Káñina, Rev. Artes y Letras, Univ. Costa Rica XXXIX (1): 201-212, 2015 / ISSN: 2215-2636

# DIFERENCIAS EN EL LÉXICO ENTRE LOS PERIÓDICOS *LA NACIÓN* Y *DIARIO EXTRA* DESDE LA LINGÜÍSTICA DE CORPUS: APLICACIÓN DE UNA METODOLOGÍA

*Differences in the lexicon between newspapers La Nación and Diario Extra from Corpus Linguistics: Implementation of a methodology*

*Mariana Cortés Kandler**

**RESUMEN**

La Lingüística de Corpus es un tipo de análisis lingüístico que se basa en herramientas tecnológicas para "reunir, organizar y procesar" datos reales del lenguaje (Villayandre Llamazares, 2010). Dos ejemplos de estas herramientas son AntConc (análisis cuantitativos), que ordena las palabras de un texto por frecuencia de aparición, de manera que se pueden extraer las palabras de más alta frecuencia, y ATLAS.ti (análisis cualitativos), que establece relaciones ente los términos para ilustrar cuáles son los principales focos de interés de un texto. Otras pruebas que se pueden aplicar son el análisis de variación léxica, el porcentaje de léxico específico de un corpus en relación con otro, el porcentaje de palabras compartidas por dos corpus y, por último, la prueba de significancia estadística. El presente artículo ejemplifica la aplicación de estas herramientas para identificar las diferencias de léxico entre las secciones de El País y Deportes del periódico *La Nación*, y Nacionales y Deportes del *Diario Extra*. Del análisis se concluye que: los corpus de cada periódico difieren en la cantidad de palabras; los análisis de las redes semánticas muestran que los focos de interés son distintos en cada periódico; la prueba de comunalidad/especificidad indica que los periódicos presentan un alto porcentaje de especificidad de léxico; el análisis de variación léxica no permite establecer diferencias entre los periódicos; y la prueba de significancia estadística demuestra una diferencia significativa en la frecuencia de aparición de ciertos términos.

**Palabras clave:** Lingüística de Corpus, léxico, *La Nación*, *Diario Extra*, redes semánticas.

**ABSTRACT**

Corpus Linguistics is a type of linguistic analysis that uses technology-based tools to analyze real linguistic data. Two examples of these tools are AntConc (quantitative analyses), which orders the words in a text according to frequency of appearance, and ATLAS.ti (qualitative analyses), which allows for semantic networks to be created among the words, in order to illustrate the foci of interest in a text. Some other tests that can be applied are: analysis of lexical variation, percentage of specific lexicon between to texts, percentage of shared words, and log likelihood significance test. The purpose of this paper is to exemplify the application of these tools to identify differences in lexicon between the National News and Sports sections from the newspapers *La Nación*, and *Diario Extra*. The analysis allows for the following conclusions to be drawn: the corpora differ in number of words; the analysis of the semantic networks shows the differences in the foci of interest between the two newspapers; the communality/specificity tests indicates a high percentage of specificity in the lexicon; the analysis of lexical variation does not allow for differences to be established; and, the log likelihood significance test shows a significant difference in the frequency of appearance of certain lexical items.

**Key Words:** Corpus Linguistics, lexicon, *La Nación*, *Diario Extra*, semantic networks.

* Universidad de Costa Rica. Departamento de Filosofía, Artes y Letras, Sede de Occidente, Costa Rica. Correo electrónico: mari.ckandler@gmail.com
*Recepción: 31/03/2014. Aceptación: 07/01/2015.*

## 0. Introducción

La Lingüística de Corpus es una metodología empírica de análisis lingüístico que se caracteriza por el empleo de herramientas computacionales. Como señalan McEnery y Hardie (2012), una perspectiva de análisis basada en los corpus sirve para hacer estudios en diversas áreas de la lingüística, e incluso podría servir para proponer nuevas teorías del lenguaje basadas en los datos reales de los corpus. Este campo se relaciona con la Lingüística Computacional en tanto emplea programas computacionales para sus análisis; sin embargo, difieren en sus objetivos y aplicaciones (ver Villayandre Llamazares, 2010 para una introducción a la Lingüística Computacional). Un aspecto que tienen en común es el uso de corpus como punto de partida para la investigación. En su introducción a la Lingüística Computacional, Villayandre Llamazares (2010) señala en relación con la Lingüística de Corpus:

> Por otra parte, hay que entender e inscribir el empleo de corpus en Lingüística dentro de una perspectiva metodológica general que adopta el empirismo como forma de concebir el estudio de la lengua. En este sentido, el empleo de datos reales, de muestras de uso lingüístico, resulta el complemento ideal y la referencia ineludible en cualquier investigación que aspire a dar cuenta de algún aspecto relacionado con el lenguaje: los datos son los que apoyan o contradicen una postura teórica, los que permiten inferir reglas y generalizaciones, los que proporcionan informaciones cuantitativas, etc. Y también constituyen el material necesario como punto de partida para el desarrollo de una aplicación práctica (Villayandre Llamazares 2010, URL: http://www3.unileon.es/dp/dfh/Milka/LCII/Corpus0.pdf).

