II Congreso Internacional en Inteligencia Ambiental, Ingeniería de Software y Salud Electrónica y Móvil
AmITIC 2018
12 al 14 de septiembre de 2018
David, Chiriquí, República de Panamá

Categorías léxicas en medios digitales de Honduras de 2009 - 2016

Part of Speech in digital media from Honduras during 2009 - 2016

Jairo Jonathán Martínez [1*], Eva Leticia Martínez [2]
[1, 2] *Universidad Nacional Autónoma de Honduras, UNAH-TEC Danlí, Honduras*
jairo.martinez@unah.edu.hn

RESUMEN– *Un recurso valioso para las empresas y personas es la información. Aunque se pueden encontrar muchos datos estructurados, gran parte del conocimiento se encuentra en formatos no estructurados, en forma de lenguaje natural. En los últimos años las tecnologías han favorecido un crecimiento constante de la producción de volúmenes de texto que están disponibles, pero que son difíciles de procesar. Estos constituyen una gran fuente de información importante para las empresas, la política y las personas que quiere aplicar tecnicas de minería de texto para encontrar información que les sea de utilidad. Sin embargo, el procesamiento del lenguaje natural es un campo de investigación en pleno desarrollo, y una tarea pendiente para los científicos lingüístico-computacionales. En Honduras también ha crecido la producción de texto digital. Como parte del procesamiento computacional de texto se realiza el etiquetamiento de la categoría léxica a la que pertenece cada palabra. Para este artículo se realizó el etiquetamiento de una colección compuesta por más de 173 mil noticias publicades entre los años 2009 y 2016 en periódicos digitales del país. Además, se realiza un análisis de la frecuencia de las palabras y de las categorías léxicas en las que fueron clasificadas.*

Palabras claves– *categorías léxicas, Honduras, lingüística computacional, periódicos digitales.*

ABSTRACT– *A valuable resource for companies and people is information. Although you can find many structured data, much of the knowledge is in unstructured formats, in the form of natural language. In recent years, technologies have favored a steady growth in the production of text volumes that are available, but are difficult to process. These are a great source of important information for companies, politics and people who want to apply text mining techniques to discover useful information. However, the processing of natural language is a field of research in full development, and a pending task for linguistic-computational scientists. In Honduras, the production of digital text has also grown. As part of the computational processing of text, the labeling of the lexical category to which each word belongs is performed. For this article, a collection of more than 173 thousand news published in honduran digital newspaper between 2009 and 2016 was part of speech tagged. In addition, an analysis is made of the frequency of the words and the lexical categories in which they were classified.*

Keywords– *Computational linguistic, Honduras, news, media, parts of speech.*

1. Introducción

El lenguaje natural es el producido por humanos con propósitos de comunicación oral o escrita. El análisis del lenguaje natural es un reto para la computación lingüística. La forma en la que los humanos procesamos el lenguaje dificulta el análisis computacional [1]. Por ejemplo, una misma palabra puede tener diferentes significados dependiendo del contexto. También a menudo, no se respetan las reglas básicas de redacción y ortografía, es decir, tendemos a cometer errores léxicos y gramaticales, sin embargo, no imposibilitan la comunicación. Además, usamos algunas figuras idiomáticas, expresiones o incluso la ironía para transmitir un mensaje diferente al aparente.

Los últimos años han marcado una creciente producción de información en forma de texto. La popularización de las páginas web, los blogs, las wikis, los foros de discusión y las redes sociales han provocado un aumento significativo de los datos disponibles [2]. Sin embargo, estos datos no son estructurados como los que vienen de sensores automáticos o bases de datos estructuradas, sino que son presentados en lenguaje natural.

En Honduras, también existe una gran producción de texto escrito. Sin embargo, poco se utiliza esta

información para producir conocimiento útil a empresas, estado y sociedad. Por ejemplo, las primeras versiones digitales de diarios hondureños aparecen en el año 2009 y han ido creciendo en los últimos años. La Tribuna produjo cerca de 113 mil noticias entre el 2009 y agosto de 2016, con un promedio de 132 noticias diarias publicadas [3], pero el estudio del lenguaje natural desde una perspectiva computacional no parece ser una prioridad.

