Estudio del vocabulario en un corpus de artículos de investigación de Ecología[1]

Elvio Dúcculi - Verónica Muñoz - Silvia Beck[2]

Resumen

Estudios recientes han demostrado la importancia de desarrollar listas de palabras específicas para cada disciplina que cumplan con las necesidades de los escritores y lectores hispano-parlantes cuando deben leer o redactar un texto en inglés. Para elaborar listas de palabras en el área de ecología, este estudio investigó, en una primera etapa, el vocabulario de un corpus de artículos de investigación de ecología, y en una segunda etapa, las palabras de alta frecuencia del corpus. En la primera etapa se utilizó metodología de corpus, una lista de palabras generales y una lista de palabras académicas disponibles en la bibliografía. En la segunda etapa se analizaron y categorizaron las palabras de alta frecuencia usando criterios semánticos y pragmáticos. El análisis reveló que muchas de las palabras generales y académicas del corpus pertenecen al campo semántico de la ecología, por lo que

1 **Este trabajo se encuadra dentro del proyecto de investigación "El vocabulario de géneros académicos en ingles en distintas disciplinas. Estudio de corpus" (2012-continúa), dirigido por la Mgter. Iliana Martínez. Aprobado y subsidiado por SECYT (UNRC).**

2 **Departamento de Lenguas, FCH. UNRC. E-Mail: elvioducculi@gmail.com**

podrían considerarse palabras especializadas. Esto determinó la recategorización de las palabras en técnicas y no técnicas según su significado y uso en el contexto específico de los artículos de ecología. A partir de esta reclasificación se construyeron dos listas de palabras que aportan datos para el diseño de materiales pedagógicos para los cursos de lectura y escritura científica y académica dictados en la UNRC

Palabras clave: Vocabulario – Corpus – Artículo de investigación – Ecología

Abstract

Recent studies have shown the importance of developing specific word lists for each discipline that meet the needs of Spanish-speaking writers and readers when they need to read or write a text in English. To compile lists of words in the field of ecology, this study investigated, firstly, the vocabulary of a corpus of research articles of ecology, and secondly, the high frequency words in the corpus. In the first stage, a corpus-based methodology was used, together with a list of general words and one of academic words available in the literature. In the second stage, the high-frequency words were analyzed and categorized using semantic and pragmatic criteria. The analysis revealed that many of the general and academic words of the corpus belong to the semantic field of ecology, which could be considered specialized words. This determined the reclassification of words into technical and nontechnical according to their meaning and use in the specific context of the articles of ecology. From this reclassification, two lists of words were built which provide data for the design of teaching materials for courses of scientific reading and writing offered at the UNRC.

Key words: Vocabulary – Corpus – Research article – Ecology

1. Introducción

El inglés se ha consolidado como lengua internacional para la comunicación global de la ciencia (Crystal, 1997; Duszak, 1997; Flowerdew y Peacock, 2001; Wood, 2001). En este contexto, los docentes-investigadores, becarios de CONICET y doctorandos de la Universidad Nacional de Río Cuarto necesitan publicar en revistas internacionales para su promoción académica y profesional. Frente a la gran competencia en el contexto de la publicación académica internacional, donde participan investigadores de diferentes nacionalidades e idiomas, la escritura científica presenta dificultades para nuestros docentes, investigadores y doctorandos, quienes escriben en inglés como lengua extranjera. En respuesta a esta situación, en nuestra universidad se ofrecen cursos de Inglés con Fines Académicos a nivel de posgrado para la formación de los docentes e investigadores de las distintas disciplinas en las habilidades de escritura que éstos necesitan para publicar sus trabajos en revistas internacionales y, de este modo, participar en el “debate académico” (Hyland, 2005) dentro de la comunidad científico-académica internacional. El objetivo principal de estos cursos es familiarizar y asistir a los docentes e investigadores en el uso estratégico de los recursos lingüísticos-discursivos necesarios para la escritura científica en inglés.

La enseñanza de la escritura científica en inglés en cursos como los que aquí se mencionan se ha visto optimizada por diversos estudios lingüísticos que han contribuido resultados de gran valor para la práctica pedagógica (Bhatia, 1993; Flowerdew y Peacock, 2001; Hyland, 2006; Swales, 1990). Muchas de estas investigaciones se han concentrado principalmente en el artículo de investigación (AI), un género primordial para la comunicación de la ciencia, ya que “constituye uno de los medios por el cual los investigadores y docentes

universitarios reciben, construyen y transmiten el conocimiento e interactúan con sus pares para el avance de la ciencia" (Beke, 2005, p. 8). Para lograr dichos propósitos comunicativos, los investigadores utilizan diferentes estrategias determinadas por las convenciones y normas de la comunidad discursiva de cada disciplina (Swales, 1990). Tal como se advierte en la bibliografía, el vocabulario es uno de los aspectos lingüísticos que presenta mayores diferencias en el discurso científico de las diferentes disciplinas (Bowker y Pearson, 2002; Cabré, 1999; Ciaspucio y Kuguel, 2002; Halliday, 1993, 1998, 2004).

