

Cabré, M. T. (2007). "Constituir un corpus de textos de especialidad: condiciones y posibilidades". En Ballard, M.; Pineira-Tresmontant, C. (ed.). *Les corpus en linguistique et en traductologie*. Arras: Artois Presses Université. 89-106. ISBN 978-2-84832-063-2

Constituir un corpus de textos de especialidad: condiciones y posibilidades

M. Teresa Cabré
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra (Barcelona)

No cabe duda de que el desarrollo de corpus textuales ha permitido a la lingüística descriptiva dar un salto cualitativo muy importante. Este avance ha abierto a los lingüistas la posibilidad de dar cuenta de forma más adecuada del funcionamiento de las lenguas ya que los análisis han podido basarse por primera vez en muestras representativas y abundantes de producciones lingüísticas, no limitadas ni sesgadas subjetivamente como sucedía anteriormente. Además de este argumento, la denominada Lingüística de corpus permite explorar exhaustivamente las producciones lingüísticas y, con ello, ofrece al lingüista muestras de datos que mediante un análisis manual no llegan a la misma profundidad.

En esta ponencia nos proponemos tres objetivos. En primer lugar, expondremos algunas de las características de los denominados "lenguajes de especialidad, que son la fuente de los textos especializados. En segundo lugar, presentaremos brevemente el corpus textual especializado del Instituto Universitario de Lingüística Aplicada como muestra de adaptación a las condiciones mencionadas en el primer objetivo. Y en tercer lugar, mostraremos, a modo de ejemplo, un estudio sobre el contraste entre los textos de especialidad y los textos no especializados a través de sus características lingüístico-gramaticales.

1. La constitución de un corpus de especialidad: cuestiones y criterios

Ante el propósito de constituir un corpus textual de especialidad la primera cuestión que se plantea es qué entendemos por texto especializado o cómo discriminamos los textos especializados. Sin una respuesta clara a esta pregunta no podemos iniciar, obviamente, la selección del material.

1.1 Establecidos los criterios que permiten discriminar del universo de los textos producidos por los especialistas en situaciones profesionales, debemos plantearnos qué tipo de textos de especialidad debemos tener en cuenta para que el corpus resulte suficientemente equilibrado.

En tercer lugar, debemos plantearnos la cantidad de producciones que formarán parte de este corpus si pretendemos o bien que sea lo suficientemente representativo de cada especialidad o bien que sea suficiente para poder analizar un tema determinado previamente. Cabe hacer pues en este punto una precisión previa a la constitución del corpus, y sobre todo a la decisión sobre su dimensión : ¿para qué se constituye el corpus que vamos a elaborar? ¿Qué finalidad pretendemos que cumpla este corpus ? ¿A qué estudios lingüísticos queremos que dé lugar?

Y tras la resolución de estas tres cuestiones ya podemos iniciar el proceso de trabajo, que, lógicamente, deberá plantearse otras cuestiones ya de tipo más técnico: sea de técnica relativa a la lingüística, sea de técnica informática.

Finalmente, una vez constituido el corpus en formato digital, debemos entrar en la programación de sus posibilidades de exploración, posibilidades que deben haberse establecido en la etapa preliminar de caracterización del corpus a elaborar.

Vayamos respondiendo paso a paso cada una de estas cuestiones.

a) ¿Qué es un texto especializado? Y ¿cómo reconocemos entre todos los textos aquellos que son especializados?

Los textos especializados son las producciones lingüísticas, orales o escritas, que se producen en escenarios de comunicación profesional y sirven exclusivamente a una finalidad profesional. Se reconocen los escenarios profesionales por los interlocutores que actúan en la situación, por el tratamiento de una temática relativa al dominio o dominios concernidos por la profesión y por la finalidad esencial de buscar la información del receptor, aunque para ello se utilicen estrategias discursivas distintas.

Analíticamente los textos especializados se definen por tres tipos de condiciones:

- Condiciones discursivas: las propias del escenario especializado de este tipo de comunicación
- Condiciones cognitivas: el tema de qué tratan y la forma precisa de tratarlo
- Condiciones lingüísticas: las condiciones textuales generales (precisión, concisión y sistematicidad, las dos últimas en grados diversos según las condiciones discursivas), la forma textual macro y micro del texto, y sobre todo las unidades léxicas propias del dominio de que trata el texto.

b) ¿Qué variables podemos considerar en un corpus textual especializado?

