Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 259-267 recibido 03-05-2011 aceptado 23-05-2011

ModeS TimeBank: A Modern Spanish TimeBank Corpus

ModeS TimeBank: un corpus TimeBank del español moderno

Marta Guerrero Nieto
Grupo Mercator-UPM
Cª de Valencia, km.7
28031 Madrid
mguerrero@topografia.upm.es

Roser Saurí
Barcelona Media
Av. Diagonal, 177
08018 Barcelona
roser.sauri@barcelonamedia.org

Miguel Ángel Bernabé
Grupo Mercator-UPM
Cª de Valencia, km.7
28031 Madrid
ma.bernabe@upm.es

Resumen: Con el objetivo de representar y analizar grandes cantidades de fuentes históricas textuales en un Sistema de Información Geográfica (SIG), se ha creado ModeS TimeBank. ModeS TimeBank es un corpus del español moderno (s. XVIII) anotado con información semántica temporal, eventiva y espacial, donde destaca el uso de los lenguajes de marcado TimeML y SpatialML. El corpus es además relevante no sólo por su datación e idioma sino por su dominio ya que está enmarcado en la temática de las redes de cooperación. El presente artículo pretende describir cómo se ha creado el corpus y qué criterios se han tenido en cuenta en su creación, además de señalar el alcance y las aplicaciones de ModeS TimeBank.

Palabras clave: TimeML, anotación semántica, información temporal y espacial, corpus diacrónico del español.

Abstract: ModeS TimeBank has been created to assist in the representation and analysis of large amounts of historical text into Geographic Information Systems (GIS). ModeS TimeBank is a corpus of Modern Spanish (18th century) annotated with temporal, eventive and spatial information, by means of the TimeML and SpatialML markup languages. The corpus is relevant not only for its time coverage and language, but also for its domain, which is focused on cooperation networks. This article aims to describe the process of corpus compilation, the criteria considered for its creation, as well as the scope and applications of ModeS TimeBank.

Keywords: TimeML, semantic annotation, temporal and spatial information, Spanish diachronic corpora.

1. Introducción

Desde sus inicios, la creación de corpus históricos ha tenido como objetivo principal el abastecimiento de datos para el estudio sistemático de las variedades históricas de la lengua. Pero los textos de otros periodos históricos no son sólo de interés en los ámbitos filológicos y lingüísticos. En tanto que hablan de tiempos anteriores y son el medio por excelencia de trasmisión de información, también son objeto de análisis históricos en el campo de las Ciencias Sociales.

Recientemente, esta área ha empezado a incorporar sistemas computacionales de tratamiento, razonamiento y visualización de la información. Desde esta perspectiva, adquiere especial relevancia disponer de niveles de anotación semántica de los textos que expliciten elementos informativos de interés para el estudio del contenido textual, más allá de los niveles lingüísticos más básicos (como tokenización o POS), frecuentemente presentes en corpus de tipo histórico.

Dentro de los estudios históricos, se ha destacado el uso de los Sistemas de Información Geográfica temporales (SIG) para el análisis espacio-temporal de la Historia, ya que permite utilizar grandes cantidades de datos y ubicar la información sobre eventos históricos en las coordenadas de espacio y tiempo (Owens, 2008). En relación con esto, es de particular interés la identificación de información de carácter eventiva, temporal y espacial presente en texto histórico. En este sentido, las áreas de Lingüística de Corpus y Procesamiento del Lenguaje Natural proporcionan recursos valiosos a estos estudios, ya que permiten el abastecimiento automatizado de

ISSN 1135-5948

información histórica presente en documentos textuales.

