LOS MÉTODOS DE APRENDIZAJE AUTOMÁTICO SUPERVISADO EN LA CLASIFICACIÓN TEXTUAL SEGÚN EL GRADO DE ESPECIALIZACIÓN

Sergio Rodríguez-Tapia

(Universidad de Córdoba. Facultad de Filosofía y Letras. Departamento de Ciencias del Lenguaje. Córdoba, España)

sergio.rodriguez@uco.es

Julio Camacho-Cañamón

(Universidad de Córdoba. Escuela Politécnica Superior. Departamento de Departamento de Informática y Análisis Numérico. Córdoba, España)

julio.camacho@uco.es

Fecha de recepción: 27-12-2017 / Fecha de aceptación: 4-5-2018

RESUMEN:

Este artículo pretende aplicar métodos de aprendizaje automático supervisado a una base de datos con información textual según el grado de especialización para comprobar la relevancia teórica de métodos numéricos de clasificación empleados *a priori*. Los resultados destacan las debilidades de trabajar con clasificaciones cuantitativas y las fortalezas de predicción de clase al usar el algoritmo J48.

Palabras clave: aprendizaje automático supervisado; terminología; texto especializado; clasificación; grado de especialización

ABSTRACT:

This paper aims to apply supervised machine learning methods to a database containing textual information on specialization degree to test the theoretical relevance of numerical classification methods used beforehand. Results highlight

the weaknesses when working with quantitative classifications and the class-prediction strengths when using algorithm J48.

Keywords: supervised machine learning, terminology, specialized text, classification, specialization degree

1. JUSTIFICACIÓN DEL ESTUDIO

Una de las cuestiones teóricas que quedan por resolver en la investigación terminológica tiene que ver con el contínuum del grado de especialización. Si bien las investigaciones coinciden en distinguir un grado máximo de especialización (el texto especializado: TE), un grado intermedio (el texto semiespecializado: TSE) y un grado mínimo (el texto divulgativo: TD) (Cabré, 2002a: 30-31, entre otros) y en que existen ciertas características atribuibles a los polos opuestos, es cierto que no es posible definir el texto semiespecializado si no es a través de su caracterización como objeto de transición.

Existen trabajos (Rodríguez-Tapia, 2015; 2016a) que esbozan propuestas para analizar el contínuum de especialización teniendo en cuenta elementos lingüísticos y extralingüísticos a partir de un corpus textual. Nuestro artículo parte de las líneas de investigación que quedan por resolver en los estudios referidos anteriormente sobre los criterios teórico-metodológicos para analizar el grado de especialización textual, que dejaba pendiente "aplicar los resultados obtenidos a la lingüística computacional y la lingüística de corpus en la caracterización del texto semiespecializado" (Rodríguez-Tapia, 2015: 29).

En este artículo, se pretende reflexionar acerca de la utilidad de aplicar modelos numéricos de clasificación a bases de datos con información acerca del grado de especialización de un corpus lingüístico. Para ello, hemos aplicado métodos de aprendizaje automático supervisado.

1.1 Primeros trabajos

En otros trabajos (Rodríguez-Tapia & Camacho-Cañamón, en prensa) se han aplicado algoritmos de clasificación no supervisada para contrastar la hipótesis de partida sobre el grado de especialización de los textos. Según los resultados de este estudio se pudo concluir que el primer análisis del algoritmo

de aprendizaje automático no supervisado *k-means,* presentaba grupos lo suficientemente homogéneos como para defender la existencia de clases diferenciadas por cada grupo y poder confirmar así la hipótesis de gradación en tres niveles.

Estos resultados apuntaban hacia la posibilidad de trabajar con algoritmos supervisados y contrastar los resultados con algoritmos no supervisados para evaluar el modelo numérico de clasificación con el que trabajamos e identificar sus fortalezas y debilidades. Se denomina *clasificación* al proceso de construcción de un modelo de clases a partir de un conjunto de datos que contengan atributos y una etiqueta de clase. Estos algoritmos generan reglas para la predicción de la etiqueta de clase a partir de los atributos de los patrones, en este caso, textos. En nuestro caso, contamos con un corpus de 60 textos del campo médico, de los cuales se han extraído manualmente una serie de características (atributos), que serán analizados posteriormente.

2. EL GRADO DE ESPECIALIZACIÓN TEXTUAL COMO OBJETO DE ESTUDIO

Cuando se pretende clasificar un texto especializado, suele recurrirse a dos criterios de tipologización principalmente (Cabré, 1993: 141; Cabré, 2002a: 30): el eje vertical y el eje horizontal. El grado de especialización hace referencia al eje vertical en el que pueden clasificarse los textos. Este eje se centra en el nivel de especialización (Cabré, 2002a: 30; 2002b: 93), es decir, en los distintos grados de abstracción, que conllevan "valores graduales de precisión, concisión y cientificidad" (Cabré & Estopà, 2002: 146).

El grado de especialización, pues, responde a la perspectiva que adopta el texto para abordar el tema del mismo. Este grado es analizable lingüísticamente puesto que queda actualizado en el estilo discursivo del emisor, y puede vincularse con condicionantes extralingüísticos. Desde la perspectiva teórica, los condicionantes y las estrategias de actualización en el discurso constituyen datos muy relevantes para la lingüística aplicada, puesto que, como defiende Cabré (2002a: 21), las "tipologías textuales […] son sistemas de organización que permiten hacer generalizaciones y establecer predicciones orientativas".

3. EL PROBLEMA DE LA CLASIFICACIÓN TEÓRICA DEL GRADO DE ESPECIALIZACIÓN

La existencia de diferentes condicionantes en la tipologización textual dificulta la clasificación de los textos, más aún cuando existe un nivel intermedio en el paradigma. Hasta ahora, las investigaciones se han posicionado: a) por una parte, a favor de una clasificación binaria en textos especializados y no especializados basándose en el perfil (especialista o no) del emisor (Sager, Dungworth & McDonald, 1980); y b) por otra parte, a favor del contínuum (trabajos citados de Cabré y otros como los de Ciaupuscio & Kuguel, 2002; y Ciapuscio, Kuguel & Otañi, 2005).

Cabré (2002a: 22-23) contrasta las características esenciales que permiten diferenciar el texto especializado del texto divulgativo, los dos polos opuestos del contínuum (el blanco y el negro). Este contraste puede servir de punto de partida para seleccionar criterios de análisis pero no garantiza evitar toparse con los límites difusos (los grises) que existen (véase Tabla 1). Estos límites se encuentran en la frontera entre texto especializado y texto semiespecializado y entre texto semiespecializado y texto divulgativo, de forma que el objeto *texto semiespecializado* a veces compartirá más características con el texto especializado y otras, con el texto divulgativo. No obstante, estas características solo supondrían una tendencia o una generalización pero no una relación de condiciones necesarias o suficientes (Cabré, 2002a, p 32).

<table>
<tr><td>Texto especializado</td><td>Objeto definido</td></tr>
<tr><td>Límite texto especializado-semiespecializado</td><td rowspan="3">Gradación y niveles de confusión</td></tr>
<tr><td>Texto semiespecializado</td></tr>
<tr><td>Límite texto semiespecializado-divulgativo</td></tr>
<tr><td>Texto divulgativo</td><td>Objeto definido</td></tr>
</table>

TABLA 1. Gradación de especialización y niveles de confusión. Fuente: elaboración propia

Han surgido multitud de modelos de propuestas de tipologización en las últimas décadas pero revisten especial atención el modelo de Cabré (2002a), que aborda la estructura textual (formal, informativa y lingüística), y el modelo de corte cognitivo-comunicativo multinivel y multidimensional de Ciapuscio & Kuguel (2002: 37-47), que se basa en cuatro niveles de análisis (funcional, situacional, semántico y formal).