Los textos de especialidad no son homogéneos, sino que se organizan en tipos distintos en función de los criterios de clasificación que se tomen en consideración. Los criterios que en nuestra opinión son los más relevantes para organizar los textos de especialidad en un corpus son los siguientes:

- El tema
- La perspectiva o dimensión disciplinar
- El nivel de especialización
- Las fuentes
- El género textual
- La clase de texto por la estrategia discursiva
- Las lenguas
- En el caso de los plurilingües (bilingües, y trilingües, etc.), por la relación entre los textos de las lenguas del corpus.

Por el tema, distinguimos entre corpus unidisciplinarios y pluridisciplinarios. El tema de un corpus puede abarcar un ámbito completo o solo una perspectiva de un ámbito. Un

ejemplo de este segundo caso podría ser el banco de derecho medioambiental desarrollado por el grupo TERMISUL de la Universidad de Porto Alegre (Brasil).

Por el nivel de especialización, un corpus puede incluir textos de un solo nivel de especialidad (por ejemplo: textos de artículos científicos procedentes de revistas homogéneas) o bien incluir estructuradamente textos de distintos niveles de especialidad¹.

Por el canal de transferencia, los textos del corpus pueden proceder de un solo tipo de fuente o de fuentes diversas. La diversidad de las fuentes puede obedecer también a una amplia diversidad de criterios, el que aquí nos interesa es el criterio del canal de transmisión, por el cual los textos de un corpus pueden ser exclusivamente orales o escritos o audiovisuales, o bien incluir muestras de todas las posibilidades.

Por el género textual un corpus puede ser homogéneo e incluir solamente textos de un género (por ejemplo, *abstracts* de revistas científicas) o bien comprender estructuradamente textos de distintos géneros textuales.

Por el tipo de texto según la estrategia discursiva, los corpus pueden ser heterogéneos u homogéneos en cuanto a la clase textual (por ejemplo, un corpus homogéneo incluiría solo textos argumentativos, o narrativos).

Según el criterio de las lenguas, los corpus pueden ser monolingües, bilingües, trilingües, etc. Y si comprenden textos de más de una lengua, estos pueden ser coincidentes solo en la temática o bien comprender textos en una lengua y su correspondiente traducción en la segunda o tercera lengua. En este último caso se denominan corpus paralelos.

c) **¿Qué dimensiones debe tener un corpus especializado?**

La respuesta a esta cuestión sólo puede ser: depende del corpus que hayamos decidido elaborar en lo que se refiere a su finalidad. ¿Para qué va a servir un corpus? ¿Para extraer datos que sean representativos del uso de una lengua en su conjunto? En este caso deberemos constituir un tipo de corpus, que se ha denominado corpus de referencia, que incluya una muestra representativa de la totalidad de la lengua, entendida en toda su variación interna y externa. Pero si de lo que se trata es de constituir un corpus para investigar sobre uno a distintos problemas, la dimensión del corpus debe adecuarse a la resolución de las finalidades que se propone. Por ejemplo, el corpus que hay que constituir para analizar el uso de un pronombre en situación enclítica será evidentemente menor que el que necesitamos para extraer la terminología de un dominio de especialidad; y este último podrá ser menor al necesario para extraer colocaciones.

1.2 El proceso de constitución del corpus

¹ La pertenencia de un texto a un nivel de especialidad alto, medio o bajo suele basarse en las características de los destinatarios, el medio en qué aparecen y las finalidades del texto. Así, un texto producido por un especialista para estudiantes será un texto de nivel medio de especialidad. Para más información puede verse Cabré (1998) o Ciapuscio (2003).