En el presente estudio, se tomaron en cuenta dos herramientas computacionales para el análisis de datos: AntConc, para los análisis cuantitativos, y ATLAS.ti, para los análisis cualitativos. Estas se emplearon para realizar distintos tipos de análisis que permitieran identificar diferencias entre dos corpus. El propósito de este trabajo en particular es aplicar estas herramientas para reconocer las diferencias de léxico entre las secciones de El País y Deportes del periódico *La Nación*, y Nacionales y Deportes del *Diario Extra*. Para este fin, se creó un corpus con 104 noticias, 56 de *La Nación* y 48 del *Diario Extra*, con lo cual se obtuvo un total de 54 602 palabras. Se utilizó el concepto de *redes semánticas* para definir los focos de interés de cada periódico. Estas redes se crearon a partir de las palabras de más alta frecuencia de los corpus de cada sección. Esta prueba, junto con otras de carácter cuantitativo, resultaron útiles para establecer la comparación entre los periódicos, pues mostraron diferencias significativas.

Se tomaron en cuenta el *Diario Extra* y *La Nación* por ser periódicos de alta difusión en el ámbito costarricense y por haber sido objeto de comparación en otras investigaciones (cf. Curvardic García y Vargas Castro, 2010). Se recolectó la primera noticia de cada sección de cada periódico durante 29 días consecutivos, del lunes 16 de abril al domingo 13 de mayo, para un total de 104 noticias (el *Diario Extra* no publica los domingos).

Siguiendo a Torruela y Llisterri (1999) para la clasificación de corpus, la tipología del corpus que se creó es la siguiente:

- Textual: se recopilaron muestras escritas de los periódicos disponibles en línea.
- Monolingüe: se trata de periódicos escritos en la variedad de español de Costa Rica.
- Equilibrado: se recogió una noticia de cada sección cada día.
- Cerrado: se incluyeron 104 noticias, 54 602 palabras en total.
- Especializado: se analiza la variedad periodística de las secciones de Nacionales y de Deportes.
- Cronológico: se recopilaron datos del 16 de abril al 13 de mayo.
- Simple: no se incluyen codificaciones (referencias bibliográficas) ni anotaciones (análisis morfológicos o sintácticos) en el corpus.

En el siguiente apartado, se define el concepto de *red semántica* que se empleó para

determinar los focos de interés de cada sección de ambos periódicos. Posteriormente, se procede a una descripción de las herramientas que se utilizaron y de los procedimientos que se llevaron a cabo para los análisis del corpus. Seguidamente, se exponen los resultados de los análisis y se describen las diferencias identificadas a partir de la comparación entre los corpus. Finalmente, se presentan diversas líneas de investigación que podrían iniciarse a partir de los resultados obtenidos.

## 1. Marco conceptual

A veces percibimos dos textos como diferentes pero no sabemos porqué. En este trabajo, se propone una comparación a nivel de léxico entre dos periódicos del ámbito costarricense, a partir de un análisis basado en herramientas computacionales propias de la Lingüística de Corpus. El propósito es identificar los focos de interés de cada sección y establecer una comparación entre los dos periódicos.

Para reconocer estos focos de interés, se trabajó con el concepto de *red semántica* como técnica de representación del conocimiento que permite establecer relaciones entre las palabras. Steyvers y Tenenbaum (2005) señalan que "las estructuras de las redes brindan representaciones intuitivas y útiles para modelar el conocimiento y la inferencia semántica" (2005: 41, traducción libre). De acuerdo con este planteamiento, los conceptos se representan como nodos, a los cuales se asocian otros conceptos de manera jerárquica, como las ramas de un árbol (Collins y Quillian, 1969). La estructura de las redes semánticas, por tanto, se crea a partir de la asociación de unas palabras con otras alrededor de un nodo (Yong, Mahmud y Woo, 2011). Esta técnica de representación del conocimiento coincide con las redes que se pueden crear utilizando el programa ATLAS.ti (ver sección 2.5), el cual permite organizar los conceptos o las ideas alrededor de códigos (nodo central de una red semántica). Este programa también ofrece la opción de asociar los conceptos de manera jerárquica; sin embargo, como en el presente estudio se trabajó con listas de palabras y no con textos, esto no fue necesario. Se procedió de la siguiente manera: una vez extraídas las palabras de más alta frecuencia de cada sección, estas se organizaron en redes semánticas utilizando el programa ATLAS.ti, alrededor de los códigos que resultaran pertinentes en cada caso. Como se mencionó al inicio de esta sección, el propósito de determinar las redes semánticas de cada sección era comparar los focos de interés entre los periódicos para identificar posibles diferencias temáticas evidenciadas a nivel del léxico.

## 2. Marco metodológico

### 2.0. La Lingüística de Corpus

Se caracteriza por la utilización de herramientas computacionales que permiten la recolección, análisis y ordenamiento de los corpus. En esta sección, se presentan las distintas herramientas, sus funcionalidades, y los análisis que posibilitan para el estudio del léxico de un texto, tal como lo presentan McEnery y Hardie (2012) en su libro en versión digital *Corpus Linguistics: Method, Theory and Practice*.