La propuesta de análisis expuesta en Rodríguez-Tapia (2016a: 995-1001) pretende aglutinar y modificar elementos de los modelos anteriores para, basándonos en la Teoría Comunicativa de la Terminología, simplificar el análisis en cuatro parámetros principalmente (véase Tabla 2) y, al mismo tiempo, tener en cuenta el mayor número posible de condicionantes.

A) Estructura textual	B) Condiciones discursivas
a) Plano semántico b) Plano formal	c) Plano situacional d) Plano funcional

TABLA 2. Parámetros de análisis en Rodríguez-Tapia (2016a)

Así, en este artículo estudiamos una base de datos que contiene 60 textos analizados, a los que se les han vinculado diferentes valores por cada atributo o parámetro de análisis, y una clase, resultado de la reflexión e interpretación de dichos parámetros. De forma esquemática, la base de datos contendría la información de la Tabla 3, con un valor para cada atributo y un valor de clase para cada texto:

ATRIBUTOS	
Función	Índice de densidad terminológica
• Representativa • Representativa-comunicativa • Comunicativa	• Valores entre 0 y 1
Relación discursiva de los interlocutores	Superestructura
• Especialista-especialista • Especialista-lego • Instruido-lego	• Artículo científico • Artículo divulgativo • Tesis doctoral • Carta al editor • Guía para el paciente • Manual para el especialista
CLASES	
• Texto especializado • Texto semiespecializado • Texto divulgativo	

TABLA 3. Atributos analizados y clases propuestas. Fuente: elaboración propia

3.1 Los atributos del texto. Criterios y condicionantes

Según Kohavi y Provost (1998: 271-274) se define como *atributo* una característica que describe una instancia. El dominio de un atributo está definido por el tipo de atributo y denota qué valores puede tomar. Los tipos de dominios más comunes son: categóricos y cuantitativos discretos o continuos.

(a) Los *dominios categóricos* tienen un número finito de valores, que pueden ser nominales, cuando el orden no es relevante (colores); u ordinales, donde existe un orden entre los distintos valores (grados en una enfermedad).

(b) Respecto a los *dominios cuantitativos*, se suelen atribuir al conjunto de los números reales, donde hay diferencias apreciables entre los posibles valores, como la estatura (que contaría con valores cuantitativos continuos).

En la Tabla 4 se ilustra con un ejemplo el dominio de diferentes atributos.

Característica	Atributo	Tipo de dominio	Dominio (magnitud)
La edad	'edad'	Cuantitativo discreto	{0-120} (años)
La altura	'altura'	Cuantitativo continuo	{0,1-2,5} (metros)
El color de ojos	'color_ojos'	Categórico nominal	{marrón, verde, azul, gris}
El grado de enfermedad	'grado_enfermedad'	Categórico ordinal	{leve, moderado, severo}

TABLA 4. Ejemplos de atributos y dominios. Fuente: elaboración propia

Así, los atributos con los que contamos en este estudio son: la función, la relación discursiva entre los interlocutores, el índice de densidad terminológica y la superestructura.

3.1.1 La función

Sin perder de vista que la lengua es principalmente, desde el punto de vista funcional, un vehículo de comunicación y que el léxico "tiene como función básica la de representar el conocimiento que constituye el esqueleto cognitivo del texto" (Cabré, 2002b: 94), cada texto podría cumplir con tres funciones lingüísticas principalmente (con respecto al conocimiento especializado o no que transmita):

(a) Una *función representativa*: resultado de la mayor abstracción en el tema tratado. Con esta función, el texto centra su objetivo en dibujar un esquema conceptual que exponga el objeto que trata. La comprensión de un texto con función principalmente representativa queda fuera del usuario medio de la lengua. Suele ser la función que

cumplen los artículos científicos o las tesis doctorales, o la que predomina en textos entre especialista y especialista.

(b) Una *función comunicativa*: que presta atención a la transmisión de información, más que a la elaboración de un esquema conceptual exhaustivo. Los textos con función primordialmente comunicativa son comprensibles para el usuario medio de la lengua, si bien algunos conceptos pueden tratarse con cierta laxitud conceptual. Suele ser la función observable entre un especialista y un lego o en guías y folletos para el usuario no especialista.

(c) Una función mixta *representativa-comunicativa*: el tema es tratado desde un enfoque especialista pero el discurso sufre una recontextualización con estrategia comunicativa (reformulaciones, explicaciones, etc.). El usuario medio de la lengua es capaz de comprender el texto, aunque, a veces, con serias dificultades por desconocimiento del sistema conceptual del ámbito temático. Es la función que cumplen ciertos textos orientados a la iniciación en alguna disciplina o a la enseñanza-aprendizaje de cuestiones relacionadas con las mismas.

Esta división no quiere decir que un texto con función representativa no sea comunicativo o no pretenda comunicar la información que codifica, sino que su *función principal* es la de *representar* el conocimiento, la de crear un mapa del conocimiento. Todos los textos representan y comunican conocimiento y discriminar sus funciones a través de la actualización lingüística es una tarea compleja, pues los límites difusos dificultan la categorización taxativa.

3.1.2 La relación discursiva

En todos los textos existe una relación entre los interlocutores que condiciona la configuración del mismo: emisor y receptor. El perfil de cada uno establece una serie de condiciones que afectan a cómo se estructura el discurso y a la función que cumplirá el texto. Así, percibimos tres tipos de perfiles de interlocutores en cuanto a la capacidad de transmisión de conocimiento especializado:

(a) *Especialista*: posee formación académica y profesional en la materia y es capaz de llevar a cabo reflexión y argumentación sobre el conocimiento que domina.
(b) *Instruido*: domina la materia que trata pero no tiene formación académica o profesional en ella, sino que ha sido fruto de procesos de documentación e información complementarias. Es el perfil que podría adquirir un mediador lingüístico como un traductor.
(c) *Lego*: no posee formación académico-profesional en la materia en cuestión y tampoco es capaz de crear nuevo conocimiento especializado sobre la materia.

Las relaciones discursivas entre estos tres tipos de interlocutores son, por tanto, tres, una simétrica y dos asimétricas (Rodríguez-Tapia, 2016a: 1000)[1]:

(a) *Especialista-especialista*: es la relación que existe entre un médico y un médico en un artículo científico.
(b) *Especialista-lego*: es la relación que existe entre un médico y su paciente en un folleto o en una consulta.
(c) *Instruido-lego*: es la relación que existe entre un periodista y su lector en un artículo divulgativo o en un folleto o guía.

3.1.3 El índice de densidad terminológica

La terminología es el conjunto de unidades léxicas con valor especializado (se trata de un valor semántico-pragmático; Cabré & Estopà, 2002: 146) que se emplea en un ámbito temático para representar y comunicar conocimiento especializado.

De acuerdo con la gran mayoría de estudios (Arntz & Picht, 1995: 45; Cabré, 2002a: 94; Monterde Rey, 2002: 115; Domènech, 2007: 250), la terminología es la característica esencial del texto especializado, y constituye la pieza clave a partir de la cual se estructura todo el esquema discursivo de un texto. El índice de densidad terminológica (IDT) permite conocer el nivel de conocimiento especializado que codifica un texto y puede calcularse dividiendo el número de unidades léxicas especializadas (o términos) entre el número total de

[1] En este trabajo no se tienen en cuenta las relaciones inversas, como lego-especialista, o adicionales, como instruido-especialista, por no encontrar ejemplos de dichas relaciones en el corpus analizado.

unidades léxicas de un texto. De esta forma, el valor obtenido puede hallarse en el intervalo entre 0 y 1 (este último valor 1 indicaría que todos las unidades léxicas usadas en el texto son términos).

