Cuadernos de Lingüística Hispánica N.° 20
ISSN 0121-053X
Julio-Diciembre 2012; pp. 83-98

Anotación y descripción de textos digitales sin formato de la base de casos médicos de la Facultad de Medicina de la Universidad Nacional de Colombia*

Tagging and description of non-format digital texts from the medical case data base of the faculty of medicine at Universidad Nacional de Colombia

GEORGE ENRIQUE DUEÑAS LUNA**
geduenasl@unal.edu.co
FABIO A. GONZÁLEZ***
fagonzalezo@unal.edu.co

Recepción: 13 de marzo de 2012
Aprobación: 30 de julio de 2012

* Este artículo corresponde a un informe de investigación presentado en el XXVII Congreso Nacional y I Internacional de Lingüística, Literatura y Semiótica (Uptc, Colombia: 9, 10, 11 y 12 de octubre de 2012).

** Lingüista, Universidad Nacional de Colombia. Webmaster Facultad de Ciencias Humanas, Universidad Nacional de Colombia, Bogotá, Colombia.

*** Ph.D. en Ciencias Computacionales. Profesor Asociado, Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Colombia.

Resumen

La Lingüística de Corpus es una metodología empírica ya que, a partir de grandes colecciones de textos -corpus o corpora- intenta describir las regularidades de las lenguas por medio de la implementación de programas computacionales, y así, simular los usos reales de ellas. Este trabajo aplica la Lingüística de Corpus a un conjunto de historias médicas electrónicas escritas en español nunca analizado lingüísticamente.

De estas historias se desconoce la forma en que están escritas por parte de los médicos y las clases de palabras que utilizan cuando describen un suceso en una subdisciplina médica.

El conjunto de datos está formado por 19 subdisciplinas médicas, las cuales contienen sus propias historias. Cada historia fue anotada en tres formas diferentes, lematización, tokenización y categoría gramatical (*part-of-speech*) por medio de *TreeTagger*. Posteriormente, las frecuencias de las anotaciones se describieron mediante *AntConc*.

Los resultados encontrados para cada subdisciplina muestran las palabras con mayor frecuencia. Las palabras de clase cerrada son las más comunes y utilizadas. Algunas partes de las historias médicas fueron anotadas erróneamente. Por otra parte, se muestran ejemplos que dan a conocer la variabilidad de uso entre expresiones y abreviaturas por parte del personal médico. Además, la escritura médica de la Universidad Nacional de Colombia corrobora la Ley de Zipf.

Palabras clave: Lingüística de corpus, AntConc, TreeTagger, Ley de Zipf.

Abstract

Corpus Linguistics is an empirical methodology which, based on great collections of text - corpus or corpora- , attempts to describe the regularities of languages by means of the implementation of computer programs, and in this way to simulate their real use. This work applies Corpus Linguistics to a series of electronic medical histories written in Spanish, which have never been linguistically analyzed before.

We do not know the precise form in which these histories were written by doctors or the types of words used when describing an event in a medical subdiscipline.

The set of data is formed by 19 medical subdisciplines, which contain their own histories. Each history was tagged in 3 different ways: lemmatization, tokenization, and grammatical part-of-speech, using TreeTagger. Afterwards, the frequencies of tags were described using AntConc.

The results found for each subdiscipline show the words that appear with greater frequency. The closed class words were the most commonly used. Some parts of the medical histories were tagged erroneously. On the other hand, examples were found that allowed us to recognize the variability of use of expressions and abbreviations in the medical staff. Also, medical writing at Universidad Nacional de Colombia corroborated Zipf's Law.

Key words: Corpus linguistics, AntConc, TreeTagger, Zipf's Law.

Introducción

El procesamiento del lenguaje natural ha avanzado aceleradamente en los últimos años. El estudio de la escritura médica por m2edio de *software* está convirtiéndose en un importante tema de estudio. El lenguaje médico ha llamado la atención de los ingenieros, médicos y lingüistas desde hace 50 años aproximadamente, ya que se ha logrado ejecutar *software* de análisis lingüístico sobre gran cantidad de textos médicos con la finalidad de extraer conocimiento importante para ayudar en el mejoramiento tanto de procesos médicos como de cuidados del paciente. El lenguaje médico ha sido analizado desde diferentes enfoques lingüísticos y con diversas herramientas informáticas.