### 2.1. Analizadores de concordancias

Para el análisis cuantitativo se utilizó una herramienta de la llamada "Tercera generación" (década de 1990 en adelante). A diferencia de las herramientas de "Segunda generación", las de "Tercera generación" pueden trabajar con grandes cantidades de datos a gran velocidad y cuentan con más funciones. Procesan además distintos sistemas de escritura, con los que antes no se podía trabajar (ver McEnery y Hardie, 2012 para más detalle). Entre los principales programas se encuentra AntConc, empleado en el presente análisis.

AntConc es un *concordance* o analizador de concordancias (McEnery y Hardie, 2012 para este y los subsiguientes conceptos). Esta es una herramienta que permite hacer análisis de concordancias, de listas de frecuencias de

palabras, de colocaciones, de *clusters* (n-gramas o agrupaciones) y de palabras clave. Las *concordancias* son una lista de instancias de una cadena de caracteres (puede ser una frase, una palabra, o una parte de una palabra) que se pueden observar en sus co-textos (hacia la izquierda y hacia la derecha). Esta función se conoce también como KWIC (Key Word in Context). Las *listas de frecuencias* son listas de las instancias de un corpus ordenadas por frecuencia o alfabéticamente. Las *colocaciones* son las listas de palabras que estadísticamente tienen una mayor co-aparición con la palabra que se escogió como nodo para la búsqueda. Esta función resulta útil para saber cómo se determina algo; es decir, analizando los co-textos en que aparece un nodo se puede observar cómo se define este nodo en el texto. Para el análisis de los corpus de *La Nación* y *Diario Extra* se utilizaron las concordancias y las listas de frecuencias de palabras.

### 2.2. Estudios de frecuencias

AntConc es una herramienta estadística de cálculo de frecuencias útil para los enfoques descriptivos cuantitativos de la lengua. A partir de estas herramientas se crean modelos probabilísticos, que se basan en la medida de frecuencia de las unidades (McEnery y Hardie, 2012). Las frecuencias de apariciones son uno de los elementos centrales dentro de este enfoque porque, desde esta perspectiva, los cálculos de frecuencias pueden mostrar lo que es común en una lengua, así como los grados de "comunalidad" (*commonality*) o de "especificidad" (*specificity*) entre dos o más corpus.

Los estudios de frecuencias permiten determinar no solo cuáles son las palabras con frecuencias más altas, sino que también muestran cuáles palabras no son tan frecuentes, es decir, las que tienen frecuencia nula, mínima o media. Se pueden identificar también cuáles palabras en un corpus son de frecuencia 1; a estas formas se les llama *hápax legomena* (palabras que ocurren solamente 1 vez en un texto). El cálculo de las frecuencias medias y altas permite obtener un índice del *vocabulario básico* de una lengua. Esta medida resulta relevante para diseñar métodos para la enseñanza de primeras o segundas lenguas, así como para la creación de diccionarios (los índices de frecuencia ayudarían a determinar cuáles palabras sería más pertinente incluir en el diccionario). Por último, cabe señalar que en los estudios cuantitativos se deben tomar en cuenta no solo las frecuencias más altas de un corpus, sino todo el rango de frecuencias, e incluso los elementos que no aparecen. Este punto resulta relevante para el presente trabajo, porque también se tomaron en cuenta las palabras que aparecían solo en un corpus y no en el otro (especificidad de los corpus).

### 2.3. La estadística descriptiva

El campo de la estadística descriptiva se centra en la información cuantitativa del corpus. Gracias a las herramientas computacionales como AntConc, esta es la información más simple que se puede extraer de un corpus. En los estudios de Lingüística de Corpus, generalmente se incluyen datos de estadística descriptiva; es decir, información cuantitativa que no se trabaja con herramientas de medición de significancia estadística. Las medidas básicas son el conteo de frecuencias de aparición de las palabras. Así, se puede observar la *frecuencia absoluta* de una palabra, que es la cantidad de veces que aparece en un corpus. Sin embargo, cuando se compara la frecuencia de aparición de una palabra entre dos corpus de tamaño desigual (como sucedió, por ejemplo, entre los corpus de la sección El País, *La Nación* y Nacionales, *Diario Extra*), no se puede utilizar la frecuencia absoluta (pues es relativa la tamaño del corpus), sino que se debe usar la prueba de significancia estadística (vid infra).

Otra medida que se puede extraer del corpus es el porcentaje de vocabulario léxico y de vocabulario funcional. Como se verá posteriormente, el vocabulario funcional comprende aproximadamente la mitad de las muestras del corpus, sin embargo, se trata de unos cuantos tipos que se repiten muchas veces. Por otra parte, para el presente análisis, resultó

pertinente extraer el vocabulario léxico pues a partir de este se crearon las redes semánticas de las secciones.