Dados el valor de la escritura del artículo de investigación (AI) en inglés para la participación de los investigadores en la comunidad científica internacional, la importancia del vocabulario para optimizar dicha escritura, y los aportes que los estudios de corpus han contribuido para la enseñanza de la escritura científica en inglés (Aktas y Cortés, 2008; Beke, 2005; Chen y Ge, 2007; Flowerdew, 2003; Hyland y Tse, 2007; Martínez, 2005; Martínez, Beck, y Panza, 2009; Wang, Liang, y Ge, 2008), en el presente trabajo nos propusimos estudiar un aspecto del AI, el vocabulario, en una disciplina específica, ecología. Específicamente, nos planteamos los siguientes objetivos: a) diseñar y construir un corpus de artículos de investigación del área de ecología, b) identificar los distintos tipos de palabras que componen el corpus, utilizando categorías y listas de palabras propuestas en la literatura, c) determinar frecuencia, cobertura, y rango de cada categoría de palabra, d) identificar y seleccionar las palabras de alta frecuencia y de alto rango, e) reclasificar las palabras de alta frecuencia y alto rango según su uso en un género específico, artículos de investigación, y una disciplina específica, ecología.

2. Marco Teórico

El estudio se sustenta en la Teoría del Género, la Lingüística de Corpus, en la literatura sobre el vocabulario en inglés, y en la teoría de la Terminología. La teoría del género provee la fundamentación teórica y los lineamientos para el análisis de artículos de investigación usados con fines pedagógicos (Swales, 1990, 2004). Este estudio se abordó particularmente desde el enfoque de la escuela de Inglés con Fines Específicos (Bhatia, 1993, 2001, 2004; Dudley-Evans, 1994; Swales, 1990, 2004), la cual define a los géneros como textos escritos y orales con características lingüístico-discursivas y propósitos comunicativos específicos que adquieren valor en contextos sociales y culturales determinados.

La lingüística de corpus provee el marco metodológico para el análisis lingüístico de los textos. Este enfoque establece los métodos y las herramientas para el análisis lingüístico del uso real de la lengua en textos auténticos (Biber, Conrad, y Reppen, 1998; Granger, 2002). El análisis se basa en un corpus, definido como un grupo de textos agrupados sistemáticamente en base a ciertos criterios (Biber, Conrad, y Reppen, 1998; Granger, 2002; Sinclair, 1991, 2005). Los procedimientos de análisis de un corpus combinan técnicas cuantitativas y cualitativas, lo que permite identificar patrones lingüísticos a partir de la evidencia empírica que éste provee (Granger, 2002). El análisis cuantitativo comprende el estudio de la frecuencia y distribución de elementos lingüísticos en un corpus, y se lleva a cabo mediante el uso de software para el procesamiento digital y automático de los datos lingüísticos (Scott, 2001). El análisis cualitativo se basa en la interpretación del uso de la lengua y la asociación entre ésta y su contexto de uso (Biber, Conrad, y Reppen, 1998).

La literatura referida al vocabulario provee las categorías para el estudio y la clasificación del léxico del corpus analizado en este estudio. Para el análisis cuantitativo del vocabulario del corpus se utilizaron las siguientes unidades de análisis: *types* (tipos de palabras), *tokens* (casos), y lemas, que definiremos a continuación, en la sección de metodología del presente trabajo. El enfoque de la terminología proporciona el marco teórico para el análisis cualitativo de las palabras de alta frecuencia del corpus, específicamente las categorías para la clasificación de estas palabras. En este estudio nos basamos, particularmente, en Cabré (1999) y en los criterios semánticos y pragmáticos propuestos por P. Meyer (1997).

3. Metodología

El estudio se basó en la metodología de Corpus y se llevó a cabo en tres etapas: 1) construcción del corpus, 2) análisis general del vocabulario del corpus, 3) identificación y recategorización de las palabras de alta frecuencia y alto rango de acuerdo a criterios semánticos y pragmáticos. Las etapas de análisis se basaron en a) la identificación del vocabulario del corpus (número total de tipos de palabras -'types'), su frecuencia (número de casos -'tokens') y rango (número de textos en los cuales aparece una palabra), y la relación entre el vocabulario y el tamaño del corpus, es decir entre los 'types' y los 'tokens'; b) clasificación del vocabulario en palabras generales, académicas y gramaticales; c) determinación de la cobertura de dichas categorías en el corpus, y d) identificación y recategorización de las palabras de alta frecuencia y alto rango.

