Posibilidades y límites del análisis cuantitativo de corpus especializados

Elena Bogomilova Lozanova
Centro de Estudios Lingüísticos y Literarios,
El Colegio de México

Las aplicaciones de los corpus textuales en terminología son múltiples y permiten conocer cómo se produce, transmite y enseña el conocimiento especializado, y también identificar unidades simples y complejas de carácter terminológico y su uso en textos especializados, detectar procesos de creación neológica y su grado de productividad, para mencionar sólo algunas. Los datos aquí obtenidos son la base para elaborar y actualizar bancos de datos terminológicos y otras herramientas de gran utilidad para los especialistas del área y los profesionales de la lengua, la informática, la información y la documentación.

Este trabajo enfoca el análisis cuantitativo de los corpus especializados y presenta una valoración de las posibilidades y los límites de este tipo de análisis. Con el fin de permitir la verificación de hipótesis sobre procesos y fenómenos lingüísticos, el estudio cuantitativo debe basarse en un corpus textual representativo que debe tener lo siguiente:

- una gran diversidad de textos que asegure la aparición del mayor número de términos del ámbito de conocimiento especializado en estudio, es decir, que permita la entrada de temas muy diferentes

- una adecuada estratificación de los textos que permita obtener buenos resultados en el campo de la dispersión y en el uso estadístico
- una longitud suficiente de los textos, que permita la recuperación del significado global en que aparezcan los términos.

La aplicación de estos criterios a textos del derecho ambiental mexicano le permitió a la autora conformar un corpus especializado con las características requeridas que comprende 83 documentos de esta área jurídica correspondientes a 576,689 palabras. El propósito de la constitución de ese corpus fue identificar los procedimientos de formación de términos en el ámbito jurídico-ambiental. Los resultados obtenidos en el estudio sistemático de los rasgos cuantitativos del corpus del derecho ambiental constituyen una base objetiva para el análisis cualitativo de índole morfosintáctica y semántica de los candidatos a términos de este ámbito especializado. Hay que resaltar que el acercamiento a esta parte del léxico es sólo el primer paso del análisis. Los datos obtenidos son señales de la probabilidad de que un vocablo adquiera valor especializado y se convierta en término; estos indicios deben de ser validados mediante el estudio cualitativo.

Lo que se deseaba obtener del análisis cuantitativo del corpus de datos era:

a) Un número elevado de unidades léxicas probables candidatas a términos.
b) Una base objetiva de selección de aquellas unidades léxicas posibles unidades terminológicas de carácter sintagmático.

Con vista en estos objetivos, se elaboraron dos hipótesis que fundamentan los criterios cuantitativos de identificación terminológica y que guían el análisis:

1. Un vocablo determinado tiene más probabilidades de ser un término técnico cuantas más veces aparezca en un texto reconocido como técnico y, en comparación con un corpus de lengua usual, tenga una mala dispersión en la población léxica,

sesgada precisamente por el carácter técnico del texto en que se encuentre,

2. Un sintagma determinado tiene más probabilidades de ser una expresión terminológica en cuanto:

a) Todas las palabras que parezcan constituir la expresión terminológica tengan una frecuencia de aparición notablemente alta (sin que se puede decir, *a priori*, qué tan alta) en un texto técnico,
b) La frecuencia de aparición de toda la expresión sea más alta de lo que se podría encontrar en un corpus de la lengua general.

Análisis

Para investigar estas hipótesis se procedió a estudiar las fuentes primarias del derecho ambiental, —leyes, reglamentos y normas oficiales mexicanas—. cuantificando con el programa KWIC la totalidad de los tipos de vocablos[1] que aparecen en éstas. Como resultado se obtuvieron 23.535 tipos de vocablos. De acuerdo con la frecuencia total de estos tipos, el conjunto se dividió en dos grupos; el primero contiene los tipos que aparecen por lo menos dos veces en el corpus jurídico-ambiental y abarca 14,138 tipos, el segundo grupo está conformado por 9,397 tipos de vocablos de una sola aparición.