A continuación, se presenta un ejemplo extraído de uno de los textos del corpus (texto 46), donde se comprueba un IDT de 0,78 (el más elevado del corpus), resultado de dividir 57 términos entre 73 unidades léxicas. En cursiva se marcan las unidades terminológicas:

> Se han identificado dos elementos principales que intervienen en el desarrollo del *daño endotelial:* el *óxido nítrico* y la *endotelina*, que a través de la *vasoconstricción* y la *proliferación celular* favorecen el *remodelado vascular* del *componente reactivo* de la *HTP*.
> El *Óxido nítrico* (*ON*) se genera en la *célula endotelial* del *lecho vascular pulmonar*. Produce una *relajación* e *inhibición* de la *proliferación* de la *célula muscular lisa* e *inhibe* la *adhesión* y *agregación* de las *plaquetas*. Las dos fuentes principales de *ON pulmonar* son el *endotelio* y el *epitelio* de la *vía aérea*. En presencia de *oxígeno*, las *ON sintetasas* son *activadas* y producen *ON* a partir de la *L-arginina*. El *ON activa* a nivel de la *membrana endotelial* la *Guanilato ciclasa*, que *sintetiza* el *Guanilato monofosfato cíclico* (*GMPc*), que a su vez *activa* la *GMPc kinasa* (Torres Macho, 2011: 28)[2].

3.1.4 La superestructura

El plano superestructural constituye el esqueleto formal y la estructura global que caracteriza a un texto. Según Dijk (1992), la superestructura es "un tipo de forma del texto". En otras palabras, la superestructura reúne todas las características de la organización de la información del texto, que, a su vez, viene determinada por el tipo de texto, que impone el tipo de estructura, regida por la forma o por la presentación de la información.

Existen diversas formas de referirse a la superestructura, según el criterio de clasificación al que atendamos: función, registro, temática, etc. De esta forma, si atendemos a la función textual y partimos del tipo *texto literario*, podríamos tratar con *cuento, novela, poema*, etc.; si partimos del registro, encontraríamos una *lectura de tesis* o un *debate de investidura*; y si nos centramos en el criterio temático, podríamos tratar con textos *médicos, jurídicos*, etc.

[2] Fuente del texto: TORRES MACHO, Juan (2011): *Factores de riesgo y pronóstico de la hipertensión pulmonar en los pacientes con insuficiencia cardiaca avanzada*. Madrid: Universidad Complutense de Madrid. Disponible en: http://eprints.ucm.es/12596/1/T32856.pdf.

En este sentido, las superestructuras de nuestro corpus parten de un criterio temático sobre medicina y, a pesar de esta complejidad a la hora de seleccionar un paradigma de clasificación tipológica textual, hemos optado por seleccionar un punto de vista desde el receptor, como interlocutor al que se dirige la información. De esta forma, los tipos textuales analizados se corresponden con:

(a) *Artículo científico*: texto en el que se reflexiona sobre conocimiento especializado desde una perspectiva especialista.

(b) *Artículo divulgativo***:** texto que pretende informar sobre algún aspecto del conocimiento especializado desde una perspectiva no especialista.

(c) *Tesis doctoral*: texto académico que persigue la reflexión y argumentación en torno al conocimiento especializado desde una perspectiva especialista.

(d) *Carta al editor*: texto que persigue la reflexión y argumentación en torno al conocimiento especializado de un artículo científico desde una perspectiva especialista.

(e) *Guía para el paciente*: texto informativo que persigue aportar información relacionada con un aspecto del conocimiento especializado desde una perspectiva no especialista.

(f) *Manual para el especialista*: texto informativo que persigue aportar información relacionada con un aspecto del conocimiento especializado desde una perspectiva especialista.

3.2 *La clase del texto: texto especializado, texto semiespecializado y texto divulgativo*

Desde el punto de vista del aprendizaje automático se define *clase* como la etiqueta que se le asigna a un subconjunto de elementos o patrones que comparten unas características similares. Por ejemplo, un correo electrónico puede ser etiquetado o no como *spam* en función de sus atributos: número de palabras, frecuencias de aparición de cada palabra, aparición de ciertas palabras clave, etc. Es posible distinguir diferentes tipos de clasificación.

(a) La más conocida y empleada sería la *clasificación binaria* o *biclase*, que asigna los valores *sí/no* a cada patrón dependiendo de si pertenece o

no a una determinada clase. El ejemplo anterior pertenecería a este tipo de clasificación.

(b) De forma más general, hallamos la *clasificación multiclase*, donde las posibles etiquetas corresponderían a las diferentes clases a las que puede pertenecer un patrón, por ejemplo: el estado civil podría ser *soltero*, *casado*, *viudo* o *divorciado*. En caso de que existiese un orden entre las etiquetas, se hablaría de *clasificación ordinal*; si no, de *clasificación nominal*.

Las etiquetas de clase con las que trabajamos corresponden con los distintos grados de especialización de los textos: especializados, semiespecializados o divulgativos. Hablamos, pues, de una clasificación ordinal, en el sentido de que van de mayor a menor especialización.

La clase también es una superestructura, que, en este caso, atiende a nuestro objeto de análisis: el grado de especialización. Los valores de la clase se corresponden con los tres niveles del contínuum, que podrían caracterizarse, a modo de propuesta, de la siguiente manera:

(a) *Texto especializado*:

> productos predominantemente verbales de registros comunicativos específicos, que se refieren a temáticas propias de un dominio de especialización, y que responden a convenciones y tradiciones retóricas específicas; por lo tanto, en dependencia del tipo de disciplina pueden ser más o menos dependientes de la cultura y la época dada [...]. Los textos especializados se realizan en clases textuales específicas del discurso de especialización (artículo de investigación, ponencia, artículo de divulgación científica, comunicados científicos a la prensa, etc.). Concebimos el ámbito de los textos especializados en términos de contínuum (Ciapuscio & Kuguel, 2002: 43).

(b) *Texto semiespecializado*:

> aquel texto redactado por un especialista y dirigido indistintamente a un público lego o especialista, en el que predominan, de forma independiente o en conjunto, las funciones representativa y comunicativa, cuyo índice de densidad terminológica no es especialmente alto (0,12 de media) y que se asocia a la superestructura de los glosarios, artículos divulgativos y artículos científicos (Rodríguez-Tapia, 2016b).

(c) *Texto divulgativo*: de forma opuesta al texto especializado, se trata del texto que no codifica conocimiento especializado, es producto de los intercambios comunicativos ordinarios entre interlocutores y se configura discursivamente hacia un receptor no especialista.

3.2.1 Método numérico de clasificación

La clase está determinada cuantitativamente por un modelo numérico, que trabaja con la relación porcentual y posterior suma de valores de la Tabla 5 (Rodríguez-Tapia, 2016a: 1000-1001)[3]:

Relación discursiva entre interlocutores (Máximo 0,3)	Especialista-especialista	0,3
	Especialista-lego	0,2
	Instruido-lego	0,1
Función principal del texto (Máximo 0,4)	Representativa	0,4
	Representativa-comunicativa	0,3
	Comunicativa	0,2
Índice de densidad terminológica (Máximo 0,3)	Valor numérico del IDT multiplicado por 100	0,3

TABLA 5. Valores porcentuales por atributos. Fuente: elaboración propia

Cada texto, por tanto, se codifica en un valor numérico que es resultado de la suma de dichos porcentajes. Los textos se calificarían con el 30 % (o 0,3 puntos) si se identifica una relación especialista-especialista; con un 20 % (o 0,2 puntos) si se identifica una relación especialista-lego; y con un 10 % (o 0,1 puntos) si se identifica una relación instruido-lego. De igual forma se opera con el atributo de la función principal del texto. Estos valores se suman a los que corresponden con el valor numérico del índice de densidad terminológica multiplicado por 100. Así, cada texto *se codifica* en un valor numérico que sirve para clasificarlo dentro de un grado de especialización.

