Algodoncillo, Sucre 1995
Fotografía, 1937 x 2566 pixeles

Clasificación de textos especializados a partir de su terminología*[1]

Mg. Ricardo Guantiva Acosta**
Dr. M. Teresa Cabré Castellví***
Dr. Josep M. Castellà Lidon****

El artículo clasifica textos en niveles de especialización, a partir de la tipología lingüística de las unidades terminológicas y su densidad dentro del texto especializado. Para ello revisamos algunos conceptos básicos de la lingüística textual en el marco de la comunicación especializada (texto general y texto especializado), para establecer las diferencias de registros comunicativos. También revisamos la noción de *unidad terminológica,* tomando como marco la teoría comunicativa de la terminología (TCT). En cuanto a los materiales empíricos, trabajamos con un corpus textual en relación con el tema "genoma humano", clasificado en niveles de especialización. Asimismo, presentamos, de manera muy sucinta, el tratamiento informático de los textos, para hacer la extracción y detección de la terminología contenida en ellos y, mediante el uso de un programa estadístico, presentamos los resultados discriminantes para la clasificación de textos en niveles de especialización.

Palabras clave: textos especializados, terminología, lingüística textual, unidades terminológicas, comunicación especializada

This article deals with the classification of specialized texts based on the linguistic typology of terminology units and their density within the texts. With this in mind, we review some basic concepts of text linguistics in the context of specialized communication (general texts and specialized texts) in order to establish the differences between communicative registers. We also examine the concept of terminology unit, using the Communicative Theory of Terminology (CTT). In terms of the materials used, we worked with a text related to the topic of "human genome", classified into different levels of specialization. To this respect we also offer a brief description of the computer analysis carried out in the extraction and detection of the terminology units within the texts, and we present the results, classifying the texts into different levels of specialization.

Key words: specialized texts, terminology, textual linguistics, terminology units, specialized communication

Cet article classe des textes en niveaux de spécialisation selon la typologie linguistique des unités terminologiques et de leur densité dans le texte spécialisé. C'est pourquoi nous revenons sur certains concepts fondamentaux de la linguistique textuelle dans le cadre de la communication spécialisée

* Recibido: 19-11-07 / Aceptado: 01-03-08

1 Proyecto de tesis doctoral dirigida por la doctora M. Teresa Cabré Castellví y por el doctor Josep M. Castellà Lidon, del Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra en Barcelona, España.

(texte général et texte spécialisé), afin d'établir les différences de registres communicatifs. Nous revoyons également la notion d'unité terminologique dans le cadre de la Théorie Communicative de la Terminologie (TCT). En ce qui concerne le matériel empirique, nous travaillons à partir d'un corpus textuel sur la question du génome humain du point de vue des niveaux de spécialisation. Nous proposons également une brève présentation du traitement de texte informatique pour en extraire et y détecter la terminologie, et grâce à l'utilisation d'un programme statistique, nous présentons les résultats discriminatoires pour la classification de textes en niveaux de spécialisation.

Mots clés: textes spécialisés, terminologie, linguistique textuelle, unités terminologiques, communication spécialisée

1. Introducción

La comunicación especializada o los discursos profesionales reflejan el vasto desarrollo de la sociedad actual. Prueba de ello son los avances o desarrollos tanto en la ciencia como en la tecnología a partir de la proliferación de publicaciones diversas con un alto uso predominante de terminología específica. Situarse en cualquiera de estos dos campos implica informarse sobre dichos avances, lo que representa, igualmente, conocer y dominar las distintas denominaciones que pueden darse a un mismo objeto o ente de conocimiento. Así pues, para que un proceso comunicativo especializado cumpla su propósito de transmitir información o conocimiento especializado, hace falta que los interlocutores conozcan de antemano cuáles son las condiciones necesarias para que el intercambio de referentes de conocimiento sea efectivo.

Los factores relevantes como el emisor, el destinatario, el tema, la situación comunicativa, entre otros, llevan a establecer tipos de *comunicación especializada,* entendida como el proceso mediante el cual se imparte o comparte conocimiento especializado; por ejemplo, la comunicación que se da entre un par de especialistas en algún campo de conocimiento no implica mayor esfuerzo cognitivo en cuanto al uso de terminología específica, caso contrario cuando se quiere hacer entre un especialista y el público en general.

Por tanto, la transmisión del conocimiento especializado admite una gradación, susceptible de ser descrita mediante las unidades lingüísticas que intervienen en el texto. La comunicación especializada responde a la propiedad de la adecuación, en función de la cual las condiciones de producción y recepción de los textos vienen dadas por el contexto comunicativo (estrategias comunicativas), lo que representa la producción de textos especializados como el medio natural de la terminología, de acuerdo con las condiciones pragmáticas que impulsan la producción de este tipo de textos.

El objeto de análisis de este artículo son los diferentes niveles de especialización textual mediante los cuales puede transmitirse el conocimiento. Esta distribución de niveles de especialización es lo que se conoce como *variación vertical del*

discurso (Hoffmann, 1998: 62) y viene determinada tanto por la cantidad de información compartida entre los interlocutores como por la finalidad del texto. De ahí la distinción entre: 1) el discurso especializado (con distintos grados de especialización), dirigido a los especialistas; 2) el discurso de tipo didáctico, orientado al personal en formación, y 3) el discurso de carácter divulgativo, enfocado al público general.

Estos distintos niveles de competencia reflejan el mayor o menor uso de terminología y la forma como ella recorre los textos. De este modo, pretendemos observar de qué manera la comunicación especializada requiere que la terminología se adapte a cada uno de estos niveles de especialización discursiva.

Para lograr este objetivo propuesto, partimos del supuesto de que el cálculo de la densidad terminológica debe tener en cuenta tanto la frecuencia de las unidades terminológicas como su naturaleza lingüística. Consideramos que estos factores permiten diferenciar y, por tanto, clasificar textos, mediante niveles de especialización, dado que, en la configuración del conocimiento de un ámbito o campo de especialidad, el léxico contribuye a la caracterización de textos.

Así pues, esperamos que el resultado de este análisis permita establecer las diferencias entre textos de diferentes niveles de especialización, lo que representaría un mejor conocimiento de la génesis y la articulación de los textos especializados como variedades funcionales o registros (variación vertical).