La Universidad de *New York* ha sido la pionera en la investigación para la lengua inglesa. En 1965 la universidad creó *The Linguistics String Project*[1] *(LSP)* con el fin de facilitar la recuperación de información específica a partir de textos en inglés para ayudar a responder las preguntas formuladas por los investigadores. Este proyecto se compone de cinco etapas de procesamiento para obtener una representación estándar del contenido de los documentos: *Parsing*, *Selection*, *Transformation*, *Regularization*, *Information Formatting*. A partir de este proyecto surge uno nuevo, *Medical Language Processor*[2] *(MLP)*, el cual toma un dominio específico de estudio, el dominio o sublenguaje médico. El MLP es un sistema que transforma documentos clínicos que están en texto-libre a documentos estructurados en formato XML. Muchos trabajos se han derivado del MLP para distintos campos médicos, al igual que para distintas lenguas. Algunos de los estudios que se han llevado a cabo utilizando el MLP se describen a continuación.

Una aplicación de las técnicas derivadas del MLP son los resúmenes de pediatría, el objetivo es la estructuración de los datos sin pérdida o deformación de la información.

---

1 http://cs.nyu.edu/cs/projects/lsp/index.html

2 http://mlp-xml.sourceforge.net/

Posteriormente, se aplican los métodos de recuperación automática de información tanto cualitativos como cuantitativos con el fin de mejorar el cuidado del paciente y mejorar la educación médica de futuros estudiantes (Sager, N. et al., 1978). La estructuración de la información en texto-libre a partir de los documentos de los pacientes para posteriormente, recuperar datos. La implementación de este proceso reduce la mayor cantidad de expresiones a formas estándares que corresponden a los tipos de oraciones básicas del dominio médico (Emile. C. Chi et. al., 1985.a).

La utilización de una técnica llamada *INFORMATION FORMATTING* basada en el LSP. La técnica clasifica la información de los documentos médicos en texto-libre en alguno de los seis formatos preestablecidos con base en relaciones sintácticas y semánticas. El objetivo es poder recuperar la información de pacientes dados de alta para aplicar la misma metodología a posteriores casos (Emile. C. Chi et. al. 1985.b). La adaptación del LSP del inglés para el francés, para crear de nuevo el diccionario médico francés. Otra de las utilidades es la recuperación de la información que contienen *Les Lettres de Sortie*. Una última aplicación es *Answering-question* (Lyman, M. et. al. 1989). TEXTOINFO plasma el perfil de un paciente basado en el análisis de texto-libre de un informe de alta hospitalario para agrupar pacientes con los mismos síntomas (Borst, F. et. al. 1992).

El "adivinador" de la categoría sintáctica (*category guesser*) para el vocabulario médico holandés, está conformado por un diccionario que anota las palabras que encuentra en las historias médicas y, por un conjunto de reglas que combina morfología y conocimiento morfológico logra identificar la forma superficial de las palabras desconocidas, puede mejorar las palabras mal escritas que analiza y corrige ciertas características sintácticas en las historias médicas (Spyns et. al. 1994). Un sistema de mapeo de documentos clínicos en texto-libre dentro de una base de datos formada por campos semánticos para posteriormente formularle preguntas.

Para analizar cada oración se recurrió a *Tke Linguistics String Proyect System* (Sager, N. et. al. 1994). El desarrollo de un procedimiento que mapea el resultado del sistema LSP MLP dentro de los códigos internacionales de *SNOMED* por medio de un algoritmo. Lo que hace primero el algoritmo, es una descomposición preliminar en las unidades morfo-semánticas de la cadena de texto de entrada y de la cadena de palabra de *SNOMED*; posteriormente, se comparan estas unidades y el algoritmo elige la mejor igualación (*the best match*) (Sager, N. et. al. 1995). La fusión entre los recursos lingüísticos médicos y el procesamiento del lenguaje natural combinados con el Lenguaje de Marcado Generalizado facilitan la extracción de datos de pacientes (Sager,

N. et. al. 1996). El desarrollo de un componente morfológico, a partir del LSP-MLP, cuyo conjunto de anotaciones acelerara el proceso de revisión de informes de alta para pacientes, resaltando las palabras claves contenidas en el informe médico (Spyns et. al. 1997).