Para poder clasificarlos, a cada nivel de especialización se le asigna un intervalo numérico que atiende a la hipótesis de contínuum, de forma que se pueda incluir cada texto dentro de una clase en función de la pertenencia de su valor a uno u otro intervalo. Estos intervalos establecen unos límites precisos, lo que conlleva ciertos problemas metodológicos, que ya aparecieron en Rodríguez-Tapia (2016b) y cuyas posibles soluciones se analizan en otros trabajos (Rodríguez-Tapia & Camacho-Cañamón, en prensa) (véase Tabla 6).

Texto especializado	$11 \leq x$

[3] Debemos recordar que, desde nuestra perspectiva, la superestructura no es condicionante esencial del grado de especialidad, por lo que no está presente en dicha tabla.

Texto semiespecializado	7 ≤ x < 11
Texto divulgativo	0,3 ≤ x <7

TABLA 6. Valores numéricos para clasificar las clases. Fuente: elaboración propia

A modo de síntesis y de forma ilustrativa, se presentan en la Tabla 7 algunos ejemplos extraídos de la base de datos:

	Atributos					Clase
Texto n.º	IDT (Máximo 0,3)	Relación discursiva (Máximo 0,3)	Función principal (Máximo 0,4)	Valor total	Superestructura	
1	0,20	Especialista-especialista	Representativa	13	Artículo científico	TE
	(0,2 × 100) × 0,3 = 6	10 × 0,3 = 3	10 × 0,4 = 4	6 + 3 + 4		11 < 13
12	0,12	Especialista-especialista	Representativa-comunicativa	9,6	Carta al editor	TSE
	(0,12 × 100) × 0,3 = 3,6	10 × 0,3 = 3	10 × 0,3 = 3	3,6 + 3 + 3		7 ≤ 9,6 < 11
21	0,06	Especialista-especialista	Comunicativa	6,8	Guía para el paciente	TD
	(0,06 × 100) × 0,3 = 1,8	10 × 0,3 = 3	10 × 0,2 = 2	1,8 + 3 + 2		0,3 ≤ 6,8 <7

TABLA 7. Ejemplo de proceso de clasificación: valores numéricos y asignación de clases. Fuente: elaboración propia

4. LA CONTRIBUCIÓN DE LAS CIENCIAS DE LA COMPUTACIÓN Y EL APRENDIZAJE AUTOMÁTICO A LA TERMINOLOGÍA

Lorente (2014) relaciona desde diferentes perspectivas la vinculación existente entre terminología y tecnología, revisando algunos estudios presentados en los congresos de RITerm (Red Iberoamericana de Terminología). Su revisión de las etapas de tecnificación de la terminología supone gran relevancia en la justificación de nuestro trabajo y queda ilustrado en las siguientes palabras:

> Así Lisboa (2000) significó la verdadera asunción del desarrollo tecnológico desde la terminología. Ya no se trataba de encontrar en el mercado o entre las escuelas politécnicas algún *software* que pudiéramos utilizar, sino que se instaló la consciencia de que los terminólogos debíamos participar directamente en la construcción de corpus textuales de la ciencia y de la técnica, conocer los sistemas de etiquetaje del

> procesamiento del lenguaje natural, concebir programas informáticos específicos para la extracción de términos, recoger neologismos con la ayuda de ordenadores, elaborar diccionarios máquina y diccionarios electrónicos, o diseñar programas informáticos de ayuda para la enseñanza, la redacción, la traducción y la documentación (Lorente, 2014: 21-22).

Desde sus inicios teóricos hasta la actualidad, la aplicación de la ingeniería lingüística a la terminología se ha centrado, además de en la creación de ontologías y bases de datos, en la detección y extracción de términos, como demuestran los trabajos sobre SEACUSE (Estopà, 2001), YATE (Vivaldi, 2001, 2009), WikiYATE (Vivaldi & Rodríguez, 2012) o investigaciones como la de Nazar y Cabré (2011), que usan algoritmos estadísticos supervisados.

La principal ventaja de la aplicación de las utilidades informáticas a la investigación terminológica no solo tiene que ver con el aumento de la eficiencia a la hora de tratar e interpretar los datos, sino con la reducción del posible error humano en la recogida e interpretación de los mismos. No obstante, la automatización de algunos procesos tropieza con problemas de corte teórico, dado el desconocimiento o confusión que existe sobre la propia entidad epistemológica de algunos objetos de la realidad (en este caso, el contínuum de especialización), lo que repercute en la selección de criterios y en la recogida, tratamiento e interpretación de los datos. Lo óptimo sería poder contar con una inteligencia artificial que permitiese indicar si un texto pertenece a una clase o a otra con tan solo acceder a él pero no es posible, puesto que teóricamente las características de los niveles intermedios no han sido distinguidas. Para enfrentarse a estas cuestiones es necesaria la investigación teórica y la generación de nuevos algoritmos. En este sentido, Nazar dirige desde 2014 el proyecto titulado *Desarrollo de un algoritmo estadístico de extracción terminológica basado en coocurrencia léxica*, cuyo objetivo es realizar un "estudio teórico para explorar las bases de una teoría cuantitativa de la terminología, que tendría por objeto definir de manera formal la diferencia entre lo que es y lo que no es un término".

En estrecha relación con nuestro principal objetivo investigador, encontramos la aplicación de las características gramaticales en el trabajo titulado "Automatic Specialized vs. Non-specialized Text Differentiation" publicado por Cabré *et al*. (2014) que se sirve de las diferencias teóricas esenciales entre texto especializado y texto divulgativo para distinguir de forma

automática dichos tipos textuales. Se trata de un trabajo novedoso, a la par que ambicioso, que arroja luz sobre la automatización en la distinción del grado de especialización. No obstante, se limita a los dos polos opuestos y no incluye el nivel intermedio, tan controvertido desde el punto de vista teórico.

En relación con estos niveles, si bien el objetivo del estudio de Guantiva, Cabré & Castellà (2008) (que titulan "Clasificación de textos especializados a partir de su terminología") no es la creación de programas ni de métodos computacionales, sí que se sirven del análisis factorial discriminante para clasificar individuos a partir de variables cuantitativas, lo cual se asemeja, desde una orientación metodológica, a la proyección que realizamos en este trabajo.

Así, nuestro artículo no trata de contribuir a la extracción terminológica, a la creación de técnicas para compilación de corpus o al diseño de programas, sino que, en este sentido, la contribución de las ciencias de la computación a la terminología no es estrictamente una aplicación práctica; su fin último tiene carácter eminentemente teórico: la distinción de características de los textos para determinar el grado de especialización textual a través de la evaluación del método numérico que proponemos (véase Figura 1).

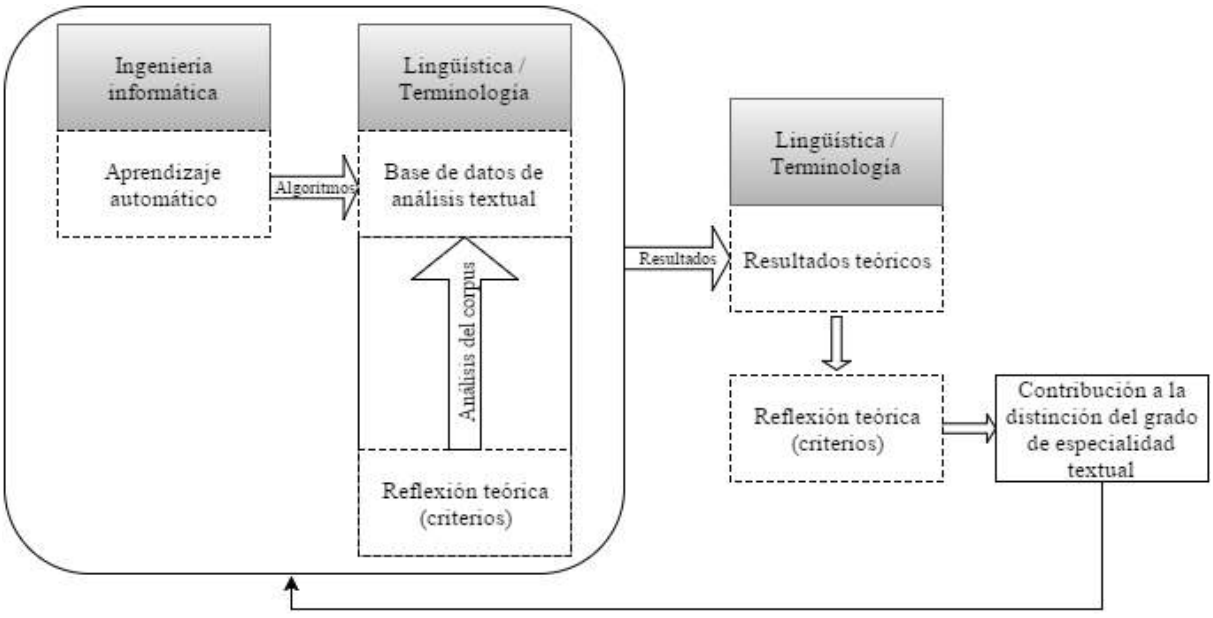


FIGURA 1. Contribución de las ciencias de la computación a la reflexión teórica de la lingüística. Fuente: elaboración propia

