Gestión terminológica, corpus especializados y extracción automática de terminología en español

Sergio Rodríguez-Tapia

EDITORIAL COMARES

Interlingua

Sergio Rodríguez-Tapia

Gestión terminológica, corpus especializados y extracción automática de terminología en español

Granada, 2024

Colección indexada en la MLA International Bibliography desde 2005

EDITORIAL COMARES

INTERLINGUA
371

Colección fundada por:
Emilio Ortega Arjonilla y Pedro San Ginés Aguilar

Comité Científico (Asesor):

Esperanza Alarcón Navío Universidad de Granada
Jesús Baigorri Jalón Universidad de Salamanca
Christian Balliu ISTI, Bruxelles
Lorenzo Blini LUSPIO, Roma
Anabel Borja Albí Universitat Jaume I de Castellón
Nicolás A. Campos Plaza Universidad de Murcia
Miguel Á. Candel-Mora Universidad Politécnica de Valencia
Ángela Collados Aís Universidad de Granada
Miguel Duro Moreno Universidad de Málaga
Francisco J. García Marcos Universidad de Almería
Gloria Guerrero Ramos Universidad de Málaga
Catalina Jiménez Hurtado Universidad de Granada
Óscar Jiménez Serrano Universidad de Granada
Ángela Larrea Espiral Universidad de Córdoba
Helena Lozano Università di Trieste
Maria Joao Marçalo Universidade de Évora
Francisco Matte Bon LUSPIO, Roma
Javier Martín Párraga Universidad de Córdoba
Antonio Raigón Rodríguez Universidad de Córdoba
Chelo Vargas-Sierra Universidad de Alicante
Mercedes Vella Ramírez Universidad de Córdoba
África Vidal Claramonte Universidad de Salamanca
Gerd Wotjak Universidad de Leipzig

ENVÍO DE PROPUESTAS DE PUBLICACIÓN:

Las propuestas de publicación han de ser remitidas (en archivo adjunto, con formato PDF) a alguna de las siguientes direcciones electrónicas: anabelen.martinez@uco.es, psgines@ugr.es

Antes de aceptar una obra para su publicación en la colección INTERLINGUA, ésta habrá de ser sometida a una revisión anónima por pares. Para llevarla a cabo se contará, inicialmente, con los miembros del comité científico asesor. En casos justificados, se acudirá a otros especialistas de reconocido prestigio en la materia objeto de consideración.

Los autores conocerán el resultado de la evaluación previa en un plazo no superior a 60 días. Una vez aceptada la obra para su publicación en INTERLINGUA (o integradas las modificaciones que se hiciesen constar en el resultado de la evaluación), habrán de dirigirse a la Editorial Comares para iniciar el proceso de edición.

Este volumen ha sido financiado por el grupo de investigación HUM-060 «Pensamiento, lenguas y textos: estudios teóricos, aplicados y didácticos» (IDEA-TEXT) de la Universidad de Córdoba.

Maquetación: Miriam L. Puerta

© Sergio Rodríguez-Tapia

© Editorial Comares, 2024

Polígono Juncaril • C/ Baza, parcela 208 • 18220 Albolote (Granada) • Tlf.: 958 465 382
https://www.comares.com • E-mail: libreriacomares@comares.com
https://www.facebook.com/Comares • https://twitter.com/comareseditor
https://www.instagram.com/editorialcomares

ISBN: 978-84-1369-743-7 • Depósito legal: Gr. 787/2024

Impresión y encuadernación: comares

Sumario

AGRADECIMIENTOS XI

Capítulo I—LA TECNOLOGÍA EN LA LINGÜÍSTICA Y LA TERMINOLOGÍA EN LA ACTUALIDAD 1
1. La tecnología en la gestión y en el tratamiento del discurso especializado 2
2. La gestión terminológica como objeto de estudio de la lingüística 4
3. Objetivos y estructura de esta obra 5
3.1. Criterios de selección de los recursos del catálogo 7
3.2. Criterios de selección para la descripción de los recursos del catálogo 11

Capítulo II—LA GESTIÓN TERMINOLÓGICA Y LAS HERRAMIENTAS INFORMÁTICAS 13
4. Las bases de datos terminológicas 14
4.1. Estructuración y presentación de los datos 18
4.1.1. *Organización onomasiológica* 18
4.1.2. *Organización semasiológica* 19
4.2. Recuperación de información: filtros y selección de información 19
4.3. Ejemplos de bases de datos terminológicas en español 22
4.3.1. *Cercaterm* 22
4.3.2. *DeCS/MeSH. Descriptores en Ciencias de la Salud* 24
4.3.3. *Dicciomed. Diccionario médico-biológico, histórico y etimológico* 26
4.3.4. *DicoAdventure* 28
4.3.5. *EuroVoc* 30
4.3.6. *IATE. Interactive Terminology for Europe* 32
4.3.7. *TERMDAT. Die Terminologiedatenbank der Bundesverwaltung* 34
4.3.8. *TERMIUM Plus* 36
5. El proceso de gestión terminológica: fases y criterios 38
5.1. Definir el alcance y los límites del recurso 39
5.1.1. *Establecer la motivación, los objetivos y los destinatarios* 39
5.1.1.1. *Definir la actividad profesional en la que se utilizará prototípicamente el recurso* 40
5.1.1.2. *Definir las necesidades prototípicas de los usuarios que suelen realizar la actividad profesional* 41