El proyecto DynCoopNet, dentro del cual se ha desarrollado el trabajo presentado a continuación, es un exponente en la utilización de los SIG dentro de la investigación histórica. Enmarcado en el estudio de la cooperación entre las redes sociales comerciales establecidas durante los siglos XIV-XIX, se propone analizar los mecanismos de cooperación de las comunidades mercantiles y las redes comerciales que sustentaban la economía mundial de la Primera Edad Global. Para esto no sólo se ha considerado información de mapas, gráficos y bases de datos, sino que se ha tenido en cuenta las distintas fuentes y datos disponibles para el estudio de la historia, como por ejemplo, documentos textuales escritos en lenguaje natural. Con esta motivación se compiló el corpus ModeS TimeBank, cuyo objetivo es contribuir a la representación y análisis de grandes cantidades de fuentes históricas textuales, donde se presta especial relevancia a la información temporal y espacial. ModeS TimeBank, pues constituye un buen ejemplo de corpus anotado donde los niveles de anotación semántica que se han tenido en cuenta han sido el eventivo, el temporal y el espacial

En las secciones siguientes se presenta una descripción del corpus así como de los resultados obtenidos. En concreto, en la segunda sección se señalan los trabajos más significativos sobre creación de corpus diacrónicos y las herramientas de PLN creadas, en la tercera sección se presenta el corpus ModeS TimeBank, donde se detalla el proceso de compilación y se hace una revisión de los lenguajes de marcado, para después describir los criterios específicos que se han tenido en cuenta para la elaboración del corpus, después se enumeran los resultados de la anotación del corpus, los cuales se comparan con los reflejados en el corpus anotado TimeBank y Spanish TimeBank.

2. Corpus diacrónicos: trabajos previos

La mayoría de los corpus diacrónicos existentes tienen una orientación filológica o lingüística, y están destinados a estudios evolutivos del lenguaje, estudios sincrónicos o consultas lingüísticas.

Los corpus diacrónicos del español más representativos y que están disponibles en internet son CORDE[1] (Corpus Diacrónico del español) (1994), Corpus del español[2] (2002) y CHEM[3] (Corpus Histórico del Español de México) (2005). El volumen de datos de los dos primeros corpus es de 125 millones de palabras en el primer caso (textos fechados hasta 1975) y de 100 millones en el segundo (donde están incluidos textos del siglo XX). El tercero consta de 320 textos transcritos. Sólo los dos últimos poseen anotación básica, aunque en el primero se está ampliando el sistema de consulta con esta información gramatical.

En otras lenguas también se han desarrollado este tipo de recursos. Para el inglés, destaca el Penn Parsed Corpora of Historical English[4]. Y para lenguas peninsulares otras que el español: el Corpus do portugués[5], el Tesouro Automatizado da Lingua Galega[6], el Corpus Xeneral de la Llingua Asturiana[7], el corpus para el mendeko Euskararen Corpus Estatistikoa [8], y el Corpus Informatitzat del Català Antic[9] .

La mayoría de los corpus diacrónicos existentes están anotados manualmente con niveles de anotación básica que típicamente incluye lematización y etiquetación morfosintáctica (POS), pero poco a poco se van desarrollado herramientas de anotación morfosintáctica automáticas. Por ejemplo, Sánchez-Marco y sus colegas (Sánchez-Marco et al., 2010) están creando una herramienta para el etiquetado de corpus diacrónicos en español basándose en la adaptación de FreeLing[10]. Para la anotación de niveles semánticos, es posible adaptar herramientas ya existentes para textos actuales de modo equivalente a la adaptación de estas para dominios

[1] http://corpus.rae.es/cordenet.html
[2] http://www.corpusdelespanol.org/
[3] http://www.iling.unam.mx/chem/
[4] http://www.ling.upenn.edu/hist-corpora/
[5] http://www.corpusdoportugues.org/
[6] http://www.ti.usc.es/TILG/
[7] http://di098.edv.uniovi.es/corpus/busqueda.html
[8] http://www.euskaracorpusa.net/XXmendea/Konts_arrunta_fr.html
[9] http://webs2002.uab.es/sfi/cica/
[10] http://nlp.lsi.upc.edu/freeling

específicos. Sin embargo, hasta donde sabemos no existen corpus históricos con la información necesaria para llevar a cabo tal proceso.

