Revista Signos ISSN 0718-0934
© 2011 PUCV, Chile
DOI: 10.4067/S0718-09342011000100004

44(75) 48-67

Estudio diacrónico de la terminología especializada utilizando métodos cuantitativos: Ejemplos de aplicación a un corpus de artículos de lingüística aplicada[1]

Diachronic study of specialized terminolgy using quantitative methods: Example with an application to a corpus of papers on Applied Linguistics

Rogelio Nazar*
rogelio.nazar@upf.edu
Universitat Pompeu Fabra
España

Recibido: 24-III-2010 / Aceptado: 15-XI-2010

Resumen: Este artículo presenta una metodología para el análisis de la evolución de la terminología de un dominio especializado, medida en variación de frecuencia de uso, aparición y desaparición de los términos. Como ejemplo, el artículo describe los resultados de la aplicación de esta metodología a un corpus conformado por las actas de los congresos organizados por la Asociación Española de Lingüística Aplicada, entre los años 1983 y 2006. La metodología se resume en un algoritmo cuantitativo independiente de lengua que acepta como entrada un conjunto de ficheros de texto organizados por años y como salida selecciona términos de ese corpus calculando cómo se distribuyen sus frecuencias a lo largo del tiempo. Las propiedades geométricas de las curvas que representan las frecuencias de uso de esas unidades terminológicas permiten identificar automáticamente aquellas unidades que se ponen de moda en el dominio así como las que dejan de utilizarse. Metafóricamente, se trata de una radiografía de los cambios de paradigma que se van dando a lo largo de la historia del campo, pero también de una neología y una arqueología de su terminología, rescatando términos que sería difícil encontrar mediante inspección manual debido a la escala del corpus. El objetivo específico del artículo es proponer una alternativa a otros modelos existentes para el estudio de unidades en la escala temporal que se limitan a seguir la curva de distribución de frecuencias en el tiempo de unidades elegidas arbitrariamente por un usuario. La alternativa en este artículo ofrece una visión distinta porque es proceder del modo inverso, en lugar de introducir unidades léxicas para ver sus curvas, introducir las curvas para obtener las unidades. La utilidad de estos conjuntos de unidades puede variar en función de las necesidades. Por ejemplo, la creación de glosarios terminológicos de distintos tipos (en papel o en formato electrónico) puede requerir ya sea una nomenclatura que incluya sólo la terminología firmemente establecida en el campo o, en otros casos, incluir también las unidades neológicas o en desuso.

Palabras Clave: Extracción de terminología, estadística de corpus, lingüística cuantitativa.

Abstract: This paper presents a methodology for analyzing the evolution of the terminology used in a specialized domain. Such terminology is measured according to its variation in the frequency of use, as well as the appearance and disappearance of the terms. As an example, the paper reports the results of the application of this methodology to a corpus made up of the 1983-2006 Spanish Association of Applied Linguistics proceedings. The methodology can be summarized in a quantitative and language-independent algorithm that accepts a set of text documents organized by years as input and offers a selection of terms as output by calculating their frequency distribution over time. The geometrical properties of the curves representing the frequency of use of the terminological units help to automatically identify those which come into use and those no longer in use. Metaphorically, the paper offers a kind of radiology of the paradigm shifts that occur in the history of the field as well as a neology and an archeology of its terminology, revealing terms that would be otherwise hard to find due to the scale of the corpus. The specific objective of this paper is to propose an alternative to other methods which only consider curves of frequency distribution of units in the time line arbitrarily selected by a user. This paper offers a new view because it is the reverse procedure: instead of introducing lexical units to study their frequency curves, the curves to obtain the units are introduced. The usefulness of these sets of units may vary according to the needs. For instance, the creation of glossaries of different types (hard copy or electronic format) may require a nomenclature that includes only the terminology firmly established in the literature or, in other cases, neologisms or terms no longer in use.

Key Words: Terminology extraction, corpus statistics, quantitative linguistics.

INTRODUCCIÓN

En este artículo se presenta un estudio de evolución en el tiempo de la terminología de un dominio científico. El interés por la terminología especializada (ver Sección 1) se da tanto desde un punto de vista teórico en lingüística como desde un punto de vista aplicado a la tarea terminográfica. Al lingüista no le interesará tanto el término en sí sino el funcionamiento de la terminología como sistema en el discurso. Al terminógrafo, en cambio, le interesará el término para la compilación de diccionarios terminológicos que representan una ayuda vital para los traductores de textos de especialidad así como para las propuestas de normalización terminológica, fundamentales para la especificidad en la designación de conceptos y la claridad en la comunicación entre especialistas.

En los últimos años ha despertado gran interés la extracción automática de terminología, como un recurso con el cual los terminólogos pueden no solo automatizar parte del proceso de compilación de diccionarios sino además justificar de manera empírica la decisión de incluir una u otra unidad terminológica en la nomenclatura. Desde la vertiente aplicada, este trabajo puede interesar por ser un método empírico y en gran medida automatizado para la selección de la nomenclatura del glosario de un ámbito especializado, por lo tanto, podría ser clasificado dentro de la familia de algoritmos de extracción de terminología. Sin embargo, este sería un subproducto de la propuesta, ya que el objetivo principal está en el estudio de la evolución de términos en una muestra diacrónica. Esta evolución se puede medir observando las tendencias de variación en la frecuencia de uso de los términos, que reflejarán los cambios de paradigma de la historia del campo. Lo fundamental de la propuesta, en comparación con otros trabajos como el de Google Ngrams Viewer, tal como se presenta actualmente (Michel, Shen, Aiden, Veres, Gray, Google Books Team, Pickett, Hoiberg, Clancy, Norvig, Orwant, Pinker, Nowak & Aiden, 2010), es que en lugar de ofrecer la distribución de unidades arbitrariamente seleccionadas por el usuario, lo que este sistema hace es el proceso inverso: obtener las unidades léxicas a partir de curvas de distribución de frecuencias arbitrariamente introducidas.

El dominio elegido para el experimento de extracción de terminología es la lingüística y la muestra elegida para el análisis son los textos de las actas de congresos que publicó la Asociación Española de

Lingüística Aplicada desde el año 1983 hasta 2006, que se encuentran disponibles en formato digital[2]. Se reporta por tanto la aplicación a este corpus de un algoritmo estadístico independiente de lengua que acepta como entrada un conjunto de ficheros de texto organizados por años y como salida selecciona términos del corpus calculando las propiedades geométricas de las curvas que representan sus frecuencias de uso a lo largo del tiempo.

La noción de término en este caso está metodológicamente sesgada por razones de conveniencia práctica. Un término es simplemente una palabra o una secuencia de palabras con una frecuencia especialmente informativa, es decir, que el criterio no es estrictamente terminológico sino estadístico. La estrategia de extracción de términos consiste en asignar a una palabra o una secuencia de palabras un valor de 'terminologicidad' basado en su rareza. La rareza de un término está dada por una frecuencia de aparición relativamente alta en el corpus de especialidad (en este caso las actas de los congresos) y relativamente baja en un corpus de referencia del lenguaje general (en este caso prensa española). También por conveniencia práctica, el corpus no es sometido a ningún tipo de procesamiento, como lematización, etiquetado morfosintáctico o agrupación de constituyentes sintácticos. Esta simplificación obviamente se hace a expensas de un grado de error en la detección terminológica, pero el resultado es suficiente a los fines prácticos de una primera descripción de la evolución de la terminología del campo. De cualquier forma, se incluye también un experimento paralelo aplicando un filtro sintáctico (generado estadísticamente) que permite cierta reducción del ruido (ver Sección 2.2.3).

El objetivo del presente artículo no es, entonces, presentar un extractor terminológico, y por esta razón no se persigue el máximo rendimiento posible en la precisión y cobertura de la selección de los términos. El refinamiento en la selección de los términos se deja como un proceso ulterior, que requerirá seguramente la combinación de distintas estrategias y la utilización de conocimiento léxico y sintáctico de la lengua analizada. Por el contrario, el objetivo perseguido es el de apoyar el trabajo del terminólogo en la creación de un glosario de especialidad con un fundamento empírico, que sirva como base sólida para la decisión de incluir una u otra unidad léxica en la nomenclatura. En función del tipo de obra terminográfica que se desee elaborar, pueden concebirse distintos perfiles de nomenclatura. Una obra en papel de tamaño reducido requerirá, típicamente, una selección de la nomenclatura que incluya la terminología más firmemente establecida en la historia del campo. Una obra especializada en la neología de un campo, en cambio, centrará la selección en las unidades más recientes. En otros casos, como obras de mayor tamaño o que no revisten dificultades para almacenar grandes cantidades de entradas –como bases de datos o demás recursos electrónicos– no encontrarán motivos para no incluir en su nomenclatura incluso aquellas unidades que han dejado de utilizarse en la disciplina.

En lo que respecta a la selección del corpus de análisis, las actas de los congresos de AESLA representan simplemente un ejemplo de aplicación, como se ha advertido ya, y su selección es meramente arbitraria. En el caso de este artículo, se trata de una muestra representativa de un dominio científico (la lingüística aplicada) que cumple con el doble requisito de tener el tamaño y la extensión a lo largo de una ventana temporal suficientes para llevar a cabo este tipo de análisis cuantitativo.

El artículo se organiza de la siguiente manera: la Sección 1 presenta un panorama muy escueto de la bibliografía sobre terminología diacrónica y extracción automática de terminología, las áreas en las que este trabajo se enmarca; la Sección 2 contiene toda la investigación desarrollada, desde el planteo de la hipótesis hasta su comprobación empírica y, finalmente, la Sección 3 presenta la discusión de los resultados y algunas líneas de trabajo futuro.

