Procesamiento de textos electrónicos para la construcción de un corpus*

Sofía N. Galicia-Haro

Centro de Investigación en Computación, Instituto Politécnico Nacional
Av. Juan de Dios Batiz S/N, 07738 México, D. F.
(+52 55) 5729-6000 Ext. 56544, Fax 5586-2936
sofia@cic.ipn.mx

Resumen. Las colecciones de textos o corpus con etiquetado en diferente proporción, de cero a la totalidad, y en diferentes niveles (morfológico, sintáctico, semántico) son recursos muy útiles que se emplean para diferentes trabajos en investigación teórica y en diversas aplicaciones de lenguaje natural. Sin embargo este recurso no existe para todos los lenguajes, principalmente por la gran cantidad de esfuerzo requerido para compilarlos y para realizar las anotaciones. En este trabajo presentamos los pasos y resultados iniciales para crear un corpus de varios millones de palabras del español en su variante mexicana, de una forma automática.

1 Introducción

La obtención de información del uso del lenguaje a partir de colecciones de textos ha sido una práctica común en la rama de la lingüística aplicada que trata con el dis e-ño y la construcción de bases de datos léxicas, es decir, en la lexicografía. En el procesamiento lingüístico de textos mediante computadora también se ha vuelto una práctica común investigar los fenómenos que principalmente aparecen en materiales sin restricciones, empleando colecciones muy grandes de textos, es decir, corpus.

Los investigadores reconocen el potencial de explotación de corpus grandes para resolver problemas en los diferentes niveles de análisis que en la lingüística general se considera componen el procesamiento lingüístico: léxico, sintáctico y semántico. Sin embargo, los corpus más útiles para el procesamiento lingüístico de textos mediante computadora requieren anotaciones o etiquetas, en cada uno de esos niveles (categoría gramatical, género, número; estructura sintáctica; marcas de significado, concepto, etc.).

Entre los usos más importantes de los corpus etiquetados está el entrenamiento de métodos para resolver distintas tareas, por ejemplo para asignación de categorías gramaticales de palabras desconocidas [17], para vincular frases preposicionales a las palabras correctas [15], para desambiguación del sentido de palabras [14], para la corrección de errores semánticos en los textos [4], etc.

La principal forma de obtención de corpus etiquetados ha sido la compilación y anotación manual. Esta es la razón principal de que no existan corpus etiquetados para

Trabajo realizado con el apoyo parcial de CONACyT, SNI, CGPI y PIFI-IPN, México.
A. Gelbukh, M. Hernández Cruz (Eds.) Avances en la ciencia de la computación en México, CORE-2003, pp. 159–170, 2003.

la mayoría de los lenguajes, entre ellos el español en su variante mexicana. Si además consideramos la construcción de corpus grandes, del orden de decenas de millones palabras, los esfuerzos manuales requeridos serían enormes. Por lo que actualmente investigan métodos automáticos tanto para la compilación como para la etiquetación de corpus grandes (en [1] se presenta una discusión sobre tiempo y costo en la construcción de corpus anotados muy grandes).

En este trabajo presentamos algunas observaciones y resultados referentes compilación automática de un corpus grande de textos para el español en su variante mexicana, a partir de textos electrónicos obtenidos de Internet. Primero presentamos las características de los corpus y enseguida los resultados hasta ahora obtenidos.

2 Corpus para el procesamiento de lenguaje natural

La compilación de corpus de textos, implica la solución de diversos problemas texto en sí mismo. Las fuentes que formarán el corpus deben analizarse respecto rios criterios: la adquisición de los textos con la información requerida, la cobertu del corpus respecto a los fenómenos lingüísticos requeridos, la confiabilidad en corpus, etc.

Para que un corpus se considere como representativo del lenguaje total debe cont ner en proporción a su uso todas las palabras y construcciones del lenguaje. Así para construir este corpus, idealmente sería deseable lograr una muestra muy grande representativa del lenguaje general. Mientras mayor es el corpus se espera un may número de palabras diferentes, lo que implicaría una mayor cobertura del diccionar del lenguaje, y principalmente supone mayor evidencia de los diversos fenóme lingüísticos (topicalización[1], oraciones subordinadas, pasiva refleja, etc.). Que sea presentativa supone diversos niveles culturales del lenguaje, diversos temas y gén ros. Sin embargo, estas cualidades no implican una a la otra, es más, en algunos se contraponen. Existe otro compromiso a considerar entre calidad y cantidad, hecho de tener un corpus grande no garantiza que posea la calidad esperada.