Otros estudios realizados sin aplicar el MLP se describen a continuación.

CAPIS extrae frases canónicas del examen físico a partir de los resúmenes de admisión para enriquecer una base de datos clínica (Lin, R., 1992). La creación de sistemas para extraer diferentes tipos de estadísticas médicas como el número de pacientes con recurrencia a metástasis, tiempo de operación a tiempo de recurrencia de la primera sospecha de metástasis, localización de nueva metástasis. Otra finalidad es crear un software para procesar preguntas con el fin de responderlas a partir de la información en la base de datos. (Hirschman, L. et. al. 1976). La utilización del sistema *SCAMP*, el cual es un gran número de encuentros entre médico y paciente, incluyendo los resultados de laboratorio, signos vitales y medicaciones cuyo objetivo es mapear los datos médicos dentro de una base de datos estructurada semánticamente para que los diferentes tipos de datos allí almacenados puedan ser recuperados, comparados y resumidos. (Carol Friedman et. al. 1983). La detección automática de palabras para completar y crear diccionarios. Lo anterior se lleva a cabo analizando las regularidades morfo-semánticas en los vocabularios médicos. (Wolff, S. 1984, 199). *SPECIALIST* retorna documentos que contengan la pregunta ingresada, también dará como respuesta documentos que contengan sinónimos al sentido de la pregunta. (McCray, A. T. et. al., 1993).

La mayoría de las investigaciones sobre la escritura médica se han centrado en la aplicación de las técnicas de MLP-LSP para extraer información de reportes médicos, crear resúmenes automáticamente, crear y/o actualizar diccionarios, almacenar la información en diferentes bases de datos, entre otros. Los análisis llevados a cabo hasta el momento han sido para lenguas diferentes a la española, por tal motivo, este trabajo se interesa en la forma de la escritura que se encuentra en el Sistema de Información del Centro de Telemedicina – SARURO[3] de la Universidad Nacional de Colombia. Hasta ahora, no existen trabajos que analicen la escritura de los médicos colombianos que colaboran con SARURO desde el punto de vista lingüístico. Por tal motivo, el objetivo de esta investigación fue anotar automáticamente cada texto con tres diferentes formas de anotación lingüística –lematización, tokenización y categoría gramatical– y observar tanto las frecuencias de palabras usadas

[3] http://www.bioingenium.unal.edu.co/pagpro.php?idp=saruro&lang=es&linea=1

como el inconstante uso entre abreviaturas y conceptos; buscando de esta manera, las tendencias al momento de escribir por parte de los médicos.

Así mismo, esta investigación pretende ser pionera en el análisis de la escritura médica, no solo en los médicos del Centro de Telemedicina sino en los médicos de Colombia.

1. Problema

Uno de los inconvenientes del análisis de la escritura médica, a diferencia de la información numérica que puede tabularse con facilidad, es la forma en que se almacena dicha información; ya que los documentos médicos de los pacientes están principalmente en texto libre, es decir, texto escrito sin ningún formato. El otro inconveniente es entenderla, ya que ésta pertenece a un dominio particular, el médico. Esta presenta características que la diferencia del lenguaje cotidiano. La escritura médica está conformada por conceptos poco usados por personas y por una gran cantidad de abreviaturas debido a que el nombre del concepto es extenso. Por otra parte, la escritura médica es de vital importancia, ya que es el medio más común, y, único hasta el momento por el cual el personal médico informa a colegas y a pacientes de los estados de salud y procedimientos, siendo esta el mecanismo por el cual se transmite la mayor cantidad de información. La Facultad de Medicina de la Universidad Nacional de Colombia posee SARURO, el Sistema de Información del Centro de Telemedicina que cuenta con más de 25000 registros médicos. En este trabajo presentamos una anotación lingüística (lema, *token* y categoría gramatical) de 22973 registros médicos que están en formato XML. Este sistema está integrado por 188 subdisciplinas médicas, aunque no todas contienen archivos. Solo 19 subdisciplinas contienen archivos. Los archivos están escritos en lenguaje natural (español colombiano), los cuales no han sido objeto de estudio lingüístico hasta la fecha.