1 La unidad de lengua que me interesa en el análisis es el vocablo al que corresponde, como sintetiza Lara, (*Investigaciones lingüísticas en lexicografía*, México: El Colegio de México, 1979, p.12) la ocurrencia en el habla y el tipo en la Σ hablas. El vocablo es, según Lara (Teoría del diccionario monolingüe, p.119), "una forma léxica abstracta, de naturaleza social y elaborada a lo largo de la historia de la comunidad lingüística"; es la representación de un conjunto de formas léxicas (los tipos de vocablos) que ocurren en el habla como palabras. Así, *aprovechamiento y aprovechamientos* son dos tipos que corresponden al vocablo **aprovechamiento**. En la elaboración del vocablo inciden la simplicidad de la forma propuesta, su brevedad en términos del número de letras que la constituyen, su frecuencia absoluta y su capacidad para construir formas en su paradigma o de derivados.

Con vista en el objetivo del presente trabajo los estudios a continuación se restringen al primer grupo mencionado debido a que el mayor uso de un tipo de vocablo en los textos de un ámbito determinado es una señal de que este tipo es un elemento significativo en la construcción de ese discurso.

Cuando un tipo está documentado una sola vez en el corpus, las conclusiones que se podrían hacer sobre su comportamiento léxico, en general, tendrán un muy bajo grado de confiabilidad, puesto que se puede tratar de un uso casual del vocablo,[2] una preferencia individual del autor del texto o una característica específica del texto que se usó como muestra que no refleja el uso en la sociedad. Pero también tenemos que pensar en la posibilidad de errores de captura o verdaderos errores ortográficos, llamados *hapax*, que se introducen de manera aleatoria en el corpus.

La reducción del umbral de validez a 2,[3] me permite no perder posibles candidatos a términos y captar un mayor número de términos diferentes que forman la riqueza del léxico jurídico-ambiental. Considerar los vocablos con una frecuencia igual o mayor que 2 me garanti-

---

2 Aún hay que añadir que la fijación de un umbral en 2 y no en un número mayor de ocurrencias obedece a factores pragmáticos; estudiamos una terminología que está surgiendo en las últimas décadas y que presenta una variación considerable. Por ello es de suponer que habrá un número significativo de vocablos que a pesar de su baja frecuencia sean términos del derecho ambiental.

3 Contrario a la opinión de muchos especialistas de considerar una frecuencia mínima igual o mayor que 4 para que un vocablo tenga valor en el estudio de la estructura léxica de una lengua. Es muy revelador para esta investigación el comentario crítico que hace Lara (1990: 91-92) respecto a estos valores de la frecuencia absoluta: "(...) si bien la documentación de cuatro o cinco ocurrencias de un vocablo parece suficiente para un estudio lexicológico, cuando se trata de los materiales que requiere la lexicografía, tal cantidad de ocurrencias de vocablo resulta, muchas veces, insuficiente si desea uno atenerse estrictamente a esos datos; pues salvo los vocablos que encuentran una dispersión homogénea en todo el corpus y que, por esa razón, suelen corresponder a un uso generalizado y no marcado social o terminológicamente, todos los demás no alcanzan a definir claramente sus usos y se convierten en accidentes aleatorios más o menos cercanos al verdadero núcleo léxico que uno pretende reconocer, y en síntomas solamente de la extensión real de un vocablo en el uso mexicano".

za poder identificar los términos representativos del derecho ambiental mexicano.

Aparte del importante valor operativo de la frecuencia absoluta de un vocablo en un *corpus* determinado, conviene tomar en cuenta también la dispersión de este vocablo entre los distintos géneros que constituyen el *corpus*; si comparamos dos vocablos del cual el primero ocurre dos veces en la misma fuente y el segundo se manifiesta con igual frecuencia, pero en dos fuentes distintas, podemos concluir que la mejor distribución del segundo vocablo tiene su origen en el mayor uso dentro de la comunidad lingüística respectiva y que el primer vocablo se usa por un grupo menor de hablantes en registros específicos de la lengua.