Idealmente el corpus debería estar balanceado entre esas cualidades. Sin embarg parece no ser posible balancear un corpus apropiadamente, al menos no sin un elev do esfuerzo. Los métodos de muestreo, por ejemplo para seleccionar calidad, son caros. Así que asumimos los problemas obvios de trabajar con datos no balanceadc ya que para construir un corpus balanceado requeriríamos mucho tiempo y un muy elevado.

Asumiendo la imposibilidad de tener un corpus con todas las cualidades deseabl limitamos las cualidades del corpus a las más importantes para los objetivos perseg dos en nuestro caso: la información requerida y el tamaño. El objetivo principal compilación de un corpus de textos mexicanos es el análisis sintáctico de textos restricciones, por lo que los textos de periódicos se consideran una buena fuente información requerida. Uno de los mayores problemas en el análisis sintáctico diante computadora es el enlace de grupos nominales y grupos preposicionales

[1] Tipo de desplazamiento por el cuál un sintagma puede aparecer en la posición oracional tópico. Por ejemplo: *Que llegues a hacerlo francamente no lo creo.*

palabras correctas[2]. Por lo que consideramos de suma importancia, que el corpus contenga uso extenso de frases preposicionales y de complementos.

Respecto a que el corpus tenga la información requerida, por ejemplo, [3] indica el diferente uso de frases preposicionales según el género de los textos. En [16] encontraron que hay diferencias significantes entre las frecuencias de los complementos que una palabra rectora puede tener y la categoría gramatical de ellos, en diferentes corpus. Los autores identificaron dos fuentes distintas para esas diferencias: la influencia del discurso y la influencia semántica. Los cambios en las formas de lenguaje que se usan en diferentes tipos de discurso originan la primera. La influencia semántica se basa en el contexto semántico del discurso. Así que la inclusión de diferentes géneros sería muy adecuada.

Respecto al tamaño del corpus, los corpus actuales están en el rango de un millón de palabras a cientos de millones, dependiendo del tipo, es decir, si son texto plano o con etiquetas de diversas clases. Por ejemplo, en [2] discuten que para obtener buenas aproximaciones de probabilidades, el corpus tiene que ser suficientemente grande para evitar los datos esparcidos y para reflejar el uso natural del lenguaje. Las autoras usaron el Wall Street Journal, un corpus de un millón de palabras. A diferencia de ellas, en otros trabajos, no emplean el corpus completo sino subcorpus con características específicas para su investigación [18], [15], [9], lo que implica un esfuerzo de selección de textos. Por lo que un corpus de varios millones de palabras sería la solución más rápida comparada con la obtención de material específico.

Para nuestro objetivo, análisis sintáctico de textos en español sin restricciones, la mayor importancia del corpus radica en la posibilidad de obtener los argumentos de verbos, adjetivos y sustantivos, para reducir el número de variantes de estructuras. Los argumentos corresponden a los complementos seleccionados semánticamente. En [16], los autores explican que conforme la cantidad de contexto circundante aumenta (yendo de una sola oración a un discurso conectado) decrece la necesidad de expresar manifiestamente todos los argumentos del verbo. Esta situación también se presenta en las oraciones muy largas. Por lo que en nuestro caso, son de gran utilidad aún las frases cortas.

2.1 La Web como fuente de textos electrónicos

Para construir rápidamente un corpus grande, de decenas de millones de palabras se requiere la disponibilidad de un vasto repositario de documentos electrónicos. Una fuente obvia de textos electrónicos es la Web que contiene millones de textos actuales aunque no soluciona por sí misma el problema de representatividad del lenguaje.

La Web comúnmente ofrece centenas de millones de páginas distintas, además de miles de nuevas páginas cada día. El problema de esta fuente es que las páginas son muy diversas en contenido y estilo. Muchas son inapropiadas para la construcción de corpus pero existen muchas otras útiles, de diversos temas y géneros.