2. Metodología

Las 19 subdisciplinas que contienen archivos con escritura médica son: Cardiología tiene 140 archivos; cuidado intermedio, 422; dermatología, 2515; ginecología y obstetricia, 81; infectología, 147; mamografía, 367; medicina interna, 522; neumología pediátrica, 1; neurología, 28; nutrición humana, 14; ortopedia, 130; otorrinolaringología, 51; patología anatómica y clínica, 10; patología, 78; pediatría, 109; psiquiatría, 5; radiología, 19.314; reumatología, 8 y urología, 31. Para un total de 23.973 archivos médicos. Cada archivo está conformado por una serie de etiquetas que albergan la información de cada paciente, varias de esas etiquetas es donde el médico almacena de forma libre la escritura de los diagnósticos y seguimientos de los casos. Para conocer la o las etiquetas que contienen dicha información escrita se obtuvo una muestra aleatoria de 10 archivos por disciplina. Para cada subdisciplina se realizó una plantilla. Esta plantilla fue

utilizada por el software Altova[4] para ubicar las etiquetas que contenían la información de cada uno de los archivos XML y extraer el texto digitado. Como ejemplo, en las figuras 1, 2 y 3 se muestran las plantillas para extraer la información de las subdisciplinas de Radiología, Dermatología y Medicina Interna.

.//caso/titulo,.//caso/texto,.//respuesta/titulo,.//respuesta-1/texto

Figura 1. Plantilla para la disciplina de Radiología

.//caso/texto,.//caracteristicas_dermatologicas/tipo_de_lesion,.//caracteristicas_dermatologicas/localizacion,.//caracteristicas_dermatologicas/dimensiones,.//caracteristicas_dermatologicas/forma,.//caracteristicas_dermatologicas/color,.//caracteristicas_dermatologicas/bordes,.//caracteristicas_dermatologicas/periferia,.//caracteristicas_dermatologicas/superficie,.//caracteristicas_dermatologicas/sensibilidad,.//caracteristicas_dermatologicas/dias_de_evolucion,.//caracteristicas_dermatologicas/sintomas,.//imagen-1/texto,.//imagen-2/texto,.//imagen-3/texto,.//imagen-4/texto,.//imagen-5/texto,.//imagen-6/texto,.//imagen-7/texto,.//respuesta-1/titulo,.//respuesta-1/texto,.//respuesta-2/entrada-1

Figura 2. Plantilla para la disciplina de Dermatología

.//caso/titulo,.//caso/texto,.//examen_fisico/cabeza,.//examen_fisico/cuello,.//examen_fisico/sistema_respiratorio,.//examen_fisico/sistema_cardiovascular,.//examen_fisico/abdomen,.//examen_fisico/extremidades,.//examen_fisico/piel_y_faneras,.//examen_fisico/otros,.//paraclinicos/radiografias,.//paraclinicos/otros,.//respuesta-1/texto,.//respuesta-1/titulo,.//respuesta-2/entrada-1,.//respuesta-2/entrada-2

Figura 3. Plantilla para la disciplina de Medicina interna

El punto y las dos plecas inclinadas al inicio son el acceso al nombre de la etiqueta principal, la palabra que sigue es el nombre de la etiqueta principal, la siguiente pleca inclinada es el acceso a la etiqueta que contiene el texto y la siguiente palabra, es el nombre de la etiqueta que contiene el texto a extraer. La coma se utiliza para pasar a la siguiente etiqueta principal. Este proceso se realiza hasta que se hayan recorrido todas las etiquetas que fueron seleccionadas para cada texto. Después de ejecutar el programa para cada subdisciplina, sólo cambiando la plantilla de extracción, los archivos de cada subdisciplina cambiaron debido a que no había texto dentro de los archivos seleccionados, quedando el nuevo corpus así: cardiología tiene 129 archivos; cuidado intermedio, 347; dermatología, 2424; ginecología y obstetricia, 65; infectología, 141;

[4] http://www.altova.com/

mamografía, 367; medicina interna, 431; Neumología pediátrica, 1; neurología, 25; nutrición humana, 10; ortopedia, 123; otorrinolaringología, 46; patología anatómica y clínica, 10; patología, 78; pediatría, 101; psiquiatría, 4; radiología, 18.635; reumatología, 8 y urología, 28. Para saber cómo descartar los archivos que no tenían texto, nos basamos en el peso del archivo, descartando los archivos que pesaban cero *bytes*. Con los datos obtenidos se realizaron las 3 anotaciones (lematización, tokenización y categoría gramatical) mediante *TreeTagger*[5]. Las frecuencias de los resultados con *TreeTagger* se tabularon con *AntConc*[6] para establecer qué anotaciones son más ocurrentes, y, a partir de los resultados, también se compararon algunas abreviaturas con sus correspondientes nombres completos para observar qué forma es la más utilizada.