En una segunda fase del análisis cuantitativo de los textos jurídico-ambientales, encaminada a comprobar la primera hipótesis postulada al inicio del trabajo, comparé el corpus jurídico con el *Corpus del español mexicano contemporáneo* (CEMC) que es un corpus de lengua general.

Un vocablo determinado tiene más probabilidades de ser un término técnico cuantas más veces aparezca en un texto reconocido como técnico y, en comparación con los usos comunes de los demás vocablos de la lengua, tenga una mala dispersión en la población léxica, sesgada precisamente por el carácter técnico del texto en que se encuentre.

Estudio comparativo entre CADAM y CEMC

El CEMC está constituido por 996 textos, de aproximadamente 2000 palabras gráficas cada uno, seleccionados de obras escritas y de grabaciones magnetofónicas originarias de toda la República Mexicana. El equipo de investigadores del proyecto del *Diccionario del Español de México* (DEM) de El Colegio de México, que realizó esta recopilación agrupó estos textos en catorce géneros que corresponden a la lengua culta, la lengua estándar coloquial, denominada también lengua sub-

culta, y la lengua no-estándar que incluye las variedades regionales del español de México.[4]

Debido a que este trabajo se centra en el comportamiento de vocablos con usos específicos en un área científica que tienen una función predominantemente referencial, supongo que éstos se usan, ante todo, en el registro sociolingüístico de la lengua culta tal y como la definen los autores del CEMC. Son los géneros de este nivel de lengua los que me interesan más y, en especial, los textos que provienen de fuentes periodísticas, científicas y técnicas. Las entidades que protege el derecho ambiental son al mismo tiempo también el objeto de estudio de distintas disciplinas científicas y técnicas. Pienso, en particular, en las ciencias naturales como geología, botánica, zoología, las ciencias fisicoquímicas, y los ámbitos científico-técnicos como son las distintas ramas de la ingeniería o las técnicas como las agropecuarias, la caza o el transporte. Los especialistas en estas áreas estudian los seres vivos, las transformaciones de sustancias, las propiedades de la materia y de la energía. El aspecto que el derecho norma y regula son las conductas humanas que pueden influir de una manera relevante en los procesos de interacción entre los sistemas de los organismos vivos y sus sistemas de ambiente e incidir en su protección, conservación, preservación y utilización. El fin que persigue el legislador es garantizar uno de los derechos fundamentales del ser humano consagrado en el artículo cuarto de la Constitución Política de los Estados Unidos Mexicanos: *Toda persona tiene derecho a un medio ambiente adecuado para su desarrollo y bienestar.* Estas reflexiones permiten suponer un número importante de vocablos de origen científico-técnico en los textos del derecho ambiental objeto de estudio.

Pero también el ámbito periodístico se considera como una fuente importante de vocablos especializados, puesto que aquí se usan por primera vez términos que hasta el momento estaban restringidos sólo a un ámbito científico-técnico. Asimismo, en los textos periodísticos se emplean denominaciones originadas en otras lenguas para reflejar problemas actuales de la sociedad que aún no han encontrado reflejo en las ciencias o técnicas del país.

---

4 Información detallada sobre la composición del CEMC puede ser consultada en Lara (1979) y Lara (1984).

De acuerdo con los objetivos de estudio de la presente investigación, se tomaron en cuenta en el análisis del CEMC los siguientes índices estadísticos.[5] a) la frecuencia absoluta, b) la frecuencia relativa y c) el índice normalizado de dispersión: C.