3.1 Construcción del corpus y software de análisis lingüístico

Para la construcción del corpus se siguieron los criterios propuestos por Sinclair (1991, 2005) y C. Meyer (2004): representatividad, balance, tamaño, período, filiación de los autores, uso de textos completos, y disponibilidad de los textos en formato electrónico. Para el análisis del corpus se utilizó un software de análisis lingüístico, *WordSmith Tools* 4.0 (Scott, 2004). Específicamente, se utilizaron las herramientas *WordList* y *Match List*. Con la aplicación *WordList* se genera automáticamente una lista de palabras de la que se obtiene el vocabulario del corpus, calculado a partir del número total de 'types' (tipos de palabras), el tamaño del corpus, determinado a partir del número total de 'tokens' (casos), y el rango de las palabras, el cual muestra la distribución de las palabras en relación a los textos del corpus. Con la herramienta *WordList* también se calcula el índice *STTR* (*standardized type-token ratio*), es decir la relación estandarizada entre los distintos tipos de palabras y los casos. Este dato estadístico mide la frecuencia promedio de cada tipo de palabra en relación al número total de palabras, lo que permite observar el grado de variedad léxica del corpus (Barnbrook, 1996).

3.2 Listas de palabras

Para clasificar las palabras del corpus se utilizaron tres listas de palabras disponibles en la literatura: la *General Service List* -GSL (West, 1953), la *Academic Word List* -AWL (Coxhead, 2000), y una lista de palabras gramaticales. La GSL contiene las 2.000 familias de palabras más frecuentes en inglés que tienen una alta probabilidad de aparecer en una amplia variedad de usos de la lengua y en diferentes tipos de textos. Estas cubren usualmente el 80% de las palabras de un texto. La AWL contiene 570 familias y usualmente tiene una cobertura del 9% de un texto académico. Esta lista incluye aquellas palabras que son recurrentes en una amplia variedad de textos académicos de

diferentes disciplinas, excluyendo las palabras de la GSL, es decir las palabras generales. La lista de palabras gramaticales que utilizamos en este estudio contiene 308 palabras: determinantes, auxiliares, pronombres, conjunciones, conectores y números en "forma alfabética" (Biber et al, 2000, p. 279).

3.3 Recolección y análisis de datos

3.3.1 Número de *types*, *tokens*, *STTR*, y rango

Se generó automáticamente una lista de palabras del corpus utilizando la herramienta *WordList*, de la que se obtuvo información sobre el número total de *types* (tipos de palabras), número de *tokens* (casos), rango, y el índice *STTR* (*standarised type-token ratio*).

3.3.2 Clasificación y cobertura de las palabras

Utilizando las listas de palabras que presentamos anteriormente - palabras generales, GSL (West, 1953), palabras académicas, AWL (Coxhead, 2000), palabras gramaticales - y la herramienta *Match List* del software, se clasificaron las palabras del corpus en palabras generales, palabras académicas y palabras gramaticales, respectivamente. Con esta aplicación se filtraron las palabras del corpus reteniendo aquellas que aparecían en las listas y excluyendo aquellas palabras que no aparecían en las listas. Como resultado, se obtuvieron diferentes listas de palabras que corresponden a cada categoría. El resto de las palabras, que no correspondieron a ninguna

de las listas, se agrupó en una lista denominada "Otras palabras". La cobertura de cada categoría de palabra se obtuvo calculando el porcentaje de '*tokens*' (casos) que cada lista cubre en el corpus (Nation, 2001b; Nation y Kyongho, 1995).

3.3.3 Identificación y análisis de las palabras de alta frecuencia

En la segunda etapa de análisis se identificaron las palabras de alta frecuencia para su análisis cualitativo. Previo a la identificación, se excluyeron los números y las palabras gramaticales de la lista del corpus obtenida en la primera etapa de análisis. De esta forma, evitamos que la alta frecuencia de los números y de las palabras gramaticales en el corpus distorsione los resultados arrojados por la mediana, la medida de tendencia central utilizada como índice estadístico para la identificación de las palabras de alta frecuencia del corpus, que explicamos a continuación.

La identificación de las palabras de alta frecuencia se llevó a cabo utilizando los criterios de frecuencia y rango. El primer criterio se basó en la mediana, calculada a partir de la sumatoria total de las frecuencias de todas las palabras de la lista dividida por dos. Dicha medida de tendencia central permitió identificar un valor medio en la lista de palabras con el que se estableció un corte para dividir la lista en dos grupos: palabras de alta frecuencia y palabras de moderada y baja frecuencia. Luego, se seleccionaron las palabras de alta frecuencia, esto es aquellas palabras cuya frecuencia era igual o mayor a la frecuencia de la palabra que marcó la división en la lista, según lo determinado a partir del cálculo de la mediana. Luego, siguiendo el segundo criterio, de las palabras de alta frecuencia se seleccionaron aquellas que tenían alto rango. Se consideraron palabras de alto

rango las que estaban presentes en la mitad o más de la mitad de los textos del corpus, es decir en 36 o más textos.