1. Antecedentes

Como se dijo en la Introducción, el estudio de la terminología especializada es un dominio de interés tanto para la teoría lingüística como para la práctica terminográfica. La terminología como disciplina surge primero como práctica normativa en el seno de los organismos de estandarización (Wüster, 1979; Arntz & Picht, 1989) y posteriormente como un campo de investigación en lingüística (Sager, 1990; Cabré, 1999; Cabré & Estopà, 2005). Desde el punto de vista lingüístico, las unidades terminológicas se consideran como parte de la lengua y son posibles por tanto de ser analizadas lingüísticamente. Como práctica, la terminología es mayoritariamente la creación de glosarios, fundamentales para la tarea de los traductores de textos de especialidad así como para la tarea de normalización terminológica.

En la bibliografía sobre terminología ocupa un lugar importante la teoría y práctica de la extracción de terminología. Desde el punto de vista práctico, se trata de automatizar la tarea del terminólogo, pero esto conlleva necesariamente una definición formal de lo que puede ser considerado un término, formalización necesaria para la implementación informática pero de importantes consecuencias teóricas. El investigador, en este punto, se ve obligado a plantearse cómo determinar el estatus de los términos. ¿Se debe preguntar acerca de las condiciones necesarias y suficientes para que una palabra o cadena de palabras sea considerada un término o debe hablarse de distintos grados de 'terminologicidad'? Desde la perspectiva de Cabré (1999), ninguna de las dos alternativas son procedentes, ya que se trata de advertir que ciertas unidades léxicas 'activan' un valor de especialidad cuando aparecen en un contexto especializado como el de la literatura científica. De esta manera, una misma palabra puede tener un uso no especializado en la lengua cotidiana y a la vez funcionar como un término en la comunicación entre especialistas.

Desde diversos puntos de vista, la literatura sobre los sistemas de extracción de terminología es abundante y solo es posible señalar algunas referencias orientativas. Para una introducción más amplia, véase los trabajos de Kageura y Umino (1996) y los reunidos en Bourigault, Jacquemin y L'Homme (2001), particularmente Cabré, Estopà y Vivaldi (2001). Existen propuestas claramente orientadas a la incorporación de conocimiento de la lengua analizada, como patrones morfológicos o sintácticos (Ananiadou, 1994; Jacquemin, 1997). Por otro lado, existe una gran profusión de algoritmos estadísticos que calculan medidas como la asociación entre los componentes de unidades poliléxicas o la forma en que se distribuyen los términos en los conjuntos de documentos (Sparck Jones, 1972; Daille, 1994; Pantel & Lin, 2001; Patry & Langlais, 2005), aunque en ambas vertientes se dan distintos grados de hibridación entre conocimiento lingüístico y estadístico, incluyendo también conocimiento ontológico del dominio de especialidad (Maynard & Ananiadou, 2000; Vivaldi, 2001; Sheremetyeva, 2009).

Lo que salta a la vista ante la gran cantidad de bibliografía sobre extracción de terminología es que los autores en general parten del supuesto tácito de que el algoritmo tiene que extraer los términos a partir de un documento o de un corpus tratado como unidad. En este sentido, uno de los aportes de este artículo es el abordar una perspectiva más amplia de manera tal que el algoritmo extractor no analice solo un documento sino una publicación de referencia en el campo. Esto aporta a su vez el eje diacrónico, no tan frecuentemente utilizado en los estudios sobre terminología en comparación con los estudios de tipo sincrónico. Recientemente, algunos terminólogos –como Temmerman (2000) o Dury & Picton (2008)– han reaccionado contra esta tendencia reivindicando el eje diacrónico entre otros principios y criticando distintos fundamentos de lo que se conoce como la teoría terminológica tradicional. En la actualidad comienza a aceptarse la idea de un estudio diacrónico de la terminología especializada como un espacio de saber diferenciado de otras aproximaciones históricas a los ámbitos de especialidad como la sociología de la ciencia (Merton, 1973), la filosofía de la ciencia (Lakatos, 1974) o la historia de la ciencia (Kuhn, 1962; Barona, 1994), aunque no por ello deberían ser considerados ámbitos del saber totalmente desconectados, puesto que la historia de los términos especializados es también parte de la historia de los conceptos de las diferentes disciplinas.

Entre los antecedentes del estudio empírico de la diacronía en terminología, cabe destacar, entre otros, los trabajos reunidos en el volumen preparado por Groult, Louis y Roger (1988) acerca de las migraciones de vocabulario científico entre diferentes ciencias, con los cambios de uso y resemantización que tales migraciones comportan. Otros autores, como Meyer y Mackintosh (2000), se interesan por los procesos de fluctuación del significado de los términos científicos a lo largo del tiempo. Concretamente, se interesan por los casos en que se produce una 'desterminologización' de las unidades que pasan de un uso especializado en la comunicación entre especialistas a un uso no especializado en círculos más amplios de la población, tal como en el caso del término inglés *bandwith* ('ancho de banda'), que inicialmente tiene un sentido técnico que hace referencia a la capacidad de un canal para transmitir información y pasa a ser utilizado de manera no especializada como la capacidad de un individuo para hacer frente a una carga de responsabilidades, como en la expresión *I'm out of bandwith* ('no me alcanza el ancho de banda') dicha por un empleado desbordado de trabajo. Algunas de estas unidades, incluso, acaban su transformación reinfiltrándose en ámbitos de especialidad (reterminologizándose), a veces también con nuevas cargas o connotaciones adquiridas durante su período de uso como palabra de léxico general.

Posiblemente, el aspecto del estudio diacrónico de la terminología especializada que haya generado

la mayor cantidad de trabajos sea el estudio de la neología especializada, tal vez por influjo de los estudios sobre neología en general (Boulanger, 1988; Cabré & Estopà, 2009). Distintos autores (Rondeau, 1984; Humbley, 2003; Desmet, 2003) justifican una precisión terminológica separando la neología, que sería el estudio de las palabras nuevas en el léxico general, de la neonimia, que sería el estudio del nacimiento (o difusión) de nuevos términos especializados. Pioneros en el análisis de la neología especializada, sin embargo, deben ser los trabajos del Office Québécois de la Langue Française (Corbeil, 1988; Célestin & Bergeron, 2003) que, presionado por la necesidad de proteger la lengua francesa del influjo de la terminología especializada en inglés, impulsó el estudio y la normalización de la actividad neológica. Respecto de estudios sobre neología producida por el paso de terminología especializada al uso en lengua general, Pozzi, Benítez y Morett (2008) presentan un estudio en prensa escrita mexicana inspirado en los criterios del Observatori de Neologia (2003) del Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra. Examinaron manualmente un conjunto de artículos en un período de tres años para identificar y posteriormente clasificar la terminología encontrada según distintas categorías de análisis, entre ellas la categoría gramatical, los procesos de formación, la afijación y también el nivel de especialización, que sería el grado en que los significados de los términos son conocidos por la población general. El enfoque de Tartier (2003), finalmente, es el estudio diacrónico de los términos dentro de los mismos ámbitos de especialidad, tal como es el caso del presente artículo. Para ello lleva a cabo un análisis sistemático de un corpus especializado diacrónico de dimensiones importantes, aunque el estudio no está orientado en su caso al seguimiento de la evolución de la terminología en función de la frecuencia de uso en las distintas épocas, como en el presente artículo, sino en los cambios formales que experimentan los términos a lo largo del tiempo, para lo cual se sirve de una medida de similitud ortográfica (la distancia de edición o distancia de Levenshtein) como medio para determinar cambios en la forma de las unidades terminológicas tanto simples como sintagmáticas.

2. La investigación

2.1. Hipótesis

Se formula la hipótesis de que un estudio que en principio podría circunscribirse a la terminología, el análisis del discurso especializado o la sociología de la ciencia, puede ser reducido a un problema geométrico si la pregunta de investigación puede ser planteada de manera también geométrica. Más específicamente, según esta hipótesis, las curvas que representan la distribución de frecuencias de los términos a lo largo del tiempo nos ofrecen una lectura de cuán informativos son los términos en función de su 'ciclo de vida'.

La pregunta que pretende responder esta hipótesis es si la distribución de la frecuencia de uso de las unidades en el eje diacrónico puede aportar una información relevante a la hora de establecer la macroestructura de un diccionario terminológico. Habrá términos cuya frecuencia de uso a lo largo del tiempo será muy volátil, o tendrá un pico concentrado en cierto período. Esos serán los términos que se ponen de moda en cierto momento y luego se dejan de utilizar. Habrá otros términos cuya frecuencia de uso tiene una tendencia descendente, los términos que caen en desuso. De la misma manera, habrá términos que comienzan a implantarse en los últimos años de la muestra, los neologismos de la disciplina. Y habrá términos cuya frecuencia de uso es relativamente constante. Estos últimos términos (y también los apellidos de los autores de la disciplina, tanto por sus propias publicaciones como por las referencias a ellos por parte de otros autores) representarían la terminología nuclear o más establecida en el campo, terminología que pueden compartir autores de cualquier período dentro de la ventana temporal estudiada.

2.2. Comprobación empírica

En esta sección, la hipótesis presentada en el apartado 2.1. es sometida a una serie de pruebas empíricas. Se explica por tanto cada uno de los pasos de los experimentos realizados y de los algoritmos utilizados. Los resultados se muestran solo parcialmente en este artículo por razones de espacio, sin embargo, los datos de los resultados en formato digital se pueden consultar en un servidor[3].