Ahora bien, para establecer el criterio de alta frecuencia en el corpus se utilizó el concepto de *frecuencia normalizada*, que es una frecuencia de 1 en 1000, o de 1 en 1 000 000 para los megacorpus (McEnery y Hardie, 2012). Siempre en un corpus las palabras de más alta frecuencia son las palabras gramaticales: en este caso, *de, el, la, que,* y *en*, pues son las que más se repiten (y conforman aproximadamente la mitad del corpus). No obstante, estas no aportan información en cuanto a los focos de interés de un texto. Por esta razón, se aplicó una lista de exclusión o *stoplist* para excluir las palabras funcionales y conservar solamente las palabras de contenido, que son las que llevan esta información.

Los corpus de cada sección tenían en promedio 12 520 palabras y como se adoptó una base de normalización de 1 en 1000, se tomaron en cuenta las palabras que tuvieran una frecuencia de aparición de 13 o más. Es decir, se utilizó un criterio de frecuencia normalizada de f ≥13.

Para comparar la frecuencia de aparición de dos términos, se utilizó la prueba de significancia estadística de Paul Rayson, en: http://corpora.lancs.ac.uk/clmtp/2-stat.php. Esta prueba indica si una palabras es significativamente más frecuente en un corpus que en otro (los corpus de las distintas secciones difieren el tamaño, así que una comparación de frecuencia normalizada no sería una medida fiable).

### 2.4. Variación léxica

Para los análisis de variación léxica o riqueza léxica en los corpus se toman en cuenta los conceptos de *types* (tipos) y *tokens* (muestras). Los tipos o formas de palabras son todas las palabras distintas de un corpus, mientras que las muestras, instancias, apariciones o casos son todas las palabras de un corpus (con sus repeticiones). Se puede establecer una relación matemática entre las dos cifras para determinar la proporción o *ratio* entre los tipos y las muestras y calcular así la variación (o variabilidad) léxica de un texto. Para obtener esta proporción, se divide el número de palabras distintas (tipos) entre el número total de palabras (muestras). Como señala Sabaj (2004), "[e]l grado de variabilidad es un coeficiente que tiene un rango que va desde 0 a 1. Si el resultado del coeficiente tiende a 1, el corpus analizado es más variable. Por el contrario, si el resultado tiende a 0, decimos que se trata de un corpus poco variable." Es importante señalar que, si se toma como parámetro esta medida, los corpus tienen que ser de tamaño similar para poder compararlos. Esto se debe a que la mayor repetición se da en el vocabulario funcional, por lo que, cuanto mayor sea el corpus, más instancias de aparición tendrá de vocabulario funcional, lo cual resultaría en una variación léxica más cercana a 0.

### 2.5. Análisis cualitativo de datos

Otro enfoque para el análisis de datos es la perspectiva cualitativa. La información no se organiza numéricamente (como en AntConc), sino que se interpretan los datos para intentar explicarlos de alguna manera. Uno de los programas que facilita la organización de los datos cualitativos es ATLAS.ti, que forma parte de los programas conocidos como QDA ("*qualitative data analysis*") (ver http://onlineqda.hud.ac.uk/Intro_QDA/what_is_qda.php). Este programa trabaja con códigos (*codes*), citas (*quotations*) y memos, los cuales permiten interpretar las relaciones entre los datos por medio de la creación de redes. En el presente estudio, se trabajó con la lista de palabras de contenido de alta frecuencia para ilustrar cuáles son los principales focos de interés de cada sección. Las palabras se asocian a códigos, los cuales se convierten en el nodo central de una red semántica.

Para crear las redes semánticas, se trabajó simultáneamente con ATLAS.ti y con AntConc. En ATLAS.ti se crearon los códigos a partir

de los campos semánticos que proponía la lista de las palabras con frecuencia igual o mayor a 13. Para verificar que una palabra pertenece a una red determinada, se utilizó el analizador de concordancias de AntConc, que muestra el co-texto en que aparece el término. Por ejemplo, para ubicar la palabra "plan", se revisaron las concordancias en AntConc y se encontró que aparece en la mayoría de los casos en la aglomeración o *cluster* "plan fiscal". De esta manera, se pudo asociar al código de "Economía". Otro ejemplo es la palabra "gobierno". En el corpus de *La Nación*, luego de revisar las concordancias, se pudo identificar que pertenece al campo de la política, mientras que este mismo término en el corpus de *Diario Extra*, se emplea para hacer referencia a las repercusiones que este tiene en el pueblo: el aumento en las tarifas, la inflación, el "buen" o "mal" gobierno; es decir, se centra más en el ámbito social.

### 2.6. Representatividad

Otro tema que cabe mencionar antes de presentar el análisis de los corpus es el la representatividad. El corpus de este trabajo es de modalidad escrita, de un registro específico: la prensa. No se podría decir que es un corpus representativo de la lengua porque no contiene todos los tipos de registros y modalidades. Se trata más bien de un corpus temático: el discurso de la prensa en las secciones de Nacionales y Deportes de *Diario Extra* y de El País y Deportes de *La Nación*. El objetivo del corpus es comparar el léxico empleado en los dos periódicos, por lo que se planteó en primera instancia la recolección de igual cantidad de noticias de cada periódico para que fuera equitativo. No obstante, *Diario Extra* no publica los domingos, así que este corpus contaba con ocho noticias menos. Para mantener el criterio de representatividad, se pensó eliminar las noticias de los domingos del periódico *La Nación*. No se procedió de esta manera, sin embargo, porque la cantidad de palabras del corpus de *Diario Extra*, a pesar de tener ocho noticias menos, era mayor que la del corpus de *La Nación*. Por esta razón, y para poder efectuar los cálculos estadísticos que requieren corpus con una cantidad similar de palabras, no se eliminaron las ocho noticias de los domingos de *La Nación*. Aún así, el corpus de *Diario Extra* resultó más grande que el de *La Nación*.