En este artículo revisamos algunos conceptos básicos de cara a la comprensión de la producción textual en el marco de la comunicación especializada. Por un lado, repasamos las nociones de *texto general* y *texto especializado,* para establecer las diferencias de registros comunicativos; y, por otro, la noción de *unidad terminológica,* tomando como marco la teoría comunicativa de la terminología (TCT) (Cabré, 1999: 2003). En cuanto a los materiales empíricos, trabajamos con un corpus textual en relación con el tema *genoma humano,* clasificado en niveles de especialización. Asimismo, presentamos, de manera muy sucinta, el tratamiento informático de los textos, para hacer la extracción y detección de la terminología contenida en ellos y, mediante el uso de un programa estadístico, mostramos los resultados discriminantes para la clasificación de textos en niveles de especialización.

2. Conceptos básicos

2.1. El texto general

La noción de *texto* ha sido discutida ampliamente en el ámbito de la lingüística textual. El texto es una unidad comunicativa, resultado de la actividad lingüística mediante la cual se transfieren significados. Dado su carácter pragmático, el emisor de un texto tiene en cuenta la intención comunicativa circunscrita a un contexto de situación y producción (Bernárdez, 1982: 82; Castellà, 1992: 49-53; Van Dijk, 1980: 9-17; 1989: 13-30; 1993: 29-46; Eggins y Martin, 2000: 335-370).

Para Beaugrande y Dressler,

> […] el texto es un acontecimiento comunicativo que cumple siete normas de textualidad o principios constitutivos: a) cohesión: dependencias gramaticales, b) coherencia: relaciones conceptuales, c) intencionalidad y modalidad: la actitud del productor del texto, d) aceptabilidad: la actitud del que recibe el texto, e) informatividad, f) situacionalidad y g) intertextualidad (1997: 35-47).

Ciapuscio propone que los textos deben considerarse como

> […] recursos primordiales de constitución del conocimiento, cuyas formas, estructuras, contenidos, funcionalidades, etc., responden a desarrollos socio-culturales y, por tanto, más allá de los sistemas lingüísticos particulares, pueden exhibir en los niveles más globales (géneros, estilos, etc.) características específicas debidas a la sociedad o comunidad particular en que son construidos y empleados (2003: 23).

Las nociones de *texto* nos dejan entrever que trabajamos con unidades comunicativas de carácter lingüístico. Éstas son herramientas clave de todo proceso de interacción social en el que convergen significados, intenciones y situaciones comunicativas concretas, para representar el conocimiento general de una sociedad.

2.2. El texto especializado

La noción de *texto especializado* lleva a la distinción entre el *lenguaje general* y *lenguaje especializado.* Con el primer término se alude al conjunto de reglas

lingüísticas que utilizan los individuos para comunicarse y, con el segundo, al conjunto de todos los recursos lingüísticos que se utilizan en un ámbito comunicativo (la temática, el tipo de interlocutores, la situación comunicativa, la intención del emisor, etc.) para garantizar la comprensión entre las personas que trabajan en este ámbito (Kocourek, 1982; Arntz y Picht, 1995; Sager, 1993; Schröder, 1991; Hoffmann, 1998a; Cabré, 1999; Ciapuscio, 2003).

Para Baldinger

> […] los límites entre el lenguaje especializado y la lengua común son fluidos tanto desde el punto de vista individual, social y geográficos, como desde el general: esto significa que entre los tres círculos se produce un constate proceso de compensación en ambas direcciones puesto que los tres círculos pueden coincidir en un mismo individuo (1952: 90, citado en Arntz y Picht, 1995: 30).

Para Kocourek (1982: 10), un texto especializado tiene una doble lectura: por un lado, la perspectiva *semiótica* considera el lenguaje especializado como un sistema de transmisión e intercambio de información que emplea diversos códigos simultáneamente; y por otro, la perspectiva *lingüística* centra la atención en la prioridad comunicativa del *lenguaje.* Desde esta perspectiva,

> […] les textes spécialisés sont-ils l'expression concrète de la connaissance approfondie dans le domaine, ils signifient cette connaissance; leur contenu c'est la connaissance approfondie que les spécialistes peuvent communiquer au public restreint de leurs pairs. Signification c'est communication virtuelle, communication c'est transmission de la connaissance signifiée (Kocourek, 1982: 33).

Para la teoría de los sublenguajes

> […] el texto especializado es el instrumento o el resultado de la actividad comunicativa ejercida en relación con una actividad socioproductiva especializada (Hoffmann, 1998b: 77).

El texto especializado se caracteriza a partir de la unidad estructural, que constituye el conjunto de sus oraciones coherentes de tipo pragmático, sintáctico y semántico. Por tanto, dada su estructura comunicativa compleja, el texto está sujeto a factores como: 1) la intención comunicativa del autor y la estrategia comunicativa que se

deriva de ella; 2) el destinatario con actitud expectante, y 3) el léxico, determinado por el uso específico de terminología que puede presentar grados de opacidad, de especialización y variación expresiva (Cabré, 1999: 156-159).

Para Ciapuscio

> […] los textos especializados son productos predominantemente verbales de registros comunicativos específicos, registros que son definidos por los usuarios de los textos, las finalidades y las temáticas. Los textos especiales se refieren a temáticas propias de un dominio de especialidad y responden a convenciones y tradiciones retóricas específicas. Los factores funcionales, situacionales y temáticos tienen su correlato en el nivel de la forma lingüística, tanto en la sintaxis como en el léxico (2003: 30).

Las propuestas anteriores coinciden en señalar que el texto especializado es un producto de naturaleza lingüística, que combina simultáneamente códigos diversos para representar la realidad de un ámbito de especialidad. Esta representación está sujeta a factores sociopragmáticos que enmarcan al texto e inciden en su tipología. Algunos de estos factores son: 1) la temática; 2) el tipo de interlocutor, y 3) el tipo de situación comunicativa, condicionada por el tipo de registro empleado.

El texto especializado, como "registro comunicativo específico", puede articular su contenido de formas condicionadas por los factores sociopragmáticos, lo que conduce a la variación discursiva.

2.3. La variación discursiva

La noción de *variación* de un texto (general o especializado) está relacionada con la propiedad de *adecuación,* por la cual el texto se acopla al contexto. El *contexto* es el campo de cultivo a partir del cual se genera el texto. En este proceso de creación intervienen factores funcionales, situacionales y temáticos, que generan registros, tipos —o clases de textos necesarios para la comunicación— y géneros textuales.