2.2.1. Preparación del corpus

La tarea de constitución y preparación del corpus ofrece cierta dificultad por la diversidad de formatos en los que dicha muestra se encuentra. La mayor parte del material está escaneado como imagen y no como texto, por lo tanto, esta porción del corpus tiene que ser sometida a un proceso

de reconocimiento óptico de caracteres. La poca definición de la imagen, más la deficiente calidad de impresión particularmente en las primeras ediciones, produce una tasa de error importante y la consecuente pérdida de datos. En el caso del primer año de la serie, casi un tercio de las páginas no pudo ser procesado debido a escasa resolución. Este porcentaje se va reduciendo en las ediciones más recientes. En el caso de los archivos que están digitalizados como texto, cada edición exige un tratamiento específico ya que los textos se encuentran en formatos diversos. Una vez convertidos los datos a ficheros de texto plano, la preparación del corpus finaliza con la ubicación de cada edición en un directorio que lleva por nombre el año correspondiente, ya que este es el formato de entrada del algoritmo desarrollado para este estudio.

2.2.2. Representación de la distribución de frecuencias

El estudio diacrónico impone una serie de restricciones que por lo general no son tenidas en cuenta en la lingüística de corpus sincrónica y esto abarca medidas tan generales como la frecuencia de aparición de las palabras. Como consecuencia de que en el año 1983 AESLA editara menos cantidad de texto para poder estudiar la evolución de la frecuencia de un término tenemos que corregir esta situación utilizando frecuencias relativas al año. Esta medida no resuelve el problema en verdad, ya que si las diferencias en tamaño de las distintas particiones del corpus son muy grandes, entonces las probabilidades de aparición de las palabras ya no serán las mismas. Una palabra tiene más oportunidades de aparecer cuando la muestra es grande. A modo ilustrativo, la Figura 1 muestra las curvas correspondientes a unidades arbitrariamente elegidas (subordinada y colocaciones) para ver cómo evoluciona su frecuencia de aparición a través del tiempo. Las dos curvas se oponen porque la de la primera unidad tiende a ser utilizada cada vez menos mientras que el uso de la segunda describe un aumento. Esta gráfica parece reflejar el cambio en el centro de gravedad en el debate lingüístico desde temas sintácticos hacia el estudio de las colocaciones.

Para poder implementar esta herramienta de representación de frecuencias de uso de los términos fue necesario indexar previamente el corpus con las frecuencias de aparición de todas las palabras y también de todas las combinaciones o cadenas de palabras (enegramas) de hasta cinco componentes. Es decir que, por ejemplo, en este índice tanto el término 'adjetivo' como el término 'adjetivo calificativo' pueden ser entradas. Existen algunas restricciones para la confección de este índice, sin embargo, no aplican las mismas restricciones que se detallan en la Sección 2.2.4.1. para la selección de la muestra de términos a estudiar. En este índice, en cambio, se registran todas las palabras, con excepción de aquellas que: a) tengan una frecuencia absoluta total inferior a 3; b) sean miembros de una lista de exclusión; c) en el caso de los enegramas, que tengan como primer o último componente un miembro de la lista de exclusión. La lista de exclusión es definida como la lista de las cien palabras más frecuentes en un corpus de referencia de lengua general conformado principalmente por artículos de periódicos y de un tamaño de dos millones de palabras. Las cien palabras más frecuentes coinciden con el segmento menos informativo del vocabulario de una lengua, el de las llamadas palabras gramaticales, es decir, preposiciones, artículos, copulativas, etc. Por lo tanto, mientras términos como 'lingüística' o

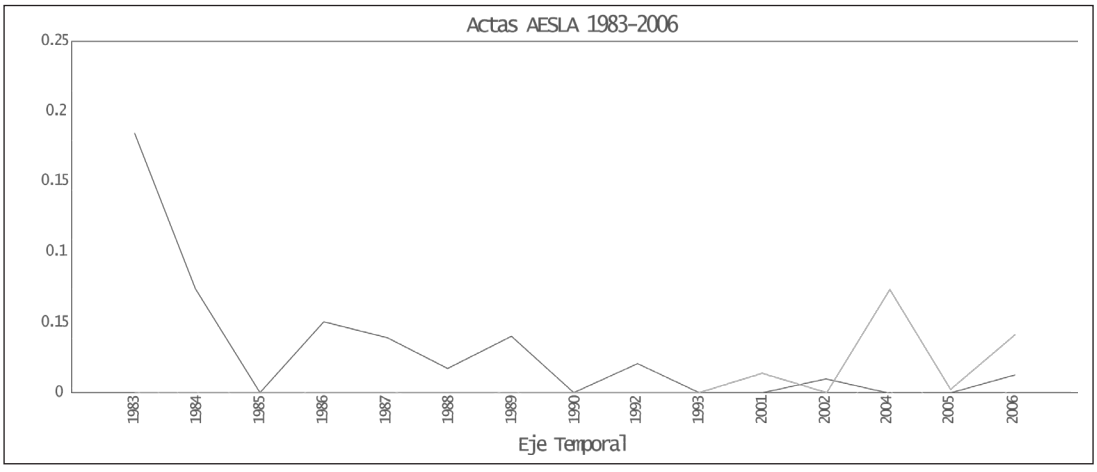

Figura 1. Frecuencia relativa de los términos 'subordinada' y 'colocaciones'.

'lingüística española' serán indexados, la secuencia 'la lingüística española' no lo será porque empieza por 'la'. Estas listas de exclusión se confeccionaron para las tres principales lenguas de las actas que son el castellano, el inglés y el francés.

Con el objeto de disponer de un punto de referencia sobre la selección de las unidades terminológicas a partir del corpus, la Figura 2 muestra (en escala logarítmica) la distribución de frecuencias de las entradas de un diccionario terminológico del área, el diccionario de lingüística del TermCat (1992) en todo el corpus de los textos de las actas de AESLA. En el caso de los adjetivos, que en el diccionario incluyen también la marca de flexión en femenino (como en el caso de 'sincrónico-a'), se buscaron en el corpus y sumaron las frecuencias de ambas formas. En esta figura podemos observar que casi dos tercios de las entradas aparece en el corpus por lo menos una vez (la comprobación no se hizo con respecto al índice del corpus sino con los textos directamente, ya que en el índice no se registran los hapax legomena y dis legomena) lo cual indica que se trata de una buena nomenclatura ya que refleja la terminología que se encuentra realmente en uso en la lingüística aplicada.

Para hacer una estimación aproximada de la cobertura del mismo diccionario, se puede tomar como referencia el porcentaje de una muestra aleatoria de términos tomados del corpus a los que corresponde también una entrada en el diccionario, porcentaje que en este caso alcanza el 32%. Es decir que, si bien la nomenclatura del diccionario está bien elegida en el sentido en que se reflejan unidades que están realmente en uso, existe todavía en el corpus una gran cantidad de términos que aún no han sido documentados.

2.2.3. Selección de las unidades terminológicas

Como un paso necesario para el ordenamiento de las unidades terminológicas, se debe hacer una selección de las unidades que conformarán la muestra sometida a análisis. Mientras en la herramienta de consulta se incluyeron todas las palabras o secuencias de palabras del corpus, ahora queremos someter a estudio no todas las palabras sino aquellas que sean interesantes desde un punto de vista terminológico. Es decir, aquellas que sean más informativas o que se acercarían más al conjunto de las unidades para ser tenidas en cuenta para la

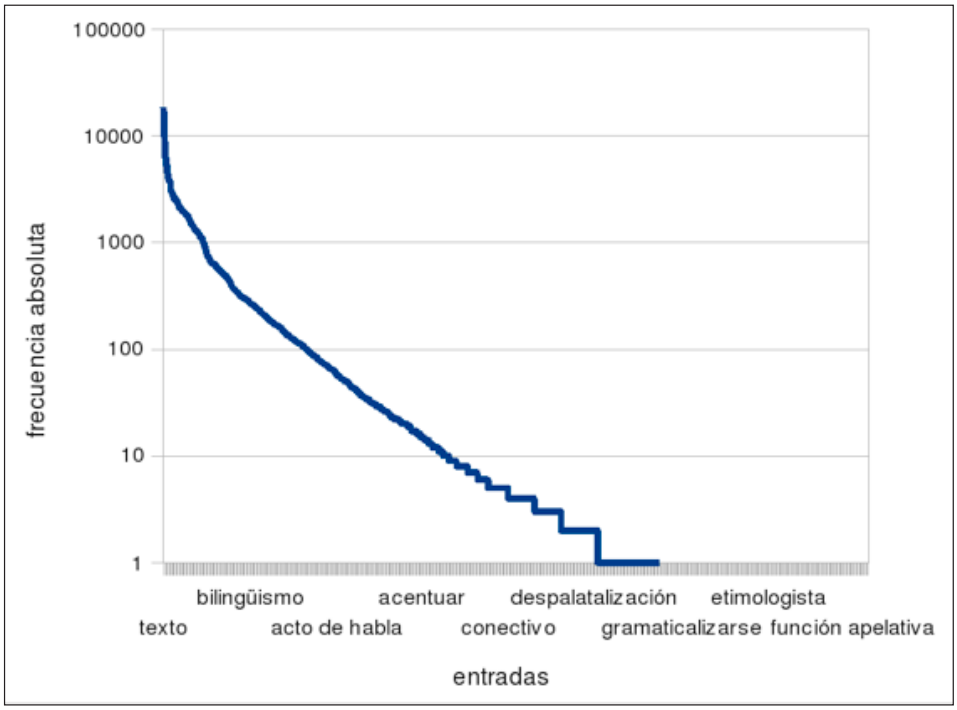

Figura 2. Frecuencias en las actas de AESLA de los 1.475 términos que aparecen en un diccionario de lingüística del TermCat (eje vertical en escala logarítimica).

confección de un glosario del dominio. Esta muestra, como listado de formas candidatas a término, será ordenada por los distintos coeficientes que se presentarán a continuación.