La construcción efectiva una vez se han establecido sus características es un proceso que se distribuye en distintas fases:

- a. Selección de fuentes
- b. Criterios de selección de los textos y decisión sobre si tomar el texto completo o fragmentos del mismo²
- c. Decisiones sobre la arquitectura de la base
- d. Decisiones sobre la infraestructura de *hardware* y *software* (sistema de gestión de corpus textuales)
- e. Selección de las convenciones para la representación de los textos
- f. Criterios, lenguaje y sistema de marcaje estructural
- g. Criterios, lenguaje y sistema de marcaje lingüístico

1.3 Herramientas de exploración

Los textos de un corpus pueden procesarse en bruto o procesados lingüísticamente. Si se opta por la segunda vía, parece lógico que debemos contar con recursos y herramientas de tratamiento automático de la información:

- Herramientas de marcaje estructural y lingüístico
- Diccionario inicial de procesamiento
- Sistema de análisis morfológico
- Sistema de lematización
- Sistema de desambiguación
- Sistema de gestión de diccionarios
- Sistema de estructuración sintáctica (*chuncker*), etc.

1.4 Posibilidades de explotación

Finalmente, las posibilidades de explotación lingüística de un corpus están condicionadas por el tratamiento que los datos han recibido en la fase de tratamiento. Las posibilidades de aplicación de los datos de corpus suelen materializarse en los ámbitos siguientes:

- En la ingeniería lingüística, para el desarrollo de herramientas y robots
- En la extracción de información para fines investigadores, docentes, industriales, editoriales, etc.
- En la recuperación de información en los servicios documentales y bibliográficos

La utilización primaria que los lingüistas hacemos de los corpus de especialidad se orientan fundamentalmente a:

- La investigación sobre discurso especializado, terminología y fraseología especializadas
- La elaboración de diccionarios especializados

² Esta decisión está condicionada por los estudios que deseamos hacer a través del corpus. Para el análisis textual (conectores, estructura informativa, géneros textuales, etc.) se requieren textos completos.

- La enseñanza de lenguas de especialidad o de lenguas para propósitos específicos

Para la enseñanza de lenguas de especialidad, los corpus ofrecen la posibilidad de programar más adecuadamente los contenidos (adecuación a necesidades y grado de conocimientos de los estudiantes), de elaborar ejercicios y de alimentar sistemas de autoaprendizaje de lenguas.

En el campo de la documentación, y concretamente para la gestión de información, los corpus proveen de información para la construcción automática o asistida de tesauros, para la indización automática y para la elaboración de sistemas de clasificación de documentos o de refinamiento de las consultas orientadas a perfiles de necesidades de usuario.

2. El corpus técnico plurilingüe del IULA

El Instituto Universitario de Lingüística Aplicada (IULA) es un centro de la Universidad Pompeu Fabra, de Barcelona, dedicado a la investigación y a la formación de postgrado. Fue creado en 1993 y organizado desde su creación por M^a Teresa Cabré. El IULA se organiza en grupos de investigación: Léxico, Terminología y discurso especializado (Grupo IULATERM, que acoge la Lingüística Computacional), Lexicografía (Grupo INFOLEX), Variación lingüística (Grupo UVAL), Documentación y edición digital (Grupo DIGIDOC) y tres laboratorios: OBNEO (Observatorio de Neología), LATEL (Laboratorio de Tecnologías Lingüísticas) y el Laboratorio de Lingüística Forense. Desde 1993 hasta la actualidad, el proyecto Corpus ha sido el proyecto de investigación común en el que han participado todos los miembros del IULA. Recopila textos escritos en cinco lenguas diferentes (catalán, castellano, 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. A su vez, el corpus multilingüe del IULA cuenta con un subcorpus de lengua general, extraído de la prensa de gran difusión y constituido como corpus contrastivo.

El objetivo de este corpus es facilitar el análisis de los datos lingüísticos a fin de poder establecer las leyes que rigen el comportamiento de cada lengua en cada área. Sus destinatarios son los investigadores y todos los usuarios que requieran consultas sobre los ámbitos de especialidad tratados. De la explotación del corpus se han derivado estudios de carácter terminológico, discursivo, morfológico, sintáctico, neológico o traductológico. Para facilitar la explotación de los datos, el IULA ha desarrollado una serie de herramientas de exploración. Una muestra de estas herramientas son un extractor automático de neología, un detector automático de terminología, un alineador de textos, un alimentador de diccionarios, etc. De hecho, este corpus es el soporte principal de las actividades de investigación y docencia de nuestro instituto.