3. ModeS TimeBank

En este trabajo presentamos ModeS TimeBank[11], un corpus de español moderno con anotación temporal y eventiva (siguiendo el lenguaje de marcado TimeML) y espacial (siguiendo el lenguaje de marcado SpatialML). ModeS TimeBank, contiene 102 documentos fechados entre 1768 y 1769, con un total de 25611 palabras. La elección de las fuentes textuales del corpus ha estado supeditada al proyecto DynCoopNet[12] dentro del que se incluye la presente investigación, tanto en lo referido a: (a) la temática, pues debía aportar información novedosa relativa a las redes de cooperación así como a los flujos de información o las rutas de navegación; (b) la datación, pues debía estar ceñida a los periodos de estudio (s. XIV-XVIII); y (c) a los recursos utilizados, ya que los Sistemas de Información Geográfica (SIG) se convirtieron en objeto de estudio del proyecto, destacando la importancia de la información espacial y temporal para el análisis de la historia (Owens, 2008) y apuntándose su uso para el estudio y visualización de textos históricos en lenguaje natural (Guerrero Nieto et al., 2010). Así, a pesar de la existencia de otros corpus históricos para el español, las exigencias temáticas y de dominio del proyecto requirieron la creación de este corpus.

ModeS TimeBank se ha anotado con dos niveles de marcado: un primer nivel básico que incluye lematización y etiquetación morfosintáctica; y una segunda capa de información semántica. Se comentan en detalle a continuación.

3.1 Procesado del corpus

La anotación de corpus con al menos el nivel de información morfosintáctica abre las puertas a la sistematización en estudios históricos y gramaticales de la lengua, además de contribuir a la creación de herramientas de PLN. Debido a la carencia de recursos lingüísticos para el etiquetado automático de corpus lingüísticos no actuales, para el marcado morfosintáctico de ModeS TimeBank se ha optado por adaptar recursos ya existentes. La elección del TreeTagger (Schmid, 1994) se ha debido a la facilidad de incorporación de un lexicón dentro del analizador.

Para la adaptación de la herramienta TreeTagger al español moderno, el primero de los pasos fue la detección del número de error que daba el etiquetador TreeTagger. Una vez detectado y analizado el error, se creó un lexicón que sirviese para enriquecer la herramienta. Al procesar el corpus con la herramienta TreeTagger el porcentaje de desconocimiento (*Unkown*) era del 14,33% debido a las características lingüísticas del corpus. Se detectaron dos causas: cobertura léxica y variación ortográfica. En cuanto a la cobertura léxica se hallan dos tipos: palabras en desuso (por ejemplo, *grandor*) y carencia de un dominio específico (ej. *proa, sotavento,* etc.). En la variación ortográfica tenemos: carencia de acentuación (*navegacion, dia, anochecio*, etc.) y falta de normalización ortográfica (*varlovento, orizonte, haver*, etc.)
Para llevar a cabo esta tarea, se creó una metodología del error (Gráfico 1), que consiste en identificar del texto de salida (*Texto_Out*) la información conocida (*Known*) y desconocida (*Unknown*). La herramienta señala automáticamente la información desconocida, de ahí que sea necesario realizar otro proceso para comprobar cuál es la información correcta (*Right*) e incorrecta (*Wrong*).

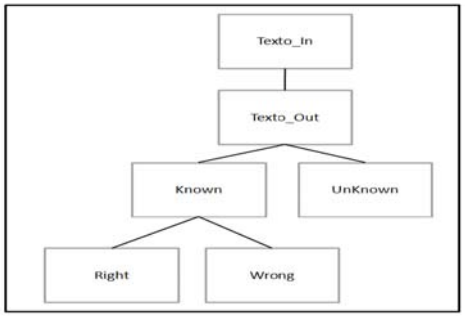

Gráfico 1: Metodología del error

11 Modes TimeBank se puede obtener en <http://redgeomatica.rediris.es/modestimebank/modestimebank_10.rar >
12 http://dyncoopnet-pt.org

En la tabla 1 se muestran los resultados de tres documentos del corpus donde aparecen los porcentajes que se establecen a partir de la metodología del error.