A partir del índice obtenido en 2.2.2., se eliminaron todas las palabras y secuencias de palabras que son más comunes en castellano, inglés y francés. Esto es posible mediante un modelo de esas lenguas elaborado a partir del mismo corpus de referencia de textos periodísticos también utilizado en la Sección 2.2.2. Estas unidades son eliminadas porque son consideradas elementos de la lengua general y no del dominio de especialidad en cuestión. En el caso del castellano, que es la lengua de la mayor parte de los textos de las actas, el modelo fue elaborado a partir de los archivos del periódico El País[4]. Todas las unidades que presentan un uso continuado a lo largo del tiempo en el corpus de este periódico son eliminadas por considerarse unidades del léxico común (ver Sección 2.2.4.5 para los detalles sobre cómo calcular una frecuencia de uso continuo en el tiempo). El grado de error que ello conlleva se ve agravado en el caso de un dominio como la lingüística, donde existe una importante cantidad de términos que tienen la misma forma de una palabra de la lengua general, como 'vocabulario', 'gramatical', 'hablante', 'verbo', 'léxico', 'oración', etc. Debido a este motivo, las exclusiones no están basadas en un sistema de reglas categóricas sino ponderando la relación entre las frecuencias de una unidad en ambos corpus, de manera tal que cuanto mayor sea la diferencia, mayor será la terminologicidad de esa unidad.

Como un experimento paralelo se implementó además un filtro sintáctico para la selección de los términos con el objetivo de reducir el grado de error en la selección de la nomenclatura, evaluada en términos de precisión y cobertura. Sin entrar en los detalles, se puede describir como un algoritmo estadístico con aprendizaje supervisado, entrenado con un diccionario terminológico del área con el objeto de identificar las secuencias de categorías gramaticales que son frecuentes en las entradas de ese diccionario. En otras palabras, un modelo sintáctico de las entradas, que luego permite segmentar un texto sometido a análisis, identificando aquellas secuencias que podrían ser terminológicas según dicho modelo sintáctico. Obviamente no se trata de reconocer en un texto analizado las unidades que se encuentran ya en el diccionario (lo cual sería una tarea prácticamente trivial), sino de reconocer unidades en el texto con una estructura sintáctica similar a las de las entradas del diccionario. Así, por ejemplo, con este método el algoritmo aprende que la categoría sustantivo o las secuencias como sustantivo+adjetivo o sustantivo+de+sustantivo son muy frecuentes en las entradas del diccionario y, por lo tanto, si encuentra esas secuencias en el texto las privilegiará como candidatos a término.

En la página web que ofrece los resultados del presente artículo (ver Nota 3) se incorpora también el resultado de la muestra de candidatos a término que ha sido obtenida con el filtro sintáctico después de haber entrenado el algoritmo con el mismo diccionario terminológico del área utilizado en la Sección 2.2.2. Después de aplicar el filtro sintáctico a una muestra de 3.000 unidades que habían sido previamente elegidas con el método presentado en esta sección, esta cantidad se reduce a un tercio (debe tenerse en cuenta que la mayoría de las unidades en inglés son automáticamente eliminadas en esta instancia). Con el objeto de estimar la precisión, un examen manual de 100 unidades seleccionadas aleatoriamente a partir de este último muestreo de unidades sintácticamente aptas revela que por lo menos 58 de ellas tendrían un estatus terminológico indudable, como 'enunciador', 'dígrafo' o 'fonema'. El resto de las unidades está conformado por vocabulario utilizado en la disciplina pero que difícilmente podría ser admitido como entradas en un diccionario terminológico del área, como 'objetividad', 'pedagógico', 'cuestionario' o 'imitación'. Determinar el estatus de algunas de estas formas es, sin embargo, sumamente difícil, incluso para un especialista. Los contextos en los que los distintos autores utilizan estas expresiones son para ello una ayuda vital. Gracias a los contextos podemos advertir que una expresión como 'imitación' no es utilizada en un sentido distinto al del lenguaje común, con lo cual podemos rechazarla como candidata a término. Para hacer una estimación de la cobertura se seleccionó una lista de 1.000 términos que están presentes tanto en el diccionario como en el corpus y se comparó esta lista con los 1.000 que habían sido seleccionados como sintácticamente aptos: la coincidencia fue del 22%, algo baja en relación a la cobertura del mismo diccionario del TermCat, estimada en un 32% en la Sección 2.2.2.

En el resto de los experimentos presentados a continuación todas las unidades son sometidas a análisis y no solo aquellas que pasaron por este último filtro sintáctico.

2.2.4. Ordenamiento de las unidades

El objetivo del trabajo descrito en esta sección es el ordenamiento de las unidades encontradas en el corpus de acuerdo con una determinada ponderación, una forma de descubrir unidades a partir del corpus que no podrían haber sido halladas por medio del examen manual del corpus o la introspección de un hablante de la lengua.

Observar la curva de distribución de frecuencias de un término puede ser interesante e informativo, sin embargo, esto no tiene el mismo valor científico que un instrumento que nos permite ir más allá de la selección de términos que haga un usuario. En otras palabras, es más interesante un algoritmo que nos permite no ya buscar un término sino descubrir, a partir del corpus, aquellas unidades cuya curva de distribución de frecuencias tiene un perfil particular. Esta diferencia entre, por un lado, comprobar (una distribución de frecuencias a partir de una unidad terminológica propuesta por un usuario) y, por otro lado, descubrir (las unidades terminológicas por medio de su curva de distribución) representa una de las diferencias más importantes entre los métodos cualitativos y cuantitativos en la investigación lingüística.

La mayoría de los coeficientes para ordenar las unidades del corpus que se presentan en esta sección se organizan en un sistema de oposiciones, de manera tal que uno representaría lo contrario del otro (por ejemplo, fugacidad frente a continuidad). Esto, sin embargo, no debe llevar a creer que en estos casos se trata del mismo coeficiente invertido, ya que no son necesariamente lo mismo que un orden inverso de los elementos de la lista.

2.2.4.1. Frecuencia relativa

Desde el punto de vista terminólogico, la frecuencia de uso de los términos no es un criterio suficiente para decidir si a una unidad debería corresponderle una entrada en un diccionario especializado.

En este sentido, las siguientes secciones aportan distintos coeficientes que pueden informar mejor esta decisión. De cualquier forma, y si bien no es un criterio suficiente, la frecuencia no deja de ser un factor importante, ya que un diccionario tiene que incluir los términos que más se utilizan.

La Tabla 1 presenta las 30 formas más frecuentes en el corpus, según frecuencia relativa ya que se debe compensar las diferencias anuales en la cantidad de texto editado, de manera tal que una palabra no parezca más frecuente solamente porque aparece mucho en un solo año o en un período de tiempo.

2.2.4.2. Fugacidad/Continuidad

Uno de los criterios más importantes para evaluar la pertinencia de una unidad terminológica es observar si el uso de un término es continuo

Tabla 1. Las treinta formas más frecuentes en todo el corpus.

nº	Unidad	Frec. Rel.
1	Aprendizaje	0,02014316
2	Lingüístico	0,01050408
3	Vocabulario	0,00814838
4	Corpus	0,0055145
5	Oral	0,00479772
6	Verbo	0,00403734
7	Léxico	0,00381777
8	Hablante	0,00365493
9	Conceptual	0,00361738
10	Lingüísticos	0,00353829
11	Aula	0,00352064
12	Verbos	0,00349989
13	Discourse	0,00329898
14	Linguistics	0,00326303
15	Oración	0,00321024
16	Adjetivo	0,00312516
17	Oraciones	0,00304161
18	*Linguistic*	0,00302079
19	Gramatical	0,00290432
20	Lingüísticas	0,0028091
21	*Grammar*	0,00276271
22	Hablantes	0,00269629
23	Lexical	0,00227442
24	Contextos	0,00226332
25	Materna	0,00223297
26	Gramaticales	0,00210058
27	Comunicativa	0,00207894
28	Interacción	0,00207112
29	*Learners*	0,00205978
30	Textual	0,00205978

o si aparece de manera fugaz en la historia de la disciplina. Es de suponer que si su uso es continuo, se trata de la terminología central, aquella que se ha consolidado en el campo y es común a la mayoría de los autores. Por eso, según las características de un diccionario (básicamente la cantidad de entradas) puede ser más interesante capturar solo aquella porción de la terminología mejor establecida. Estos términos tendrán una presencia continua a lo largo del tiempo y con variaciones interanuales menos pronunciadas. Cuando los términos son fugaces, es decir, tienen una frecuencia importante en un año y muy baja o nula en el resto de los años, se trata de términos que en cierta forma representan las modas, el signo de cada tiempo o bien el tema al que se dedica cada edición, y su presencia en el diccionario de especialidad quedaría sujeta al criterio del terminólogo y las características del proyecto terminográfico. La Tabla 2 muestra algunos de estos ejemplos con una presencia prácticamente exclusiva en alguno de los años.

En el caso de los términos que se utilizan en más de un año pero que tienen una tendencia ascendente o descendente en el uso, este artículo dedica una sección especial (2.2.4.8.) para su estudio y modelado. El que se presenta allí es un algoritmo más adecuado para la detección de neologismos y arcaísmos. El término 'arcaísmo' es utilizado en un sentido técnico en este contexto para referir al subconjunto de unidades dentro de la muestra que presenta una tendencia a la baja en el uso. Lo mismo puede decirse del uso del término 'neologismo', ya que puede referir a palabras que no son nuevas en la lengua general pero que comienzan a tener una vigencia o un sentido técnico específico en la disciplina.