Mientras que la frecuencia absoluta cuenta el número de ocurrencias de cada vocablo dentro del *corpus,* las frecuencias relativas miden el porcentaje de ocurrencias de cada vocablo respecto al total dentro de cada género y entre los géneros. La frecuencia absoluta refleja la intensidad con la cual se emplea el vocablo analizado. Para los propósitos de la elaboración del DEM, mientras más alta fuera la frecuencia de aparición absoluta en el *corpus,* mayor sería la importancia del vocablo y más merecería ser tomado en consideración. Sin embargo, si se trata de identificar los vocablos con significados especializados y con un uso limitado a determinados ámbitos del conocimiento científico-técnico, sucede lo contrario: interesan mucho más los vocablos ubicados en los niveles bajos de la escala de los valores de frecuencia absoluta. La relación entre la frecuencia absoluta y la probabilidad de ser término científico-técnico es inversamente proporcional: a una frecuencia absoluta alta corresponde una probabilidad baja.

Varios aspectos que cuestionan la adecuación de la frecuencia absoluta como ordenadora exclusiva de los vocablos son la influencia de los tipos de género en el uso de los vocablos y del tamaño del CEMC en cada género.[6] Así, los vocablos especializados pueden aparecer en cualquiera de los catorce géneros,[7] pero su frecuencia será siempre

---

5 El cálculo de los índices estadísticos se realizó con el Analizador gramatical automático desarrollado para el proyecto del DEM y que es un parser morfosintáctico con un componente estadístico y un productor de concordancias (García Hidalgo en Lara/Ham 1979: 85-155).

6 Lara, L.F. et al. (1979). *Investigaciones lingüísticas en lexicografía*, Jornadas 89. México: El Colegio de México. pp.51-52.

7 **Lengua culta**[7] **:**

1. Literatura (g_1): obras de literatura, cuentos y ensayos aparecidos en revistas y suplementos culturales
2. Periodismo (g_2): reportajes de autores mexicanos, editoriales, reseñas polí-

más elevada en el género especializado de tal vocablo. Si se considerara únicamente la frecuencia absoluta se le estaría atribuyendo a una palabra la misma importancia que a otra que tiene la misma frecuencia absoluta aunque su aparición no fuera exclusiva de un género y se repartiera entre todos. En el segundo caso, el mayor tamaño de un género dentro del corpus también le da mayor oportunidad a la aparición de sus vocablos propios y especializados. Las observaciones anteriores han llevado al equipo lexicográfico del DEM a incluir el índice normalizado de dispersión C propuesto por Roberto Ham Chande que mide la importancia y el orden de los vocablos tomando en cuenta su frecuencia de aparición en cada género y el tamaño relativo de cada género. Este índice queda establecido en un rango que varía entre 0 y 1 donde 0 indica la distribución más desigual y 1 la más uniforme. Con vista en que C me dirá cuándo un vocablo está concentrado en uno o varios géneros, cuándo es propio del vocabulario general al encontrarse bien distribuido entre todos los géneros, o cuándo hay una situación intermedia y en qué grado, consideré indispensable incluirlo junto con la frecuencia absoluta y la frecuencia relativa de un vocablo como parámetro en la comparación cuantitativa entre CEMC y CADAM.

---

ticas, reseñas sociales, reseñas culturales, reseñas deportivas, reseñas policiacas, reseñas taurinas

3. Ciencias (g_3): bibliotecología, filosofía, historia, culturas indígenas, educación y pedagogía, psicología, antropología, arqueología, derecho, economía, geografía, política, sociología, astronomía, matemáticas, electrónica y electricidad, física, geofísica, computación, biología, química, administración, contabilidad, comercio, medicina y veterinaria, medicina humana, arquitectura, artes coreográficas, artes plásticas, artes gráficas, arte dramatico, música, cine y fotografía.
4. Técnicas (g_4): correos y filatelia, periodismo, publicidad, radio y televisión, transporte, mercadotecnia, ingeniería civil, ingeniería industrial, ingeniería química, ingeniería automotriz, ingeniería aérea, ingeniería de ferrocarriles, ingeniería naval, ingeniería de minas, carpintería, electricidad, mecánica, dibujo técnico, enfermería, corte y confección, albañilería, plomería, herrería, agropecuarias, caza y pesca, ejército, charrería, textos del hogar.