Aunque es posible obtener una gran cantidad de textos al azar y seleccionar manualmente los más adecuados, no sería la mejor forma de construir rápidamente un

[2] Por ejemplo, en la frase, *compró contra su voluntad un traje nuevo en la tienda de la esquina*, se presentan tres grupos preposicionales: *contra su voluntad, en la tienda, de la esquina*. Mientras los dos primeros se enlazan al verbo, el tercero se enlaza a *tienda*. El grupo nominal *un traje nuevo* también se enlaza al verbo.

corpus. Para nuestro objetivo es posible obtener automáticamente miles de páginas periódicos mexicanos en sus diversas secciones.

2.2. Etiquetado de corpus

Existen diversos niveles para el etiquetado de los corpus, entre ellos: léxico, sintácti- co, semántico. En cada uno de estos niveles pueden existir a su vez otros niveles etiquetado. En el nivel léxico se considera usualmente el etiquetado del lema y las tegorías gramaticales. Por ejemplo, la palabra *textos* se etiquetaría: *texto* (lema), *tantivo masculino plural* (categoría gramatical). Estas asignaciones pueden tener versos grados de detalle. Por ejemplo, en el Penn Treebank [10] se utilizan etiquetas de categorías gramaticales, y 12 para puntuación y otros símbolos; mientras en el Brown Corpus [7] se distinguen 87 etiquetas simples y se permite a partir ellas la formación de etiquetas compuestas.

En el nivel sintáctico, las etiquetas muestran la estructura de la oración, es decir, describen cómo las palabras de la oración se relacionan y cuál es la función que cada palabra realiza en la oración. Usualmente se muestra agrupando las palabras mediante paréntesis, y adicionalmente etiquetando esos grupos. Existen diferentes grados realización de la estructura jerárquica de la oración, debido a que una estructura com- pleta requiere mayor tiempo de aprendizaje del esquema por parte de los anotadores más tiempo para etiquetar las oraciones. Por ejemplo, en el desarrollo del Penn Tree- bank se ignoró, en la primera etapa, la distinción de argumentos y adjuntos en la ción. Aunque el etiquetado de argumentos es crucial para la interpretación semántic del verbo. Los adjuntos corresponden a los complementos que expresan las cunstancias en las que se da la acción. Por ejemplo, en la frase, *compró contra su voluntad un traje nuevo*, el grupo preposicional *contra su voluntad* es un junto que expresa un modificador a la acción *comprar*, pero no es un partic pante de la acción del verbo.

En el nivel semántico, por ejemplo, se ha considerado el etiquetado de significac y de tipo (o concepto) en el desarrollo del *Italian Syntactic-Semantic Treebank* [5].

La mayor parte del etiquetado de corpus en los niveles sintáctico y semántico realizado de forma manual. Considerar esta tarea en un corpus muy grande requeri demasiados recursos humanos. Sin embargo, hay que considerar que se están expl rando numerosos métodos para aprovechar la situación donde se dispone de algu cantidad de datos etiquetados y una cantidad mucho más grande de datos sin etiqu tar, por ejemplo: clasificación de documentos por temas [13], [8].

3. Construcción del corpus

Seleccionamos cuatro periódicos mexicanos que diariamente sitúan en la red una siderable parte de su publicación y que por el tipo de organización permitía una tracción automática por períodos mensuales y anuales. Coleccionamos los textos los artículos de diferentes secciones (economía, política, cultura, deportes, etc.) dur- te los años de 1998 al 2000, todos en formato HTML.