Las palabras de alta frecuencia y alto rango se analizaron cuantitativa y cualitativamente. Para el análisis cuantitativo se determinó su cobertura en el corpus siguiendo el mismo procedimiento que se llevó a cabo para calcular la cobertura de las listas de palabras en el corpus en la primera etapa del estudio. Con el objetivo de reducir el número de palabras para su análisis cualitativo, se agruparon las palabras en lemas, definidos por Nation (2001a) como grupos de palabras que presentan variaciones en su forma gramatical, tales como las inflexiones que marcan singulares y plurales (Ej. área y áreas), pero que representan el mismo significado. Para el análisis cualitativo de las palabras de alta frecuencia, se clasificaron las palabras según su uso específico en el corpus de artículos de investigación de ecología. Utilizando criterios semánticos y pragmáticos (Cabré, 1999; Bowker y Pearson, 2002; P. Meyer, 1997; Pearson, 1998), las palabras se clasificaron en técnicas y no técnicas, siguiendo a Cabré (1999) y P. Meyer (1997), respectivamente. Específicamente, el análisis consistió en observar, por un lado, el significado de las palabras de acuerdo a su pertenencia a la terminología propia de la ecología y, por otro lado, el uso que las palabras adquieren en el contexto especializado del artículo de investigación como género de comunicación científica. Para este fin, se observó el contexto lingüístico inmediato de las palabras, utilizando la herramienta Concord del software *WordSmith Tools* (Scott, 2004).

4. Resultados y discusión

4.1 Construcción del corpus

Se construyó un corpus especializado (Hunston, 2002; Sinclair, 1991) de artículos de investigación escritos en inglés en el área de ecología. El corpus es representativo del género artículo de investigación, del registro científico, y de una disciplina específica, ecología. Tal como anticipamos en el apartado 3.1, se construyó el corpus, que describiremos a continuación, siguiendo los criterios propuestos por Sinclair (1991, 2005) y C. Meyer (2004), atendiendo principalmente a los criterios de representatividad, balance, tamaño, período, filiación de los autores, uso de textos completos y disponibilidad de los textos en formato electrónico.

El balance del corpus está dado por su estructura homogénea ya que, como se mencionó anteriormente, está constituido por textos de un mismo género, de un mismo registro y una disciplina específica. En cuanto al criterio de tamaño, se construyó un corpus de tipo pequeño, de 329.819 palabras, considerándose suficiente para los objetivos planteados en este estudio y para que, sumado a otros dos corpus que se encontraban en proceso de construcción en el marco del proyecto de investigación que contextualiza este estudio, se construya un corpus de 1 millón de palabras. En cuanto al criterio de período, se seleccionaron artículos publicados en los años 2012, 2011 y 2010 para representar la lengua de un momento determinado y así evitar el efecto del cambio lingüístico a través del tiempo. Es importante destacar que se seleccionaron sólo artículos que respondían a la estructura IMRD (Introducción, Materiales y métodos, Resultados, Discusión), típica de artículos de las ciencias experimentales. En lo que refiere a la filiación de los escritores, se seleccionaron artículos

cuyos autores declaraban pertenecer a una institución de un país de habla inglesa. En relación a los criterios de uso de textos completos y disponibilidad de los textos en formato electrónico, se seleccionaron 71 artículos de 3 revistas científicas publicadas online y disponibles en la Biblioteca electrónica del MinCyT (http://www.biblioteca.mincyt.gob.ar/). Para asegurar la calidad y el nivel de especialización de las publicaciones, se seleccionaron revistas que, al momento de recolección de los textos, tenían factor de impacto mayor a 2. La tabla 1 sintetiza la estructura del corpus.

Finalmente, los textos fueron preparados para ser procesados por el software *WordSmith Tools* (Scott, 2004). De cada texto se eliminaron tablas, gráficos, figuras, información sobre autores, referencias y agradecimientos. Cada texto se guardó en un archivo con formato ".txt". Los textos fueron codificados con números y letras para facilitar su identificación en el corpus.

Tabla 1 – Estructura del corpus

	Corpus de Ecología		
Genero	Artículos de investigación		
Registro	Científico		
Disciplina	Ecología		
Fuentes de publicación	Tres revistas de investigación (*journals*) publicadas online y disponibles en la biblioteca digital de la Secretaria de Ciencia y Técnica de la UNRC		
Revistas y Factor de Impacto	*Acuatic Toxicology* 3.33	*Ecological Engineering* 2.80	*Journal of Hazardous Materials* 3.72
Editorial	ELSEVIER		
Período (fecha de publicación)	2010 – 2011 – 2012		
Tamaño	71 textos – 329.819 palabras		
Escritores	Científicos especializados en la disciplina		
Lectores	Pares (científicos especializados en la disciplina)		
Contexto	Científico-académico		
Canal	Escrito		

4.2 Primera etapa de análisis: estudio del vocabulario del corpus

4.2.1 Número de *types* y *tokens*, *STTR*, rango

La lista generada con el software mostró que el corpus contiene 329.819 *tokens* (número total de casos) y 18.221 *types* (distintos tipos de palabras). Debido a que el corpus incluye textos de diferentes tamaños, para calcular el índice *STTR*, es decir la relación estandarizada entre los tipos de palabras (*types*) y el número total de casos (*tokens*), se programó al software para que computara la relación entre *types* y *tokens* en segmentos de 2.000 palabras, los cuales corresponden a la extensión de los textos más pequeños, tal lo sugerido por Malvern et al. (2004). Los resultados revelaron que la relación entre los tipos de palabras y el número total de casos es de 29%, lo que indica que hay un promedio de 29 nuevas palabras cada 100 palabras en cada texto del corpus. Esto indicaría una baja variabilidad de tipos de palabras en el corpus, es decir mucha repetición de palabras.