La herramienta que permite acceder a los datos del corpus a través de Internet es BwanaNet, que puede encontrarse en la página principal de la web del IULA (<www.iula.upf.edu>), en el apartado denominado «Portal de recursos del IULA».

El corpus del IULA contiene textos escritos en cinco lenguas diferentes (catalán, castellano, inglés, francés y alemán) de las áreas de especialidad de economía, derecho, medio ambiente, medicina e informática, además de documentos paralelos sobre estas materias. Cada una de las áreas fue estructurada en diferentes subáreas por un especialista, a fin de que los textos pudieran recuperarse con mayor precisión temática

Véase a continuación cómo está estructurada el área de la medicina:

Anatomía	(AN)
Organismos	(OR)
Enfermedades	(MA)
Productos químicos y fármacos	(PQ)
Técnicas y equipamientos analíticos, diagnósticos y terapéuticos	(TE)
Psiquiatría y sicología	(PS)
Ciencias biológicas	(CB)
Ciencias físicas	(CF)
Antropología, educación, sociología y fenómenos sociales	(FS)
Tecnología, industria, agricultura	(TI)
Humanidades	(HU)
Información científica	(IC)
Grupos nominales	(GN)
Planificación y gestión sanitaria	(GS)
Asesor:	Toni Valero

El procesamiento de los textos del corpus sigue los siguientes pasos:

a) Fase de selección de los textos

Los especialistas en cada materia seleccionan aquellos textos que consideran pertinentes y los clasifican temáticamente dentro de una estructuración del dominio previamente consensuada por especialistas de la materia.

b) Fase de anotación y registro de la información del documento

Los documentos se marcan de acuerdo con el estándar SGML y siguiendo las directrices marcadas por el *Corpus Encoding Standard* (CES) de la iniciativa EAGLES. Posteriormente se registra la información documental de los textos (autor, título, edición, páginas seleccionadas, subdominio al cual pertenece, idiomas en que ese mismo documento se encuentra en el corpus).

c) Fase de procesamiento lingüístico

El procesamiento lingüístico de los documentos está automatizado y consta de un preproceso, a través del cual se tratan lingüísticamente aquellas entidades que admiten una detección automática previa al análisis morfológico (fechas, números, locuciones, nombres propios, abreviaturas), un análisis morfológico, mediante el cual se lematizan todas las palabras de los documentos y se les da una o más etiquetas morfológicas, de acuerdo con los etiquetarios morfosintácticos diseñados en el IULA, y una posterior desambiguación lingüística y estadística, de forma que a cada palabra le acabe correspondiendo un solo lema y una sola etiqueta.

d) Almacenamiento en una base de datos textual

Finalmente, cuando ya cada palabra tiene el lema y la categoría gramatical que le corresponde, los textos se almacenan en una base de datos textual, que contiene toda la información que se ha generado sobre el documento. El resultado de todo el proceso de tratamiento de los textos puede consultarse actualmente en línea en <[brangaene upf es/bwananet/index htm](http://brangaene.upf.es/bwananet/index.htm)>.

Área	Catalán	Español	Inglés	Francés	Alemán	Total
Derecho	1 463 000	2 085 000	431 000	44 000	16 000	4 039 000
Economía	1 776 000	1 091 000	274 000	78 000	27 000	3 246 000
Medio ambiente	1 506 000	1 062 000	599 000	230 000	429 000	3 826 000
Informática	655 000	1 227 000	338 000	194 000	83 000	2 497 000
Medicina	2 619 000	4 077 000	1 555 000	27 000	198 000	8 476 000
Total:	8 019 000	9 542 000	3 197 000	573 000	753 000	22 084 000

Cuadro 1 Número de palabras por lengua y ámbito

El corpus de medicina incluye un subcorpus de genoma humano, elaborado por el grupo Iulaterm, que contiene 945 000 palabras en catalán, 1 447 000 en español y 1 119 000 en inglés. Los datos en relación con el corpus paralelo de las parejas lingüísticas más significativas catalán-español, catalán-inglés, español-inglés, se presentan en el cuadro 2.