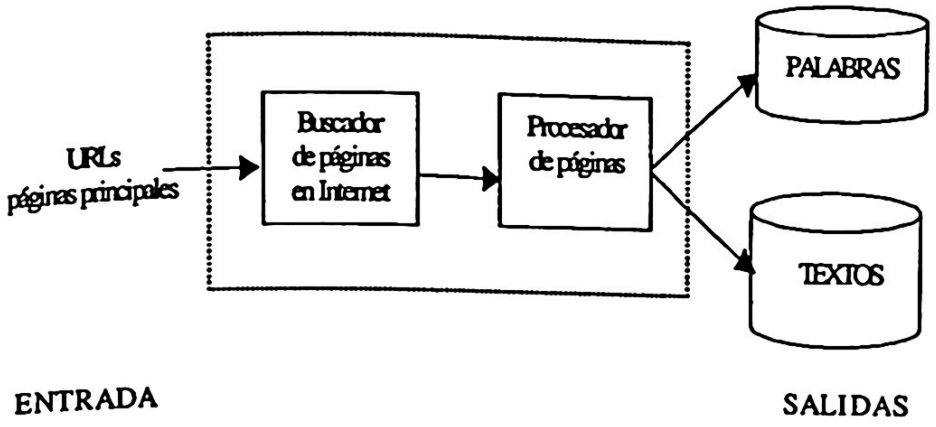

Fig. 1. Primera etapa de la construcción del corpus

En este trabajo presentamos la compilación de los textos, su formateo, su normalización y la etiquetación de grupos de palabras. La realización de estas tareas se llevó a cabo en dos etapas que se describen a continuación.

Primera etapa. En esta etapa (Fig. 1) mediante las direcciones de las páginas principales de cuatro periódicos nacionales coleccionamos 1540 MB en textos en formato original y mediante una serie de pasos que se listan a continuación obtuvimos 1092 MB de texto plano. Todos estos pasos fueron realizados mediante programas en lenguaje PERL.

1. **Buscador de páginas.** La entrada corresponde a las direcciones de las páginas iniciales de cuatro periódicos mexicanos. A partir de estas cuatro direcciones se buscan las direcciones correspondientes a cada día de los años considerados y en forma recursiva se obtienen las páginas incluidas.

2. **Filtro de textos.** Se eliminan los anuncios, los pies de fotos, porciones de código de lenguajes de programación, etc. Se normalizaron algunos signos de puntuación, por ejemplo: comillas (´¨'|"|¨`).

3. **Eliminación de marcas HTML.** La razón para eliminar estas marcas es que su uso no es consistente aún dentro del mismo periódico. Las marcas HTML tampoco aportan información relevante del texto para la construcción del corpus (como palabras extranjeras, marcas semánticas, etc.).

4. **Asignación de estructura.** Se marcan títulos, subtítulos, cuerpo del texto, párrafos y oraciones. Los párrafos se respetaron tal y como aparecen en la versión electrónica. Las oraciones se separaron mediante heurísticas del uso: de punto, signos de admiración e interrogación, números, y abreviaturas comunes.
   Un ejemplo de estructura es el siguiente:

```
<ARTICULO>
2001/dic01/011229/004n1pol
<TITULO>
```

```
PRD y PT votaron en contra, mientras que el Verde Ecologist
se abstuvo.
<SUBTITULO>
Priistas y panistas alcanzaron acuerdo sobre ISR
<AUTOR>
ROBERTO GARDUÑO Y CIRO PEREZ SILVA
<PARRAFO>
Diputados de PAN y PRI aprobaron esta madrugada, con el vot
en contra de PRD, PT y la abstención del PVEM, modificacion
a la Ley del Impuesto sobre la Renta (ISR), destrabando así
uno de los temas fundamentales de la miscelánea fiscal para
allegar mayores recursos al gobierno foxista. El coordinador
de los diputados perredistas, Martí Batres, declaró que el
articulado -con 286 páginas y 86 transitorios- llegó casi
cinco horas después de iniciado el debate, por lo que era
posible analizarlo, y en él se introdujeron elementos desco-
nocidos para los legisladores, que no fueron consensuados
```

Ejemplos de heurísticas para separar oraciones son las siguientes:

- Si se presenta el patrón H. K dónde H es número y K comienza mayúscula:
  Separar en dos oraciones con K iniciando la segunda oración si antes de encuentran varias palabras y la precedente no está en la lista ELEMENTOS *ciso, apartado, capítulo,* etc.)

- Si se presenta el patrón H. K dónde H es una palabra de la ABREV (*sra, Sra, sr, Sr, Dr,* etc.) mantener el grupo unido.

5. **Palabras totales.** Se obtuvo adicionalmente el total de las palabras diferentes contradas en el corpus.
   Las estadísticas de los dos últimos pasos se presentan en la siguiente tabla:

PERIÓDICO	# 1	#2	#3	#4
Número de oraciones	2,927,723	1,328,157	208,298	1,696,358
Número de palabras	87,597,168	38,387,767	5,652,358	45,702,200

Segunda etapa. En esta etapa (Fig. 2) se obtienen las palabras correctas de acuerdc los siguientes pasos.