5.1.1.3. *Determinar el valor de las unidades seleccionadas para incluirse en el proyecto* 41
5.1.1.4. *Determinar las lenguas del proyecto* 41
5.1.2. *Establecer las características del recurso: los campos de la entrada terminológica* 42
5.1.2.1. *Información fonético-fonológica* 42
5.1.2.2. *Información morfológica* 43
5.1.2.3. *Información sintáctica* 44
5.1.2.4. *Información semántica* 45
5.1.2.5. *Información pragmática* 50
5.1.2.6. *Información administrativa* 52
5.1.2.7. *Algunas reflexiones sobre los campos* 56
5.1.3. *Determinar los recursos humanos y técnicos y planificar el proyecto* 56
5.2. Elaborar el recurso 57
5.3. Revisar el recurso 57
5.4. Editar, actualizar y dar difusión al producto 58
6. Los sistemas de gestión terminológica 59
6.1. Ejemplos de sistemas de gestión terminológica 60
6.1.1. *MultiTerm* 60
6.1.2. *TermStar NXT* 62
6.1.3. *WebTerm* 64
7. Las ontologías como bases de conocimiento terminológico 65
7.1. La ingeniería del conocimiento y la terminología 66
7.2. El concepto de ontología 67
7.3. La formalización en las ontologías 69
7.4. Componentes de una ontología 72
7.5. Aplicaciones de una ontología 74

Capítulo III—EL DISEÑO Y USO DE LOS CORPUS ESPECIALIZADOS 77
8. Motivación de la lingüística de corpus 77
8.1. El corpus: el objeto de la lingüística de corpus 78
8.2. Consideraciones epistemológicas y ventajas instrumentales 79
8.3. La relevancia de los corpus en el análisis del discurso especializado y sus aplicaciones 81
9. Criterios de clasificación de los corpus 83
9.1. Los corpus específicos y especializados: denominaciones y ventajas 89
10. Fases y criterios en la compilación de corpus especializados 91
10.1. Compilar el corpus: alcanzar la representatividad del corpus especializado 93
10.1.1. *Criterios cualitativos de representatividad* 94
10.1.1.1. *La fiabiliad de las fuentes* 94
10.1.1.2. *Los límites del ámbito de especialidad* 96
10.1.1.3. *El grado de especialización* 100
10.1.1.4. *Diversificación de las muestras: homogeneidad y heterogeneidad* 101
10.1.2. *Criterios cuantitativos de representatividad* 103
10.1.2.1. *Tamaño del corpus* 103
10.1.2.2. *Distribución de las muestras del corpus* 105

10.2. Codificación y etiquetado 106
10.3. Recuperación y análisis de información. Las funciones de los gestores de corpus 113
11. Ejemplos de corpus especializados en español 119
11.1. *Corpus de las Sexualidades de México* (CSMX) 119
11.2. *Corpus Tècnic* 121
11.3. *DIACOM-es* 123

Capítulo IV—LA EXTRACCIÓN TERMINOLÓGICA 127
12. Delimitación conceptual: extracción terminológica manual y automática 127
13. Ventajas de la extracción automática de terminología 128
14. Aplicaciones de la extracción automática de terminología 130
15. Extracción manual de terminología: la relatividad del valor terminológico 132
16. Sistemas de extracción automática de terminología 135
16.1. Sistemas lingüísticos de extracción automática de terminología 135
16.2. Sistemas estadísticos de extracción automática de terminología 137
16.3. Sistemas híbridos de extracción automática de terminología 138
17. Criterios para la extracción automática de terminología 139
17.1. La unicidad 139
17.1.1. *Identificación de la unicidad mediante método lingüístico* . 139
17.1.2. *Identificación de la unicidad mediante método estadístico.* 141
17.2. La terminologicidad 142
17.2.1. *Identificación de la terminologicidad mediante método estadístico* 142
17.2.1.1. *Método distribucional. Relación TF-IDF* 142
17.2.1.2. *Método contextual: algoritmo C-value* 143
17.2.1.3. *Método contextual: algoritmo NC-value* 143
17.2.2. *Identificación de la terminologicidad mediante método lingüístico* 144
17.2.3. *Identificación de la terminologicidad mediante método contrastivo* 144
17.3. La variación terminológica 145
17.4. La evaluación y la validación 146
18. Ejemplos de sistemas para la extracción automática de terminología 147
18.1. *OneClick Terms* de *Sketch Engine* 148
18.2. *TermoStat Web 3.0* 150