En cuanto al rango -distribución de los distintos tipos de palabras (*types*) a lo largo de los 71 textos del corpus-, se observó que del total de *types* en el corpus (18.221), sólo 32 aparecen en todos los textos. De estos *types*, 29 representan palabras gramaticales, por ejemplo '*the*' (el, los, la, las), '*and*' (y), 'of' (de), 'in' (en), y '*to*' (para/a), y solo 2 representan palabras léxicas o de contenido, '*using*' (usando/uso/usar), y '*study*' (estudio/estudiar). Se observó que hay 369 palabras que tienen un rango igual o mayor a 35, es decir que aparecen en la mitad o más de los textos del corpus. Las palabras que predominan con este rango son léxicas, dado que se observaron 266 palabras léxicas y 103 palabras gramaticales. Por el contrario, la mayoría de los *types* tienen un rango bajo, ya que del total de *types* en el corpus (18.221), 17.180 aparecen en sólo 16 textos o menos. En síntesis, las

palabras gramaticales poseen en general una distribución más amplia (alto rango), esto es, un alto porcentaje de palabras gramaticales (50%) ocurre en la mitad o más de los textos. Por el contrario, se puede observar que menos de un 2% de las palabras léxicas ocurre en la mitad o más de los textos. La Tabla 2 sintetiza los resultados en cuanto al rango de las palabras.

Tabla 2 – Rango de las palabras en el corpus

Número de textos	Palabras gramaticales	Palabras léxicas	Número total de palabras
71	29 *14,15 %*	3 *0,02 %*	32
70 -35	74 *36,10 %*	t263 *1,46 %*	337
34-17	32 *15,61 %*	640 *3,55 %*	672
16-1	70 *34,15 %*	17.110 *94,97 %*	17.180
Total	**205**	**18.016**	**18.221**

4.2.2 Clasificación de las palabras

Se identificaron 2.998 palabras (*types*) generales, es decir que pertenecen a la lista de palabras generales, GSL, construida por West (1953), 1.610 palabras académicas, es decir que pertenecen a la lista de palabras académicas, AWL, construida por Coxhead (2000), y 205 palabras gramaticales. Las palabras restantes, 13.408, se agruparon en una cuarta lista, que fue denominada "Otras palabras" (*ver Fig. 1*).

Estos resultados indican que el 73% de los distintos tipos de palabras del corpus están incluidos en esta última categoría.

Figura 1 - Clasificación de las palabras

4.2.3 Cobertura de las listas de palabras

Luego de clasificar las palabras en las cuatro listas, se procedió a calcular la cobertura de cada categoría en el corpus. Se encontró que las 2.998 palabras generales tienen una cobertura del 23% de la totalidad de palabras del corpus, las académicas del 10% y las gramaticales del 37%. Se observó que las palabras generales y las palabras gramaticales combinadas cubren el 61% del corpus. La relación entre los tipos de palabras (*types*), los casos (*tokens*) y la cobertura de cada lista se presenta en la Tabla 3.

Tabla 3 – Tipos, casos y cobertura de las listas

Listas de palabras	Tipos	Casos	Cobertura
Gramaticales	205	125.261	37,98%
GSL	2.998	77.407	23,47%
AWL	1.610	33.474	10,15%
Otras Palabras	13.408	93.677	28,40%
Total	**18.221**	**329.819**	**100,00%**

Tal como se advierte en la Tabla 3, la lista de palabras generales de West (1953) en combinación con las palabras gramaticales tiene una cobertura de un 61% del corpus, lo cual muestra una diferencia importante con el 75-80% de cobertura usualmente sugerido en la literatura (Coxhead y Nation, 2001; Nation, 2001a, 2001b). Con

respecto a las palabras gramaticales, la relación entre los tipos de palabras de esta lista y su cobertura en el corpus indica que, como se esperaba, un bajo número de palabras, 205, provee una alta cobertura, 37%. Esto se explica por el hecho de que las palabras gramaticales siempre ocurren con alta frecuencia en una amplia variedad de textos. Por el contrario, el alto número de tipos de palabras, 13.408, de la categoría “Otras palabras” provee una cobertura del 28% del corpus, menor a la de las palabras gramaticales (*ver Figura 2*). Estos resultados indican que las palabras agrupadas en la lista "Otras palabras" no tienen tantas repeticiones en el corpus como las palabras generales, académicas y gramaticales.