El porcentaje medio de error de la lematización y POS del corpus ha sido del 20,95%. El porcentaje de *Unkown* es del 14,33% y el de *Wrong* del 6,62%. En términos de acierto (*accuracy*) en el uso del TreeTagger en Modes TB es de 79,05%.

Documento	Tokens	Unkown	% Unkown	KnownWrong	% KnownWrong	Unkown + Known_Wrong
1	622	111	17,78%	35	5,62%	23,47%
2	759	134	17,64%	45	5,92%	23,58%
3	290	43	14,82%	12	4,14%	18,96%

Tabla 1: Porcentajes de error del procesamiento

A partir de la metodología de corrección, se ha creado un lexicón nuevo con la finalidad de enriquecer el TreeTagger y así poder incorporar información nueva y etiquetar con información morfosintáctica el corpus. El lexicón nuevo está compuesto de la unión de ambos lexicones y está formado por 822 palabras, con sus correspondientes lemas y POS.

El resultado de procesar el corpus con este lexicón nuevo se ha validado manualmente. La validación ha sido efectuada aleatoriamente sobre un total de 20 textos, es decir, un 20% del total del corpus. Se ha obtenido un porcentaje de acierto del 99,11%, sin embargo hay que tener en cuenta que los documentos que han servido para la evaluación están dentro del corpus utilizado.

3.2 Niveles de anotación semántica

Debido a su encaje dentro de un proyecto de orientación histórica, ModeS TimeBank se ha anotado manualmente con información relativa a los eventos narrados, así como sus propiedades temporales y espaciales. Para estos niveles se han utilizado los lenguajes de marcado TimeML y SpatialML, presentados en las siguientes secciones. En este artículo, se hará mayor énfasis en el componente temporal.

3.2.1 Anotación eventiva y temporal

El corpus se ha anotado con información temporal y eventiva a partir del lenguaje de especificación TimeML (Pustejovsky et al., 2005). Las propiedades más destacadas de este lenguaje son: normalización de las expresiones temporales, anotación temporal de eventos y ordenación de los eventos a través del anclaje temporal de estos. Por el momento se tiene constancia de seis corpus basados en TimeML en distintas lenguas como por ejemplo inglés, francés, español, catalán, coreano, italiano y chino, algunos de los cuales han sido utilizados en TempEval-2 (Verhagen et al., 2010). Sin embargo, no hay referencias de corpus creados con este lenguaje de anotación en variedades lingüísticas no actuales.

TimeML tiene tres etiquetas básicas: TIMEX3, EVENT, SIGNAL, y tres etiquetas relacionales: TLINK, ALINK y SLINK.

- TIMEX3 se usa para anotar expresiones de tiempo: *17 de enero de 2011, ayer, las 12 de la mañana, el próximo año.*
- EVENT se usa para anotar eventos mencionados en el texto: *ocurrir, creer, estudiar, empezar.*
- SIGNAL se usa para anotar marcadores de relaciones temporales: *antes, después, durante.*
- TLINK se usa para anotar relaciones temporales entre eventos (ej. *irá*) y expresiones temporales (ej. *el 6 de septiembre*): *Luisa irá a Huelva el 6 de septiembre.*
- ALINK se usa para anotar la relación entre un predicado aspectual (ej. *empezará*) y el evento al que este califica (ej. *presentación*): *María empezará su presentación en seguida.*
- SLINK se usa para anotar relaciones de evidencialidad o modalidad entre dos predicados: *Juan dijo que iría a Huelva en septiembre*.

Para la tarea de anotación, se ha usado la herramienta BAT (Brandeis Annotation Tool)[13]. A continuación se señalan los

[13] http://www.timeml.org/site/bat/

criterios lingüísticos que se han seguido de acuerdo a las guías de anotación (Saurí et al., 2009-2010), ejemplificados a través del corpus ModeS TimeBank y ampliados con criterios específicos.