Registro, tipo y género textuales

La noción de *registro* permite desarrollar, para este trabajo, la noción de *variación vertical* en el texto especializado, pues ésta

> […] es a la vez muy simple y muy importante: se refiere al hecho de que la lengua que hablamos o escribimos varía de acuerdo con el tipo de situación. [...] La noción de registro constituye así una forma de predicción: dado que conocemos la situación (el contexto social de utilización del lenguaje), podemos predecir mucho respecto del lenguaje que se producirá con probabilidades razonables de estar en lo cierto (Halliday, 1994: 46-47).

Los registros responden a formas lingüísticas determinadas, condicionadas por funciones comunicativas concretas. Cada una de las características de los registros admiten una gradación con límites imprecisos, lo que origina que un texto pueda ser más o menos específico o más o menos formal. Los registros orientan tanto al emisor en la producción de los textos (adecuación al medio y a su interlocutor), como al receptor (creación de conjunto de expectativas en función del tipo de texto que está dispuesto a interpretar). Estos factores intervienen en las condiciones de producción del texto y en la configuración de su forma lingüística final.

Los elementos de la realidad contextual que conforman los registros pueden agruparse en tres apartados:

1. El *campo* (tema): se refiere al marco institucional en el que se produce un trozo de lenguaje e incluye no sólo el tema de que se trata, sino también la actividad del hablante o del participante en un determinado marco. El campo permite, para nuestro caso, la distinción entre temática general (campos diversos de estudio) y temática especializada (grados o niveles de abstracción o especialización con un único campo).
2. El *tenor* (la relación interpersonal entre los interlocutores): permite la relación emisor-receptor, a partir del grado de formalidad o de especialización textual y del propósito o intención del autor del texto, de acuerdo con un objetivo determinado: informar, argumentar, etc.
3. El *modo* (el canal de producción, transmisión y recepción del texto): establece el medio a través del cual se emite, se transmite y se recibe el mensaje, y también determina las condiciones de producción de los textos (Halliday, 1994: 49).

En cuanto a las nociones de *tipo* y *género discursivos,* Ciapuscio plantea que

> […] el término "género discursivo" ha perdurado especialmente en la lingüística francesa. […] la revista *Langue Française* […] distingue explícitamente entre

> *géneros* y *tipos discursivos*. El género discursivo se relaciona con una dimensión histórico-cultural más general, que incluye la competencia sobre tipos discursivos; estos últimos hacen referencia a una dimensión estrictamente lingüística.
>
> En el marco de la lingüística del texto, los términos que se utilizan son *tipo* y *clase textual*. Heinemann y Viehweger (1991) subrayan el gran consenso que se comprueba en las publicaciones lingüístico-textuales: clase textual se aplica hoy a estratificaciones empíricas, tal cual son realizadas por los miembros de una comunidad lingüística, es decir, clasificaciones cotidianas que pueden mencionarse por medio de determinados lexemas condensadores del saber sobre una determinada clase textual: por ejemplo, "esto es un *cuento",* "esto es un *chiste",* "esta es una *descripción",* "esto es un *diálogo",* etc. Por el contrario, *tipo textual* se concibe como una categoría ligada a una teoría para la clasificación científica de los textos. Por lo tanto, los hablantes de una comunidad tienen un saber sobre las clases textuales o un saber sobre las estructuras globales, pero no un saber sobre los tipos textuales (1994: 25).

Así, pues, asumimos que el tipo textual alude, principalmente, a la estructura organizativa del texto, siguiendo una clasificación teórica, y que el género textual se refiere a la representación de la competencia lingüística de los hablantes, en términos de producción de textos.

La noción de variación vertical

La variación en el texto, referida a los elementos de registro (campo, tenor y modo), se establece a partir de dos tipos: 1) la *variación horizontal:* determinada por la temática, y 2) la *variación vertical:* determinada por el grado o nivel de especialización (Hoffmann, 1998a: 62-69).

El elemento diferenciador de la variación vertical es el léxico, representado por un número mayor de unidades terminológicas. Este uso específico de léxico especializado permite la precisión en el lenguaje a partir del perfeccionamiento del conocimiento, lo que conlleva a establecer niveles de comunicación especializada, desde el plano concreto hasta el abstracto, desde el plano particular hasta el general.

De acuerdo con la propuesta de Wichter (1994, citado por Ciapuscio, 2003: 36-37) sobre lexicología de la verticalidad, el *léxico* es el bagaje o la suma de conocimientos que atesora una comunidad. Dentro de esta escala imaginaria que

supone la distribución vertical del léxico, el punto más alto de la verticalidad lo ocupan los expertos en una disciplina específica, y el punto más bajo, los legos. El objetivo de la lexicología vertical consiste en describir, analizar y explicar el vocabulario de expertos y legos, a partir de la distribución en los distintos niveles de especialización.

La variación vertical del léxico permite una aproximación a las competencias comunicacionales de los diferentes tipos de usuario. Desde esta perspectiva, personas y textos son fuentes potenciales para la investigación de la lexicología vertical.

3. El léxico especializado

La noción y tipos de *unidades terminológicas,* tomando como marco teórico la teoría comunicativa de la terminología (TCT) (Cabré, 1999: 2003), y la noción de *densidad léxica* (*densidad terminológica* para este trabajo) propuesta por Halliday (1987, citado por Martín, 2003: 160), quien la ha definido como la "proporción de elementos léxicos con relación a la totalidad del discurso", permiten el análisis del texto especializado.

3.1. La unidad terminológica

El estatus científico de la terminología proviene de una larga tradición sobre estudios basados en datos terminológicos pertenecientes a distintos ámbitos del conocimiento. El referente histórico clásico sobre los inicios de la terminología es el trabajo de Wüster (1998), conocido como la teoría general de la terminología (TGT). Para éste, el objeto de la terminología son los conceptos y las relaciones entre ellos. No obstante, dada la repercusión que ha tenido la TGT,

> […] las posiciones críticas no la invalidan como teoría, sino que simplemente subrayan su limitación conceptual y funcional y su falta de generalización, lo que la hacen devenir insuficiente para explicar las unidades terminológicas en toda amplitud (Cabré, 1999: 114).