3. Resultados

De los 23973 registros médicos en formato XML, se redujeron a 22973 debido a que algunos registros no contenían texto escrito por los médicos. A continuación se muestran las frecuencias (F) de los 10 primeros lemas, *tokens* y categorías gramaticales para 3 de las 19 tablas de las subdisciplinas con más registros médicos. El comportamiento de las anotaciones a lo largo de las 16 subdisciplinas restantes ocurrió de manera similar a las mencionadas anteriormente.

Para LEMA en Radiología (ver Tabla 1) y Dermatología (ver Tabla 2), a excepción de "normal", "observar" y "aplicar" las demás pertenecen a palabras de clase cerrada. En

Tabla 1. Anotaciones con *TreeTagger* para la subdisciplina Radiología.

RADIOLOGÍA					
F	**LEMA**	**F**	***TOKEN***	**F**	**POS**
43537	de	36148	de	158352	NC
34519	el	17077	y	89341	ADJ
18057	y	13687	se	60719	PREP
18016	se	10160	La	40081	VLfin
11135	normal	7386	DE	38548	FS
10389	ser	7126	es	35794	ART
8204	n	6506	Los	23435	NP
7457	observar	6311	del	18932	CC
7168	del	5753	la	18255	SE

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6] http://www.antlab.sci.waseda.ac.jp/antconc_index.html

cuanto a LEMA en Medicina Interna (ver Tabla 3), las palabras pertenecen a la clase cerrada. Los 10 primeros *tokens* de estas tres subdisciplinas pertenecen a palabras de clase cerrada. Finalmente, para la categoría gramatical (POS) en las tres subdisciplinas se observa que predominan las palabras de clase abierta.

Tabla 2. Anotaciones con *TreeTagger* para la subdisciplina Dermatología.

DERMATOLGÍA					
F	**LEMA**	**F**	***TOKEN***	**F**	**POS**
8344	de	5133	de	42743	NC
6617	el	3198	DE	18638	PREP
4237	en	2457	en	16129	NP
3439	y	2189	y	13725	ADJ
2184	por	2181	la	9754	CARD
1414	con	1861	SE	8189	VLfin
1292	CON	1748	EN	7396	ART
1279	SE	1718	QU	7159	CM
1265	se	1414	por	6468	VLinf
1246	aplicar	1389	con	6322	FS

Tabla 3. Anotaciones con *TreeTagger* para la subdisciplina Medicina Interna.

MEDICINA INTERNA					
F	**LEMA**	**F**	***TOKEN***	**F**	**POS**
6034	de	4491	de	28203	NC
3094	el	1935	con	13556	PREP
2508	y	1891	y	10734	ADJ
2066	en	1785	SE	10521	NP
1975	con	1528	DE	7107	CM
1467	por	1441	en	6537	VLfin
1443	no	1145	a	6254	CARD
1427	se	1090	la	4769	FS
1421	SE	1051	por	3665	ART
1166	a	906	que	3041	CC

Para mostrar unos ejemplos de los que se puede encontrar en SARURO mediante la aplicación de *AntConc*, se escogieron los siguientes pares de términos debido a su cercanía en significado, ya que el segundo término es usado como abreviatura del primero.

El objetivo es hallar las distintas apariciones de las parejas de palabras seleccionadas con sus correspondientes frecuencias; es decir, observar cómo escriben estos términos los médicos y cuál es la forma más utilizada. Primero buscamos la palabra "radiografía", nos dimos cuenta que se escribe de diversas formas por el personal médico. Se encontraron 69 realizaciones de este término. Por otra parte buscamos "rx" ya que es la abreviatura de la palabra "radiografía". Se encontró que los médicos la escriben de 4 formas diferentes. Se hallaron 8662 realizaciones de este término. Cada palabra con su correspondiente frecuencia es mostrada en las tablas 4 y 5:

Tabla 4. Realizaciones para la palabra "radiografía".