1. **Separación de las palabras en correctas y erróneas.** Mediante el empleo de corrección ortográfica de un procesador de palabras y la búsqueda en la lista de labras de dos diccionarios para el español, obtuvimos automáticamente las palabr marcadas como "correctas", es decir, reconocidas en dichos recursos. Inicialmen de 751360 palabras diferentes, 60% estaban marcadas como incorrectas.

2. **La corrección ortográfica** usando el procesador de palabras Word se utilizó diante una macro en Visual Basic que marca automáticamente las palabras inc rrectas de la lista completa de palabras obtenidas en el último paso de la etapa ant rior. Los diccionarios se convirtieron de sus archivos electrónicos a un forma texto que considera la palabra y su categoría gramatical. Un programa en PER

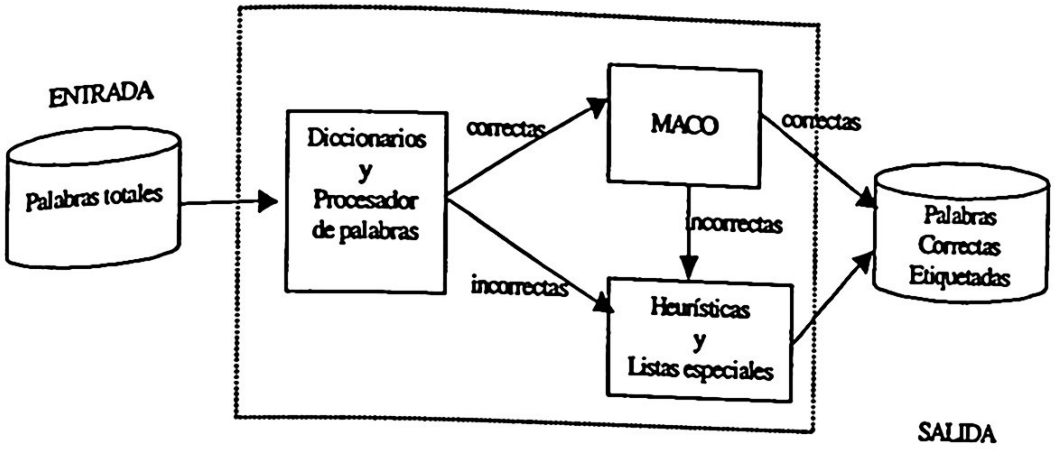

Fig. 2. Segunda etapa de la construcción del corpus

que utiliza esos archivos marca au las palabras correctas y erróneas conforme a esos recursos.

Entre las palabras inicialmente incorrectas encontramos los siguientes casos:

- *Algunas palabras mexicanas específicas*
  Una revisión manual no exhaustiva nos permitió identificar palabras como *ámpula* que no aparecen en DRAE[3], ni en el Diccionario de María Moliner, pero que si aparece en el DEUM[4].
- *Palabras de origen indígena*
  Existen palabras de origen náhuatl, maya, otomí, etc. Por ejemplo: *zotehuela, xochimilca, xochiteca, xoconostle*, etc. Para reconocer estas palabras obtuvimos de Internet y de manuales especializados una lista de palabras de origen indígena para comparar las palabras obtenidas de los textos electrónicos de periódicos. De esta forma se obtuvo un nuevo grupo de palabras correctas.
- *Palabras extranjeras*
  Los textos contienen palabras extranjeras, principalmente del inglés. Usamos la herramienta del procesador de palabras para detectar palabras "correctas" inglesas y francesas, también mediante una macro en Visual Basic.
- *Formas de palabras no contenidas en los recursos empleados*
  Mediante reglas lingüísticas se obtuvo un nuevo grupo de palabras correctas, reconociendo los sufijos de diminutivos, plurales y otras formas regulares.
  Un ejemplo de estas reglas es:
  Si una palabra terminada en "itos" (*abrigaditos*) se encuentra en su forma adjetivo o sustantivo masculino (*abrigado*) en el diccionario, se marca como correcta y diminutivo.
- *Palabras compuestas mediante guiones*
  Los grupos de palabras conectados mediante guiones son muy comunes en el inglés. En cambio, en los manuales de estilo del español se indica que su uso es