Bajo estas condiciones, se reorienta la teoría de la terminología hacia la consolidación de un nuevo modelo teórico más flexible y de carácter multidisciplinar:

la *teoría comunicativa de la terminología* (TCT) o *teoría de las puertas* (Cabré, 1999, 2002, 2003). El objetivo principal de la TCT es la descripción y explicación de las unidades terminológicas *in vivo,* es decir, la descripción real de los términos en sus contextos naturales de aparición: los textos especializados.

Para llevar a cabo este trabajo, la TCT incorpora un análisis multidimensional de la unidad terminológica, que integra tres enfoques: el *cognitivo* (el concepto), el *lingüístico* (el término) y el *comunicativo* (la situación), bajo el supuesto de que la naturaleza de estas unidades, así como su función, comparte muchas de las características de las unidades de la lengua general.

Las *unidades terminológicas* o *términos* designan los conceptos propios de un ámbito de conocimiento o de especialidad. El tratamiento que reciben estas unidades se determina a partir de los factores externos e internos de la comunicación especializada, lo que posibilita su mayor o menor presencia en el discurso especializado. El carácter *poliédrico* de las unidades terminológicas comporta un doble valor significativo para la terminología: por un lado, son "unidades léxicas" o palabras, en tanto forman parte del lenguaje natural, y por el otro, son "unidades terminológicas" que, bajo las condiciones semánticas y pragmáticas, obligan a la especificación entre estos dos tipos de unidades.

El conjunto de unidades especializadas está determinado por: 1) las *unidades de conocimiento especializado* (UCE), unidades de distinto nivel descriptivo que transmiten un tipo de conocimiento específico (condiciones cognitivas y semánticas), bajo un uso restringido en el discurso (condiciones pragmático-discursivas); y 2) las *unidades terminológicas* (UT), unidades léxicas con un valor semántico específico (valor especializado), asociado a un ámbito que consolida la estructura conceptual del dominio del que forman parte (Cabré y Estopà, 2005: 78-83).

Así, pues, siguiendo a Cabré y Estopà, las UCE y, en concreto, las UT, pueden clasificarse de acuerdo con diferentes criterios:

1. El *sistema al que pertenecen:* las UCE o UT pueden ser unidades del lenguaje natural o pertenecer a una gran multiplicidad de sistemas artificiales. Por ejemplo: *gen, cultivar, antisuero; AC, A, K,* H_20.

2. La *estructura:* desde el punto de vista interno, las UCE o UT pueden coincidir con morfemas, con unidades léxicas, simples, derivadas y compuestas, con sintagmas, ya sean terminológicos o fraseológicos, o con oraciones (que son muy escasas). Por ejemplo: *-itis, -genia; célula, caldo; cortar enzimas, alteración cromosómica; el ADN cromosómico permanece doblemente enhebrado.*
3. El *proceso de gramaticalización:* las UCE léxicas o UT pueden pertenecer a cuatro categorías gramaticales: nominal, verbal, adjetival y adverbial. Por ejemplo: *almidón, adenina; alogenético, antiviral; cultivar, clonar; biológicamente, por vía oral.*

3.2. La densidad terminológica

Para describir el proceso de cómo operan las UT en los contextos comunicativos especializados y dar cuenta de su mayor o menor presencia, se propone este estudio sobre la densidad de estas unidades dentro de los textos, a la hora de distinguir entre diferentes niveles o grados de especialización.

Los estudios de la densidad léxica trabajan sobre la base de muestras textuales de distintos ámbitos, autores, idiomas, etc., lo que permite discriminar textos de forma automática, a partir de las relaciones matemáticas entre *types* (formas) y *tokens* (ocurrencias), que constituyen un modelo de regresión adecuado que puede ayudar a diferenciar tipos de texto (Cantos, 2000: 74-80. Castellà, 2002: 183-184; Martín, 2003: 159-161).

El concepto de *densidad terminológica* hace referencia al número de UT en relación con el número total de unidades léxicas contenidas en un texto especializado. Esta conjunción de unidades está condicionada por los interlocutores de la comunicación y el nivel de especialización del discurso (la variación vertical). Desde esta perspectiva, las líneas de trabajo tienen que ver con procesos para la extracción terminológica, entre los cuales destaca la *densidad de términos,* entendida como la medición del promedio de términos por frase o párrafo, mediante la selección previa de un corpus con marcaje estructural, así como la definición de qué palabras pueden tener valor especializado (Yzaguirre, 1996: 69-71; Cabré, 1999).

4. El corpus y la metodología

El corpus textual especializado para esta investigación está constituido por un conjunto de textos escritos en español, que corresponden a un ámbito temático del Corpus Tècnic del IULA:[2] el genoma humano. La selección de este ámbito dentro del corpus está determinada por un doble propósito. Por un lado, la gran acogida de este nuevo ámbito temático en la medicina y el gran número de publicaciones recientes en relación con él; y, por otro, la variedad de niveles de especialización a través de los cuales se ha tratado el tema, debido al gran interés que ha despertado en la sociedad.

4.1. El corpus textual

Para la selección de los textos que constituyen el corpus de este estudio, se han tenido en cuenta dos criterios: los externos —1) los interlocutores (emisor-receptor) y 2) el tipo y el género textual (tipos: textos descriptivos, informativos; géneros: tesis, libros de texto, artículos de divulgación, etc.)—; y los internos —1) la valoración realizada por especialistas,[3] a partir del tratamiento que se hace del tema y 2) el uso de terminología según el tipo de interlocutor (véase tabla 1)—.

Tabla 1. Niveles de especialización

Nivel	Interlocutores	Muestras de textos
Alto	De especialista a especialista	Tesis doctorales
Medio	De especialista a aprendiz de especialista	Manuales de texto, artículos científicos
Bajo	De especialista a público general o lego	Artículos de carácter divulgativo

2 El Corpus Técnico del IULA (Bwananet, s. f.) recopila textos escritos en cinco lenguas diferentes (catalán, español, inglés, francés y alemán) de las áreas de especialidad de la economía, el derecho, el medio ambiente, la medicina y la informática. El corpus comprende, además, documentos paralelos, con el objetivo de facilitar estudios de traducción, así como el análisis de los datos lingüísticos, a fin de poder establecer las leyes que rigen el comportamiento de cada lengua en cada área (Cabré y Bach, 2004: 173; Bach et al., 1997: 6-11).