## 3. Resultados

El corpus tiene un total de 54 602 palabras. Se seleccionó la opción "Treat all data as lowercase" (tratar todos los datos como minúsculas) porque no se consideraron las diferencias entre términos escritos con mayúscula y con minúscula (por ejemplo Gobierno/gobierno). Además, si no se seleccionaba esta opción, las palabras que aparecían con mayúscula (por ejemplo por estar a inicio de una oración) quedaban registradas como entradas diferentes, lo cual alteraba los datos. En este caso particular, los titulares de *Diario Extra* están escritos todos en mayúscula, lo cual creaba un desfase en los conteos.

Se codificaron los archivos de texto de cada noticia según: 1) el periódico (sigla LN para *La Nación* y DE para *Diario Extra*), 2) la sección (N para *Nacionales* y D para *Deportes*) y la fecha de la noticia. Todas las noticias se guardaron en formato .txt (formato de texto plano) para que se pudieran analizar en las herramientas AntConc y ATLAS.ti. Se contó con dos corpus principales, el de *La Nación* y el de *Diario Extra* (disponibles a petición), compuestos por las noticas correspondientes a cada periódico. Se aplicó una lista de exclusión (205 tipos) con el vocabulario funcional (disponible a petición) porque lo que interesaba extraer era el vocabulario léxico e identificar las palabras de contenido de más alta frecuencia. La variación léxica del corpus es la siguiente (Tabla 1):

TABLA 1

Variación léxica del corpus

| | Tipos | % | Muestras | % |
|---|---|---|---|---|
| Vocabulario léxico | 8034 | 97,51% | 27 693 | 50,71% |
| Vocabulario funcional | 205 | 2,49% | 26 909 | 49,28% |
| Total | 8239 | 100% | 54 602 | 100% |

*Hápax legomena*: 4255 palabras

### 3.1. Análisis de los periódicos

El léxico de los periódicos *La Nación* y *Diario Extra* se analizó tanto cuantitativa como cualitativamente. Para el análisis cuantitativo, se tomó en cuenta la cantidad de noticias, las muestras, los tipos, la proporción entre muestras y tipos, el *hápax legomena* y las palabras de alta frecuencia, determinadas como las palabras que cuentan con una frecuencia mayor o igual a 13, luego de haber aplicado la lista de exclusión. Se obtuvieron los siguientes resultados (Tabla 2):

TABLA 2

Datos cuantitativos de las secciones analizadas de *La Nación* y *Diario Extra*

| | **El País, *La Nación*** | **Deportes, *La Nación*** | **Nacionales, *Diario Extra*** | **Deportes, *Diario Extra*** |
|---|---|---|---|---|
| Cantidad de noticias | 28 | 28 | 24 | 24 |
| Muestras | 16 382 (30,0%) | 8942 (16,4%) | 10 598 (19,4%) | 18 680 (34,2%) |
| Tipos | 3553 | 2428 | 2763 | 3670 |
| Proporción tipos/muestras | 0,22 | 0,27 | 0,26 | 0,20 |
| Hápax legomena | 2021 | 1482 | 1608 | 2008 |
| Frecuencia ≥ 13 (con lista de exclusión) | 81 palabras | 29 palabras | 28 palabras | 100 palabras |

En cuanto al análisis cualitativo, se extrajeron códigos a partir de las palabras de más alta frecuencia. Ya que estas podían no coincidir de un periódico a otro, los códigos no necesariamente coincidían, a pesar de tratarse de la misma sección. Se utilizaron los siguientes códigos para crear las redes semánticas:

El País, *La Nación*: "Economía", "Partidos políticos", "Poder Ejecutivo", "Poder Judicial", "Poder Legislativo", "Política" (incluye códigos de "partidos políticos", "Poder Ejecutivo", "Poder Judicial" y "Poder Legislativo"), "Relaciones exteriores" y "Social".

Nacionales, *Diario Extra*: "Economía" y "Social". El resto de los códigos relacionados con "Política" aplicados al corpus de El País, *La Nación*, no resultaron pertinentes.

Deportes, *La Nación*: "Fútbol nacional", "Deportes en Europa" y "Juegos Olímpicos". Estos se pudieron agrupar en una sola red por la conexión entre los nodos que fueron utilizados en relación con los distintos códigos.

Deportes, *Diario Extra*: "D.T. (director técnico), presidentes de los equipos, árbitros, etc.", "Equipos de fútbol", "Fútbol nacional", "Jugadores de fútbol", y "Partidos de fútbol". Se necesitaron muchos más códigos para tratar de explicar el campo de Deportes, *Diario Extra* que el de Deportes, *La Nación*.