4.1 El aprendizaje automático

El aprendizaje automático es una rama de la inteligencia artificial dedicada al desarrollo de técnicas que permitan a los ordenadores aprender. Se entiende por *aprender* la capacidad de generalizar comportamientos a partir de información no estructurada suministrada como ejemplo. Es decir, se pretende que un ordenador sea capaz de reconocer patrones similares dentro de un conjunto determinado (Bishop, 2010). El aprendizaje automático tiene una amplia gama de aplicaciones científicas como el diagnóstico médico de una enfermedad, los movimientos de la bolsa financiera, el procesamiento del lenguaje natural, etc. (Witten & Frank, 2005). En el ámbito de la lingüística, el aprendizaje automático ha colaborado estrechamente, por ejemplo para medir la complejidad de las lenguas (Becerra-Bonache & Jiménez-López, 2014) o para categorizar de forma automática los textos (Sebastiani, 2002).

4.2 Tipos de algoritmos y enfoques

A partir de una serie de datos, utilizando algoritmos de aprendizaje automático, es posible interpretar la información para transformarla en conocimiento. Algunas de las formas de representación del conocimiento pueden ser las tablas de decisión, los árboles de decisión, las reglas de clasificación o los *clusters*, entre otros (Witten & Frank, 2005: 61-82).

(a) La forma más rudimentaria de representar el conocimiento obtenido a partir de un problema es mediante el uso de las *tablas de decisión* (Tabla 9). Estas tablas se construyen utilizando los patrones por filas y los atributos por columnas. Para minimizar el tamaño de estas tablas se suelen eliminar atributos que no sean suficientemente relevantes para la decisión.

	Atributo 1 *Valor especializado*	Atributo 2 *Relación discursiva*	Atributo 2 *Función lingüística*	Decisión *Clase*
Patrón 1 *Texto 1*				
Patrón 2 *Texto 2*				
Patrón n *Texto n*				

TABLA 9. Ejemplo de tabla de decisión. Fuente: elaboración propia

(b) Otra forma, algo más elaborada, son los *árboles de decisión*. Basta con seleccionar un patrón que se desee clasificar y recorrer el árbol por sus

ramas hasta llegar a la hoja final, es decir, la clase asignada a dicho patrón. En cada nodo, o intersección de ramas, se evalúan uno o varios atributos, bien con un valor fijo, bien atendiendo a una función establecida (véase Figura 2).

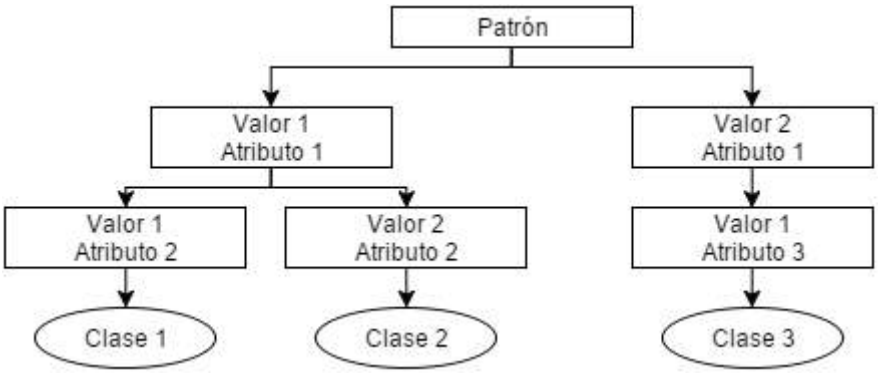


FIGURA 2. Ejemplo de árbol de decisión. Fuente: elaboración propia

(c) Otra forma de representar el conocimiento consistiría en el uso de *clusters* o agrupaciones, donde se concentran patrones cuyas características son comunes (Véase Figura 3).

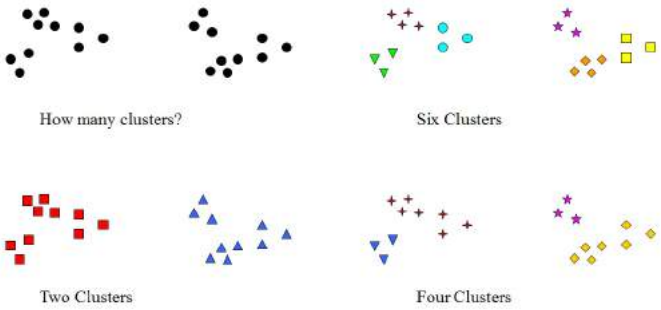


FIGURA 3. Ejemplo de agrupamiento en diferente número de clusters
Fuente: Pandre (s. f.)

Para generar conocimiento a partir de datos, es necesaria la aplicación de algoritmos de aprendizaje automático, que pueden clasificarse en supervisados o no supervisados, en función del uso de la etiqueta de *clase* para realizar el aprendizaje. Las diferencias principales y los resultados de aplicar dichos algoritmos a nuestro estudio se detallan a continuación.

5. PRIMEROS RESULTADOS: RESULTADOS DEL APRENDIZAJE NO SUPERVISADO

La primera aproximación al análisis de clases según el grado de especialización empleando métodos de aprendizaje automático no supervisado

(Rodríguez-Tapia & Camacho-Cañamón, en preparación) permitió trabajar con tres agrupaciones de patrones (siguiendo la hipótesis del contínuum). El corpus empleado fue el mismo que en los trabajos de Rodríguez-Tapia (2015 y 2016b) y los atributos sí corresponden con los de esta investigación. El primer análisis de los tres *clusters* propuestos por *k-means,* presentaba grupos lo suficientemente homogéneos como para defender la existencia de clases diferenciadas por cada *cluster.*

No obstante, hallamos una agrupación indeterminada: la perteneciente a la clase TSE. La mayoría de los textos de tipo TSE fueron agrupados en un único *cluster* pero ciertos patrones se infiltraron en los grupos donde se concentraban principalmente los de tipo TE y TD respectivamente. También ocurre esto con los textos TE y TD: quedan agrupados en clusters individuales pero ciertos patrones aparecen agrupados en el cluster correspondiente al TSE. A la luz de estos datos, se concluyó que la clase TSE cuenta con unos límites más difusos de lo que se consideró *a priori*.

Con este primer experimento se comprobaba si las agrupaciones realizadas por *k-means* y las clases propuestas manualmente por el experto coincidían. Tras la agrupación de los textos por el algoritmo, se ha revisado la clase que predomina en cada grupo, de forma que pudiese ser la clase *representante* del *cluster*. Así, es posible analizar qué textos han sido vinculados con un *cluster* cuya clase predominante coincide o no con la clase del texto (véase Figura 4).