3 Los doctores Fernando Giráldez y José Francisco Aramburu (miembros de la Unidad de Biología del Desarrollo de la Universitat Pompeu Fabra) colaboraron en la determinación de los niveles de especialización de los textos del corpus constituido para esta investigación.

El corpus de referencia para el estudio piloto consta de nueve textos (muestras de textos), entre los cuales hay tesis doctorales, manuales de textos y artículos tanto científicos como de carácter divulgativo, con un total de 41.418 palabras. La longitud de cada uno de éstos es de cinco mil palabras por muestra. Además, son textos escritos en español, tanto originales como traducciones. Estos textos forman parte del Corpus Tècnic del IULA y están clasificados en el ámbito temático del genoma humano. Este ámbito se divide, a la vez, en los siguientes subdominios temáticos: estructura interna (EI), con 51 textos; investigación genética (RG), con 10; ingeniería genética (EG), con 7; enfermedades (MA), con 2; ciencias biológicas (CB), con 2, y farmacogenómica (FA), con 1.

En cuanto a la distribución de los textos del corpus de referencia, éste contiene tres textos por cada uno de los tres niveles de especialización propuestos, lo que forma tres subcorpus (nivel alto: a, b, c; medio: d, e, f; y bajo: g, h, i), para un total de nueve textos.

El primer grupo de textos (nivel alto) recoge tres muestras de tesis doctorales sobre algún aspecto del genoma humano, en cuanto a su descripción y aplicación; el segundo grupo (nivel medio) corresponde a manuales universitarios y revistas de investigación clínica, y el tercero (nivel bajo) contiene muestras de artículos de carácter divulgativo (véase tabla 2).

Tabla 2. Corpus textual de análisis del estudio piloto

	Niveles de especialización						
	Alto		**Medio**		**Bajo**		
	Subcorpus	Palabras	Subcorpus	Palabras	Subcorpus	Palabras	**Total palabras**
	a	3.001	d	3.738	g	2.160	
	b	10.365	e	1.799	h	3.138	
	c	10.324	f	4.179	i	2.714	
Subtotal palabras		**23.690**		**9.716**		**8.012**	**41.418**

El cálculo promedio de palabras por textos y por niveles de especialización difiere, porque los textos seleccionados contienen una o tres muestras de cinco

mil palabras. Para el momento de hacer este trabajo no habíamos reestructurado el corpus de referencia con el total de palabras descrito anteriormente.

Para describir la incidencia de los tipos de unidades y la densidad terminológica en la clasificación de textos en niveles de especialización (NE), presentamos el estudio piloto, el cual consta de dos partes: 1) la selección de siete variables lingüísticas; y 2) el análisis discriminante a partir del programa Statgraphics.

En cuanto a la primera parte, se recogen los nueve textos del corpus clasificado en niveles de especialización. Para cada texto se han aplicado siete variables, que corresponden a patrones sintácticos diferentes de unidades terminológicas: nombre [N], adjetivo [Adj], verbo [V], adverbio [Adv], siglas [SG] y unidades lexicalizadas o sintagmáticas [N+P+N] o [N+Adj+N+N…]. Cada variable recoge la información siguiente: total de *types* (número de unidades terminológicas) y total de *tokens* (número total de ocurrencias). En segundo lugar, para la determinación de la distinción entre unidades terminológicas y unidades léxicas, han intervenido las tres herramientas con las que se ha trabajado el corpus.

En aquellos patrones en los que las herramientas de extracción y detección muestran todavía muy poca robustez para el reconocimiento de unidades (patrones del tipo [N+P+N], [V] y [Adv]), se ha utilizado *Bwananet* como herramienta complementaria para hacer búsquedas sobre el conjunto de documentos seleccionados. Asimismo, para cada uno de los patrones dentro de cada texto, se presenta el cálculo de la densidad media. Dicho cálculo se obtiene dividendo el número total de unidades terminológicas reconocidas dentro del texto entre el número total de unidades léxicas. El cálculo se aplica tanto a los *types* como a los *tokens,* con el objetivo de utilizar estos resultados en el análisis discriminante aplicado a los resultados.

En cuanto a la segunda parte, la aplicación del análisis discriminante[4] a los datos, se toman los *types* en cada una de las variables trabajadas. La densidad media

4 Este análisis se hace con base en el paquete estadístico Statgraphics Plus (2002), lo que permite llevar a cabo las acciones siguientes: 1) realizar un estudio descriptivo de una

detectada para cada una de ellas en los tres niveles de especialización, sirve para determinar qué variables son las que intervienen de manera más clara a la hora de distinguir entre los diferentes grados o niveles de especialización.

Para realizar el análisis discriminante se ejecutan algunas etapas previas, entre las que se incluyen las siguientes:

1. Selección de la variable dependiente (niveles de especialización) y de la variable independiente (el tipo de unidades terminológicas, en cuanto a su categoría gramatical dentro de los textos).
2. Determinación del comportamiento de cada variable, es decir, cada una de las unidades terminológicas monoléxicas o poliléxicas, para distinguir los conceptos de *población* (el corpus de referencia) e *individuos* (los nueve textos clasificados en niveles de especialización).
3. Definición del tamaño de la muestra textual: la muestra debe ser representativa en relación con las variables definidas para el análisis discriminante. La representatividad hace referencia al grado en el que una muestra incluye un rango pleno de variabilidad en una población (Biber, 1993: 243).

4.2. Procesamiento y extracción de los términos

Una vez constituido el corpus textual, se procede a la detección[5] y extracción[6] de las unidades terminológicas, para su análisis. Aquí se utilizan tres herramientas, diseñadas en el IULA:

1. *Yate* (Yet another Term Extractor) es un sistema de extracción de candidatos a términos nominales (CAT) en los textos de medicina que han sido procesados previamente con las herramientas del Corpus Tècnic del IULA, diseñado por Vivaldi (2003a).

---

o varias variables, a partir de un conjunto de características precisas; 2) trabajar con modelos de distribución de probabilidad, y 3) realizar descripciones para parámetros como la media, la varianza o proporciones.