No existe una única manera de calcular la continuidad de un término en el tiempo. La que se muestra en las Ecuaciones 1 y 2 está motivada básicamente por su simplicidad. Dado un vector V que representa el vocabulario de la muestra y un vector T que registra la frecuencia de cada unidad i de ese vocabulario en cada año j, la Ecuación 1 expresa que el coeficiente de continuidad (cont) de un término Vi será mayor

Tabla 2. Ejemplos de formas utilizadas en un solo año.

Año	Términos
1983	Aculturación; asimetría interlingual; delimitación tonal; desviación referencial; equivalencia interlingüística; hiper generalización; interferencia léxica; negación transferida; neurofuncional; nexo de subordinación; permutación; oraciones atributivas; oraciones intransitivas; realizaciones translémicas; tonemicidad; translemas; translémico; verbo subordinado; vernacular.
1984	Agramaticalidad; automaticidad; autosegmentación; biculturalismo; disimetría; dislocación; disociación; encabalgamiento; experimentadores; inmiscusión; materialización; oraciones coordinadas; reciprocidad inherente; significación partitiva; tematización léxica; unilingües.
1993	Adjetivos participios; apódosis; aprendizaje receptivo; clasemas aspectuales; codas compuestas; contextos narrativos; descripciones definidas; enfoques comunicativos; enunciado asertivo; enunciados contextuales; ergatividad léxica; gramaticalizado; ilocucionario; indeterminación; interindividualmente; léxicos mitigadores; metodología comunicativa; micro ordenador; patrones fonéticos; postestructuralismo; proto agente; univocidad.
2006	Alumnado de origen inmigrante; alumnado extranjero inmigrante; alumnado inmigrante; anticausativa; argumentatividad; basados en corpus; bilingüismo cíclico; corpus anotado; corpus etiquetado; corpus paralelos; deontológico; dialectología hispanoamericana; dígrafo contextual; etiquetado del corpus; interacción conversacional; literacidad; mediador lingüístico; mediadores interculturales; métrica fenomenológica; minimalista; minimizadores; no palabra; ontología terminológica; pausas comunicativas; preinterpretación; procesabilidad; pronombre resumptivo; reconocedor; reformulación explicativa; relexificación; sordera fonológica; superestrato; supraoracionales; sustantivos postverbales; terminografía; trilingüismo;

según la cantidad de veces en que la diferencia de frecuencia entre un año y el siguiente sea inferior a un parámetro arbitrario k.

(1) $$cont(V_i) = \sum_{j=1}^{|t_i|} \begin{cases} 1 & \text{if}(|t_{i,j} - t_{i,(j+1)}| < k) \\ 0 & \text{otherwise} \end{cases}$$

La Figura 3 refleja las curvas de distribución de frecuencias de dos unidades que recibieron un alto puntaje por el coeficiente cont, y son en efecto curvas de frecuencia relativamente continuas en el tiempo. La Tabla 3, por su parte, ofrece otros ejemplos de formas cuyas curvas de frecuencia muestran una forma similarmente constante. Como en el caso de los demás coeficientes, estos listados no siempre contienen unidades terminológicas. Se aprecian numerosas palabras de la lengua general muy utilizadas en el corpus y nombres propios, entre los que destaca el de Halliday, por las constantes referencias que hacen los lingüistas españoles a este autor.

2.2.4.5. Concentración /Dispersión

En la línea de la sección anterior, el valor opuesto a la concentración podría ser otra vez la continuidad. Sin embargo, podemos definir también dos comportamientos opuestos que serían la concentración de los términos en un período de años frente a una aparición esporádica o discontinua. Si bien son opuestos, los dos coeficientes están emparentados con el rango (la diferencia entre el mayor y menor valor en una muestra), ya que nos hablan de la volatilidad de un término o de su capacidad de cambiar su frecuencia de uso en el tiempo. Sin embargo, estos coeficientes no miden lo mismo, ya que una unidad puede tener un rango muy alto y además tener poca volatilidad, es decir, puede tratarse de un término que en un período se utilizaba muy poco pero una vez que se instaló en la comunidad mantuvo una presencia estable en el tiempo. La concentración y la dispersión de los términos nos ayudarán a encontrar justamente lo contrario, es decir, los términos que no han conseguido todavía estabilizarse en la disciplina.

Dados una unidad Vi y, por un lado, max(ti) que es su frecuencia relativa máxima en una partición j del corpus y, por otro lado, la variable Zi definida en la Ecuación 2 como la cantidad de veces en que en una partición del corpus la unidad tiene una

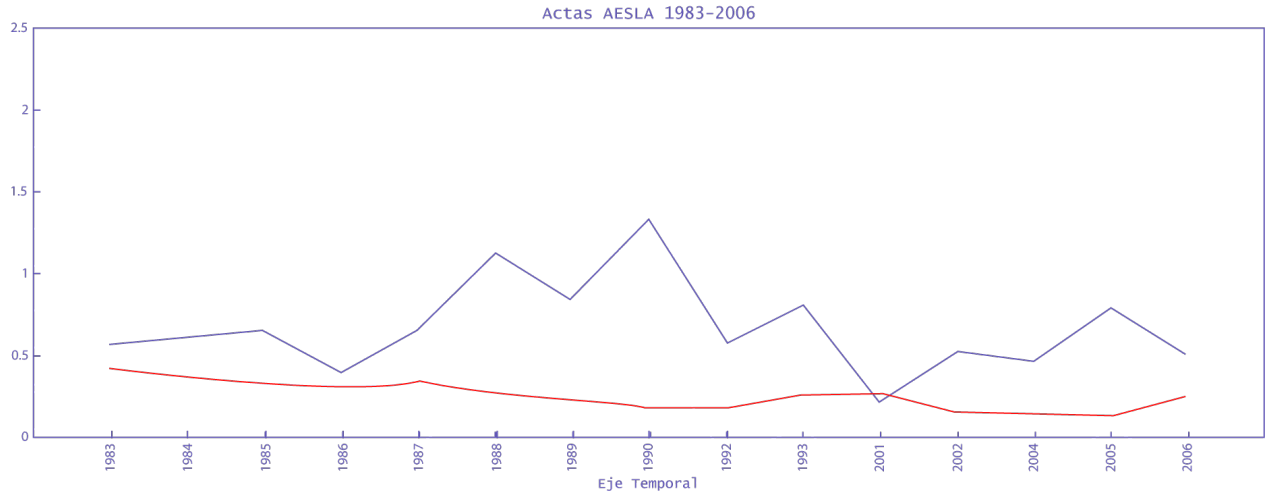

Figura 3. Distribución de frecuencias de las unidades 'aprendizaje' y 'léxico'.

Tabla 3. Las 40 formas con uso más constante en las actas.

Adjetivo,AESLA; aprendizaje; aula; comunicativa; conceptual; contrastivo; entonación; estructuración; funcional; gramatical; gramaticales; hablada; hablante; hablantes; Halliday; interacción; lexical; léxico; lingüistas; linguistic; lingüísticas; lingüístico; lingüísticos; linguistics; Longman; materna; metodología; motivación; nativos; oración; oraciones; oral; pronombres; pronunciación; secuencia; sociolingüística; variables; verbales; vocabulario

frecuencia inferior al parámetro arbitrario k, la Ecuación 3 define la concentración (conc) como la multiplicación de estos dos valores y del coeficiente de continuidad introducido en la Ecuación 1. Este último coeficiente es el que informa el grado de aglutinación en el tiempo de las apariciones de un término.

(2) $$Z_i = \sum_{j=1}^{|t_i|} \begin{cases} 1 & \text{if}(t_{i,j} < k) \\ 0 & \text{otherwise} \end{cases}$$

(3) $$\text{conc}(V_i) = max(t_i) \,.\, Z_i \,.\, \text{cont}(V_i)$$

La forma 'Drae' (Figura 4) obtiene una de las ponderaciones más altas según este coeficiente por su concentración en el año 1992, coincidente con la vigésimo primera edición del DRAE. Utilizando las variables ya introducidas, la Ecuación 4 define la dispersión de forma similar a la concentración pero dejando de lado el valor cont(Vi), correspondiente a la continuidad del término.

La Figura 5 muestra la curva correspondiente a la forma 'predicciones', que es una de las que obtienen mayor ponderación, lo cual puede ser reflejo de que este ámbito científico no se caracteriza por elaborar teorías con poder predictivo.

(4) $$\text{conc}(V_i) = max(t_i) \,.\, Z_i \,.\, \text{cont}(V_i)$$

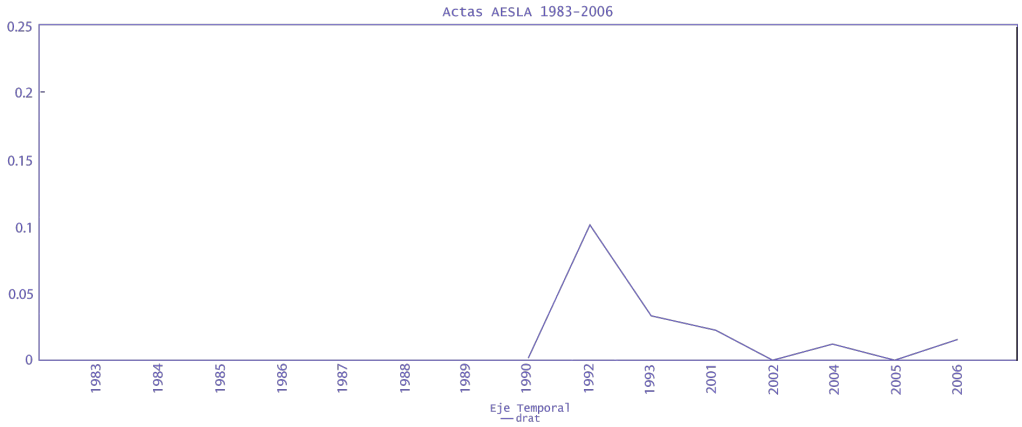

Figura 4. Distribución de frecuencias de 'Drae', una forma con alta concentración.