Para ejemplificar el tratamiento cualitativo de los datos y las redes semánticas que se pueden crear, a continuación se presentan la red de "Economía" de El País, *La Nación* (Figura 1).

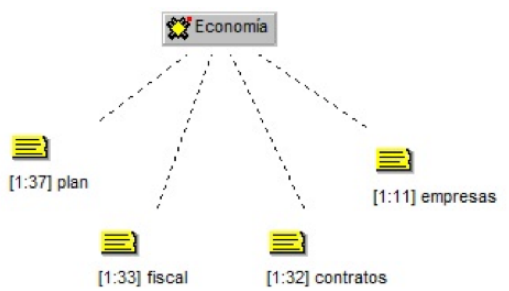


FIGURA 1.

Red de "Economía" de El País, *La Nación*

No todas las redes semánticas son así de sencillas. En el caso de Deportes, *Diario Extra*, por ejemplo, los códigos se encontraban interconectados, por lo que se pudo extraer una red global de esta sección, la cual estaba compuesta por códigos que a su vez se ramifican en nodos; estos nodos, por su parte, pueden incluso estar relacionados con más de un código. La complejidad de las redes se eleva cuanto más se profundice en un solo tema, pues se agregan nodos que de una u otra manera van a estar interconectados. En el caso de El País, *La Nación*, algunos códigos estaban interrelacionados, pero no todos. De este modo, se obtuvieron algunas redes más complejas, por ejemplo la que conectaba al gobierno con los diferentes poderes, pero también se pudieron extraer otras redes independientes que no estaban directamente relacionadas con un único tema central.

### 3.3. Comparación entre los dos periódicos

#### *3.3.1. Consideraciones generales*

En primer lugar, llama la atención que el corpus de *Diario Extra* es más grande que el de la Nación, aun cuando tenía ocho noticias menos. Sin embargo, la proporción de cantidad de palabras está invertida: en el periódico *La Nación*, la sección de El País tiene una mayor cantidad de palabras que la sección de Deportes, mientras que en el *Diario Extra*, es la sección de Deportes la que tiene más palabras que la sección de Nacionales.

La variación léxica, indicada por la proporción tipos/muestras en la Tabla 2, indicaría que las secciones de Deportes, *La Nación* y Nacionales, *Diario Extra* tienen una variación léxica más elevada que las otras secciones; sin embargo, puesto que los corpus tienen un tamaño diferente, esta comparación no sería válida para determinar si un periódico tiene mayor o menor variación que el otro. Las secciones que presentan una variación ligeramente más elevada son las secciones con menos palabras; esto se explica porque cuanto más extenso el texto, mayor cantidad de muestras de vocabulario funcional va a tener (está compuesto por pocos tipos

que se repiten con alta frecuencia y constituye aproximadamente el 50% de las muestras totales). Aun así, cabe señalar que la proporción tipos/muestras es similar entre las secciones que tienen una cantidad parecida de palabras. Se podría sugerir, por tanto, que no hay gran diferencia de variación léxica entre los dos periódicos.

### *3.3.2. "Comunalidad" (commonality) y "especificidad" (specificity)*

Para extraer las palabras propias de *La Nación* y de *Diario Extra* se copiaron todas las palabras de cada periódico y se aplicaron como lista de exclusión al otro; de esta manera, quedan solo las palabras que no se encuentran en el otro periódico. Es decir, se aplicó el corpus de *La Nación* como lista de exclusión al corpus de *Diario Extra*, con lo cual quedaban las palabras que aparecen exclusivamente en el corpus de *Diario Extra*. Con la cantidad de palabras propias de *Diario Extra*, se calcula el porcentaje de estas palabras sobre el total de palabras del corpus del periódico para calcular el porcentaje de vocabulario específico. El mismo procedimiento se realizó aplicando el corpus de *Diario Extra* al de *La Nación* para calcular el porcentaje de vocabulario específico de este periódico.

Para determinar el porcentaje de léxico común (prueba de "comunalidad"), se aplicaron como lista de exclusión las listas de palabras propias de cada periódico al corpus entero. El porcentaje se calcula a partir de la cantidad de palabras restantes sobre el total de palabras del corpus (Tabla 3).

TABLA 3.

Prueba de "comunalidad" / "especificidad"

| | Tipos | Muestras |
|---|---|---|
| Total de palabras del corpus | 8239 | 54 602 |
| Total de palabras con lista de exclusión | 2311 | 44 953 |
| Porcentaje de "comunalidad" | 28,0% | 82,3% |
| Porcentaje de "especificidad" | 72,0% | 17,7% |

Este análisis muestra una diferencia significativa entre los dos periódicos: presentan un alto grado de especificidad en cuanto a los tipos. Como se mencionó anteriormente, las muestras no son pertinentes porque en un corpus la mayor repetición se da en el vocabulario funcional. Sin embargo, es notorio que comparten solamente el 28% del vocabulario de contenido. Los análisis de las redes semánticas demostraron que los focos de interés de cada sección difieren de un periódico a otro. La prueba de comunalidad/especificidad vendría a reforzar este resultado, pues revela que los periódicos usan palabras diferentes, presumiblemente para tratar temas distintos.