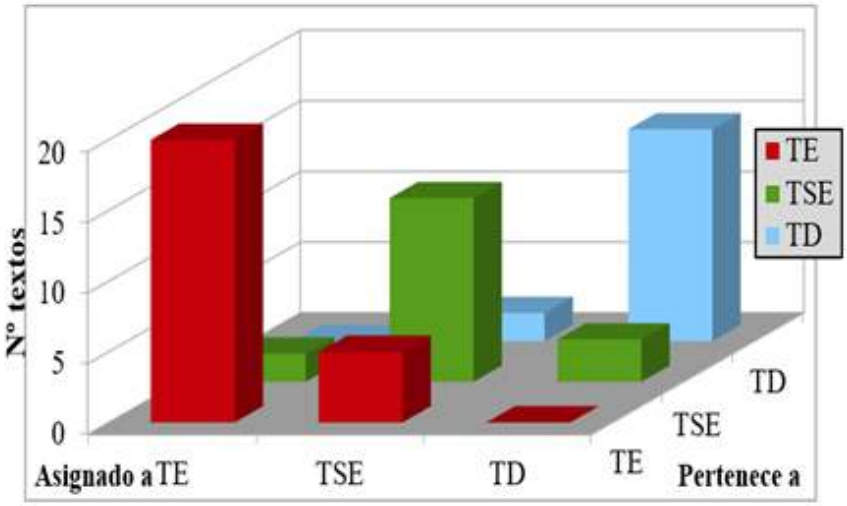


FIGURA 4. Representación de los *clusters* generados por el algoritmo *k-means* con las etiquetas de clase. Fuente: elaboración propia

6. RESULTADOS CON ALGORITMOS SUPERVISADOS

Las técnicas de aprendizaje automático supervisado utilizan el resultado esperado de clasificación como atributo implicado en el aprendizaje. Este resultado debe ser facilitado por el experto en la clasificación de patrones del problema (Marsland, 2015: 7-8). En otras palabras, los algoritmos supervisados usan la clase que facilita el experto en la base de datos como un atributo con el que trabaja el algoritmo.

Para el estudio de los resultados obtenidos por los distintos algoritmos se ha utilizado un método de validación cruzada (*cross-validation*) en la ejecución de los algoritmos (Witten & Frank, 2005: 149-152). En concreto, se ha empleado el método *k-fold*, que consiste en la división de la totalidad de los datos en *k* grupos de igual tamaño. Tras esto, se han realizado *k* ejecuciones de cada algoritmo donde se ha entrenado con *k*-1 de los grupos y se ha generalizado con uno de los *k* grupos. En cada una de las *k* ejecuciones, se ha ido variando el grupo elegido para generalizar, y al concluir los experimentos se ha realizado una media aritmética de los resultados obtenidos en las *k* ejecuciones. En nuestro caso *k* ha tomado el valor de 10.

A continuación, presentamos la Tabla 11 con los resultados de los algoritmos utilizados. En ella, se puede observar el porcentaje de acierto y de error en la clasificación textual de acuerdo con los valores de los atributos proporcionados en la base de datos.

	Textos bien clasificados		Textos mal clasificados	
Función logística	51	85 %	9	15 %
Naive Bayes	52	86,67 %	8	13,33 %
Red bayesiana	53	88,33 %	7	11,67 %
Perceptrón multicapa	53	88,33 %	7	11,67 %
Red RBF	54	90 %	6	10 %
Clasificación con regresión usando árbol M5P	54	90 %	6	10 %
Clasificador ordinal	54	90 %	6	10 %
Best-First Decision Tree	54	90 %	6	10 %
J48 pruned tree	57	95 %	3	5 %

TABLA 11. Resultados de los algoritmos supervisados. Fuente: elaboración propia

Destaca por encima del resto el último algoritmo (*J48 pruned tree*, de aquí en adelante: algoritmo J48) que ha logrado clasificar los textos

satisfactoriamente en un 95 % de los casos. A continuación, se ofrece la Figura 5, que muestra de forma gráfica los resultados del algoritmo:

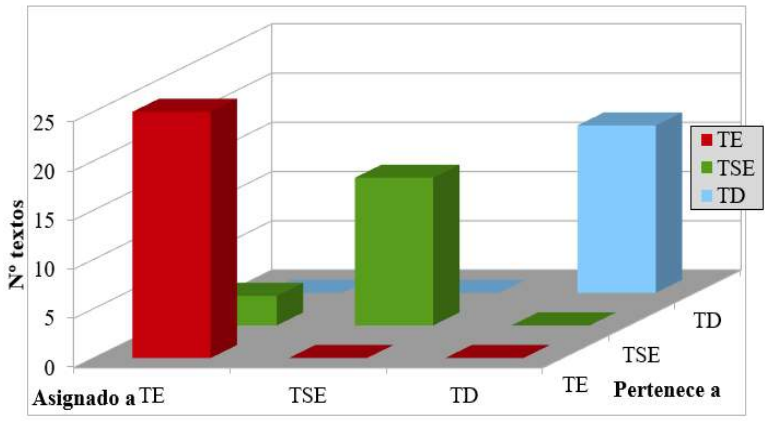
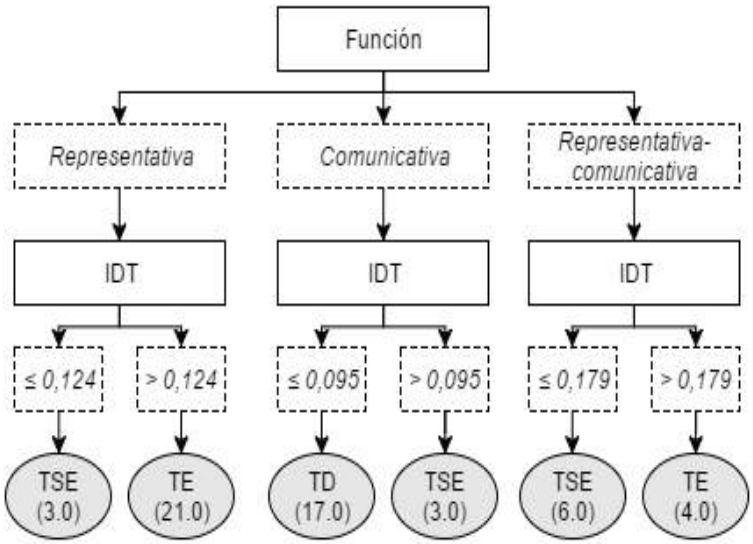


FIGURA 5. Resultado de la distribución de *clusters* usando el algoritmo J48. Fuente: elaboración propia

En el esquema de árbol con poda que muestra la clasificación de dichos textos se puede comprobar cómo la función es el factor determinante en la mayoría de los casos y que el índice de densidad terminológica es el segundo factor discriminador[4]. De esta forma, el árbol se interpretaría de la siguiente forma: un texto cuya función sea comunicativa, será incluido dentro de la clase TD si su IDT es igual o menor que 0,095 y dentro de la clase TSE si presenta un IDT mayor que 0,095 (véase Figura 6).



[4] Lo que concuerda con el método empleado para clasificar los textos: recuérdese que la función textual es el atributo con más peso en la división por porcentajes.

FIGURA 6. Árbol de decisión resultado de usar el algoritmo supervisado J48. Fuente: elaboración propia

Aun contando con dichos aciertos, existen tres textos mal clasificados que convendría analizar (véase Tabla 12):

Texto n.º	7	8	37
IDT	0,172	0,19	0,122
Relación discursiva	Especialista-lego	Especialista-lego	Especialista-especialista
Función	Representativa-comunicativa	Comunicativa	Representativa
Superestructura	Artículo divulgativo	Guía para el paciente	Carta al editor
Clase según el algoritmo	TE	TE	TE
Clase según el experto	TSE	TSE	TSE

TABLA 12. Textos mal clasificados usando J48. Fuente: elaboración propia

Cabe recordar que los sistemas de aprendizaje automático aprenden a realizar el trabajo de la forma más próxima posible a cómo lo realizaría un experto (en este caso, el lingüista), sin valorar si la clasificación proporcionada por el experto es acertada o no. Ante esta clasificación errónea, es posible encontrar dos justificaciones: a) que, efectivamente, el algoritmo haya clasificado incorrectamente los textos; o b) que el experto no haya clasificado adecuadamente los textos.