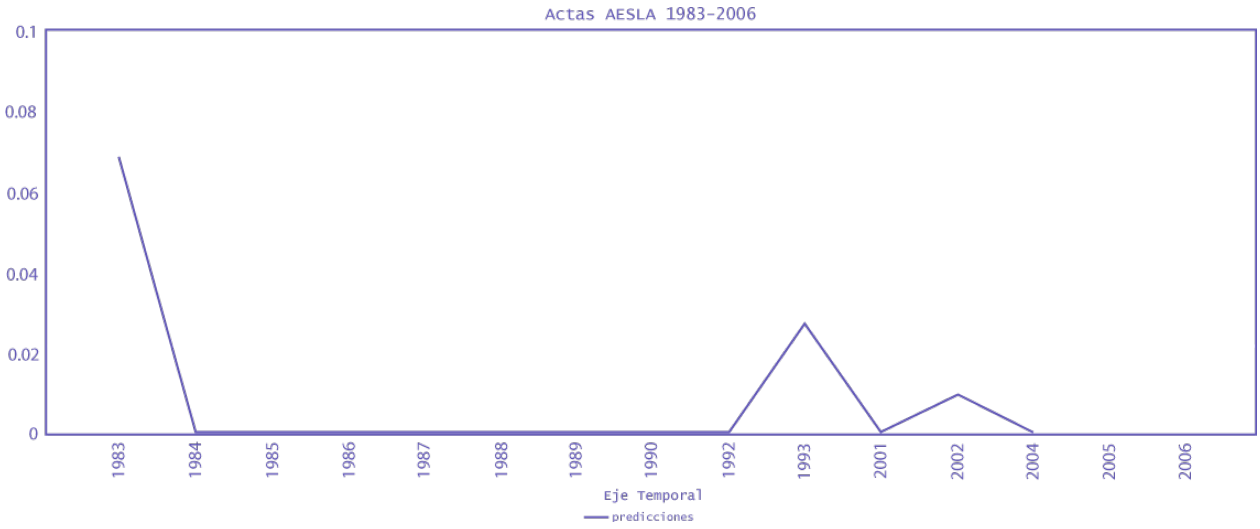

Figura 5. Distribución de frecuencias de la palabra 'predicciones' con una alta dispersión.

Tabla 4. Las 15 unidades con mayor concentración en el corpus.

nº	Unidad	Concentración
1	Corpus	0,012911
2	*Grammar*	0,006367
3	*Syllabus*	0,000549
4	Psicolingüística	0,000386
5	*Lingua*	0,000309
6	Drae	0,000302
7	Aplicadas	0,000276
8	Discursivos	0,000259
9	Electrónico	0,000201
10	Asigna	0,000138
11	Uned	0,000110
12	Wordsmith	0,000100
13	Explorar	0,000100
14	Actante	0,000000
15	Codificar	0,000000

Tabla 5. Las 15 unidades con mayor dispersión en el corpus.

nº	Unidad	Dispersión
1	*Paralanguage*	0,00018460
2	Predicciones	0,00006889
3	Posteriori	0,00004503
4	Correspondido	0,00004100
5	Racismo	0,00002894
6	*Skimming*	0,00002529
7	Especia	0,00001851
8	Copiar	0,00001655
9	Motiva	0,00001655
10	Progra	0,00001655
11	Zapatos	0,00001613
12	Documenta	0,00001613
13	Macro	0,00001335
14	*Gramma*	0,00001286
15	Valverde	0,00001000

Las Tablas 4 y 5 muestran las unidades que obtienen mayor ponderación de concentración y de dispersión, es decir, formas que tienen un comportamiento opuesto. Mientras las unidades de la Tabla 4 concentran su uso en un período de tiempo relativamente corto, en la Tabla 5 aparecen aquellas unidades cuyas apariciones, en lugar de concentrarse, se reparten de manera más heterogénea en la línea del tiempo. Debido a que estos coeficientes se aplicaron a toda la muestra y no solo las unidades seleccionadas por el filtro sintáctico de la Sección 2.2.3., se debió excluir manualmente de las Tablas 4 y 5 algunos artefactos producidos probablemente por errores de tipeo bastante frecuentes (como 'lingüistica', 'lexico' y 'termino', los tres escritos sin acento) o bien de segmentación de palabras durante el reconocimiento de óptico de caracteres (como 'cons', 'univer', 'inter' o 'apli'). Es preciso notar que es en la Tabla 4 donde se concentran las unidades más significativas desde el punto de vista terminológico, ya que, al contrario de una distribución dispersa, una distribución concentrada es más improbable que sea debida al azar. En la Tabla 5, en cambio, abundan los ejemplos de formas no terminológicas, como 'correspondido' o 'zapato'. Es decir, que en este caso el coeficiente de dispersión podría funcionar como un factor penalizador a la hora de admitir o rechazar la inclusión de una determinada unidad en un diccionario terminológico.

2.2.4.6. Tendencia ascendente/Tendencia descendente

En la historia de una disciplina científica se aprecian distintas tendencias ascendentes o descendentes en la frecuencia de uso de un término, lo cual refleja la evolución de esta ciencia y la puesta en vigor o caída en desuso de diferentes conceptos. Parte de este comportamiento ya se observó en la Sección 2.2.4.4., en la que seleccionamos términos cuyas apariciones se concentran en un solo año. Lo que queremos estudiar y modelar en esta sección es ahora la selección de los términos que tienen curvas ascendentes y descendentes. Estos perfiles de la distribución de los términos nos pueden informar acerca de la inclusión o exclusión de un determinado término en función de las características del diccionario que se proyecte. Un diccionario de mayor cobertura incluirá tanto los términos que se ponen de moda como los que dejan de usarse. Uno de nomenclatura reducida, debido a condicionamientos materiales por ejemplo, se centrará en una nomenclatura estrictamente sincrónica.

La metodología para la extracción de arcaísmos (en el sentido técnico ya explicitado) y neologismos consiste en plantear una curva modelo de la frecuencia de unos y otros y a partir de entonces calcular la similitud que tienen con ellos las curvas

de frecuencias de cada uno de los términos de la muestra. En el caso de los arcaísmos, utilizamos la Ecuación 5 para definir este arcaísmo ideal, función que es representada en la Figura 6. Para calcular la similitud que tienen cada uno de los términos con este arcaísmo ideal utilizamos la distancia euclideana. La Ecuación 7 define la distancia euclideana entre dos vectores X e Y. Para poder llevar a cabo esta comparación entre curvas previamente tenemos que normalizar los valores (Ecuación 8), es decir, llevar los términos de distinta frecuencia a la misma escala. La Figura 7 muestra la curva de frecuencias de la forma 'generativa', que es una de las que muestra mayor similitud con el arcaísmo ideal, y la Tabla 6 muestra las 20 formas con mayor similitud a este ideal. Entre las formas cada vez menos usadas destacan los términos relacionados con el generativismo o el nombre de Noam Chomsky, aunque, de nuevo, otras palabras resultan menos significativas, como 'adulta', 'delimitación', 'pasiva' y posiblemente su distribución de frecuencias en forma descendente se deba simplemente al azar.

(5) $$f(x) = x^{-1.5}$$

(6) $$d(X,Y) = \sqrt{\sum_{j=1}^{n}(X_j - Y_j)^2}$$

(7) $$t'_{i,j} = \frac{t_{i,j}}{max(t_i)}$$

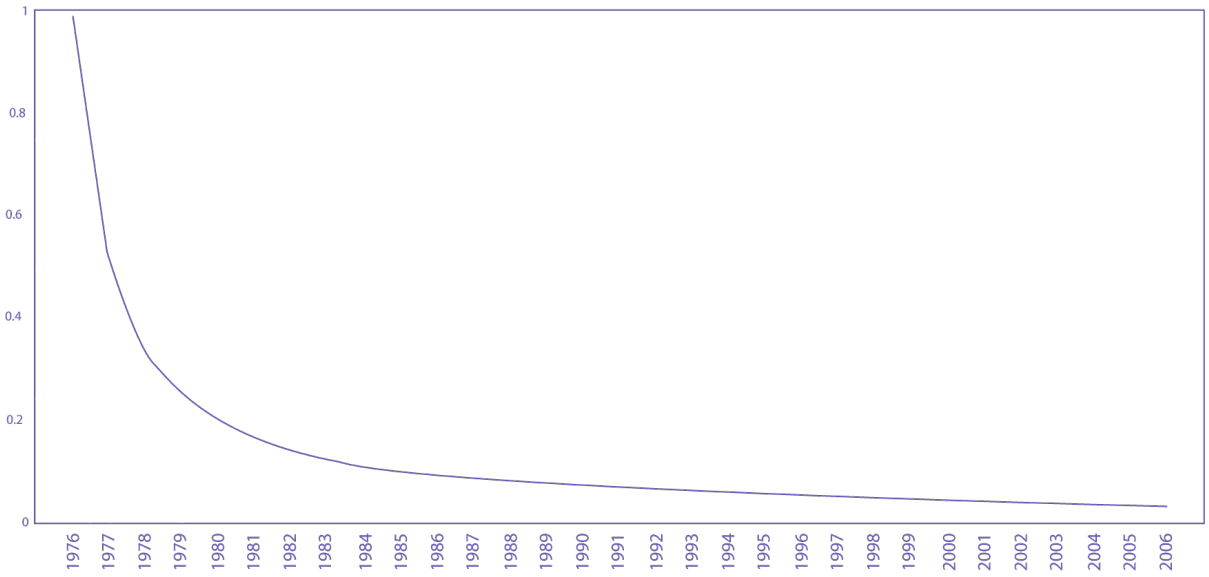

Figura 6. Distribución de frecuencias de un 'arcaísmo ideal'.