[3] Diccionario del español de la Real Academia Española. Espasa, Calpe, 21 ed. 1995
[4] Diccionario del Español Usual en México. Ed. Colegio de México. México, 1996.

muy restringido. Sin embargo, los textos de periódicos muestran el uso de guiones para diversos propósitos:

a) Para enfatizar un grupo de palabras. Ejemplo: *única-y-mejor-ruta*

b) Para evocar consignas. Ejemplo: *sí-se-puede, duro-duro-duro*

c) Para indicar la pronunciación lenta de una palabra. Ejemplo: *é-x-i-t-o, zoo-l gi-co*

d) Para expresar varios atributos diferentes. Ejemplo: *étnico-nacionalista-soci*

e) Para mostrar los vínculos entre corporaciones o nombres reales. Ejempl *ABB-Alsthom, ACDH-ONU, ADM-Dreyfus-Novartis-Maseca*

f) Para indicar una ruta. Ejemplo: *Durango-Mazatlán*

g) Para indicar competencias deportivas (fútbol, básquetbol, etc.). Ejempl *Atlas-Tecos, Alemania- Francia*

h) Para algunas palabras de origen inglés. Ejemplo: *e-mail, e-business, zig-z* etc.

También se encontraron algunos errores introducidos por la copia de los textos ginales en su forma impresa, que contienen el uso principal del guión en el españ para separar sílabas, por ejemplo *de-rechos*

Todos los casos anteriores se procesaron primero como compuesto de palabras, decir, separándolas y revisando que cada una fuera una palabra "correcta". Al esta suposición se procesaron como partes de una sola palabra, es decir, uniéndola revisando que fuera una sola palabra correcta. En general, una marca de sustantivo adjetivo puede asignarse. Esta clasificación se utilizará en el futuro para una asigr ción semántica a cada componente.

Para obtener una mayor cantidad de textos correctos se requerirá de alguna clase corrección de errores puesto que permanecen más del 50% de palabras erróneas: gráficos, ortográficos, etc. La cantidad de errores en las palabras de origen indíger también es alta ya que la mayoría de los hablantes del español desconocen su ortog fía correcta.

Hasta estos resultados no se ha llevado a cabo una evaluación del sistema. evaluación de diagnóstico, un perfil del funcionamiento del sistema respecto a un pacio de posibles entradas, implicaría la construcción de una prueba representati grande y confiable. Esta evaluación requiere a su vez de una inversión significante ra crear: conjuntos de pruebas, procedimientos bien documentados y programas; más de implementar y depurar esos procedimientos.

4 Etiquetado morfológico

Para el etiquetado morfológico consideramos las etiquetas del corpus LEXESP[5] nivel léxico, que ha sido empleado para investigaciones de nuestro grupo de traba

[5] El corpus LEXESP nos fue proporcionado amablemente por Horacio Rodríguez de la versidad Politécnica de Cataluña, en Barcelona, España.

Utilizamos las 275 etiquetas que se emplean en él. Esta cantidad de etiquetas se debe principalmente a la concordancia en género, número y persona que existe en el español.

El corpus LEXESP tiene las categorías PAROLE[6] [11]. La clasificación de categorías gramaticales en PAROLE la presentamos a continuación, donde sólo se detallan las claves completas para los determinantes (Tabla 1), indicando los rasgos considerados.

Adjetivo (A). Ejemplo: *frágiles* <AQ0CP00>
Adverbio (R). Ejemplo: no <RG000>
Artículo (T). Ejemplo: *la* <TDFS0>
Determinante (D), véase ilustración 1. Ejemplo: *tal* <DD0CS00>
Sustantivo (N). Ejemplo: *señora* <NCFS000>
Verbo (V). Ejemplo: *acabó* <VMIS3S0>
Pronombre (P). Ejemplo: *ella* <PP3FS000>
Conjunciones (C). Ejemplo: *y* <CC00>
Numerales (M). Ejemplo: *cinco* <MCCP00>
Preposiciones (SPS00). Ejemplo: *a* <SPS00>
Números (Z). Ejemplo: *5000* <Z>
Interjecciones (I). Ejemplo: *oh* <I>
Abreviaturas (Y). Ejemplo: *etc.* <Y>
Puntuación (F). Todos los signos de puntuación (.,:;-¡!'¿?"%). Ejemplo: "." <Fp>
Residuales (X). Las palabras que no encajan en las categorías previas. Ejemplo: *sine* <X>