5 Reconocimiento de términos (previamente validados como tales) en un texto.

6 Localización de candidatos a términos nominales dentro del texto.

2. *Bwananet* (s. f.) es la herramienta general desarrollada por Vivaldi dentro del marco de diferentes proyectos del IULA para la interrogación del *Corpus Tècnic*. Esta herramienta permite hacer búsquedas sobre el conjunto de documentos seleccionados.
3. *Mercedes* es un sistema de reconocimiento de unidades terminológicas, compuesto por dos módulos: un programa de reconocimiento y un módulo de referencia terminológica. Esta herramienta ha sido desarrollada específicamente para reconocer términos del genoma humano en español, catalán e inglés, y es fácilmente adaptable a cualquier otro dominio (Vivaldi, 2003b).

De acuerdo con las características descritas para estas herramientas, su aplicación al corpus permitirá obtener:

1. Un subcorpus de unidades candidatas a término de base monoléxica —nombre [N], adjetivo [Adj], verbo [V], adverbio [Adv], siglas [SG])— y poliléxica —nombre+preposición+nombre [N+P+N] o unidades que responden a patrones como [N+Adj+N(nombre)+N(nombre), etc.]— (*Yate*).
2. Un subcorpus de unidades léxicas, tanto generales como especializadas, en función de patrones de búsqueda compleja (*Bwanane*t).
3. Un subcorpus de UT de base monoléxica y poliléxica (*Mercedes*).

En cuanto al análisis de los subcorpus de las UT obtenidas a partir de la integración de las herramientas de extracción y detección, se tienen en cuenta los aspectos siguientes:

1. Descripción de las UT en relación con la unidad léxica de la lengua general, lo que permite dar cuenta de su naturaleza lingüística (aspectos morfológicos y sintácticos).
2. La incidencia de las unidades terminológicas en la distinción de niveles de especialización, a partir de la distribución de frecuencias (frecuencia absoluta, relativa y cálculo promedio de unidades), el análisis discriminante,[7] y la densidad terminológica.

7 El análisis factorial discriminante (AFD) es un método de análisis estadístico que permite clasificar individuos a partir de variables cuantitativas. Este análisis también posibilita

5. Resultados

En este apartado presentamos los resultados obtenidos a partir del tratamiento del corpus textual de referencia, tal como lo hemos indicado en § 4.1 y 4.2.

El valor de *types* de la variable [N] en cada uno de los niveles de especialización refleja un resultado gradual (véase tabla 3).

Tabla 3. Valor de *types* de las unidades terminológicas

	Patrones (%)					
NE	[N]	[N+Adj]	[N+P+N]	[V]	[Adv]	[SG]
Alto	0,37	0,54	0,47	0,18	0,02	0,87
Medio	0,30	0,24	0,52	0,16	0,07	0,95
Bajo	0,23	0,17	0,37	0,06	0,02	1,0

Estos resultados permiten evidenciar que las unidades terminológicas de base nominal ([N]) representan un porcentaje considerable a la hora de transmitir el conocimiento especializado de un ámbito temático. Por tanto, este conjunto de unidades terminológicas corresponde al grupo de unidades más significativo, en relación con las demás que constituyen esta tabla de resultados.

El valor de *types* para la variable [N+Adj], en cada uno de los niveles de especialización, también refleja un resultado gradual. No obstante, se ha de tener presente que, a pesar de obtener este resultado gradual, la detección de este conjunto de unidades representa un problema, pues los diccionarios del programa *Mercedes* no cuentan con la suficiente capacidad para reconocer el mayor número posible de estas unidades.

predecir o determinar la pertenencia de individuos a un grupo, en función de los valores obtenidos en las variables definidas. Para abordar este análisis discriminante se utiliza el programa Statgraphics Plus, para la manipulación de datos en una hoja de cálculo con diferentes funciones, con el fin de generar, transformar y recodificar dichos datos (véase StatGraphics. Tutorial, s. f.).

En cuanto a las variables [V], [Adv], [SG] y [N+P+N] en los tres niveles de especialización, el valor de *types* obtenido es heterogéneo, dado que para el reconocimiento de este tipo de unidades se ha utilizado únicamente la herramienta de consulta general *Bwananet,* pues tanto el *Yate* como el *Mercedes* no las reconocen y los resultados no corresponden con el nivel especializado definido.

El análisis discriminante permite desarrollar una serie de funciones discriminantes, que pueden ayudar a predecir el grado o nivel de especialización, basado en los valores de otras variables cuantitativas. Se utilizaron nueve textos para diferenciar entre los tres niveles de especialización. Se introdujeron dos variables [N] y [N+Adj] (véase tabla 4).

Tabla 4. Función discriminante entre [N] y [N+Adj]

Variables	Dependiente	Nivel de especialización
	Independiente	UT [N] y UT [N+Adj]
Corpus	Número de textos	9
	Número de población	3 grupos (alto, medio y bajo)
P-valor	UT [N]	types: 0,0233
	UT [N+Adj]	types: 0,5111

La función discriminante con P-valor[8] inferior a 0,05 es estadísticamente significativa, con un 95% de nivel de confianza, de acuerdo con los parámetros del programa estadístico empleado.

En la figura 1 se observa cómo el valor de [N] discrimina claramente los tres niveles de especialización de acuerdo con el P-valor obtenido. Los tres textos del nivel de especialización bajo (o) se agrupan en el margen izquierdo del gráfico, mientras que los textos del nivel de especialización alto (□) se agrupan en el margen derecho. Los textos del nivel de especialización medio (x) se agrupan entre los dos niveles anteriores.

8 Es una nota promedio de nivel de significación empírico a partir del valor estadístico del contraste entre las variables analizadas.

En cuanto a las demás variables, [V], [Adv], [SG] y [N+P+N], los resultados obtenidos tras la comparación con la [N] no son heterogéneos, porque el P-valor no corresponde con variables discriminantes (véase tabla 5).