Área	Catalán-español	Catalán-inglés	Español-inglés
Derecho	460 000	12 000	57 000
Economía	600 000	250 000	283 000
Medio ambiente	214 000	213 000	144 000
Informática	28 000	-	300 000
Medicina	118 000	40 000	640 000
Total	420 000	515 000	1 424 000

Cuadro 2 Número de palabras en corpus paralelos por ámbito y parejas de lenguas

Finalmente, los datos del corpus de contraste se muestran en el cuadro 3.

Área	Catalán	Español	Total
General	1 526 000	3 230 000	4 756 000

Cuadro 3 Número de palabras en el corpus de lengua general

La consulta del corpus del IULA se realiza vía Internet a través de BwanaNet, una interfaz desarrollada en el IULA. El Corpus Técnico del IULA (CT-IULA) está indexado con un paquete de herramientas desarrolladas por el *Institut für Maschinelle Sprachverarbeitung*, de la Universidad de Stuttgart (Corpus Workbench). El IULA ha desarrollado la herramienta que permite la interrogación del CT-IULA en línea ([brangaene.upf.es/bwananet/index htm](http://brangaene.upf.es/bwananet/index.htm)).

3. Una aplicación de la lingüística de corpus : Contraste gramatical entre textos especializados y textos no especializados

Gracias a este corpus se han podido realizar más de veinte tesis de doctorado³.

Además de las tesis, el corpus ha permitido desarrollar una base de conocimiento (GENOMA) que puede verse en www.iula.upof.edu/genoma.

Actualmente se está llevando a cabo un proyecto de investigación sobre las características específicas de los textos especializados en relación a los no especializados. Presentamos a continuación una breve síntesis del proyecto y algunos de sus resultados.

El proyecto ESPETEX, que forma parte de un proyecto más amplio financiado por el Ministerio de Educación y Cultura del gobierno español (TEXTTERM-2. Fundamentos, estrategias y herramientas para el procesamiento y extracción automáticos de la información especializada. N° REFERENCIA: BFF2003-02111) al que contribuyen una veintena de investigadores y colaboradores, se propone dos objetivos:

- Comprobar a través de un corpus suficientemente representativo si se confirman las características gramaticales que los manuales de lenguajes especializados atribuyen a los textos de especialidad.
- En caso de que no se confirmen en parte o totalmente, intentar encontrar y establecer algunos de los factores gramaticales específicos que diferencian los textos especializados.

Para llevar a cabo el proyecto hemos partido de la lista de características de los textos especializados expuesta en alguno de los dos manuales siguientes:

- Kocourek, R. (1991). *La langue française de la technique et de la science. Vers une linguistique de la langue savante*, Wiesbaden: Oscar Branstetter.

³ Las tesinas y tesis realizadas sobre la base de los datos del corpus son las siguientes: **Araceli Alonso**: Descripción y análisis de los sufijos nominalizadores en el área del medio ambiente. **Rosanna Folguerà**: Adjectius en el discurs especialitzat: una primera descripció deis adjectius en els textos del genoma humà. **Vanesa Vidal**: Aproximación al fenómeno de la combinatoria verbo-nominal en el discurso especializado en Genoma Humano. **Gabriel Quiroz**: Las unidades sintagmáticas extensas especializadas en inglés y en español: descripción y clasificación en un corpus de genoma. **John Jairo Giraldo**: Análisis y descripción de las siglas en el discurso especializado de Genoma humano y Medio ambiente. **Iria da Cunha**: Hacia un modelo lingüístico de resumen automático de artículos médicos en español. **Rogelio Nazar**: Aproximación cuantitativa al mapeo conceptual. **Carles Tebé**: La representació conceptual en terminologia: l'atribució temàtica en els bancs de dades terminològiques. **Ricardo Guantiva**: Terminología y variación vertical: clasificación de textos en niveles de especialización a partir del análisis del tipo y la densidad de las unidades terminológicas. **Ona Domènech**: Textos especialitzats i variació vertical: la diversitat terminològica com a factor discriminant del nivell d'especialització d'un text.

- Sager, J.C.; Dungworth, D. (1980) *English Special Languages*. Wiesbaden, Oscar Brandstetter Verlag.

Estos manuales se han basado en corpus de pequeña talla. En el proyecto ESPETEX. Se ha constituido un doble corpus: un primer corpus de textos especializados y un segundo corpus de textos de carácter general.