Palabra	Frecuencia
RADIOGRAF	1
RADIOGRAFIA	22
Radiografia	2
radiografia	10
RADIOGRAFIAS	3
Radiografías	2
radiografía	15
radiografias	2
RADIOGRAFÍAS	3
Radiografía	4
Radiografias	1
RADIOGRAFÍA	3
RADIAGRAFIA	1

Tabla 5. Realizaciones para la abreviatura "rx".

Palabra	Frecuencia
RX	8311
Rx	268
rx	83

Otro ejemplo realizado fue buscar la palabra "tórax", nos dimos cuenta que se escribe de diversas formas por parte de los médicos. Se hallaron 8091 realizaciones de este término. Por otra parte, buscamos "tx" ya que es la abreviatura de la palabra "tórax".

Nos dimos cuenta que los médicos la escriben de 3 formas diferentes. Se hallaron 121 realizaciones de este término. Las tablas 6 y 7 muestran las correspondencias de cada palabra con su frecuencia:

Tabla 6. Realizaciones para la palabra "tórax".

Palabra	Frecuencia
TORAX	2084
torax	308
Tórax	60
Torax	54
Tórax	3600
TÓRAX	77
TORÁX	1908

Tabla 7. Realizaciones para la abreviatura "tx".

Palabra	Frecuencia
TX	103
Tx	8
tx	10

El último ejemplo para observar lo que se hace con *AntConc* fue el siguiente, buscamos la cadena "enfermedades de transmisión sexual" y la cadena "infecciones de transmisión sexual", debido a que hacen referencia a una mismo concepto, nos dimos cuenta que sólo está escrita una sola vez la cadena "enfermedades de transmisión sexual" y cero veces la cadena "infecciones de transmisión sexual". Adicionalmente se buscó "ETS" y "ITS", ya que son las abreviaturas más utilizadas de las cadenas "enfermedades de transmisión sexual" e "infecciones de transmisión sexual". Se hallaron 14 realizaciones para "ETS", 10 para "ITS". La tabla 8 muestra la correspondencia de cada palabra con su frecuencia:

Tabla 8. Realizaciones para "enfermedades de transmisión sexual" y para las abreviaturas "ETS" E "ITS"

Palabra	Frecuencia
ETS	14
ITS	10
enfermedades de transmisión sexual	1
infecciones de transmisión sexual	0

El análisis de las tres anotaciones y de las palabras con sus correspondientes abreviaturas mostró claramente la variabilidad entre mayúsculas y minúsculas, es decir, palabras totalmente escritas o en mayúscula o en minúscula o entre ambas. Además, mostró que los errores más frecuentes son la escritura de palabras sin el espacio de división entre ellas, palabras incompletas y la falta del acento diacrítico. Igualmente, se observa el gran uso de las abreviaturas en lugar de las palabras.

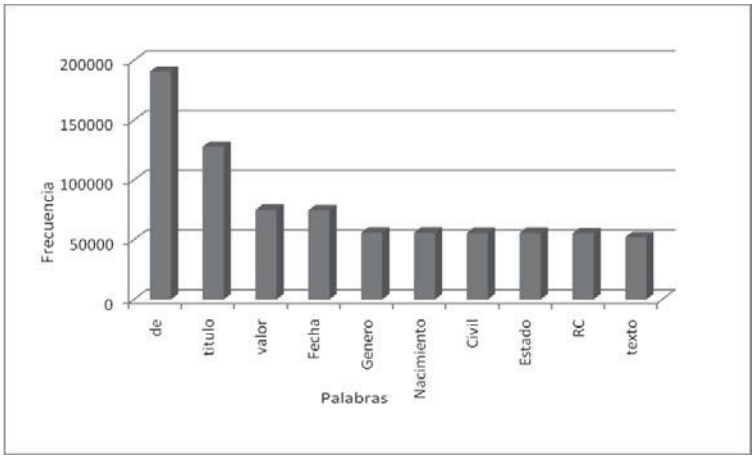

Figura 4. *Tokens* para la subdisciplina de Radiología.