Tabla 5. Función discriminante entre [N] y [V], [Adv], [SG], [N+Adj], [N+P+N]

UT	[N] - [V]		[N] - [Adv]		[N] - [SG]		[N] – [N+Adj]		[N] - [N+P+N]	
P-V	0,0188	0,3308	0,0439	0,8472	0,0351	0,7326	0,0233	0,5111	0,0715	0,6534

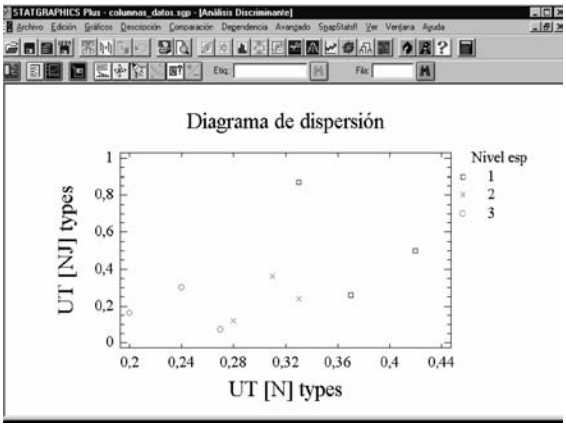

Figura 1. Función discriminante entre [N] y [N+Adj o NJ]

6. Consideraciones finales

El estudio preliminar sobre la clasificación de textos en niveles de especialización a partir del análisis de los tipos de unidades terminológicas, así como de la densidad terminológica, conduce a las siguientes consideraciones: en relación con los criterios de selección del corpus textual, éstos se adecuan a las características propuestas para cada uno de los niveles de especialización.

Para la extracción y detección de los distintos tipos de unidades terminológicas, en trabajos posteriores, deberá tenerse en cuenta que:

1. Las unidades terminológicas plenamente reconocidas por las tres herramientas de explotación del corpus son las unidades monoléxicas con valor nominal [N].
2. Para la detección de las unidades terminológicas monoléxicas [V], [Adv] y [SG], se ha utilizado tan sólo una de las tres herramientas propuestas (*Bwananet*), pues de momento este tipo de unidades no las trabajan los otros dos programas con los que se interroga el corpus.
3. Pese a que las unidades terminológicas poliléxicas [N+Adj] y [N+P+N] tienen un número altamente representativo en cada uno de los subcorpus de análisis (nivel alto, medio y bajo), los resultados obtenidos son escasos, debido a la limitación que supone el hecho de que los diccionarios del programa *Mercedes* no contengan demasiadas unidades terminológicas correspondientes a este patrón. Para poder llevar a cabo la investigación propuesta, será necesario alimentar los diccionarios del programa con unidades que vayan más allá del patrón [N] y [N+Adj].

En cuanto al número y tipos de unidades terminológicas de cada uno de los subcorpus seleccionados para el estudio preliminar, ha de tenerse en cuenta que:

1. La unidad terminológica más representativa en los tres niveles de especialización, tanto cualitativa como cuantitativamente, es la unidad monoléxica con valor nominal. Este hecho permite corroborar que esta unidad es la más prototípica para la representación del campo conceptual de un ámbito especializado.
2. Para el grupo de unidades terminológicas poliléxicas [N+Adj] y [N+P+N], aunque adquieren un valor significativo en los tres niveles de especialización, los resultados obtenidos reflejan los problemas de detección para estos patrones, lo que hace más difícil precisar su cuantificación.

El análisis preliminar sobre la densidad terminológica, en relación con el análisis discriminante propuesto, implica que:

1. La unidad terminológica que mejor permite diferenciar o discriminar textos, según niveles de especialización, es la unidad monoléxica nominal [N]. Este hecho permite considerar que, para el análisis del corpus total, este

grupo de unidades debe ser el más representativo, en comparación con las demás unidades terminológicas propuestas.

2. A partir de la comparación de los resultados obtenidos para el P-valor, es posible replantear el análisis de los diferentes tipos de unidades terminológicas, en relación con el número de variables propuestas.

En la medida en que el desarrollo de este trabajo arroje los resultados de la extracción, la detección y el análisis de las unidades terminológicas, se comprobará si las estrategias utilizadas son aplicables a diferentes ámbitos de la lingüística aplicada. Por ejemplo, estrategias de selección o legibilidad de textos para la enseñanza de lenguas con propósitos específicos, criterios para la clasificación en niveles de especialización de los textos del Corpus Tècnic del IULA, etc.

Bibliografía

Arntz, R. y Picth, H. (1995). *Introducción a la terminología.* [Traducción al español de A. Irazazábal *et al.*]. Madrid: Fundación Germán Sánchez Ruipérez, Pirámide, Biblioteca del libro.

Bach, C., Saurí, R., Vivaldi, J. y Cabré, M.T. (1997). *El corpus de l'IULA: descripció.* [El corpus del IULA: descripción]. Sèrie de Informes, 17. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

Beaugrande, R. y Dressler, W. (1997). *Introducción a la lingüística del texto.* [Traducción al español de S. Bonilla]. Barcelona: Ariel.

Bernárdez, E. (1982). *Introducción a la lingüística del texto.* Madrid: Espasa-Calpe.

Biber, D. (1993). Representativeness in Corpus Design. En: *Literary & Linguistic Computing* 8 (4), 243-257. UK: Oxford University Press.

Bwananet (s. f.). Herramienta de consulta general para la interrogación del Corpus Tècnic del IULA. *Bwananet.* Recuperado 1.º de diciembre, 2007, de http://bwananet.iula.upf.edu

Cabré, M.T. (1999). *La terminología. Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos.* (Sèrie Monografies 3). Barcelona: Universitat Pompeu Fabra / Institut Universitari de Lingüística Aplicada.

_ (2002). Terminologie et linguistique: La théorie des portes. En *Terminologies nouvelles. Terminologie et diversité culturelle 21,* 10-15. En R. Folguerà (Trad.), Terminología y lingüística: la teoría de las puertas. *Estudios de Lingüística Española,* 16. Recuperado 8 de abril, 2006, de http://elies.rediris.es/elies16/Cabre.html

_ (2003). Theories of terminology. Their description, prescription and explanation. En *Terminology,* 9 (2), 163-200.