En el segundo caso, conviene revisar el método de clasificación empleado o contar con un grupo de control que (in)valide la propuesta del primer experto, como podría ser un conjunto de expertos que revise dichos textos y los reclasifique.

A continuación, proponemos diferentes razones por las cuales el sistema ha errado en la clasificación:

(a) En el texto 7: el sistema indica que cuando el texto tiene función representativa-comunicativa, depende del IDT. Si este es mayor que 0,179, se clasifica como TE, si no, como TSE. Aun cuando el texto 7 tiene un IDT de 0,172, ha sido clasificado en la clase TE, por lo que, al no encontrar mayor justificación, consideramos que se trata de un error del sistema.

(b) En el texto 8: el sistema indica que cuando el texto tiene función comunicativa, depende del IDT. Si este es mayor que 0,095, se clasifica como TSE, si no, como TD. El texto ha sido clasificado dentro

del grupo de TE por poseer un IDT con un valor considerablemente alto (0,19) para un TD.

(c) En el texto 37: el texto ha sido clasificado como TE, aun cuando por su IDT (0,122) y función debería haber sido clasificado como TSE. Esta clasificación por parte del algoritmo se justifica teniendo en cuenta que los atributos del texto 37 son más cercanos al TE que al TSE. De hecho, lo cierto es que, por los valores de sus atributos, cabría reconsiderar su clase como TE (coincidiendo precisamente con lo que propone el sistema automático). Su clasificación dentro de la clase TSE solo se justifica por el método numérico que planteamos, lo que invita a considerar de nuevo que los límites numéricos conllevan numerosos problemas al discriminar algunos patrones.

Como puede comprobarse, los errores de clasificación tienen al TSE como protagonista, lo que es justificable desde el punto de vista del contínuum teórico, así como metodológico. Por ejemplo, en el caso del texto 37, queda clasificado como TSE de acuerdo con el valor resultante del método de clasificación empleado (es decir, se considera TSE porque su valor numérico es un 9,1), siendo 11 el límite para considerarse TE. De esta forma, podemos demostrar con este sistema la rigidez numérica de los límites (Rodríguez-Tapia, 2016a: 1002) y la restricción dentro del abanico de posibilidades (2016b: 247).

La alta tasa de acierto del algoritmo supervisado J48 (recordemos que se trata de un 95 %) puede deberse a que el propio experto utiliza un conjunto de restricciones para crear de forma manual un árbol (es decir, el método numérico expuesto en Rodríguez-Tapia, 2016a). De este modo, el algoritmo intenta generar un modelo muy próximo al elaborado por el experto. De hecho, si los patrones de la base de datos fuesen más numerosos, el algoritmo J48 llegaría a un 100 % de acierto, pues conseguiría simular el mismo modelo matemático que se ha utilizado para clasificar los textos (que, recordemos, diferencia la relevancia de los atributos a través de diferentes porcentajes).

7. CONCLUSIONES Y LÍNEAS FUTURAS

En primer lugar, cabe subrayar que con este trabajo ha sido posible demostrar el carácter teleológico de las ciencias de la computación para evaluar la metodología de análisis propuesta. En otras palabras, los métodos empleados

por las ciencias de la computación permiten a la lingüística trabajar con eficacia sobre los fundamentos teóricos de los métodos que propone para analizar sus objetos de estudio. En concreto, la interpretación de los resultados nos lleva a proyectar conclusiones que afectan al valor de los algoritmos empleados, por una parte, y, a partir de estos, conclusiones sobre el método de clasificación propuesto en Rodríguez-Tapia (2016a).

Así, en relación con los métodos de aprendizaje automático, ha sido posible comprobar cómo el algoritmo no supervisado *k-means*, basándose en los atributos que proponemos, ha distinguido tres grupos de textos que se aproximan estrechamente a los tres propuestos en nuestra hipótesis del contínuum. Por otra parte, el algoritmo supervisado J48, un árbol de decisión, permite elaborar modelos matemáticos que nos ayudan a esbozar la probabilidad de que un objeto textual se clasifique dentro de una clase (texto especializado, semiespecializado o divulgativo) según la relación que exista entre los valores de sus atributos. Debido a las fortalezas de los dos algoritmos mencionados, consideramos que los futuros trabajos deberían seguir explotando al máximo los resultados de los mismos. Para ello, habría que aumentar el número de textos analizados, es decir, habría que trabajar con un corpus mayor y aumentar el número de características de los textos para contar con nuevos datos que permitan ampliar las conclusiones. De igual forma, sería necesario volver a etiquetar cada uno de los textos que forman el corpus sin utilizar el mencionado método numérico. De hecho, beneficiaría al estudio la colaboración de varios expertos que etiqueten de forma independiente e individual cada uno de los textos. Así, gracias al aprendizaje automático, se podría conseguir un modelo que clasifique los textos de la forma más parecida posible a como lo harían los expertos.

Los resultados de los algoritmos con los que hemos trabajado hacen posible distinguir algunas de las debilidades del método numérico propuesto en Rodríguez-Tapia (2016a). Así, los problemas planteados sobre el método en Rodríguez-Tapia (2016b), es decir, el apriorismo de los valores porcentuales y los problemas que conlleva trabajar con límites numéricos, que, en cierta forma, anulan los límites difusos, han vuelto a ser confirmados en este trabajo. Estos problemas nos hacen reconsiderar trabajar con un método matemático para clasificar las clases de textos. Proponemos, por tanto, rechazar el uso de los

porcentajes de caracterización de los textos y de los límites numéricos de discriminación de clases.

Con todo, por otra parte, los atributos seleccionados y sus valores parecen que se constituyen como una vía útil de análisis. Por tanto, proponemos reflexionar en la elaboración de criterios teóricos suficientes para profundizar en dichos atributos, permitir la creación de nuevos valores para los mismos, que sean más descriptivos y abarcadores de la realidad del contínuum. Esta amplitud en los atributos podría conseguirse aumentando el número de valores de la relación discursiva (por ejemplo, distinguiendo seis en lugar de tres tipos de relación emisor-receptor), de tipos de texto o de funciones lingüísticas.

Otras de las líneas de investigación futuras que se desprenden de estas conclusiones afectan a la hipótesis de trabajo. Dado que, en definitiva, nuestro interés último se halla en caracterizar los niveles del contínuum y determinar la probabilidad de pertenencia a una determinada clase de un texto, según sus atributos, y debido a que el nivel semiespecializado es el que conlleva mayores dificultades teóricas y metodológicas, consideramos pertinente subdividir dicho nivel en dos, de forma que las clases con las que trabajemos en futuros trabajos sean cuatro. Así, proponemos distinguir las siguientes clases:

a) texto especializado

b) texto especializado con aproximación al texto divulgativo

c) texto divulgativo con aproximación al texto especializado

d) texto divulgativo.

Con esta aproximación y con un análisis adicional sobre la clasificación tripartita como llevamos trabajando tradicionalmente, sería interesante contrastar dichos datos para comprobar los grados de coincidencia y divergencia entre la clase *texto semiespecializado* y las nuevas clases esbozadas en b) y c), para lo cual podría contarse con dos nuevos algoritmos que presentan amplias expectativas en el análisis de resultados, como pueden ser los métodos de clasificación ordinal (Sáez *et al*. 2016) o los modelos de campo aleatorio condicional (Wallach, 2004). Además, una vez planteado este análisis contrastivo, sería muy recomendable conocer el *ideal* de cada clase con respecto a la caracterización de sus atributos, empleando para ello métodos de aprendizaje automático para la obtención de los centroides de cada clase.