El Corpus de lengua general (prensa) consta de 5.002.121 palabras en 155 documentos del *Corpus de l'IULA*.

El corpus de especialidad se compone de 5.018.193 palabras en 251 documentos del *Corpus de l'IULA* (Derecho, Economía, Informática, Medio ambiente, Medicina: 1.000.000 palabras de cada dominio).

Las características gramaticales no léxicas que los manuales atribuyen a los textos de especialidad se distribuyen, siguiendo a Kocourek (1982, 1991), en cuatro grupos⁴:

1. Selección de las categorías gramaticales
2. Complejidad de la estructura
3. Condensación sintáctica
4. Impersonalidad de la frase

En lo que se refiere a la selección de las categorías gramaticales se subrayan los siguientes fenómenos:

- Predominio de los nombres
- Empleo especial de categorías gramaticales, sobre todo en relación al verbo (y por tanto a los pronombres personales):
 - Ausencia de la 2ª persona singular y plural
 - Raramente, uso de la 1ª persona singular a favor de *nosotros*
- La ausencia de ciertas palabras o morfemas gramaticales de la morfología verbal:
 - Predominio de la 3ª persona del singular
 - Predominio del presente Indicativo
 - Frecuencia de la 2ª persona plural del imperativo
 - Predominio de las frases declarativas
 - El uso reduce frases interrogativas directas

Respecto a la complejidad estructural, se distinguen como específicos de los textos especializados los siguientes rasgos gramaticales:

- Escasa longitud de la frase
- Abundancia de sintagmas nominales
- Nominalización de los verbos
- Frecuencia de expansiones de nombres y SN
- Abundancia de frases relativas

⁴ Otros autores, además de Kocourek han realizado estudios sobre el tema. Entre ellos destacamos los siguientes: Phal (1968), Vigner et Martin (1976), Kocourek (1982, 1991), Loffler-Laurian (1980, 1982, 1983, 1985, 1986), D. Candel (1984), Hoffmann (1985) y L'Homme (2005)

- Construcciones de participio y de infinitivo
- Diversidad de conjunciones circunstanciales
- Construcciones insertadas en la frase

En cuanto a la condensación sintáctica destacan los fenómenos siguientes:

- Uso abundante de la pronominalización
- Uso de frases de infinitivo y participio
- Nominalización de formas verbales

Y, finalmente, la impersonalidad de la frase en los textos de especialidad se proyecta en los siguientes fenómenos:

- Pronombre de modestia: *nosotros*
- Uso del impersonal *uno/una* como sujeto del verbo
- Giros impersonales tales como es + adjetivo (probable, cierto, sorprendente, etc), resulta que, conviene que, se ha dicho que, etc.
- Abundancia de la voz pasiva

Además de todas estas características gramaticales, se han subrayado en el plano textual:

- Falta de ciertos géneros (cartas, piezas teatrales, etc.)
- Abundancia de ciertos géneros: dependencia de dominio (derecho, medicina, genómica, etc.)
- Control de la estructuración de la información (marcadores discursivos y meta-discursivos, tablas, listas, etc.)

Y en el plano léxico:

- La abundante presencia de terminología
- La evitación de unidades polisémicas
- La tendencia a usar sistemáticamente la misma unidad para un concepto evitando así el uso de sinónimos

Y, para terminar, en el plano gráfico, la presencia de símbolos, fórmulas, representaciones icónicas o unidades léxicas híbridas: *comando-c*, etc.

El análisis realizado sobre nuestro doble corpus se ha limitado a los siguientes fenómenos:

- Clases gramaticales: N, V, Adj, Adv, Prep, Conj
- Nombres propios y nombres comunes
- Género y número de los nombres
- Nombre precedido de determinante definido
- Adjetivos calificativos
- Pronombres relativos
- Persona, modo y tiempo de los verbos
- Formas verbales no personales
- Preposiciones

- Conjunciones

Dentro de los nombres y pronombres:

- N + Adj
- N + SP
- Pronombres 1ª, 2ª, 3ª persona singular y plural
- Forma impersonal *se*
- Pronombres relativos: *que, quien(es), cuyo(s)*

En el apartado de las formas verbales, se han analizado:

- Tiempos: presente/pasado
- Persona: 1ª, 2ª, 3ª
- Nombre: singular/plural
- Formes en 1ª, 2ª, 3ª persona en activa/pasiva
- Modo indicativo/subjuntivo/imperativo/condicional

Se han observado además algunas preposiciones, conjunciones simples y complejas, concretamente las siguientes :

- Preposición *de*
- Conjunciones coordinativas: *y, o, ni, pero*
- Conjunciones subordinadas: *porque*
- Conjunciones subordinadas complejas: *por consiguiente, puesto que, de forma que, a menos que, si bien, ni siquiera, aun cuando, tanto más cuanto, a menos que*
- Algunos marcadores metadiscursivos
- Lema = *aludir, definir, designar, llamar, sobreentender*
- Lema = *conocer, definir, entender + como*
- Lema = *entender + por*
- Lema = *querer + Lema = decir*
- Lema = *recibir + el nombre de*
- *es decir*
- *esto es*
- *o sea*

Los resultados a los que hemos llegado se muestran en los siguientes cuadros:

	LG	LE
noms	1.218.815	1.302.211
Adj qualificatifs	<i>381.813</i>	<i>430.576</i>
Verbs	684.530	624.766
determinants	612.499	659.823
Préposition <i>de</i>	<i>366.827</i>	<i>457.584</i>
conjonctions	239.865	235.434
Adverbes	231.341	202.956

	TG	TE
Adj qualificat.	<i>381.813</i>	<i>430.576</i>
N+Adj	<i>150.386 (38,07%)</i>	<i>225.856 (42,68%)</i>
N+SP	<i>244.635 (61,93%)</i>	<i>303.469 (57,33%)</i>
N+participe	--	--

	TG	TE
Formes personnelles	497.278	454.947
Formes non personnelles	187.252	169.819

	TG	TE
présent	287.983	312.423
passé	<i>148.318</i>	<i>40.079</i>

	TG	TE
Indicatif	313.992	219.648
Subjonctif	9.437	8.315
Ambigues Imperatif-Indicatif	115.917	120.258
Ambigues Imperatif-Sbjonctif	29.614 (0,72%)	41.202 (0,88%)
Conditionnel	9.378	7.612

	TG	TE
1ère personne	36.243 (12,47%)	26.190 (11,61%)
2ème personne	4.525 (1,56%)	3.316 (1,47%)
3ème personne	249.989 (85,9 %)	196.049 (86,9 %)
1 ^a singulier/pluriel	23.270/12.973	12.472/13.718
2 ^a singulier/ pluriel	4.214/311	3.210/106
Total formes sing	174.904 (63,08%)	102.389 (36,92%)
Total formes plur.	115.853 (48,48%)	123.166 (51,52%)

	TG	TE
passive	3.469	3.562
active		
1 ^a sing/plur	16/17	0/0
2 ^a sing/plur	0/0	1/0
3 ^a sing/plur	1.892/1.544	1.570/1.991

	TG	TE
Total	120.453	105.222
que	114.204	97.391
cual, cuales	1.216	3.948
quien, quienes	1.103	387
cuyo,-a, cuyos, -as	1.743	2.973
se impersonnel	69.867	97.418

	TG	TE
Total conj	239.895	235.434
ni	4.496	2.087
o	13.240	35.690
pero	15.574	7.412
que (completif)	42.116	26.305
porque	6.028	2.092

	TG	TE
puesto que	272	863
de forma que	99	334
a menos que	33	209
si bien	212	587
aun cuando	17	173
tanto más cuanto	11	96
a menos que	33	209
Por consiguiente	21	400

	TG	TE
Total	3.092	8.067
V type <i>llamar, denominar + (det) N</i>	2.620	4.858
<i>Ventender + por</i>	27	97
querer decir	168	199
<i>Recibir el nombre de</i>	5	64
<i>es decir</i>	500	1.552
<i>o sea</i>	85	307
<i>esto es</i>	180	449

5. A modo de conclusión

Partíamos del principio de que las denominadas *lenguas* de especialidad forman parte del conjunto de la lengua como globalidad y en ella pueden constituir conjuntos únicamente virtuales. Si compartimos este principio, las lenguas de especialidad serían únicamente variedades o estilos de la lengua como totalidad. Sería sobre la base de los textos producidos en las situaciones de comunicación especializada que podríamos extraer sus características discriminantes en relación de contraste con las producciones no especializadas. Estas características comprenden tanto recursos léxicos, como morfológicos, sintácticos y gráficos.