Una parte de este análisis del análisis lingüístico comprende el marcado e identificación de la categoría a la que pertenece una palabra, por ejemplo si es un sustantivo, pronombre o verbo, entre otros. En este artículo de investigación se utiliza el corpus de noticias generado en [3] para estudiar las categorías léxicas utilizadas en el español de las noticias de diarios digitales hondureños, durante el periodo de 2009 a 2016. Generando la pregunta de investigación: **¿Cuáles son las categorías léxicas utilizadas en periódicos digitales de Honduras durante los años 2009 a 2016?.** Se analizan las categorías léxicas en periódicos digitales de Honduras durante los años 2009 a 2016, identificando las palabras utilizadas y las categorías léxicas a la que pertenecen las palabras de la colección en estudio.

El artículo se estructuró presentando primero los conceptos relativos al procesamiento de lenguaje natural. Luego se describen las categorías léxicas que se analizaron en el proceso de investigación. Seguido, se especifican los elementos metodológicos para finalmente presentar los resultados y las conclusiones.

2. Procesamiento de lenguaje natural

El procesamiento de lenguaje natural (NLP, por sus siglas en inglés) es un área de estudio que busca la forma en que las computadoras pueden ser usadas para procesar el texto en lenguaje natural [4] y la ciencia que estudia el NLP es la lingüística computacional [5]. El NLP busca técnicas computacionales para analizar y representar el texto que luego será analizado a uno o más niveles lingüísticos tratando de procesar el lenguaje de forma similar a como lo hace el humano [6].

Parte del trabajo realizado fue la evaluación de las herramientas de procesamiento de lenguaje natural que están disponibles. Sin embargo, aunque existen muchas, la mayoría de ellas se centra en el análisis de textos en inglés. Entre las herramientas consideradas están: Apache OpenNLP [4], GATE (*General Arquitecture for Text Engineering*) [5] y NLTK (*Natural Language Toolkit*) [6].

3. Categorías léxicas

Las palabras son consideradas como los elementos más pequeños que son capaces de tener un significado único. De acuerdo a la función que cumplen dentro de una oración, cada palabra se puede clasificar en diferentes tipos. Han sido tradicionalmente conocidas como partes de la oración o clases de palabras [7]. Los problemas de la categorización de las palabras ha sido un campo de estudio durante de mucho tiempo. En una visión clásica, las categorías léxicas deben definirse para cada lengua porque cada una de ellas tiene propiedades que no se puede asegurar sean universales [11]. En los siguientes apartados se definen las categorías léxicas que se analizan durante esta investigación.

3.1 Nombre propios

Los nombres propios se utilizan para designar sustantivos con un nombre particular. Algunos ejemplos de nombre propio serían: Honduras, Motagua, Sula.

3.2 Verbos

Es la categoría léxica utilizada para denotar acción, movimiento, existencia, consecución, condición o estado del sujeto. Una palabra se clasifica en esta categoría sin importar su conjugación. En español, el verbo se puede modificar para concordar con el tiempo, la persona, el número y el modo del sujeto de la oración. Ejemplos de verbos son: correr, jugaba, escribimos.

3.3 Determinantes

Los determinantes son palabras que sirven para expresar a qué objeto se refiere una frase expresada. Los determinantes más conocidos son: los artículos: el, la, lo, las, los, un, una, unos, unas. Pero también se incluyen los demostrativos, posesivos, indefinidos, interrogativos y exclamativos.

3.4 Sustantivos

Un sustantivo se define como una palabra que sirve para designar a personas, animales, lugares, sentimientos o cosas. Los sustantivos en español tienen género y número, es decir, pueden ser masculinos o femeninos, por ejemplo: niño – niña. El número se refiere a que puede ser singular o plural, por ejemplo: niño – niños.

3.5 Adposición

La adposición es un término general utilizada para englobar las preposiciones, posposiciones y circumposiciones. En el idioma castellano, las preposiciones son las más frecuentes. Entre las

preposiciones se encuentran: a, ante, bajo, cabe, con, contra, de, desde, en, entre, hacia, hasta, para, por, según, sin, so, sobre, tras.