El hecho de que la muestra del CEMC se haya obtenido de documentos del español usado en México entre 1921 a 1974, sin que se reflejen las décadas de los ochenta y noventa, y de que no se cuenta con muestras más recientes de tal extensión y posibilidades de aplicación, limita la efectividad de la comparación.

El análisis comparativo comprende varios pasos realizados, en parte, con la ayuda de distintas herramientas computacionales, como Informix, Keywords in context (KWIC), Word y Excel:

a) confrontación de los tipos del corpus especializado (CADAM) con el corpus de lengua general (CEMC),
b) agrupación de los tipos de CADAM según su aparición o no aparición en CEMC,
c) agrupación de los tipos de CADAM que aparecen en CEMC con un índice de dispersión menor que 0.6 y
d) agrupación de los tipos recogidos bajo c) según la clasificación genérica de la muestra de textos de CEMC.

La fijación del índice de dispersión en los valores menores de 0.6 se fundamenta en la hipótesis de que en cuanto más restringida es la distribución de un vocablo dentro del CEMC tanto mayor es la probabili-

---

5. Discursos políticos (g5): discurso político.
6. Religión (g6): religión.
7. Habla culta (g7): habla culta de la Ciudad de México.

Lengua sub-culta[7] **:**

8. Literatura popular (g8): novela rosa, telenovela, fotonovela, historieta, novela popular.
9. Habla media (g9): habla media de la Ciudad de México.
10. Lírica popular (g10): habla media, habla regional.

Lengua no-estandar[7] **:**

11. Textos dialectales (g11): textos dialectales.
12. Documentos antropológicos (g12): documentos antropológicos.
13. Textos jergales (g13): textos jergales.
14. Textos del hampa y conversación popular (g14): textos del hampa, conversaciones populares.

dad de que se trate de un candidato a término. Los tipos de vocablos que tengan un mayor índice de dispersión que 0.6; adolecen de una restricción genérica y tienen muy poca probabilidad de convertirse en términos técnicos, mientras que los tipos con un bajo índice de dispersión están limitados a ciertos géneros del *corpus*, por lo que tienen una probabilidad más alta de adquirir usos específicos.

De la primera cuantificación del CADAM resultó una lista de 23.535 tipos de vocablos que muestra las 100 palabras más frecuentes en el *corpus*. Depuré de esta lista todas las palabras gramaticales, números y combinaciones de letras que corresponden a partes de distintos vocablos y que por limitaciones del programa computacional kwic no fueron identificados como tales (kwic interpreta el acento como marca delimitativa de una palabra gráfica) y obtuve las 100 palabras plenas con mayor número de ocurrencias en el *corpus* jurídico.

Realicé este recuento con el fin de obtener los vocablos de mayor importancia en los textos del derecho ambiental; es decir, las palabras clave que describen el contenido del los textos en estudio. Identifiqué seis grandes grupos temáticos relacionados con:

a) las instituciones involucradas en la elaboración y el cumplimiento de la legislación ambiental mexicana (p.ej. *secretaría(s), estados*)
b) los distintos tipos de disposiciones jurídicas en la materia y su estructura interna (p.ej. *artículo (abreviatura: art), norma(s), reglamento, ley, diario, fracción, acuerdo, disposiciones*)
c) el bien jurídico protegido (p.ej. *ambiente, naturales, pesca, recursos, especies, agua(s), ambiental, forestal(es), medio, área(s), mar, zona, equilibrio, silvestre*, etcétera)
d) las actividades humanas que tienen consecuencias negativas para el ambiente (p.ej. *aprovechamiento, uso, desarrollo, residuos, peligrosos*)
e) las actividades humanas que tienen consecuencias positivas para el ambiente (p.ej. *protección, manejo, control, conservación, almacenamiento*)
f) el léxico general del discurso jurídico (p.ej. *objeto, materia, plazo, términos, condiciones, fecha, autorización, especifi-*

caciones, medidas, cumplimiento, procedimiento, debe(n), establece, realizar, refiere)

Estos temas sintetizan la estructura conceptual básica del derecho ambiental y me facilitan, como no especialista en la materia, una primera visión panorámica de los contenidos del mismo. Además, estas altas frecuencias de uso pueden estar basadas en el carácter terminológico de los vocablos; por el momento, me sirven sólo como indicios que habría que confirmar o invalidar en los siguientes pasos del análisis cuantitativo y cualitativo.