De todos los fenómenos que los analistas habían considerado discriminantes, en este estudio empírico sobre un corpus de especialidad amplio hemos podido comprobar que solamente algunos de estos rasgos aparecían con suficiente frecuencia en los textos especializados, pero otros no podían considerarse representativos por su falta de frecuencia. En contraste, se han podido observar otros fenómenos que no habían descrito las obras sobre los lenguajes de especialidad.

De entre los fenómenos no descritos podemos subrayar los siguientes:

- Nombres propios menos representativos en LE
- Predominio de N+Adj en LE
- Pronombres de 1ª persona singular y plural más presentes en LG
- Distribución complementaria de las formas del pronombre relativo (salvo *que*)
- Conjunciones complejas en LE
- *Que* completivo en LG
- Conjunción *o* en LE
- Conjunción *pero, porque, ni* en LG
- Marcadores metadiscursivos en LE, etc.

Y en cambio los datos han confirmado que los siguientes rasgos aparecen como significativos de los textos de especialidad:

- Predominio de nombres (respecto a otras categorías; no más que en LG)
- Empleo especial de categorías gramaticales, sobre todo en relación al verbo (y por tanto a los pronombres personales):
 - Ausencia de la 2ª persona del singular y del plural
 - Raramente, uso de la 1ª persona singular a favor de *nosotros*
 - Uso considerable de la 3ª persona del singular, reforzada con el sujeto impersonal
- Predominio del presente de indicativo (respecto al tiempo pasado)
- Expansión adjetival de los nombres
- Nominalización de formas verbales
- Predominio de la voz pasiva
- *Nosotros*
- *Uno*

Con estos resultados pensamos poder contribuir a la caracterización gramatical de los textos especializados y facilitar así su tratamiento automático.

4. Bibliografía

Beaugrande, R. de; Dressler, W. (1997) *Introducción a la lingüística del texto*. Barcelona, Ariel

Cabré, M.T. (1998) Variació pel tema. El discurs especialitzat o la variació funcional determinada per la temàtica: noves perspectives. En: *Caplletra, Revista Internacional de Filologia*, Tardor, 1998, pp. 137-194.

Cajolet-Laganière, H. and N. Maillet (1995). « Caractérisation des textes techniques québécois », *Présence francophone* 47, pp. 113-147.

Ciapuscio, G. (2003). *Textos especializados y terminología*. Barcelona: IULA.

Coulon, R. (1972). « French as it is written by French sociologists », *Bulletin pédagogique des IUT18*, pp. 11-25.

Harris, Z. (1952) Discourse Analysis. En: *Language*, 28, 1-30, pp. 474-494.

Hoffmann, L. (1976). *Kommunikationsmittel Fachsprache – Eine Einführung*, Berlin: Sammlung Akademie Verlag.

Kocourek, R. (1991). *La langue française de la technique et de la science. Vers une linguistique de la langue savante*, Wiesbaden: Oscar Brandstetter.

L'Homme, M.C. (1993). *Contribution à l'analyse grammaticale de la langue de spécialité : le mode, le temps et la personne du verbe dans quelques textes scientifiques écrits à vocation pédagogique*. Québec: Université Laval.

L'Homme, M.C. (1995). « Formes verbales de temps et texte scientifique », *Le langage et l'homme*, 31(2-3), pp. 107-123.

Lauffler-Laurian, A.M. (1983) Typologie des discours scientifiques : deux approches. En : *Études de Linguistique Appliquée*, 51

Lauffler-Laurian, A.M. (1984) Vulgarisation scientifique: formulation, reformulation, traduction. *Langue Française*, 64, pp. 109-125

Opitz, K. (1980). "Language for Special Purposes. An intractable presence", *Fachsprache* 2(2), pp. 21-27.

Sager, J.C.; Dungworth, D. (1980) *English Special Languages*. Wiesbaden, Oscar Brandstetter Verlag.