3.6 Conjunción subordinada

Las conjunciones subordinadas dejan uno de los términos sujeto o dependiente del otro. Además siempre unen dos proposiciones. Ejemplos de palabras en esta categoría son: mientras, como, según.

3.7 Verbo auxiliar

Los verbos auxiliares se utilizan para crear formas verbales compuestas. Un verbo auxiliar puede ir seguido de uno, dos o más verbos.

3.8 Puntuación

En esta categoría aparecen los signos de puntuación encontrados en los textos estudiados. Estos se usan como delimitantes de frases y párrafos. Estos permiten estructurar el texto y ordenar las ideas. Esta estructura lograda con los signos de puntuación le permiten al lector obtener una mayor comprensión de los textos.

3.9 Adjetivo

Un adjetivo es una clase de palabra utilizada para complementar el significado del sustantivo, asignándole una calificación. Los adjetivos expresan atributos del sustantivo al que determinan. El adjetivo puede hacer referencias a características propias del sustantivo.

3.10 Conjunción

Una conjunción es una categoría de palabras sin contenido significativo, pero sirven para enlazar palabras u oraciones.

3.11 Pronombre

La categoría de pronombres está constituida por las palabras que se pueden utilizar en el lugar del nombre, con el fin de evitar la repetición. Estos hacen referencia a otros términos que ya han sido mencionados en el texto y por lo tanto, el lector puede establecer una referencia.

3.12 Adverbio

Un adverbio es una parte invariable de la oración. Los adverbios sirven para calificar o determinar el verbo o el adjetivo, incluso, otro adverbio.

3.13 Partícula gramatical

La categoría de partícula gramatical tiene comúnmente un significado difuso. Así, cuando se define una partícula se refiere a un conjunto de palabras heterogéneas que carecen de un significado léxico preciso.

3.14 Interjección

A las interjecciones pertenecen las palabras utilizadas para expresar un sentimiento profundo, o una emoción súbita. Es un enunciado de un solo término, es decir, no pertenecen al entramado de una oración. Ejemplos de interjecciones serían: ¡Ah!, ¡ay! ¡bravo!.

4. Metodología

La investigación es de carácter cuantitativo siguiendo el paradigma positivista. El diseño es no experimental, siendo que por las condiciones propias del fenómeno a describir no es posible la manipulación intencionada de variables. Se logra un alcance descriptivo. La recolección de datos se hizo a través de un análisis computacional del corpus descrito en secciones anteriores. La investigación es de corte transversal. Aunque las noticias fueron publicadas en un periodo de tiempo de 7 años, la variable de tiempo no fue analizada durante la investigación.

Para el análisis de los datos se utilizaron herramientas estadísticas. Se hizo un análisis de estadística descriptiva, analizando la frecuencia y tendencias centrales. Además, se presentan los datos en forma gráfica y tabular.

Se estudiaron las noticias publicadas en los diarios Deportivo Más, La Tribuna y Tiempo de Honduras, que pertenecen a la colección UTD-MB-2016. Para esta colección se recuperaron 178,125 noticias de los diferentes diarios. En promedio, por cada noticia se tienen 1714 bytes. No se seleccionó muestra específica, sino que se analizó la totalidad de las noticias disponibles en la colección. Por tanto se tiene una muestra universo.

5. Resultados

En esta sección se presentan algunos de los resultados encontrados en relación con los objetivos de la investigación. Primero se analiza la frecuencia de las palabras, verificando la ley empírica de Zipf. Luego se presenta un análisis de las palabras más utilizadas en los diferentes diarios, para finalizar presentando el análisis de las categorías léxicas a las cuales pertenecen.

5.1 Ley de Zipf

La ley de Zipf establece una relación matemática entre el ranking de una palabra y su frecuencia. Se establece que la palabra más frecuente tiene el ranking 1, la segunda más frecuente tiene ranking 2, y sucesivamente. La ley de Zipf define que la palabra con

ranking 1 será el doble de frecuente que el ranking 2 y el triple que el ranking 3, etc. Esta relación se ve como una línea recta cuando se grafica el ranking en el eje horizontal y la frecuencia en el eje vertical, ambos ejes con escala logarítmica.

