). Seattle WA: ACM.

L'Homme, M. C. (2018). Maintaining the balance between knowledge and the lexicon in terminology: a methodology based on Frame Semantics. *Lexicography*, 4(1), 3-21. https://doi.org/10.1007/s40607-018-0034-1

L'Homme, M. C., B. Robichaud y C. Subirats Rüggeberg. (2020). Building Multilingual Specialized Resources Based on FrameNet: Application to the Field of the Environment. En T. Timponi Torrent, C. F. Baker, O. Czulo, K. Ohara, y M. R. L. Petruck (Eds.), *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet* (pp. 85-92). Marseille: ELRA.

L'Homme, M. C., C. Subirats Rüggeberg y B. Robichaud. (2016). A Proposal for combining 'general' and specialized frames. En M. Zock, A. Lenci y S. Evert (Eds.), *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon* (CogALex-V) (pp. 156-165). Osaka, Japan: ACL.

Lossio- Ventura, J. A., C. Jonquet, M. Roche y M. Teisseire. (2014). Towards a mixed approach to extract biomedical terms from text corpus. *International Journal of Knowledge Discovery in Bioinformatics* (IJKDB), 4(1), 1-15. https://doi.org/10.4018/ijkdb.2014010101

Malm, P., V. Mumtaz, B. Shafqat, Lars y A. Saxena. (2018). LingFN: Towards a framenet for the linguistics domain. En T. Timponi Torrent, L. Borin y C. F. Baker*,11th edition of the Language Resources and Evaluation Conference* (pp. 37-43). Paris, France: ELRA.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross y K. Miller (Eds.). (1993). *Five Papers on WordNet, cls report 43. Tecnical report*. New Jersey: Cognitive Science Laboratory. Princeton University.

Pérez Hernández, C. (2002). Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. *Estudios de lingüística del español*, 18. Disponible online en http://elies.rediris.es/elies18/

Potęga, P. (2017). Frame Based Modelling of Specialist Languages. *Studia Anglica Resoviensia 14*: 121–31. http://dx.doi.org/10.15584/sar.2017.14.10.

Ramírez Salado, M. (2019). *Terminología y lingüística forense: usos terminológicos relacionados con los ámbitos de actuación de la lingüística forense y su interfaz con otras disciplinas*. (Tesis doctoral, Universidad de Cádiz).

Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck, C. Johnson y J. Scheffczyk. (2016). *FrameNet II: Extended Theory and Practice*. https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf

Triola, M. (2007). *Elementary statistics.* 10th ed. Boston: Addison-Wesley.

Venturi, G. (2013). A semantic annotation of Italian legal texts. A FrameNet-based approach. En M. Fried y K. Nikiforidou (Eds). *Advances in Frame Semantics* (pp. 51-84). Amsterdam / Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/bct.58.02ven

Verdaguer, I. (2020). Semantic frames and semantic networks in the Health Science Corpus. *Estudios de lingüística del español*, (Anejo 1), 117-155.

Vossen, P. (Ed.). (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-017-1491-4

Vossen, P. (Ed.). (2002). *EuroWordNet: general document*. URL: http://vossen.info/docs/2002/EWNGeneral.pdf

Wang, L. L., k. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick y S. Kohlmeier. (2020). CORD-*19: The Covid-19 Open Research Dataset*. ArXiv. https://aclanthology.org/2020.nlpcovid19-acl.1.pdf

Witschel, H. F. (2005). Terminology extraction and automatic indexing - comparison and qualitative evaluation of methods. En B. Nistrup Madsen y H. Erdman Thomsen (Eds.), *Proceedings of 7th International Conference on Terminology and Knowledge Engineering.* (pp. 363-374). Copenhagen: Association for Terminology and Knowledge Transfer.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. En H. Uszkoreit, *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (pp. 189-196). Cambridge, Massachusetts: ACL.