Cabré, M. T. y Bach, C. (2004). El Corpus Tècnic del IULA: corpus textual especializado plurilingüe. En *PANACE@ V, 16,* 173-176. Recuperado 5 de noviembre, 2006 (actualizado 24 de marzo, 2008) de http://www.medtrad.org/panacea/PanaceaPDFs/Panacea16_Junio2004.pdf

Cabré, M. T y Estopà, R. (2005). Unidades de conocimiento especializado: caracterización y tipología. En Cabré, M. T.y Bach, C. (Eds.), *Coneixement, llenguatge i discurs especialitzat.* [Conocimiento, lenguaje y discurso especializado]. Sèrie Monografies 7. (pp.78-83). Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

Cantos, P. (2000). Investigating type-token regression and its potential for automated text discrimination. En: Cantos, P. y Sánchez, A. (Eds.), *Corpus-based Research in English Language and Linguistics* (pp. 71-92). Monográfico. Cuadernos de Filología Inglesa. Murcia: Servicio de Publicaciones de la Universidad de Murcia.

Castellà, J. M. (1992). *De la frase al text. Teories de l'ús lingüístic.* [De la frase al texto. Teorías del uso lingüístico]. (pp. 49-53). Barcelona: Empúries.

_ (2002). La complexitat lingüística en el discurs oral i escrit: densitat lèxica, composició oracional i connexió textual. [La complejidad lingüística en el discurso oral y escrito: densidad léxica, composición oracional y conexión textual]. Tesis doctoral. *Universitat Pompeu Fabra.* Recuperado 16 de julio, 2007 de http://www.tdx.cesca.es/TDX-0311102-134928/index.html

Ciapuscio, G. (1994). *Tipos textuales*. Buenos Aires: Oficina de Publicaciones Ciclo Básico Común, Universidad de Buenos Aires.

_ (2003). *Textos especializados y terminología.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Eggins, S. y Martin, J.R. (2000). Géneros y registros del discurso. En Van Dijk, T. A. (Ed.) *El texto como estructura y como proceso. Estudios del discurso: Introducción multidisciplinaria* (pp. 335-370). Barcelona: Gedisa / SAP.

Guantiva, R. (2005). Terminología y variación vertical: clasificación de textos en niveles de especialización a partir del análisis del tipo y de la densidad de las unidades terminológica. Proyecto de tesis doctoral no publicado, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona, España.

Halliday, M. (1994). *El lenguaje como semiótica social. La interpretación social del lenguaje y del significado* [Traducción al español de Fondo de Cultura Económica, México]. Colombia: Fondo de Cultura Económica.

Hoffmann, L. (1998a). Característiques dels llenguatges d'especialitat. (Trad., al catalán) [Características de los lenguajes de especialidad]. En Brumme, J. (Ed.), *Llenguatges d'especialitat. Selecció de textos.* [Lenguajes de especialidad. Selección de textos] (pp.21-69) Barcelona: Institut Universitari de Lingüística Aplicada.

—. (1998b). Conceptes bàsics de la lingüística dels llenguatges d'especialitat. (Trad., al catalán), [Conceptos básicos de la lingüística de los lenguajes de especialidad]. En Brumme, J. (Ed.), *Llenguatges d'especialitat. Selecció de textos.* [Lenguajes de especialidad. Selección de textos] (pp.71-78). Barcelona: Institut Universitari de Lingüística Aplicada.

Kocourek, R. (1982). *La langue française de la technique et de la science. Vers une linguistique de la langue savante.* Wiesbaden: Oscar Brandstetter.

Martín, P. (2003). Análisis contrastivo de los componentes estructurales y gramaticales de los resúmenes de los artículos científicos. *Revista Española de Lingüística (RIEL),* 33 (1), 153-183.

Sager, J. C. (1993). *Curso práctico sobre el procesamiento de la terminología.* Madrid: Pirámide.

Schröder, H. (1991). Linguistic and text-theoretical research on languages for special purposes. A thematic and bibliographical guide. En Schröder, H. (Ed.), *Subject-oriented Texts. Languages for Special Purposes and Text Theory* (pp. 1-49). Berlin-New York: De Gruyter / Walter.

Statgraphics Plus, versión 5.1. (2002). Edición en español CD-ROM. EE.UU.: M & INC.

Statgraphics. El programa Statgraphics. [Versión electrónica] Recuperado 15 de noviembre, 2004 de http://www.etsii.upm.es/ingor/estadistica/docencia/prac2000/prac2001.pdf

Statgraphics Plus. Clásico de los programas estadísticos en castellano. *Addlink.* Recuperado 17 de noviembre, 2006 de http://www.addlink.es/productos.asp?pid=138.

Statgraphics. Tutorial (s. f.). ¿Como puede ayudarte STATGRAPHICS PLUS en tus investigaciones, cálculos e informes? *Statgraphics.* Recuperado 10 de noviembre, 2006 de http://www.statgraphics.net/Recursos/CPASPICI.htm#Indice

Van Dijk, T.A. (1980). *Estructuras y funciones del discurso. Una introducción a la lingüística del texto y a los estudios del discurso.* (pp.9-17). Madrid: Siglo XXI.

_ (1989). *La ciencia del texto.* [Traducción al español de Hunzinger, S.]. (pp.13-30). Barcelona: Paidós Ibérica.

_ (1993). *Texto y contexto, semántica y pragmática del discurso.* [Traducción al español de Moyano, J.]. (pp. 29-46). Madrid: Cátedra-Lingüística.

Vivaldi, J. (2003a). *Sistema de extracción de candidatos a términos YATE.* Manual de utilización (1.ª versión). Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

—. (2003b). *Sistema de reconocimiento de términos MERCEDES.* Manual de utilización (2.ª versión). Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Wüster, E. (1998). *Introducción a la teoría general de la terminología y a la lexicografía terminológica.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Yzaguirre, L. de. (1996). Ingeniería lingüística y terminología. En *Terminómetro. Monográfico: La Terminología en España* (pp. 69-71). París: Unión Latina-IULA.

Los autores

**Ricardo Guantiva Acosta es Magíster en Lingüística Española del Instituto Caro y Cuervo y profesor de lingüística general de la Universidad de Bogotá Jorge Tadeo Lozano. Correo electrónico: ricardo.guantiva @upf.edu

***M. Teresa Cabré Castellví es doctora en filología catalana y catedrática en lingüística aplicada de la Facultad de Traducción e Interpretación de la Universitat Pompeu Fabra. Correo electrónico: teresa.cabre @upf.edu

****Josep M. Castellà Lidon es doctor en lingüística, profesor de la Facultad de Humanidades, área Filología Catalana y miembro del IULA de la Universitat Pompeu Fabra. Correo electrónico: josep.castella @upf.edu