En la Figura 1 se puede ver el gráfico. Se nota que, aunque no es una recta perfecta, la curva se acerca a la linealidad establecida. Se puede concluir que la colección de estudio sigue la distribución de frecuencias referida como Ley de Zipf.

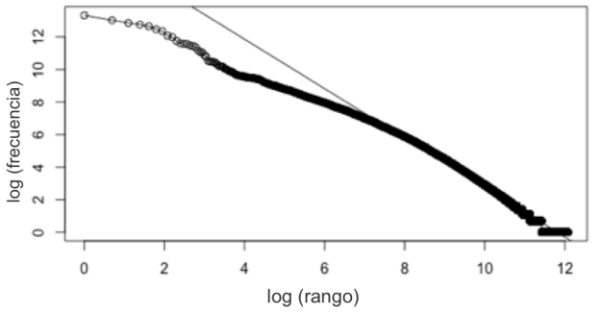

Figura 1. Ley de Zipf para la colección de noticias estudiada.

5.2 Palabras utilizadas.

En esta sección se muestran las palabras más utilizadas en los diferentes medios de la colección y el análisis por cada uno de los tres diarios que componen la colección utilizada.

5.2.1 Diario Tiempo

La Figura 2 muestra las palabras más frecuentes en el Diario Tiempo. Como se puede notar, la palabra más común es "de". En el resto de las palabras que alcanzan a leerse son palabras que no se pueden relacionar con un tema específico. Estas palabras normalmente son conocidas como palabras vacías o *stopwords*.

Figura 2. Palabras más frecuentes en diario Tiempo, incluyendo las palabras vacías.

Para tener una mejor comprensión de las temáticas tratadas en este diario, se suprimieron de la lista las que son consideradas *stopwords* y se rehízo el gráfico de nube de palabras. El resultado se muestra en la Figura 3. Primero, se puede notar que el tamaño de las palabras presente es más equitativo que en la Figura 2. Esto se debe a que la distribución de frecuencias de las palabras restantes es más equitativo.

Figura 3. Palabras más frecuentes en Diario Tiempo al retirar las palabras vacías.

Se pueden destacar algunas palabras que parecen responder a la temática del periódico. Honduras es mencionando frecuentemente, también "Tegucigalpa", "presidente", "partido", "Hernández", "policía". Y con menos cuantía, pero siempre entre las 150 palabras más frecuentes se encuentra "justicia", "seguridad", "dólares", "investigación", "México", entre otras.

5.2.2 Diario La Tribuna

La Figura 4 muestra las palabras más frecuentes en el Diario La Tribuna. La palabra "de" sigue siendo la más frecuente y como era de esperarse, las *stopwords*, siguen siendo las más frecuentes.

Figura 4. Palabras frecuentes en Diario La Tribuna, incluyendo las palabras vacías.

Siguiendo el procedimiento del diario anterior, se volvió a realizar el gráfico luego de eliminar las *stopwords*. El resultado se ve en la Figura 5. Como era de esperar, "Honduras" sigue apareciendo frecuentemente en las noticias de la Tribuna. Así como "nacional", "presidente" y "autoridades". La palabra "unidos" que aparece en este y en el gráfico del diario Tiempo, puede ser una referencia a "Estados Unidos",

aunado al “unidos” con otra connotación que puede aparecer en el texto en español.

Figura 5. Palabras más frecuentes en diario La Tribuna removiendo las palabras vacías.

5.2.3 Diario Deportivo Más

La Figura 6 muestra las palabras más frecuentes en el Diario Deportivo Más. Como se puede notar, la palabra más común es “de”. Aquí se alcanza a ver palabras relacionadas con la temática del diario deportivo, como: equipo, partido, fútbol, balón. Sin embargo, su frecuencia es opacada por la presencia de las *stopwords*.