Tabla 6. Las 20 formas cuya curva de distribución de frecuencias se parece más a la del arcaísmo ideal.

Palatal; Nemser; generativa; lingüista; subordinada; intransitiva; maximalista; inserción; delimitación; lexicalización; translémicas; vocativo; Noam Chomsky; interlingual; adjetival; adulta; pasiva; adverbiales

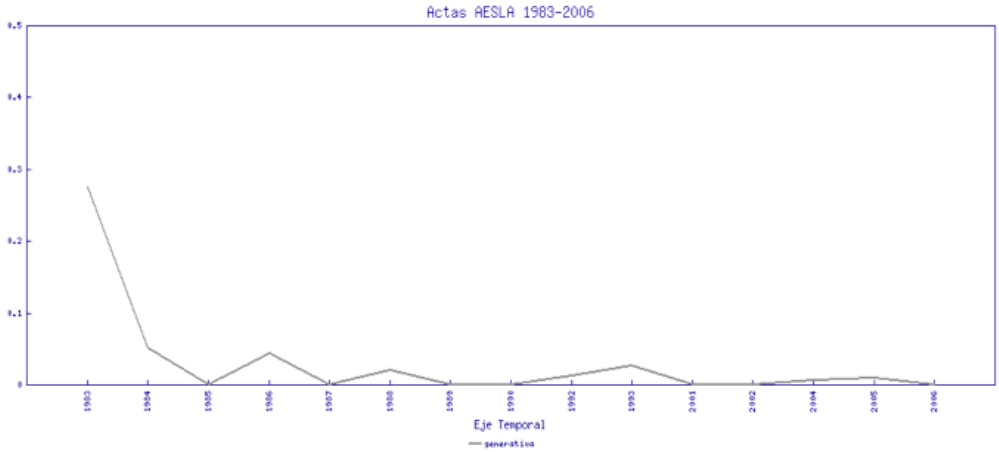

Figura 7. Distribución de frecuencias de la forma 'generativa'.

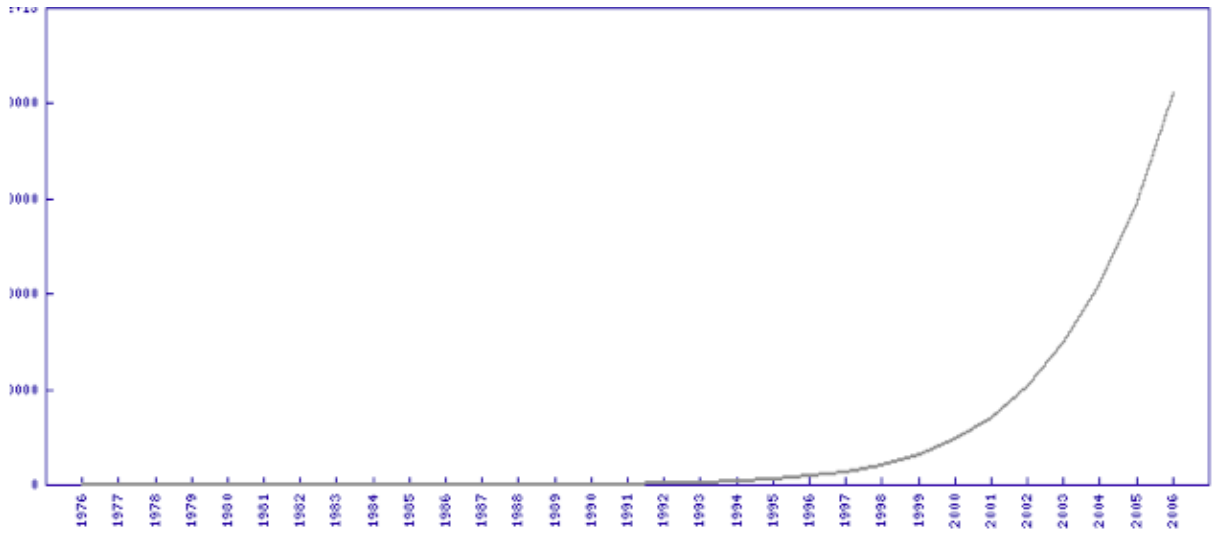

Figura 8. Representación del neologismo ideal.

En el caso de los neologismos, el procedimiento es similar al utilizado originalmente para la extracción de neología en lengua general a partir de archivos de prensa (Nazar & Vidal, 2008). Para la extensión temporal de este corpus, el neologismo ideal está definido en la Ecuación 8 y su representación en la Figura 8. El procedimiento de normalización y cálculo de similitud es el mismo que en el caso de los arcaísmos.

(8) $f(x)=x^{10}$

La Tabla 7 presenta la lista de los 20 términos cuya curva de distribución de frecuencias ofrece la mayor similitud con la del neologismo ideal. La gran mayoría de los términos con frecuencia de uso ascendente son términos en inglés, lo cual refleja la tendencia a la internacionalización que se produce en los últimos años en las comunicaciones de AESLA, y tal es así que las palabras en castellano recién comienzan a aparecer alrededor del puesto número 300 de la lista de neologismos. Entre estas palabras encontramos neologismos ya conocidos como Internet o emails, nombres de algunos autores que se han visto favorecidos con un aumento importante en la cantidad de citaciones, como el de Joaquim Llisterri, y también formas entre

las cuales encontramos términos provenientes de la jerga de teorías lingüísticas más recientes, como 'anotación', 'gramaticalización' o 'padecedor' (este último representado en la Figura 9). En algunos casos, como 'inmigración', no se trata de términos de la disciplina sino de temas o referentes que han cobrado importancia en los estudios lingüísticos de los últimos años.

conceptual entre los términos, como en el caso de 'subjuntivo-indicativo' o *figurative-metaphorically* (Tabla 8). Se producen, además, frecuentes apareamientos de las distintas formas flexivas de un mismo término. No pasa de ser un fenómeno curioso, consecuencia de no haber llevado a cabo un procedimiento de lematización de los textos, por lo cual para el sistema las distintas formas flexivas

Tabla 7. Las 20 formas cuya curva de distribución de frecuencias se parece más a la del neologismo ideal.

Conceptual metaphor; français; Llisterri; pronominal subjects; padecedor; phraseology; ong; richness; synaesthesia; perales; consejería; lexical grammar; Cascadilla; argumentative; directness; mitigation; lexical grammar model; Pragglejaz; ecuatoriana; phraseological units;

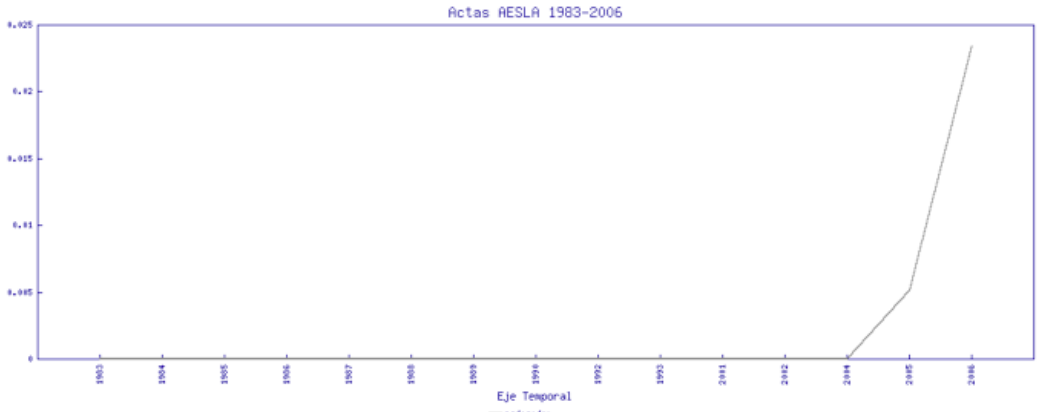

Figura 9. Distribución de frecuencias del término 'padecedor'.

2.2.4.7. Similitud

El cálculo de similitud de los términos consiste en comparar la curva de distribución de frecuencias con cada uno con la de los demás, de manera tal que se pueda elegir, para cada término del corpus, aquel término que tenga la curva de distribución de frecuencia más similar. En muchos casos se producen apareamientos de términos que tienen una distribución similar debido al azar, por lo tanto no son significativos. Sin embargo, muchos de los apareamientos son motivados por la relación

son unidades distintas. Recuerda al experimento de reconocimiento de sinónimos presentado por Grefenstette (1994) en el que apareaba pseudosinónimos, que eran palabras cuya ortografía él había alterado intencionalmente para evaluar si el sistema podía reconocerlas como sinónimos. El método que utilizó Grefenstette (1994), sin embargo, es diferente al de este artículo ya que en su caso consiste en comparar vectores de coocurrencia, es decir, que reconoce la similitud entre palabras porque estas aparecen en contextos parecidos y no por similitud en la distribución en la serie temporal.

Tabla 8. Ejemplos de parejas de formas con mayor similitud en las curvas de distribución de frecuencias.