La próxima fase de la investigación centrada en los 14,138 tipos de vocablos que aparecen, por lo menos dos veces en el CADAM, arrojó que 7,296, es decir, el 51.60%, aparecen también en el *Corpus del español mexicano contemporáneo.* No se reflejan en la lengua ordinaria 6,842, esto es, el 48.39%, de los tipos del CADAM. Estos datos, sistemáticamente recogidos, permiten suponer que la mayoría del léxico empleado en los textos jurídico-ambientales es parte del español mexicano contemporáneo. Al mismo tiempo, hay que subrayar que más de la mitad de los tipos de vocablos que comparten ambos corpus muestran un índice de dispersión más bajo que 0.6 en las fuentes textuales del corpus general, lo que indica que su distribución es muy desigual; el conjunto restante de 3,290 tipos de vocablos está definido por un índice de dispersión mayor que 0.6. El primer grupo está conformado por 4,006 tipos que reducidos a 3,907 tipos al eliminar los tipos con un error de ortografía, corresponden finalmente a 2,987 vocablos. Como se verá a continuación, el uso de estos vocablos es muy restringido considerando el amplio abanico de los géneros de la lengua usual.

Estudio de los vocablos por género

Con el objetivo de comprobar la hipótesis planteada acerca del uso restringido de vocablos con significado especializado, estudio a continuación la distribución genérica dentro del CEMC de aquellos tipos[8] del

8 A continuación utilizaré la forma corta tipo(s) para referirme a tipo(s) de vocablo(s).

corpus jurídico que muestran un índice de dispersión menor que 0.6 en el corpus de lengua general.

Los resultados finales de este estudio confirman los supuestos anunciados al inicio del capítulo acerca de los géneros de CEMC más importantes para la presente investigación.

Identifiqué una concentración de las ocurrencias de los tipos analizados en los textos de las fuentes periodísticas, científicas y técnicas; 3,578 del total de 3,907 tipos; eso es, 91.58%, ocurren en uno o varios de estos géneros. Al lematizar estos 3,578 tipos obtuve 2,729 lemas.

Con vista en estos resultados decidí enfocar el análisis en los tipos de los tres géneros arriba mencionados, suponiendo que es más probable que de estos géneros provengan los candidatos a términos del derecho ambiental.

La comparación de la distribución de los usos entre estos tres géneros me indica, por un lado, que la mayoría de los tipos analizados aparecen en uno o dos géneros, mientras que sólo alrededor de 18% de los tipos se usan en los tres géneros y, por el otro, que los textos provenientes de las ciencias y técnicas parecen ser una fuente más abundante que los textos periodísticos para vocablos de uso limitado en el español mexicano contemporáneo y que posiblemente sean términos del derecho ambiental mexicano.

Respecto a su pertenencia a una categoría gramatical, el análisis me indica que el mayor número de vocablos es una forma nominal. Este dato confirma mi hipótesis inicial con respecto a la naturaleza nominal de la gran mayoría de unidades léxicas posibles candidatas a términos jurídico-ambientales. Estos 2,729 vocablos constituyen junto con los 168 vocablos limitadas al *corpus* jurídico la base de los estudios de carácter cualitativo realizados con el propósito de identificar los términos propios del derecho ambiental.

Confrontación de los tipos de vocablos exclusivos de CADAM con un diccionario de lengua general —DRAE—

La confrontación con la 22ª edición del DRAE tiene el objetivo de recabar aún más datos objetivos e imparciales sobre si los tipos de vocablos del *corpus* del derecho ambiental que no pudieron ser identifica-

dos en el *Corpus del español mexicano contemporáneo*, son parte de la lengua común o se usan con una acepción técnica; el hecho de que un vocablo aparezca sólo en el corpus jurídico y no en el *Corpus del español mexicano contemporáneo* y además no se relaciona con una entrada en el DRAE, eleva la probabilidad de que este vocablo sea término del derecho ambiental.