Capítulo V—RECAPITULACIÓN Y REFLEXIONES FINALES 155

Capítulo VI—REFERENCIAS BIBLIOGRÁFICAS 161

Capítulo VII—CATÁLOGO DE RECURSOS TECNOLÓGICOS SOBRE GESTIÓN TERMINOLÓGICA EN ESPAÑOL 175

Agradecimientos

Este proyecto nace por una motivación personal y docente. Esta monografía ha sido fruto de un proceso de introspección que ha pretendido responder a muchas de las preguntas que me he ido haciendo durante mi etapa investigadora, así como a las cuestiones planteadas por el alumnado de Terminología de la Universidad de Córdoba a lo largo de los años. Esta obra ha sido resultado de la consulta a muchos colegas y compañeros que admiro, quienes han ofrecido su tiempo y dedicación para brindarme su opinión.

En primer lugar, agradezco a Laia Vidal e Iria Da Cunha los comentarios recibidos a la génesis de este proyecto y a una primera versión de este volumen. Gracias a sus sugerencias, este trabajo ve la luz con nueva estructura, planteamientos y preguntas. Asimismo, agradezco a Judit Freixa el hecho de facilitarme la información administrativa sobre algunos de los proyectos del IULA, que han sido integrados en este texto.

En mi universidad doy las gracias a Adela González y a María Martínez-Atienza por sus precisos comentarios acerca de la lingüística de corpus y de la fraseología respectivamente. A Eduardo J. Jacinto le agradezco especialmente la atenta lectura de partes de este volumen, pues esta obra contiene muchas de las conversaciones mantenidas desde hace siete años, porque se ha ido construyendo mediante el diálogo, el intercambio de opinión, de intereses y, especialmente, de la ilusión que compartimos por el discurso especializado. A Alfonso Zamorano le debo mi interés científico por el metalenguaje, que he pretendido plasmar en esta obra. Le agradezco todo su apoyo académico e institucional, su paciencia y sus constantes ánimos y alegrías compartidas.

A Carmen Oliva, Sergio Martínez y Yordan Apostolov les agradezco sus sugerencias y comentarios porque, con buen ojo clínico, han demostrado ser los discípulos con los que cualquier profesor querría contar. Les agradezco su atenta lectura a capítulos de este trabajo y las notas que han permitido precisar ambigüedades o confusiones, o acabar con gazapos propios de mi falta de cuidado.

En lo personal, querría agradecer las ideas de Carmen Moreno, quien me ha ayudado a reflexionar sobre mis objetos de estudio y sobre mis objetivos vitales más de lo que cree. A Rafa y Javi me gustaría agradecerles la inmensa paciencia demostrada, su incuestionable talante y su curiosidad por este mundo de «términos impronunciables y cosas difíciles que nadie entiende».

A todos ellos, doy las gracias por su tiempo e inestimable dedicación.

Sergio Rodríguez-Tapia

colección:

INTERLINGUA

371

Dirigida por:

Ana Belén Martínez López y Pedro San Ginés Aguilar

La gestión terminológica consiste en una serie de procedimientos de carácter técnico que involucran el empleo de la tecnología para aumentar la eficiencia en la recopilación, documentación, almacenamiento y recuperación del vocabulario especializado. Constituye, pues, un trabajo esencial para el tejido industrial, especialmente para las empresas y las instituciones. El resultado de la gestión terminológica suele ser una base de datos, donde se consignan los términos, sus usos, recomendaciones o vinculaciones a ciertos productos o empresas, aclaraciones, observaciones o definiciones, sinónimos, etc. La utilidad real de estas bases de datos terminológicas reside en que permiten consultar rápidamente información validada y contrastada, ya sea por los usuarios externos o por el personal de la empresa o institución. Este volumen presenta un estado de la cuestión donde se aborda la fundamentación teórica y metodológica en torno a la gestión terminológica, que se presenta como una de las actividades relacionadas con el tratamiento del discurso especializado. Para ello, se introduce su necesaria relación con las bases de datos terminológicas, con la lingüística de corpus y con la extracción terminológica, aplicaciones para las cuales el presente volumen recopila un catálogo de 103 recursos informáticos diferentes. El fin último de este volumen es servir como instrumento de apoyo a la investigación teórica y a la docencia universitaria de los campos relacionados con los estudios lingüísticos; en concreto, terminológicos.

COMARES
editorial

ISBN 978-84-1369-743-7