Al retirar las *stopwords* se construye la nube de palabras que se ve en la Figura 7. Aquí, la palabra “equipo” supera la frecuencia de “Honduras”. Se destaca la aparición de las palabras “Olimpia” y “Motagua” que son los equipos más representativos del deporte nacional.

Figura 6. Frecuencia de las palabras en Diario Deportivo más, incluyendo las palabras vacías.

Sin embargo, la frecuencia de estas palabras es menor a las de “Madrid” y “Barcelona”. Lo que hace suponer que la temática del futbol español supera incluso al deporte nacional en las publicaciones de diario Deportivo Más. También, hace pensar que como hondureños consumimos más el producto extranjero.

Figura 7. Palabras frecuentes en Diario Deportivo más, removiendo las palabras vacías.

En las secciones siguientes se hace un análisis por la categoría léxica de cada palabra. Inicialmente se hace un análisis global, para luego ver la distribución de frecuencias por cada categoría léxica en los tres diferentes periódicos que se están usando como fuente.

5.3 Análisis por cada categoría léxica

En la Tabla 1 se muestran los descriptivos básicos para cada categoría. Los valores son referencias al conteo de cada categoría por noticia analizada. Se destaca que la categoría más utilizada son los sustantivos, con una media de 58.74 sustantivos por noticia. En el otro extremo, las partículas gramaticales es la categoría menos utilizada con una media de 0.05 por noticia.

En las siguientes secciones se presentan las estadísticas principales para cada categoría. Así mismo, la Tabla 1 hace referencia a la media de apariciones por noticia de cada categoría y del máximo encontrado.

Tabla 1. Estadísticos descriptivos para las diferentes categorías léxicas en la colección de estudio. Fuente: datos propios.

	Máximo	Media
Nombre propio	1870	31.66
Verbos	1868	30.17
Determinante	1764	41.82
Sustantivo	2483	58.74
Adposición	2039	48.27
Conjunción Subordinadas	683	7.62
Verbo Auxiliar	716	7.04
Puntuación	1715	34.54
Numeral	577	5.96
Adjetivo	1141	18.30
Conjunción	515	8.98
Pronombre	1412	12.96
Adverbio	710	9.41
Partícula	21	.05
Interjección	28	.09

5.3.1 Nombres propios

Los nombres propios son bastante frecuentes en las noticias analizadas. Lo más común es que hayan alrededor de 13 nombres propios por noticia. Se hizo un análisis por los diferentes diarios estudiados. En el Diario Deportivo más la media por noticia es de 42 nombres propios. En La Tribuna, se encuentra un promedio de 28 nombres propios por noticia. En el Tiempo se tiene una media de 37 nombre propios por cada noticia. Se puede notar que el Diario Más utiliza mayor cantidad de nombres propios en sus noticias, lo que indica que sus noticias están más centradas en personas y organizaciones.

5.3.2 Verbos

La moda de verbos es de 11 por noticia. En el Diario Deportivo más la media por noticia es de 31 verbos. En La Tribuna, se encuentra un promedio de 29 nombres propios por noticia. En el Tiempo se tiene una media de 33 nombre propios por cada noticia. Se puede notar que el Diario Tiempo utiliza mayor cantidad de verbos.

5.3.3 Determinantes

Lo más común es que hayan alrededor de 21 determinantes por noticia. Se tiene un mínimo de 0 y un máximo de 1764 determinantes por noticia. En el Diario Deportivo más la media por noticia es de 41 determinantes. En La Tribuna, se encuentra un promedio de 41 determinantes por noticia. En el Tiempo se tiene una media de 46 determinantes por cada noticia. El diario Tiempo utiliza mayor cantidad de determinantes.

5.3.4 Sustantivos

La moda es de 26 sustantivos por noticia. Se tiene un mínimo de 0 y un máximo de 2483 sustantivos por noticia. La media global es de 58.74 sustantivos por noticia. En el Diario Deportivo más la media por noticia es de 54 sustantivos. En La Tribuna, se encuentra un promedio de 58 sustantivos por noticia. En el Tiempo se tiene una media de 65 sustantivos por cada noticia. El diario Tiempo utiliza mayor cantidad de sustantivos en sus noticias.