Tabla 1. Descripción del marcado morfológico para los determinantes

Tipo		Persona	Género		Número		Caso	Poseedor
Valor	Clave		Valor	Clave	Valor	Clave		
Demostrativo	D	1	Femenino	F	singular	S	0	0
Posesivo	P	2	Masculino	M	Plural	P		
Interrogativo	T	3	Común	C	Invariable	N		
Exclamativo	E							
Indefinido	I							

Un ejemplo de etiquetas en el corpus, es el siguiente, para la palabra *bajo* que puede ser tanto una forma verbal, como preposición, adverbio, sustantivo o adjetivo:

bajar<VMIP1S0> bajo<SPS00> bajo<RG000> bajo<NCMS000> bajo<AQ0MS00>.

El valor común de género se emplea tanto para femenino como para masculino, por ejemplo: *alegre*. El valor invariable en número se emplea tanto en singular como en plural, por ejemplo, el pronombre *se*.

[6] www.ub.es/gilcub/castellano/proyectos/europeos/parole.html

El etiquetado se realizó automáticamente mediante el programa MACO [6] sarrollado por el grupo de Procesamiento de lenguaje natural de la Sección de Inte gencia artificial, Departamento de software de la Universidad Politécnica de Catal en colaboración con el Laboratorio de lingüística computacional de la Universidad Barcelona.

Esta última etapa considera el marcado de palabras especiales, entre las cuales tán las preposiciones compuestas y los nombres propios.

1. Preposiciones compuestas

Existen muchas preposiciones compuestas en el español, además de las preposicio simples. Grupos de palabras como *al cabo de*, *a fin de*, *con respecto a*, requieren manejo conjunto. Conforme a la lista de preposiciones compuestas de [12] se etiq taron automáticamente, mediante un programa en PERL, las preposiciones compu tas en el corpus de textos electrónicos de periódicos.

2. Nombres propios

Se encontraron 168,333 compuestos diferentes de palabras con mayúsculas (inicia zando las palabras o de forma generalizada). Estos compuestos se repiten tenien 1,804,959 ocurrencias en los textos. Los grupos que se encontraron corresponden

- *Acrónimos.* Por ejemplo: PRD, PT, ONU.
- *Frases de nombres.* Por ejemplo: *Convergencia por la Democracia, Ley del puesto sobre la Renta, Luz y Fuerza del Centro.*
- *Nombres.* Por ejemplo: *San Lázaro, Benito Juárez.*

En esta asignación se presentan los siguientes problemas para los cuales se desarrollando heurísticas:

- *Coordinación.* Por ejemplo: *Luz y Fuerza del Centro y Lotería Nacional* (Luz Fuerza del Centro, Lotería Nacional), *Margarita Diéguez y Armas y Carlos gilio Ferrer* (Margarita Diéguez y Armas, Carlos Virgilio Ferrer.)
- *Nombres compuestos donde solamente la primera palabra tiene mayúscula.* ejemplo: *cantando México lindo, Cucurrucucú, La malagueña y Amorcito coraz como La boa, El mudo, Luces de Nueva York, Perfume de mujer.*
- *Varios nombres compuestos ligados.* Por ejemplo: *Hospital de Traumatología Lomas Verdes* (Hospital de Traumatología, Lomas Verdes), *Comisión de Prot ción al Empleo y Previsión Social de la Asamblea Legislativa del Distrito Fede* (Comisión de Protección al Empleo y Previsión Social, Asamblea Legislativa Distrito Federal).

Las heurísticas consideran los casos de uso de preposición como "por la" que une sistemáticamente grupos de nombre de persona como en : *Juan Ramón Fuente por la Federación de Colegios de Personal Académico* (Juan Ramón Fuente, Federación de Colegios de Personal Académico) pero si une grupos de bras comunes como en *Alianza por la Ciudad de México.*

5 Conclusiones

Indicamos la utilidad de las colecciones de textos con etiquetado en diferentes niveles, es decir, su empleo para diferentes trabajos en investigación teórica y en diversas aplicaciones de lenguaje natural.