Figura 2 – Relación entre número de tipos de palabras y su cobertura

Una observación superficial de las palabras generales y académicas del corpus nos permitió detectar intuitivamente la existencia de palabras que estaban muy relacionadas a la disciplina (ecología) y que además tenían una alta frecuencia, como por ejemplo las palabras generales ‘*fish*’ (pez), ‘*water’* (agua), ‘*soil*’” (suelo), ‘*plant’* (planta), y ‘*root’* (raíz), todas con una frecuencia mayor a 250. En el caso de las palabras que fueron inicialmente clasificadas como académicas y que, sin embargo, reflejan conceptos relacionados al área de ecología, podemos mencionar como ejemplos las palabras “*site”* (sitio), ‘*environmental’* (ambiental), ‘*chemical’* (químico), ‘*compounds’* (compuestos), y ‘*energy’* (energía), todas con una frecuencia mayor a 100. Estas observaciones muestran que si bien las palabras que mencionamos como ejemplos fueron inicialmente clasificadas como generales y académicas, utilizando la lista de West (1953) y la lista de Coxhead (2000), respectivamente, es posible advertir que pertenecen

a un campo semántico específico que refleja la disciplina (ecología), por lo que podrían considerarse palabras especializadas. Estas observaciones confirman la crítica de otros autores sobre el uso de la lista de palabras generales - GSL (West, 1953) y la lista de palabras académicas - AWL (Coxhead, 2000) para describir el vocabulario de géneros específicos en diferentes disciplinas (Hyland y Tse, 2007; Martínez et al, 2009). Dichos autores argumentan que ambas listas no reflejan la especificidad del vocabulario en los diferentes tipos de textos de cada disciplina.

4.3 Segunda etapa de análisis: estudio de las palabras de alta frecuencia

4.3.1 Identificación y análisis cuantitativo de las palabras de alta frecuencia

El cálculo de la mediana permitió determinar que el punto de corte en la lista de palabras, que en esta etapa de análisis no incluyó los números y las palabras gramaticales, correspondía a la palabra (*type*) número 569 en la lista, con una frecuencia de 63 casos. A partir de este resultado, se seleccionaron aquellas palabras cuya frecuencia era igual o mayor a 63. Como señalamos en la sección de metodología, de las palabras de alta frecuencia se seleccionaron las palabras de alto rango, es decir aquellas con un rango mayor o igual a 36 textos. Finalmente, se obtuvo una lista de palabras de alta frecuencia y alto rango que contenía 240 tipos de palabras (*types*) y 59.535 casos (*tokens*), los cuales proporcionaban una cobertura del 30.83% del corpus.

Utilizando el lema (Nation, 2001a) como unidad de análisis, se lematizó la lista de palabras de alta frecuencia y alto rango, agrupando

las palabras que presentaban variaciones gramaticales, es decir variaciones marcadas por las flexiones, y que designaban el mismo significado (Ej. *treatment*, *treatments*). Como resultado, el número de tipos de palabras (*types*) de la lista original de alta frecuencia, 240, se redujo a 223.

4.3.2 Análisis cualitativo de las palabras de alta frecuencia

Tal como mencionamos anteriormente, advertimos que la categorización inicial de las palabras del corpus no refleja los significados y usos de las palabras en el contexto de los artículos de investigación analizados, dado que algunas palabras generales y académicas representan palabras especializadas propias del área de la ecología. Por lo tanto, decidimos recategorizar las palabras de alta frecuencia según su significado y uso en el contexto específico de los artículos de ecología. Específicamente, las palabras se clasificaron en técnicas (Cabré, 1999) y no técnicas (P. Meyer, 1997). Así pues, es importante señalar que la categorización de las palabras de alta frecuencia no se llevó a cabo utilizando listas de palabras ya establecidas en la bibliografía, como procedimos en la primera etapa de análisis, sino que se consideró el significado y el uso específico de las palabras en el corpus especializado construido y analizado para los objetivos de este estudio.

Las palabras técnicas se identificaron en base a los criterios semánticos y pragmáticos propuestos por Cabré (1999). Siguiendo criterios semánticos, las palabras técnicas se definieron como aquellos términos que hacen referencia a conceptos específicos al área de ecología, relacionados semánticamente en una red conceptual que refleja el léxico especializado del campo disciplinar. Siguiendo criterios pragmáticos, las palabras se clasificaron como técnicas

considerando que los textos analizados en este estudio, artículos de investigación, representan situaciones de comunicación técnica, ya que son publicados en revistas altamente especializadas (*journals*) y representan un área específica de conocimiento (ecología). Los criterios semánticos y pragmáticos nos permitieron incluir en la lista de palabras técnicas no sólo palabras con un significado altamente especializado y único al área de ecología sino también palabras cuyo significado no es altamente especializado pero claramente refleja conceptos específicos de la disciplina. Ejemplos de palabras técnicas en el corpus de artículos de investigación de ecología son '*cell' (*célula*)*, '*flow' (*flujo*)*, '*site' (*sitio*)*, y '*organic' (*orgánico*)*. En el Apéndice A se presenta la lista completa de palabras técnicas.