Las listas del total de palabras de cada periódico se aplicaron a cada sección del otro periódico para determinar cuáles eran las palabras específicas de esa sección de ese periódico. En otras palabras, para identificar las palabras específicas de El País y de Deportes, *La Nación*, se aplicó, en cada caso, la lista de la totalidad de palabras del corpus del *Diario Extra* como lista de exclusión. De igual manera, para las secciones

de Nacionales y de Deportes, *Diario Extra*, se aplicó como lista de exclusión la lista de la totalidad de palabras de *La Nación*.

Las siguientes palabras son específicas (siguiendo el criterio de frecuencia normalizada ≥ 13) de la sección de El País, *La Nación*: pln, conavi, notas, relacionadas[1], tregua, trocha, mopt, legislativo, directorio, presidencia, iv, tse. Todas corresponden al código de "política", el cual que está ausente en el corpus de Nacionales, *Diario Extra*. Incluso términos en común, como "gobierno", se refieren a asuntos sociales cuando se trata en el *Diario Extra*, y a asuntos políticos cuando se trata en *La Nación*.

En cuanto a la sección de Deportes, *La Nación*, las palabras específicas: londres, futbol, notas, relacionadas[2]. El periódico *La Nación* contempla más noticias sobre los deportes en el exterior, como lo demuestra la referencia a Londres, que no aparece en el *Diario Extra*. En cuanto a la palabra "futbol", en el corpus del *Diario Extra* aparece con tilde: "fútbol".

En la sección de Nacionales, *Diario Extra*, se encuentran las siguientes palabras específicas: mil, nuevas, placas. Aparentemente, cuando hablan de cantidades, escriben la palabra "mil" en vez de poner los ceros, como se hace en *La Nación*. Además, es de importancia el tema de las placas nuevas, que atañe a todos los que tengan carro. De nuevo, se puede observar que el foco es en el área social.

La sección de Deportes, *Diario Extra*, tiene las siguientes palabras específicas: lagos y pci. "Lagos", es Cristian Lagos, delantero del Santos de Guápiles, y "pci" es Provident Capital Indemnity, nombre de la compañía de Minor Vargas Calvo, presidente del Brujas F.C., que fue arrestado en Estados Unidos por fraude. Como mencionan Cuvardic García y Vargas Castro (2010), el *Diario Extra* es un periódico sensacionalista, por lo que no es de extrañar que una noticia como el arresto del presidente de un equipo de primera división aparezca repetidas veces en el corpus.

### *3.3.3 Pruebas de significancia estadística*

En este apartado se comparó la frecuencia de aparición de distintos términos en los corpus del *Diario Extra* (Corpus 1) *y* de *La Nación* (Corpus 2) para determinar si la diferencia era significativa. Se utilizó la prueba de Paul Rayson que se encuentra en el sitio: http://corpora.lancs.ac.uk/clmtp/2-stat.php (McEnery & Hardie, 2012, para más detalle). Como se señaló anteriormente, esta prueba permite determinar si la diferencia en la frecuencia de aparición de una palabra entre dos corpus es significativa aun cuando los corpus son de distinto tamaño. Esta prueba resulta necesaria porque no se puede comparar la cantidad de veces que aparece una palabra (frecuencia absoluta) entre corpus de distinto tamaño, pues como es lógico, cuantas más palabras tenga un corpus, más probabilidades tendría una palabra de aparecer. Por tanto, esta fórmula compara la frecuencia de aparición en relación con el corpus mismo, lo cual sí es un resultado válido. Se presenta a continuación el resultado de la palabra "saprissa" (Figura 2), para ejemplificar la aplicación de la prueba

**Prueba de significancia (LL>3.84):**

| | Corpus 1 | Corpus 2 |
|---|---|---|
| Frequency of X (e.g. freq of word) | 133 | 21 |
| Total opportunities for X (e.g. Corpus size) | 29278 | 25325 |

Calculate LL | Clear form

| Item | O1 | %1 | O2 | %2 | LL |
|---|---|---|---|---|---|
| Word | 133 | 0.45 | 21 | 0.08 + | 75.38 |

FIGURA 2. Muestra: *saprissa*

Como se puede observar, hay diferencias significativas en las frecuencias de uso de esta palabra, pues se hace alusión a "saprissa" significativamente más en *Diario Extra* que en *La Nación*. Un análisis más extenso aplicando la prueba de significancia estadística podría arrojar luz sobre qué temas son tratados más extensamente en cada periódico y si esa diferencia es significativa.