Presentamos el desarrollo de un recurso para el procesamiento lingüístico de textos en español: un corpus de textos de decenas de millones de palabras con etiquetado morfológico. Presentamos el proceso que requiere la colección de textos obtenida para la construcción del corpus. Detallamos el método empleado para la compilación del corpus y para el anotado léxico.

La principal ventaja de nuestro método de obtención es que la mayor parte del trabajo se ha realizado en forma automática, lo que reduce tiempos y costos en la compilación.

Referencias

[1] Banko, M. and Brill, E. Mitigating the Paucity of Data Problem. In Proceedings of the Conference on Human Language Technology, San Diego, Ca. 2001

[2] Berthouzoz, C. and Merlo, P. *Statistical ambiguity resolution for principle-based parsing.* In Proceedings of the Recent Advances in Natural Language Processing. Pag. 179-186, 1997

[3] Biber, D. Using Register. *Diversified Corpora for general Language Studies.* Computational Linguistics 19 (2) pp. 219—241, 1993.

[4] Bolshakov, I. A., A. Gelbukh. On Detection of Malapropisms by Multistage Collocation Testing. Proc. of NLDB-2003, 8th International Workshop on Applications of Natural Language to Information Systems. Lecture Notes in Computer Science, Springer-Verlag, 2003, to appear.

[5] Calzolari, N. Corazzari, O. & Zampolli, A. *Lexical-Semantic tagging of an Italian Corpus.* Second Conference on Intelligent text processing and Computational linguistics. CICLing-2001.

[6] Carmona, J., S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé & J. Turmo. *An Environment for Morphosyntactic Processing of Unrestricted Spanish Text.* First International Conference on Language Resources and Evaluation (LREC'98). Granada, Spain, 1998.

[7] Francis, W. N. and Henry Ku•era. *Frequency Análisis of English Usage: Lexicon and Grammar.* Houghton Mifflin. 1982

[8] Gelbukh, A., G. Sidorov, A. Guzman-Arenas. *Use of a weighted topic hierarchy for document classification.* In Václav Matoušek *et al.* (Eds.). Text, Speech and Dialogue. Proc. TSD-99. Lecture Notes in Artificial Intelligence, N 1692, Springer-Verlag, 1999, pp. 130–135.

[9] Gelbukh, A., G. Sidorov, *et al. Compilation of a Spanish representative corpus.* In Intelligent Text Processing and Computational Linguistic. Proc. CICLing-2002. Lecture Notes in Computer Science, N 2276, Springer-Verlag, 2002, pp. 285–288.

[10] Marcus, M., Santorini, B. and Marcinkiewicz, M. *Building a large annotated corpus of English The Penn Treebank.* Computational Linguistics 19, 2, 1993.

[11] Martí M.A., Rodríguez H., Serrano J. *Declaración de categorías morfosintácticas.* Documento interno ITEM n-o2. Universidad Politécnica de Cataluña, España, UB, 1997.

[12] Nañez Fernández, E. *Diccionario de construcciones sintácticas del español*. Preposiciones. Ed. de la Universidad Autónoma de Madrid, España 1995.

[13] Nigam, K. et al. Using Maximum Entropy for Text Classication. In Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61-67. 1999

[14] Pedersen, T. *An Ensemble Approach to Corpus-based Word Sense Disambiguation*. Conference on Intelligent text processing and Computational linguistics. CICLing-2000

[15] Ratnaparkhi, A. *Statistical Models for Unsupervised Prepositional Phrase Attachment*. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. Montreal, Quebec, Canada, 1998 http://xxx.lanl.gov/ps/cmp-lg/9807011

[16] Roland, D. and D. Jurafsky. *How Verb Subcategorization Frequencies are Effected Corpus Choice*. In Proceedings International Conference COLING-ACL'98. August 10-14 Quebec, Canada, pp. 1122-1128, 1998.

[17] Weischedel, Ralph; Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeoe Palmucci. *Coping with ambiguity and unknown words through probabilistic models*. Computational Linguistics, 19(2): 359-382, 1993.

[18] Yeh, Alexander S., M. B. Vilain. *Some Properties of Preposition and Subordinate Conjunction Attachments*. In Proceedings International Conference COLING-ACL'98. August 10-14 Quebec, Canada, pp. 1436-1442, 1998. http://xxx.lanl.gov/ps/cmp-lg/9808007.