Una primera revisión del listado de los tipos exclusivos del corpus jurídico sugirió confrontarlos con un diccionario de lengua general de reciente edición suponiendo que aquellos tipos que no aparecen en este diccionario, tengan mayor probabilidad de ser términos del área estudiada. Elegí el DRAE y hallé del conjunto inicial de 6,842 tipos de vocablos restringidos al *corpus* jurídico-ambiental, 2,618 no figuran en el DRAE y 4,224 tipos corresponden a una entrada de este diccionario. El número inicial de tipos 2,618 se elevó a 2,699, puesto que algunos tipos corresponderían a distintas categorías gramaticales; este número de 2,699 tipos al aplicar los criterios de exclusión se redujo a 198 tipos de vocablos, que a su vez corresponden a 168 lemas; es decir, se eliminaron 2,501 tipos que pertenecen a uno de los conjuntos definidos con base en los criterios de exclusión[9]. Entre los lemas identificados se encuentran *abulonero* (adj.; frq.: 3), *acreditamien-*

---

9 La primera revisión de los datos llevó a establecer criterios que permitan excluir además aquellas expresiones que por razones obvias no pueden ser parte del vocabulario propiamente jurídico-ambiental. Las expresiones finalmente excluidas pertenecen a uno de los grupos abajo mencionados:

a) Denominaciones comunes y del latín de taxonomías científicas, ante todo, de biología y química; denominaciones de especies biológicas y de elementos, sustancias y procesos químicos, como por ejemplo, *guayacón, guerrerensis, germanio, grafito, sulfonación,* etc.

b) Expresiones que aparecen únicamente en uno de los dos apartados *fuentes bibliográficas o grado de concordancia con normas y recomendaciones internacionales* de los textos del corpus del derecho ambiental.

c) Denominaciones de los numerales, fórmulas matemáticas, símbolos matemáticos y científicos, en general: *r = radio de la centrífuga = spindle to the center of the bracker,* etc.

d) Denominaciones de nombres propios (personas, lugares, instituciones, organizaciones, empresas, lugares): *Guadalajara, Galindo, Guzmán.*

e) Denominaciones de publicaciones periódicas, e.g.: *Gaceta Ecológica, Diario Oficial de la Federación.*

to (sust.; frq.: 11), *acuacultur* (sust.; frq.: 2), *acuacultura* (sust.: frq.: 111), *acuitardo* (sust.; frq.: 3), *agroforestería* (sust.; frq.: 8), *antropogénico* (adj.; frq.: 2), *autocero* (sust.; frq.: 3), *bioacumulables* (adj.; frq.: 2), *corraleo* (sust.; frq.: 4), *desforestación* (sust.; frq.: 3), *ecotono* (sust.; frq.: 2), *entrecara* (sust.; frq.: 2), *redoblamiento* (sust.; frq.: 10), *traslocación* (sust.; frq.: 5), *zoosanitaria* (adj.; frq.: 4).

Conclusiones del análisis cuantitativo

Serán objeto del análisis cualitativo sólo aquellos vocablos que aparecen tanto en el corpus jurídico-ambiental como en el *Corpus del español mexicano contemporáneo* que muestran en este último *corpus* un índice de dispersión menor que 0.6 y que fueron identificados en textos de los ámbitos periodísticos, científicos y técnicos. Asimismo forman parte del conjunto de datos identificados con base en el estudio cuantitativo los vocablos que figuran sólo en el corpus especializado y que no corresponden a una entrada de la 22ª edición del *Diccionario de la Real Academia Española* (DRAE).