5.3.5 Adposición

Aparecen 26 adposiciones en promedio por noticia. Se tiene un mínimo de 0 y un máximo de 2039 adposiciones por noticia. La media global es de 48.27 adposiciones por noticia. En el Diario Deportivo Más la media por noticia es de 48 adposiciones. En La Tribuna, se encuentra un promedio de 47 adposiciones por noticia. En el Tiempo se tiene una media de 53 sustantivos por cada noticia.

5.3.6 Conjunción subordinada

Las conjunciones subordinadas son menos frecuentes que las categorías presentadas hasta ahora. Se tiene un mínimo de 0 y un máximo de 683 conjunciones subordinadas y la media global es de 7.62 por noticia. En el Diario Deportivo más la media por noticia es de 7 conjunciones subordinadas. En La Tribuna, se encuentra un promedio de 8 conjunciones subordinadas por noticia. En el Tiempo se tiene una media de 8 conjunciones subordinadas por cada noticia. El Diario Tiempo utiliza mayor cantidad de conjunciones subordinadas en sus noticias, por una diferencia decimal.

5.3.7 Verbo auxiliar

Los verbos auxiliares no son muy comunes en las noticias escritas. Se tiene un mínimo de 0 y un máximo de 716 determinantes por noticia, para una media global de 7.04 verbos auxiliares por noticia. En el Diario Deportivo Más la media por noticia es de 7 verbos auxiliares. En La Tribuna, se encuentra un promedio de 7 verbos auxiliares por noticia. En el Tiempo se tiene una media de 8 verbos auxiliares por cada noticia.

5.3.8 Puntuación

La moda es que haya alrededor de 15 signos de puntuación por noticia. Se tiene un mínimo de 0 y un máximo de 1715, y con una media de 34.54 signos de punta por noticia. En el Diario Deportivo Más la media es de 37 signos de puntuación. En La Tribuna, se encuentra un promedio de 33 signos de puntuación. En el Tiempo se tiene una media de 38 signos de puntuación por cada noticia. El diario Tiempo utiliza en promedio mayor cantidad de signos de puntuación en sus noticias.

5.3.9 Numeral

La moda es que haya uno o más números por noticia. Se tiene un mínimo de 0 y un máximo de 577 números por noticia, con una media global de 5.96. En el Diario Deportivo Más la media por noticia es de 9 números. En La Tribuna, se encuentra un promedio de 5 números por noticia. En el Tiempo se tiene una media de 7 números por cada noticia.

5.3.10 Adjetivo

La moda es de 7 adjetivos por noticia. Se tiene un mínimo de 0 y un máximo de 1141 adjetivos por noticia y una media global de 18.3 adjetivos por noticia. En el

Diario Deportivo Más la media por noticia es de 18 adjetivos. En La Tribuna, se encuentra un promedio de 18 adjetivos por noticia. En el Tiempo se tiene una media de 19 adjetivos por cada noticia.

5.3.11 Conjunción

La moda es de tres conjunciones por noticia. Se tiene un mínimo de 0 y un máximo de 515 conjunciones y una media global de 8.98 conjunciones. En el Diario Deportivo Más la media por noticia es de 9 conjunciones. En La Tribuna 9 conjunciones y en el Diario Tiempo se tiene una media de 10 conjunciones por cada noticia.

5.3.12 Pronombre

En relación a lo pronombres, la moda es de 4 pronombre por noticia. Se tiene un mínimo de 0, un máximo de 1412 y una media de 12.96 pronombres por noticia. En el Diario Deportivo Más la media por noticia es de 13 pronombres. En La Tribuna se encuentra un promedio de 13 y en el Tiempo se tiene una media de 14 pronombres por cada noticia.

5.3.13 Adverbio

Existe una moda de 3 adverbios por noticia. Se tiene un mínimo de 0 y un máximo de 710 determinantes por noticia y una media global por noticia de 9.41. En el Diario Deportivo Más la media por noticia es de 10 adverbios. En La Tribuna, se encuentra un promedio de 9 adverbios por noticia. En el Tiempo se tiene una media de 10 adverbios por cada noticia. El diario Tiempo y el Deportivo más una cantidad similar de adverbios por noticia.