La clasificación de las palabras no técnicas se basó en los criterios semánticos y pragmáticos propuestos por P. Meyer (1997). Siguiendo criterios semánticos, las palabras no técnicas se definieron como aquellas palabras que describen relaciones entre los conceptos específicos del área de ecología, a través de la organización discursiva de los textos y la evaluación del contenido, y aquellas palabras que se refieren a aspectos relacionados con el proceso de investigación. En cuanto a los criterios pragmáticos, se identificaron como palabras no técnicas aquellas palabras que se utilizan para evaluar el contenido proposicional, organizar los textos, y guiar al lector. Ejemplos de palabras no técnicas en el corpus de artículos de ecología son '*compared' (*comparados*), 'exposure' (*exposición*), 'well' (*bien*),* y *'respectively' (*respectivamente*)*. En el Apéndice B se presenta la lista completa de palabras no técnicas.

Como se observa en el Apéndice A y el Apéndice B de este trabajo, y contrario a lo esperado debido a la especificidad de la disciplina, es de destacar que se identificó un número notablemente mayor de palabras no técnicas, 192, que de palabras técnicas, 31. Esto podría atribuirse al

tipo de corpus analizado. Por un lado, el corpus es pequeño (329.819 palabras) y, por otro lado, consta de artículos que representan una diversidad de sub-áreas de la ecología, y por lo tanto muchas de las palabras especializadas no tienen una amplia distribución a lo largo de todos los textos del corpus, es decir tienen bajo rango.

A partir de los resultados obtenidos, se construyeron dos listas de palabras de alta frecuencia: una lista de palabras técnicas y una lista de palabras no técnicas que, como anticipamos anteriormente, presentamos en el Apéndice A y en el Apéndice B, respectivamente. Dichas listas sin dudas tienen un potencial pedagógico muy interesante para responder a la especificidad requerida en los cursos de lectura y escritura científica de la Universidad Nacional de Río Cuarto.

5. Conclusiones

Los resultados obtenidos en este trabajo han aportado información descriptiva sobre el vocabulario del artículo de investigación del área de ecología en inglés, contribuyendo de este modo a los estudios de corpus sobre el léxico del artículo de investigación en diferentes disciplinas. A partir de los resultados obtenidos en relación a la clasificación de las palabras del corpus, se observó que muchas de las palabras generales y académicas más frecuentes en el corpus claramente denotan conceptos relacionados a la disciplina (ecología). Estos resultados nos llevaron a cuestionarnos la efectividad para la descripción de géneros y disciplinas específicas de dos listas de palabras pre-establecidas y disponibles en la bibliografía: la lista de palabras generales (West, 1953) y la lista de palabras académicas (Coxhead, 2000). Por lo tanto, resultó necesario proponer una reclasificación de las palabras que respondiera a la especificidad del género y de la disciplina estudiados en este trabajo, esto es el artículo

de investigación de ecología. Dicha reclasificación de las palabras del corpus aporta datos para el diseño curricular y de materiales pedagógicos para los cursos de lectura y escritura científica y académica de la UNRC. Se espera que los resultados obtenidos a partir del presente estudio permitan mejorar la práctica pedagógica en los cursos de inglés con fines académicos que se dictan en nuestra universidad.

6. Referencias

Aktas, R., y V. Cortés (2008). Shell nouns as cohesive devices in published and ESL student writing. En *Journal of English for Academic Purposes*, 7(1), 3-14.

Barnbrook, G. (1996). *Language and computers. A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Beke, R. (2005). El metadiscurso interpersonal en artículos de investigación. En *Revista Signos*, 38(57), 7-18.

Bhatia, V. (1993). *Analyzing genre: language use in professional settings*. Essex: Longman.

Bhatia, V. (2001). Analysing genre: some conceptual issues. En S. M. Hewing (Ed.), *Academic writing in context* (pp. 79-92). Birmingham: University of Birmingham Press.

Bhatia, V. (2004). *Worlds of written discourse. A genre-based view*. London: Continuum.

Biber, D., S. Conrad y R. Reppen (1998). *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad y E. Finegan (2000). *Longman grammar of spoken and written English*. Essex: Longman.

Bowker, L., y J. Pearson (2002). *Working with specialized language. A practical guide to using corpora*. London: Routledge.

Cabré, T. (1999). *Terminology. Theory, methods and applications*.

Amsterdam: John Benjamins Publishing Company.

Chen, Q. y G. Ge (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). En *English for Specific Purposes*, 26(4), 502-514.

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

Coxhead, A. (2000). A new academic word list. En *TESOL Quarterly*, 34(2), 213-238.

Coxhead, A. y P. Nation (2001). The specialized vocabulary of English for academic purposes. En J. Flowerdew, & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 252-267). Cambridge: Cambridge University Press.

Crystal, D. (2003). *English as a global Language*. Cambridge: Cambridge University Press.

Dudley-Evans, T. (1994). Genre analysis: an approach to text analysis for ESP. En M. Coulthard (Ed.), *Advances in written text analysis* (pp.219-228). London: Routledge.