BIBLIOGRAFÍA

Arntz, R. & Pitch, H. (1995). *Introducción a la terminología*. Madrid: Fundación Germán Sánchez Ruipérez.

Becerra-Bonache, L. & Jiménez-López, M. D. (2014). Un modelo de aprendizaje automático para medir la complejidad de las lenguas. En: *Congreso Internacional de la Asociación Española de Lingüística Aplicada*, 1-10. Recuperado el 26 de diciembre, 2017 de: http://cvc.cervantes.es/lengua/eaesla/pdf/01/53.pdf.

Bishop, C. M. (2010). *Pattern recognition and machine learning*. New York: Springer.

Cabré, M. T. (1993). *La terminología: teoría, metodología, aplicaciones*. Barcelona: Antártida/Empúries.

Cabré, M. T. (2002a). Textos especializados y unidades de conocimiento: metodología y tipologización. En: J. García Palacios y M. T. Fuentes Morán (Eds.), *Texto, terminología y traducción* (pp. 15-36). Salamanca: Almar.

Cabré, M. T. (2002b). Análisis textual y terminología, factores de activación de la competencia cognitiva en la traducción. En: A. Alcina Caudet y S. Gamero Pérez (Eds.), *La traducción científico-técnica y la terminología en la sociedad de la información* (pp. 87-105). Castelló de la Plana: Publicacions de la Universitat Jaume I.

Cabré, M. T. & Estopà, R. (2002). El conocimiento especializado y sus unidades de representación: diversidad cognitiva. *Sendebar*, 13, 141-153.

Cabré, M. T.; Da Cunha, I.; Sanjuan, E.; Torres-Moreno, J.-M. & Vivaldi. J. (2014). Automatic Specialized vs. Non-specialized Text Differentiation: The Usability of Grammatical Features in a Latin Multilingual Context. En E. Bárcena; T. Read & J. Arús (Eds.) *Languages for specific purposes in the digital era* (pp. 223-241). Cham/Heidelberg: Springer.

Ciapuscio, G. E. & Kuguel, I. (2002). Hacia una tipología del discurso especializado: aspectos teóricos y aplicados. J. García Palacios & M. T. Fuentes Morán (Eds.), *Texto, terminología y traducción* (pp. 37-73). Salamanca: Almar.

Ciapuscio, G. E.; Kuguel, I. & Otañi, I. (2005). El conocimiento especializado: el texto de especialización y los criterios para su tipologización. En M. T.

Cabré & C. Bach (Eds.) *Coneixement, llenguatge i discurs especialitzat* (pp. 95-110). Barcelona: IULA. Universitat Pompeu Fabra.

Dijk, T. (1992). *La ciencia del texto.* Barcelona: Paidós.

Domènech, Ona (2007). La noció de text especialitzat des de la perspectiva de la teoría comunicativa de la terminologia. En: M. Lorente; R. Estopà; J. Freixa; J. Martí & C. Tebé (Eds.) *Estudis de lingüística i de lingüística aplicada en honor de M. Teresa Cabré Castellví* (pp. 241-254). Barcelona: IULA. Universitat Pompeu Fabra.

Estopà, R. (1999). *Extracció de Terminologia. Elements per a La Construcció D'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. Tesis doctoral. Barcelona: IULA. Universitat Pompeu Fabra.

Guantiva, R; Cabré, M. T. & Castellà, J. M. (2008). Clasificación de Textos Especializados a Partir de Su Terminología. *Íkala. Revista de lenguaje y cultura* 13(19), 15-39. Recuperado el 26 de diciembre, 2017 de: http://seer.ufrgs.br/riterm/article/download/20712/11980.

Kohavi, R. & Provost, F (1998). Glossary of terms. *Machine Learning* 30(2-3), 271-274.

Lorente, M. (2014). Hacia la terminología 3.0: evolución del uso de las tecnologías en la terminología. En: C. Vargas (Ed.) *TIC, trabajo colaborativo e interacción en terminología y traducción* (pp. 9-26). Granada: Comares.

Marsland, S. (2015). *Machine learning: an algorithmic perspective*. Boca Raton: CRC.

Monterde Rey, A. M. (2002). *Ejercicios de Introducción a La Terminología*. Las Palmas de Gran Canaria: Universidad de Las Palmas de Gran Canaria.

Nazar, R. & Cabré, M. T. (2011). Un experimento de extracción de terminología utilizando algoritmos estadísticos supervisados. *Debate Terminológico*, 7, 36-55. Recuperado el 26 de diciembre, 2017 de: http://seer.ufrgs.br/riterm/article/download/20712/11980.

Pandre, A. (s. f). Cluster Analysis: see it 1st. *Data Visualization*. Recuperado el 26 de diciembre, 2017 de: https://apandre.wordpress.com/visible-data/cluster-analysis/.

Rodríguez-Tapia, S. (2015). Estrategias de traducción inglés-español basadas en el análisis cuantitativo de procedimientos de reformulación formal y

conceptual del texto semiespecializado. *Tonos digital: Revista electrónica de estudios filológicos*, 29(2), 1-34. Recuperado el 26 de diciembre, 2017 de: http://www.tonosdigital.com/ojs/index.php/tonos/article/viewFile/1330/805.

Rodríguez-Tapia, S. (2016a). Los textos especializados, semiespecializados y divulgativos: una propuesta de análisis cualitativo y de clasificación cuantitativa. *Signa: Revista de la Asociación Española de Semiótica*, 25, 987-1006. Recuperado el 26 de diciembre, 2017 de: http://revistas.uned.es/index.php/signa/article/view/16926/14512.

Rodríguez-Tapia, S. (2016b). Clasificación cuantitativa de los textos según su grado de especialidad: parámetros para la elaboración de los índices de densidad terminológica y de reformulación de un corpus sobre insuficiencia cardíaca. *Anuario de Estudios Filológicos*, 39, 227-250.

Rodríguez-Tapia, S. & Camacho-Cañamón, J. (en prensa). La contribución de los métodos de aprendizaje automático no supervisado a la clasificación textual según el grado de especialización, *Sintagma*.

Sáez, A., Sánchez-Monedero, J., Gutiérrez: A., & Hervás-Martínez, C. (2016). Machine Learning Methods for Binary and Multiclass Classification of Melanoma Thickness From Dermoscopic Images. *IEEE transactions on medical imaging*, 35(4), 1036-1045.

Sager, J. C.; Dungworth, D. & Mcdonald, P. F. (1980). *English special languages: principles and practice in science and technology*. Wiesbaden: Brandstetter.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys* 34(1), 1-47. Recuperado el 26 de diciembre, 2017 de: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf.

Vivaldi, J. & Rodríguez, H. (2012). Using Wikipedia for Domain Terms Extraction. En: T. Gornostay (Ed.): *Proceedings of CHAT 2012: The 2nd Workshop on the Creation, Harmonization and Application of Terminology Resources: co-located with TKE 2012* (pp. 3-10). Universidad Politécnica de Madrid. Linköping, Sweden: Linköping University Electronic Press, Linköpings Universitet.

Vivaldi, J. (2001). *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Tesis doctoral. Barcelona: IULA. Universitat Pompeu Fabra.

Vivaldi, J. (2009). Corpus and exploitation tool: IULACT and bwanaNet. En: P. Cantos & A. Sánchez (Ed.) *A survey on corpus-based research = Panorama de investigaciones basadas en corpus (Actas del I Congreso Internacional de Lingüística de Corpus (CICL-09)* (pp. 224-239). Murcia: Asociación Española de Lingüística del Corpus. pp. 224-239.

Wallach, H. (2004). *Conditional random fields: An introduction*. Technical Report MS-CIS-04-21. Pennsylvania: University of Pennsylvania.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Amsterdam: Morgan Kaufmann.