5.3.14 Partícula gramatical

Lo más común es que no hayan partículas gramaticales, y cuando hay, solo aparezca una. Se tiene un mínimo de 0 y un máximo de 21 partículas gramaticales por noticia. En La Tribuna, se encuentra un promedio de 0.04 partículas gramaticales por noticia. En el Tiempo se tiene una media de 0.05 partículas por cada noticia. El diario Más utiliza mayor cantidad de palabras en esta categoría.

5.3.15 Interjección

Se encontró un mínimo de 0 y un máximo de 28 interjecciones por noticia. En el Diario Deportivo Más la media por noticia es de 0.06 interjecciones. En La Tribuna, se encuentra un promedio de 0.08 interjecciones por noticia. En el Tiempo se tiene una media de 0.17 interjecciones por cada noticia.

5.3.16 Otros

También existió un grupo de palabras que no pudieron clasificarse en alguna de las categorías estudiadas. Por cada noticia fue común encontrar al menos una palabra no clasificada.

5.4 Diario Más

Se realizó también un análisis por diario, comparando la media de cada una de las categorías. El resultado se muestra en la Figura 8. Por la estructura gramatical del español, los textos escritos deberían tener una composición similar. En el diario Más las tres categorías gramaticales más frecuentes en su respectivo orden son los sustantivos, las adposiciones y los nombres propios.

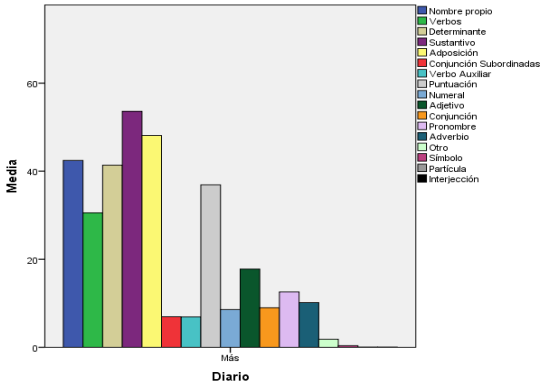

Figura 8. Frecuencia de utilización de las diferentes categorías léxicas en Diario Deportivo Más. Fuente: elaboración propia.

5.5 Diario La tribuna

En el diario La Tribuna, las categorías más frecuentes en su orden son: los sustantivos, las adposiciones y los determinantes, como se ve en la Figura 9. Con muy pocas variaciones, se puede notar que la relación entre las frecuencias de las diferentes categorías es similar a la presentada en la Figura 8 de diario Más.

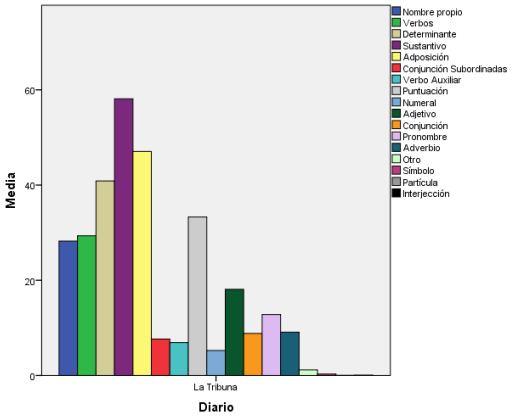

Figura 9. Frecuencia de utilización de las diferentes categorías léxicas en Diario La Tribuna. **Fuente**: elaboración propia.

5.6 Diario El Tiempo

Al igual que en La Tribuna, en el diario Tiempo las categorías más frecuentes son los sustantivos, las adposiciones y los determinantes. La figura 10 muestra el gráfico de barras para la media de frecuencias por categoría encontradas en el Diario El Tiempo. Como se ha venido notando en los diferentes análisis realizados, El Tiempo tiene media de frecuencias considerablemente más altas que los demás diarios.