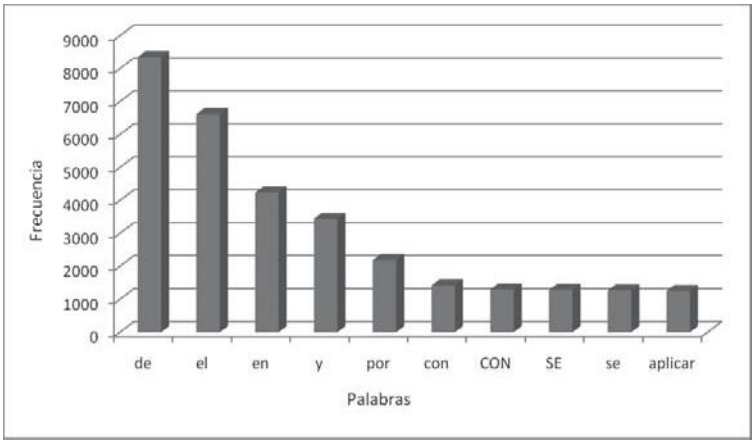

Figura 5. *Tokens* para la subdisciplina de Dermatología.

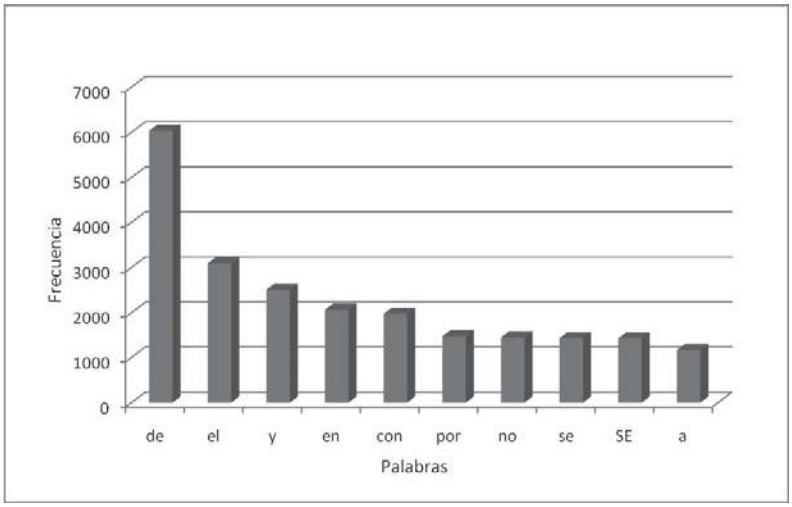

Figura 6. *Tokens* para la subdisciplina de Medicina Interna.

En las figuras 4, 5 y 6 se muestra la corroboración de la Ley de Zipf para las tres subdisciplinas elegidas. Lo anterior ocurrió de la misma manera para el resto de las subdisciplinas. Para estas subdisciplinas, graficamos la frecuencia de los primeros 10 *tokens* para que la visualización de la figura sea legible.

4. Discusión

El contenido escrito en cada registro médico no se convirtió a minúscula, tampoco se corrigieron los errores de acentuación, ortografía, división de palabras; por eso encontramos por una parte, ya sea en lema, *token* o categoría gramatical la misma palabra tanto en mayúscula como en minúscula; por otra parte, encontramos palabras sin anotación ya que dos palabras no están separadas por espacio y *TreeTagger* no supo cómo anotarlas. Asimismo, encontramos las anotaciones y frecuencias de todos los signos de puntuación escritos por los médicos.

Además, *TreeTagger* no asignó correctamente el lema a algunas palabras, es el caso de "areas" para la subdisciplina de dermatología; de "VASCULARES", "AREOLAS", "PEZONES", "MICROCALCIFICACIONES", "RETROMAMARIOS", para la subdisciplina

de Mamografía; de "AÑOS" para la subdisciplina de Neurología; de "IMAGENES", "GENES" para la subdisciplina de Patología; "Anovulatorios", "Parenterales", "GESTACIONES", "IMAGENES" para la subdisciplina de Patología anatómica y clínica; de "sintomas" para la subdisciplina de Psiquiatría; de "óseas", "cardiofrénicos", "pulmonares" para la subdisciplina de Radiología; de "DETALLES" para la subdisciplina de Reumatología; los lemas adecuados son las formas en singular. Una posible explicación a lo anterior, es que *TreeTagger* no tiene almacenadas estas palabras en sus bases de datos.