Duszak, A. (1997). Cross-cultural academic communication: a discourse-community view. En A. Duszak (Ed.), *Culture and styles of Academic Discourse* (pp. 11-39). Berlín/New York: Mouton de Gruyter.

Flowerdew, J. (2003). Signalling nouns in discourse. En *English for Specific Purposes*, 22(4), 329-346.

Flowerdew, J. y M. Peacock (2001). Issues in EAP: a preliminary perspective. En J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 8-24). Cambridge: Cambridge University Press.

Granger, S. (2002). A Bird's-eye view of learner corpus research. En S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Amsterdam: John Benjamins Publishing Company.

Halliday, M.A.K. (1993). On the language of physical science. En Halliday, M.A.K. y J. R. Martin *Writing Science. Literacy and discursive power* (pp. 59-68). London: The Falmer Press.

Halliday, M.A.K. (1998). Things and relations. Regrammaticising experience as technical knowledge. En J.R.Martin y R. Veel (Eds.), *Reading science. Critical and functional perspectives on discourses of science* (pp. 185-236). London: Routledge.

Halliday, M.A.K. (2004). *The language of science*. London: Continuum.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hyland, K. (2005). *Metadiscourse: exploring interaction in writing*. London: Continuum.

Hyland, K. (2006). *English for academic purposes. An advanced resource book*. London: Routledge.

Hyland, K. y P. Tse (2007). Is there an "Academic Vocabulary"? En *TESOL Quarterly*, 41(2), 235-253.

Malvern, D., B. Richards, N. Chipere y P. Durán (2004). *Lexical diversity and language development: Quantification and assessment*. Hampshire: Palgrave Macmillan.

Martinez, I. (2005). Native and non-native writers' use of first person pronouns in the different sections of biology research articles. En *English Journal of Second Language Writing*, 14(3), 174-190.

Martínez, I., S. Beck, y C. Panza (2009). Academic vocabulary in agriculture research articles: A corpus-based study. En *English for Specific Purposes*, 28(3), 183-198.

Meyer, C. (2004). *English corpus linguistics. An introduction*. Cambridge: Cambridge University Press.

Meyer, P. G. (1997). *Coming to know: studies in the lexical semantics and pragmatics of academic English*. Tübingen: Gunter Narr Verlag.

Nation, P. (2001a). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P. (2001b). Using small corpora to investigate learner needs. En M. Ghadessy, A. Henry, & R. Roseberry (Eds.), *Small corpus studies in ELT. Theory and practice* (pp. 31-45). Amsterdam: John Benjamins Publishing Company.

Nation, P. y H. Kyongho (1995). Where would general service vocabulary stop and special purposes vocabulary begin? En

System, 1(23), 35-41.

Pearson, J. (1998). *Terms in context*. Amsterdam: John Benjamins Publishing Company.

Scott, M. (2001). Comparing corpora and identifying key words, collocations, frequency distributions through WordSmith Tools suite of computer programs. En M. Ghadessy, A. Henry, & R. Roseberry (Eds.), *Small corpus studies in ELT. Theory and practice* (pp. 47-67). Amsterdam: John Benjamins Publishing Company.

Scott, M. (2004). WordSmith Tools (Versión 4.0) [Software]. Oxford: Oxford University Press.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (2005). Corpus and text: basic Principles. En M. Wynne (Ed.), *Developing linguistic corpora: a guide to good practice* (pp. 1-16). Oxford: Oxbow Books: Consultado el 9 de Diciembre, 2009, de http://ahds.ac.uk/linguistic-corpora/.

Swales, J. M. (1990). *Genre analysis. English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (2004). *Research Genres. Exploration and Applications*. Cambridge: Cambridge University Press.

Wang, J., S. Liang, y G. Ge (2008). Establishment of a medical academic word list. En *English for Specific Purposes*, 27(4), 442-458.

West, M. (1953). *A general service list of English words*. London: Longman.

Wood, A. (2001). International scientific English: The language of research scientists around the world. En Flowerdew, J. y M. Peacock (Eds.). *Research Perspectives on English for Academic Purposes*, pp, (71-83). Cambridge: CUP.

Apéndice A

Palabras técnicas

1.	AQUATIC
2.	BIOLOGICAL
3.	C (*chemical element*)
4.	CA (*Calcium/ California*)
5.	CELL
6.	CHEMICAL
7.	DISSOLVED
8.	ENVIRONMENT
9.	ENVIRONMENTAL
10.	FIELD
11.	FLOW
12.	GROWTH
13.	LENGTH
14.	M (*species of plants*)
15.	MASS
16.	MG (magnesium)
17.	N (nitrogen, north)
18.	NATURAL
19.	ORGANIC
20.	P (phosphorus, type of plant, type of fish)
21.	PH
22.	S (Spartina)
23.	SITE
24.	SIZE
25.	SOLUTION
26.	SPECIES
27.	SURFACE
28.	TEMPERATURE
29.	VOLUME
30.	WATER
31.	WEIGHT