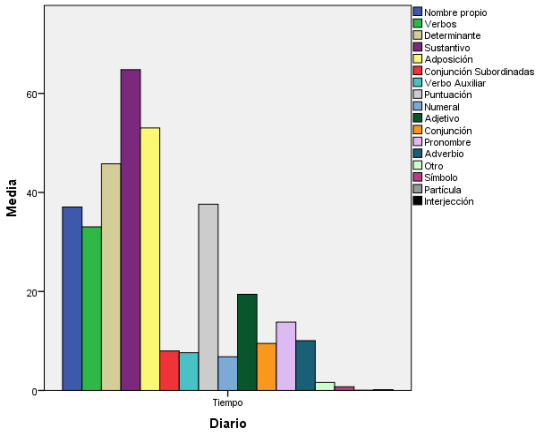

Figura 10. Frecuencia de utilización de las diferentes categorías léxicas en Diario Tiempo. Fuente: elaboración propia.

6. Conclusiones

Las palabras vacías (stopwords) son muy frecuentes, como era de esperarse. En todos los periódicos la palabra más frecuente es "de". También destacan "la", "el", "que" y "en". Al quitar esas palabras aparecen las palabras relacionadas con la temática del periódico. En el diario Tiempo destacan "Honduras", "nacional" y "Tegucigalpa". En La Tribuna "Honduras", "nacional" y "país". El diario Deportivo Más es de corte deportivo, por tanto, destacan las palabras "equipo", "partido" y "futbol". En relación con el diario Más, se nota que cubren más noticias de la liga española de futbol. Siendo que las palabras "Madrid" y "Barcelona" son más frecuentes que las contrapartes de "Olimpia" y "Motagua" que son los equipos referentes del ámbito nacional.

Las categorías gramaticales más frecuentes son los sustantivos en todos los casos. Seguidos en todos los casos por las adposiciones. En tercer lugar, aparecen los nombres propios en Diario Más y los determinantes en La Tribuna y El Tiempo. El software utilizado no fue capaz de encontrar una clasificación para al menos una palabra por cada noticia analizada. La composición de las noticias en los diferentes diarios (en relación a las categorías gramaticales utilizadas) es similar. Al graficar las medias de cada categoría, la relación de tamaño entre las diferentes barras es similar. De los casos es destacable la categoría Interjección, estás son muy escasas en los diarios digitales estudiados.

7. Referencias

[1] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider y N. A. Smith, «Improved part-of-speech tagging for online conversational text with word clusters,» de *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.

[2] P. Ranganathan, «From microprocessors to nanostores: Rethinking data-centric systems,» *IEEE Computer Society*, vol. 44, nº 3, pp. 6-10, 2011.

[3] J. Martínez y L. Bográn, «Construcción de un corpus de noticias hondureñas para análisis lingüístico por medio de técnicas de procesamiento de lenguaje natural,» de *IV Congreso de economía, administración y Tecnología*, Tegucigalpa, 2016.

[4] G. G. Chowdhury, «Natural language processing,» *Annual review of information science and technology*, vol. 37, nº 1, pp. 51-89, 2003.

[5] A. Gelbukh, «Procesamiento de lenguaje natural y sus aplicaciones,» *Komputer Sapiens*, vol. 1, pp. 6-11, 2010.

[6] E. D. Liddy, «Natural Language Processing,» de *Encyclopedia of Library and Information Science*, 2nd Edition ed., New York, Marcel Decker, 2001.

[7] Apache, «Open NLP,» 2017. [En línea]. [Último acceso: 10 agosto 2017].

[8] University of Sheffield, «GATE: a full-lifecycle open source solution for text processing,» 2017. [En línea]. Available: https://gate.ac.uk/overview.html. [Último acceso: 12 agosto 2017].

[9] S. Bird, E. Klein y E. Loper, Natural Language Processing with Python, O'Reilly, 2009.

[10] J. M. García-Miguel, «Categorías léxicas en Tipología Lingüística,» *Verba: Anuario Galego de Filoloxía*, vol. 40, pp. 355-388, 2013.

[11] J. M. García-MiGuel, «Categorías léxicas en Tipología Lingüística.,» *Verba: Anuario Galego de Filoloxía*, vol. 40, pp. 355-388, 2013.