En relación con el grado de objetividad del inventario de vocablos candidatos a términos conviene señalar la diferencia entre juzgar los resultados a partir de una consideración intuitiva de la "realidad" léxica con lo que la evaluación se torna imposible al quedar sujeta a la experiencia de cada especialista en materia jurídico-ambiental, juzgarlos en comparación con trabajos realizados bajo muy diferentes enfoques (por ejemplo, con diccionarios especializados elaborados con objetivos distintos a los presentes en esta tesis) y evaluarlos tras un análisis

---

f) Denominaciones de las distintas disposiciones jurídicas, como son reglamento, Norma Oficial Mexicana en materia de ecología, ley, etc.

g) Nombres de medidas y equipos e instrumentos de medición: *centímetros*,

Se excluyen estas expresiones, debido a que por razones obvias no pueden ser parte del vocabulario propiamente jurídico-ambiental. La razón fundamental para ello es el hecho de que con estos vocablos se denominan los objetos de la legislación jurídico-ambiental en sí y no nos dicen nada sobre el cómo de la actividad legislativa. Este trabajo debe permitir a detectar los términos sobre que se ha legislado y que se restringen al ámbito medioambiental.

cuantitativo basado en criterios objetivos e imparciales. El hecho de que el corpus es representativo para el ámbito especializado en estudio y no era necesario limitarlo a un muestra de un determinado número de párrafos de las fuentes y la forma en que se realizó el estudio cuantitativo, me permite esperar un gran acercamiento a las realidades del léxico propio a esta área del derecho.

Los resultados obtenidos me permiten constatar que el estudio cuantitativo ha sido de gran utilidad debido a los siguientes datos:

1. Pude limitar con métodos objetivos y confiables el conjunto de vocablos candidatos al derecho ambiental.

2. La comparación de un *corpus* especializado con un *corpus* de lengua general con el fin de obtener información confiable sobre el primero, es un método de análisis aún poco usual en el ámbito de la terminología. Es a partir de la primera década del siglo XXI cuando aparecen los primeros estudios terminológicos basados en una comparación con textos de la lengua común.[10]

3. El estudio confirmó las hipótesis planteadas al inicio del capítulo.

Llegué a la conclusión de que las hipótesis planteadas sirvieron para filtrar de una considerable muestra de textos del derecho ambiental aquellos vocablos que con mayor probabilidad pueden ser términos o parte de una unidad terminológica compleja.

Al mismo tiempo debo anotar que no existe un índice numérico que indique con certeza si se trata de una expresión terminológica o no. Es decir, no es posible decir, por ejemplo, si la formación terminológica en derecho ambiental comienza a partir de un índice de dis-

---

10 En el año 2003 Patrick Drouin esboza en *Terminology*, la revista más renombrada en esta área, por primera vez, una nueva técnica para la extracción de términos de corpus especializados basada en la comparación con un corpus no especializado; su objetivo es "to reduce the amount of noise in the list of candidate terms (CTs) by restricting the lexical items that can appear incide candidate terms".

persión de 0.3 o 0.4, o que una expresión que aparezca por lo menos 3 o 4 veces en el corpus jurídico, es candidata a término. Junto con esta limitante del análisis cuantitativo hay que aludir a las restricciones temporales del CEMC y a la falta de un corpus de fecha reciente.

El carácter terminológico de una expresión simple o compleja se puede corroborar sólo a través del estudio semántico de los contextos de uso de la expresión; no basta con la cuantificación de los significantes, sino que hay que estudiar también el plano del significado de cualquier signo lingüístico. Es decir, tenemos que abandonar el ámbito de la estadística para adentrarnos en la sustancia del contenido.

Bibliografía

Drouin, Patrick, "Term extraction using non-technical corpora as a point of leverage", en: *Terminology* 9:1, 2003, 99-115.

Lara, Luis Fernando *et al.* (1979), *Investigaciones lingüísticas en lexicografía*, Jornadas 89, México, El Colegio de México, pp.51-52.

Lozanova Bogomilova, Elena, *Estudio de los procedimientos de formación de términos en el derecho ambiental mexicano*, Tesis de doctorado, El Colegio de México, (en proceso).

Real Academia Española, *Diccionario de la lengua española,* 22ª edición.