Término	Término más similar	Coef. Similitud
Linguísticos	Linguísticas	0,9120035
Subjuntivo	Indicativo	0,8220303
Figurative	*Metaphorically*	0,7935322
Analítica	Analítico	0,6615516
Lexemas	Lexema	0,6029185
Fonemas	Fonema	0,5883786
Collocations	*Collocation*	0,5726545
Sema	Semas	0,5193972
Linguística	Linguísticos	0,5090957
Informantes	Informante	0,5029455

Conclusiones y trabajo futuro

Este artículo ha presentado un enfoque estadístico para el estudio diacrónico de la terminología especializada, y ha mostrado y evaluado una serie de coeficientes que pueden ser de utilidad a los terminólogos a la hora de generar material de partida para la nomenclatura de un glosario a partir de un corpus diacrónico. Las posibilidades que se abren a partir de este punto son muchas y variadas. Sería sumamente útil llegar a estructurar, además de un glosario, la forma en que se relacionan los términos entre sí para formar el mapa conceptual entero de la disciplina como resultado de un análisis cuantitativo. En esta línea el presente artículo ha querido proponer un análisis complementario a estudios sincrónicos (Nazar, 2010) en los que se utiliza grafos de coocurrencia que representan las relaciones entre términos como nodos que incrementan su interconexión en la medida en que estos términos coocurren en una ventana de contexto (en una misma oración, párrafo o documento). Sin embargo, estas vías de investigación ya trascenderían el tema del presente artículo, en el que se intenta promover una visión holística para superar el acuerdo tácito acerca de que la extracción de terminología es el vaciado de unidades terminológicas a partir del documento o del corpus especializado tomado como unidad. Se trata de pasar entonces de una terminología orientada hacia el documento a una terminología orientada hacia el dominio de especialidad en su conjunto.

Líneas de trabajo futuro se abren en distintas direcciones. Una posibilidad es el estudio comparativo de un campo similar utilizando datos de otras organizaciones que dispongan de actas en formato digital. En este sentido, el corpus liberado por el N-gramsViewer de Google (Michel et al., 2010) representa una posibilidad sumamente interesante. Otra posibilidad puede ser replicar el experimento en el mismo campo pero en distintos países, el mismo campo en distintas lenguas (abriendo una vía más para la extracción de terminología bilingüe) y diferentes dominios de especialidad en diferentes lenguas que dispongan de corpus, aprovechando la facilidad de reutilización de un algoritmo que no necesita conocimiento de lengua.

REFERENCIAS BIBLIOGRÁFICAS

Ananiadou, S. (1994). *A methodology for automatic term recognition.* Ponencia presentada en el 15th International Conference on Computational Linguistics, Kyoto, Japón.

Arntz, R. & Picht, H. (1989). *Introducción a la terminología.* Madrid: Fundación Germán Sánchez Ruipérez.

Barona, J. (1994). *Ciencia e historia: Debates y tendencias en la historiografía de la ciencia*. Madrid: Godella, Seminari d'Estudis sobre la Ciència.

Boulanger, J. (1988). L'evolution du concept de néologie de la linguistique aux industries de la langue. En C. de Schaetzen (Comp.), *Terminologie diachronique, actes de colloque organisé à Bruxelles les 25 et 26 mars* (pp. 193-211). Bruselas: Centre de terminologie de Bruxelles-Institut Libre Marie Haps.

Bourigault, D., Jacquemin, C. & L'Homme, M. C. (2001). *Recent advances in computational terminology*. Amsterdam: John Benjamins.

Cabré, M.T. (1999). *La terminología: Representación y comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada.

Cabré, M. T., Estopà, R. & Vivaldi., J. (2001). Automatic term detection: A review of current systems. En D. Bourigault, C. Jacquemin & M. C. L'Homme (Eds.), *Recent Advances in Computational Terminology* (pp. 1-28). Amsterdam: John Benjamins.

Cabré, M. T.; Domènech, M.; Estopà, R.; Freixa, J. & Solé, E. (2003). L'Observatoire de néologie: conception, méthodologie, résultats et nouveaux travaux. En *L'innovation lexicale* (pp.125-147). Paris: Honoré Champion

Cabré, M. T. & Estopà, R. (2005). Unidades de conocimiento especializado, caracterización y tipología. En T. Cabré & C. Bach (Eds.), *Coneixement, llenguatge i discurs especialitzat* (pp. 69-94). Barcelona: Institut Universitari de Lingüística Aplicada.

Cabré, M.T. & Estopà, R. (2009). *Les paraules noves. Criteris per detectar i mesurar els neologismes*. Vic/Barcelona: Eumo Editorial/Universitat Pompeu Fabra.

Célestin, T. & Bergeron, M. (2003). *Le phénomène de la néologie technique et scientifique au Québec- bilan et perspectives*. Colloquio Internazionale: La neologia scientifica e tecnica: Bilancio e prospettive. Accademia di Romania, Roma, Italia.

Corbeil, J. (1988). Quinze ans de politique terminologique au Québec. En C. de Schaetzen (Comp.), *Terminologie diachronique, actes de colloque organisé à Bruxelles les 25 et 26 mars* (pp. 186-192). Bruselas: Centre de terminologie de Bruxelles Institut Libre Marie Haps.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie: Statistiques Lexicales et filtres linguistiques*. Tesis doctoral, Universidad Paris 7, París, Francia.

Desmet, I. (2003). *Évolutions théoriques et méthodologiques dans la recherche en néologie scientifique et technique.* Colloquio Internazionale: La neologia scientifica e tecnica: Bilancio e prospettive. Accademia di Romania, Roma, Italia.

Dury, A. & Picton, A. (2009). Terminologie et diachronie: Vers une réconciliation théorique et méthodologique? *Revue Française de Linguistique Appliquée, 2,* 14.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Norwell, MA: Kluwer Acad.

Groult, M., Louis, P. & Roger, J. (1988). *Transfert de vocabulaire dans les sciences*. Paris: Éditions du

Centre National de la Recherche Scientifique.

Humbley, J. (2003). La néologie en terminologie. En J. F. Sablayrolles (Ed.), *L'Innovation Lexicale* (pp. 260-278). Paris: Champion.

Jacquemin, C. (1997). *Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus.* Mémoire d'Habilitation à Diriger des Recherches en Informatique Fondamentale, Université de Nantes, Francia.

Kageura, K. & Umino, B. (1996). Methods of automatic term recognition. *Terminology*, *3*(2), 259-290.

Kuhn, T. (1962). *La estructura de las revoluciones científicas.* Madrid: Fondo de Cultura Económica.

Lakatos, I. (1974). *Historia de la ciencia y sus reconstrucciones racionales.* Madrid: Tecnos.

Maynard, D. & Ananiadou, S. (2000). TRUCKS: A model for automatic multi-word term recognition. *Journal of Natural Language Processing, 8*(1), 101-125.

Merton, R. (1973). The sociology of science: Theoretical and empirical investigations. Chicago: University of Chicago Press.

Meyer, I. & Mackintosh, K. (2000). "L'Étirement" du sens terminologique: Aperçu du phénomène de la déterminologisation. En H. Béjoint & P. Thoiron (Eds.), *Le Sens en Terminologie* (pp. 198-217). Lyon: Presses Universitaires de Lyon.

Michel J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. & Aiden, E. L. (2010). Quantitative analysis of culture using millions of digitized books science. *Science, 331*(6014), 176-182.

Nazar, R. (2010). *A quantitative approach to concept analysis*. Tesis doctoral, Universidad Pompeu Fabra, Barcelona, España.

Nazar, R. & Vidal, V. (2008). *Aproximación cuantitativa a la neología*. Ponencia presentada en el I Congreso Internacional de Neología en las Lenguas Románicas, Universidad Pompeu Fabra, Barcelona, España.

Pantel, P. & Lin, D. (2001). *A statistical corpus-based term extractor*. Ponencia presentada en el 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, Londres, Inglaterra.

Patry, A. & Langlais, P. (2005). *Corpus-based terminology extraction.* Ponencia presentada en el 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, Dinamarca.

Pozzi, M.; Benítez, V.; Morett, S. (2008). *Neologismos científicos y técnicos en la prensa mexicana.* Actas del XI Simposio Iberoamericano de Terminología. Lima: RITerm.

Rondeau, G. (1984). *Introduction à la terminologie*. Québec: Gaëtan Morin.

Sager, J. (1990). *A practical course in terminology processing*. Amsterdam/Philadelphia: John Benjamins.

Sheremetyeva, S. (2009). *On extracting multiword NP terminology for MT*. Ponencia presentada en el EAMT Conference, Barcelona, España.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11-21.

Tartier, A. (2003). A method for observing terminological evolution. En G. Angelova, K. Bontcheva, R. Mitkov & N. Nikolov (Eds.), *Proceedings of "Recent Advances in Natural Language Processing"* (pp. 467-471). Bulgary: Borovets.

Temmerman, R. (2000). *Towards new ways of terminology description: The socio-cognitive approach*. Amsterdam: John Benjamins.

TermCat (1992). *Diccionari de Lingüística.* Barcelona: Fundació Barcelona.

Vivaldi, J. (2001). *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Barcelona: IULA, Sèrie Tesis 9.

Wüster, E. (1979). *Introducción a la Teoría General de la Terminología y a la Lexicografía Terminológica.* Barcelona: IULA, Sèrie Monografies 1.

NOTAS

1.-Este artículo es una versión extendida de la comunicación "Evolución de la terminología lingüística en las Actas de Congresos de AESLA entre 1983 y 2006", presentada en el XXVIII Congreso Internacional de AESLA, en la Universidad de Vigo del 15 al 17 de abril de 2010.

2.-http://www.aesla.uji.es/publicaciones

3.- http://melot.upf.edu/aesla2010/ (con acceso octubre 2010)

4.-http://www.elpais.es (con acceso octubre 2010)

* Este artículo ha sido posible gracias al financiamiento del proyecto RICOTERM3 (Ministerio de Educación y Ciencia: HUM2007-65966-C02-01/FILO. Investigadora principal: Dra. Mercè Lorente). Querría agradecer además a AESLA por facilitar los archivos de las actas de los congresos y al TermCat por facilitar la versión electrónica del diccionario utilizado.