5. Conclusiones

Todas las frecuencias mostradas dependen de la cantidad de textos o archivos que contenga cada subdisciplina. En este trabajo se observa la desigualdad entre las subdisciplinas, lo ideal es que tengan la misma cantidad de archivos para confrontarlas y no obtener resultados sesgados.

La escritura médica muestra un campo de creación de términos, ya que el personal médico fusiona dos o más palabras en una sola, por lo que muchas de las anotaciones deben ser revisadas para determinar si la anotación es errónea o se ha creado una nueva palabra. En algunas de las anotaciones no se sabe qué función gramatical está desempeñando la palabra; qué significado representa la abreviatura, ya que de las abreviaturas encontradas no se tiene certeza si son abreviaturas en realidad de dichos conceptos o estamos en frente de un problema de confusión de uso de abreviaturas. En ocasiones, un concepto o abreviatura posee diversas acepciones en la literatura médica. Una tarea será pasar todos los textos a minúscula, seguida de la revisión silábica de las palabras para determinar si dos palabras deben ser separadas o no por espacio o guión ya que *TreeTagger* no puede identificar esos fenómenos.

Referencias bibliográficas

McCray. A. T. et al. (1993). *UMLS® Knowledge for Biomedical Language Processing*. Bulletin of the Medical Library Association, 81, 2, 184-194.

Borst, F. et al. (1992). *TEXTINFO: A Tool for Automatic Determination of Patient Clinical Profiles Using Text Analysis*. Proceedings of the Annual Symposium on Computer Application in Medical Care, 63-67.

Friedman, C. et al. (1983). C*omputer Structuring of Free-Text Patient Data*. Proceedings of the Annual Symposium on Computer Application in Medical Care, 688-691.

Emile. C., Chi et al. (1985.a). A *Database of Computer-Structured Narrative: Methods of Computing Complex Relations*. Proceedings of the Annual Symposium on Computer Application in Medical Care, 221-226.

Emile. C., Chi. et al. (1985.b). *Processing Free-Text Input to Obtain a Database of Medical Information*. AFIPS Joint Computer Conferences. Proceedings of the June 7-10, national computer conference and exposition, 45, 267-275.

Hirschman, L. et al. (1976). *From Text to Structured Information: Automatic Processing of Medical Reports*. Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, 82-90.

Lin, Richard, et al. (1992). *A Free-Text Processing System to Capture Physical Findings: Canonical Phrase Identification System (CAPIS)*. Proceedings of the Annual Symposium on Computer Application in Medical Care, 843–847.

Lyman, M. et al. (1989). *Medical Language Processing for Knowledge Representation and Retrievals*. Proceedings of the Annual Symposium on Computer Application in Medical Care, 548-553.

Sager, N. et al. (1978). *Computerized Language Processing for Multiple Use of Narrative Discharge Summaries.* Proceedings of the Annual Symposium on Computer Application in Medical Care, 330-343.

Sager, N. et al. (1994). *Natural Language Processing and the Representation of Clinical Data*. Journal of the American Medical Informatics Association, 1, 2, 142-160.

Sager, N. et al. (1995). *Medical Language Processing: Applications to Patient Data Representation and Automatic Encoding*. Methods of Information in Medicine, 34, 140-146.

Sager, N. et al. (1996). *Medical Language Processing with SGML Display*. AMIA Symposium Proceedings, 547-551.

Spyns, P. et al. (1994). *A Robust Category Guesser for Dutch Medical Language*. Proceedings of the fourth Conference on Applied Natural Language Processing 94 (ANLP94), 150-155.

Spyns, P. et al. (1997). *Dutch Sublanguage Semantic Tagging combined with Mark-Up Technology*. Proceedings of the fifth Conference on Applied Natural Language Processing 97 (ANLP97), 182-189.

Wolff, S. (1984). *The Use of Morphosemantic Regularities in the Medical Vocabulary*. Methods of Information in Medicine, 23, 4, 195-203.