## Conclusiones

Los análisis que resultaron de la aplicación de las distintas herramientas computacionales de la Lingüística de Corpus permitieron identificar varias diferencias entre los corpus de *La Nación* y del *Diario Extra* en distintos ámbitos. Primero, los corpus correspondientes a cada periódico diferían en la cantidad de palabras, y se encontró que en cada caso una sección tenía más palabras que la otra, pero en relación inversa: en *La Nación*, el corpus de El País era más grande que el de Deportes, mientras que en *Diario Extra*, el corpus de Deportes tenía una cantidad mucho mayor de palabras que el de Nacionales. Segundo, las redes semánticas de cada sección no coincidían de un periódico a otro. Para organizar los conceptos de El País, *La Nación*, se necesitaron más códigos que para analizar la sección de Nacionales, *Diario Extra*. Por el contrario, al clasificar las palabras de Deportes, *Diario Extra*, se tuvieron que crear más códigos que los que fueron necesarios para ordenar las palabras de la misma sección del periódico *La Nación*. En la misma línea, se encontró una gran especificidad del léxico, lo que muestra que además de concentrar el vocabulario alrededor de nodos distintos, las palabras que eligen son en general distintas. En cuanto a la variedad léxica, esta no demostró ser un criterio distintivo para comparar los periódicos. Por último, la frecuencia de uso de ciertas palabras difiere de un periódico a otro, lo cual refuerza la propuesta de discrepancia entre las redes semánticas de cada sección.

Ahora bien, cabría preguntarse porqué se necesitaron más o menos códigos para definir las secciones. En este punto, resulta pertinente indicar que las secciones más grandes de ambos periódicos (El País, *La Nación* y Deportes, *Diario Extra*), fueron las que requirieron la mayor cantidad de códigos. Esto no es de extrañar, pues si un corpus tiene más palabras, también sería de esperar que sean más los temas que se traten. Las secciones de menor tamaño, siguiendo este razonamiento, tendrían menos focos de interés sencillamente por contar con un número menor de palabras. Sin embargo, no todos los códigos coincidían, y precisamente con la prueba de comunalidad/especificidad se pudieron identificar cuáles eran los temas que se trataban en un periódico que no se mencionaban en le otro. Futuros estudios podrían investigar los campos semánticos de otras secciones para evaluar si también se da esta diferencia entre los periódicos.

Por otra parte, el hecho mismo de que el corpus de *Diaro Extra* fuera de mayor tamaño resulta interesante. Sus lectores ¿leen más?, ¿tienen más tiempo para leer?, ¿tienen un período de atención más largo?, o sus periodistas ¿usan más palabras?, o, más bien, ¿dan más detalle y elaboran más las ideas? Para responder a estas preguntas, probablemente habría que evaluar los datos desde otras disciplinas, lo cual dará lugar, espero, a nuevas investigaciones.

## Notas

1. Tanto en las secciones de El País como de Deportes, *La Nación*, aparecen las palabras "notas" y "relacionadas" como palabras específicas, porque en las noticias de este periódico siempre se incluyen notas relacionadas con la noticia en cuestión; esto no se presenta en *Diario Extra*.

2. Ver nota anterior.

## Bibiliografía

Collins, A. M. & M. R. Quillian. 1969. Retrieval Time from Semantic Memory. *Journal of Verbal Learning and Verbal Behavior.*

8,(2): 240-247. [WWW] URL: http://cds.unibas.ch/~hills/cogsci1/readings/Chapter25_Collins.pdf

Cuvardic García, D. & E. Vargas Castro. 2010. Recursos lingüísticos en la titulación periodística costarricense: el caso de *La Nación* y el diario *La Extra*. *Filología y Lingüística*. 36(1): 207-232. [WWW]. URL http://www.latindex.ucr.ac.cr/filologia-36-1/filologia-36-1-10.pdf

McEnery, T & A. Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press. Suplemento web: http://corpora.lancs.ac.uk/clmtp/

Quesada Villalobos, P. 2006. Léxico del habla costarricense: estudio del capo semántico de la vivienda. *Revista Pensamiento Actual, Universidad de Costa Rica*. 6(7): 125-133. [WWW]. URL http://www.latindex.ucr.ac.cr/pnsac003/012_capitulo.pdf

Sabaj, O. 2004. Especificidad, especialización y variabilidad verbal: Una aproximación computacional en estadística lexica. *Revista Signos*. 37(56): 75-89. [WWW] URL: http://www.scielo.cl/scielo.php?pid=S0718-09342004005600006&script=sci_arttext#joha

Steyvers, M, y J. B. Tenenbaum. 2005. The large-scale structure of semantic networks: statistical analysis and a model of semantic growth. *Cognitive Science*. 29(1): 41-78. [WWW] URL: http://web.mit.edu/cocosci/Papers/03nSteyvers.pdf

Torruela, J. & J. Llisterri. 1999. Diseño de corpus textuales y orales. En Blecua, J.M. *et al*. (eds.). *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Seminario de Filología e Informática, Universidad de Barcelona, Editorial Milenio: 45-77. [WWW]. URL: http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf

Villayandre Llamazares, M. 2010. *Lingüística Computacional II. Lingüística de corpus*. Universidad de León. [WWW]. URL: http://www3.unileon.es/dp/dfh/Milka/LCII/LC1.htm

Yong, K. K., R. R. Mahmud, & C. S. Woo. 2011. Lexical Database for Multiple Languages: Multilingual Word Semantic Network. *World Academy of Science, Engineering & Technology*. 80: 229-234.