Expresiones temporales del corpus. Las expresiones temporales en TimeML pueden clasificarse en DATE (septiembre de 2011), TIME (a las 4 de la tarde), DURATION (los próximos días) y SET (todos los días). En concreto, en la guía de anotación temporal para la identificación de expresiones temporales, son considerados expresión temporal los nombres propios y sintagmas nominales con significado temporal: (1a) *Viernes Santo*; (1b) *la mañana;* (1c) *algunas horas.* Así mismo son anotadas de manera separada las expresiones que puedan dividirse con respecto a su granularidad (mayor o menos de 24 horas) o su naturaleza (instantes o duraciones): (2a) *ayer tarde a las 4*; (2b) e*l 31 de julio a las 11 de la mañana*; (2c) *hoy todo el dia.* En el corpus ModeS TimeBank es muy común encontrar dos expresiones temporales que forman un intervalo cerrado. Sin embargo cada una de las expresiones quedará anotada de manera separada: (3a) *dia 17 de dicho hasta el 18.* Los sintagmas que integran estas estructuras contienen con frecuencia distintos constituyentes o con distinto orden que los del español actual (3a).

Las expresiones referenciales que necesitan de una fecha de anclaje para poder establecer un significado, son de dos tipos: en relación con el tiempo del discurso (4a) *estos dos días*); o con respecto a otra expresión del texto (4b) *el 27 del mismo (mes)*; (4c) *a esta de 12 (hora)*; (4d) *a esta hora de las 6.* Al tratarse de un corpus del español no actual, es frecuente utilizar expresiones como las anteriores (4b) y (4c) tanto en lo referente a las estructuras como a las palabras que las componen, así como a la omisión de las unidades temporales. De esta manera, construcciones que en español actual se construyen con el sustantivo o nombre propio en español moderno se utiliza un demostrativo o adjetivo sustantivado, es decir, podría esquematizarse de esta manera: preposición + adjetivo / demostrativo+ unidad de tiempo > preposición + sustantivación (4b). Las construcciones verbales compuestas por los verbos *haber* o *hacer* seguido un sintagma nominal temporal son consideradas expresiones temporales (*hace tres días*). Habitualmente expresan el tiempo en el que se produce la acción o a partir del cual se produce la acción en relación a otro tiempo o en relación al tiempo del discurso. De ahí que necesiten otro timex como anclaje temporal: (5a) *havia tres días*. Podría resumirse como: *Havia/ Habia/ hace / hacia* + SN temporal, donde se incorpora el uso del verbo haber para este tipo de enunciados.

Las construcciones de infinitivo que conforman una subordinada temporal han sido marcadas como timex: (6a) *al romper el alva;* (6b) *al clariar el día* (Al + infinitivo+ SN temporal).

También son expresiones temporales aquellos verbos que hacen referencia a una división del tiempo, quedando la información temporal embebida en el verbo: (7b) *amanece el dia*; (7b) anocheció. Estos eventos suelen funcionar como Complemento Circunstancial Temporal (CCT), como segundo miembro de la comparación o como núcleo de una construcción impersonal o intransitiva.

En la guía TimeML se matiza que expresiones temporales como en cualquier momento, al instante, etc., que hacen referencia a la rapidez en la que se va a producir el acontecimiento y no al tiempo en el que se produce no deben ser anotadas. Este uso también se observa en ModeS TimeBank: (8b) *por instantes se esperava llegase de la Habana.*

Eventos del corpus. La mayoría de los eventos vienen expresados por verbos finitos o no finitos como por ejemplo: (1a) *saltó el viento al Norte*; (1b) *a las nueve avisté una vela al OSO*; (1c) *por no haver viento alguno*; (1d) *vendo para Montevideo.* Pero también se pueden expresar con adjetivos, sustantivos, preposiciones, etc…

Verbos. En lo que respecta a los que están expresados por verbos, cuando los eventos constan de un verbo auxiliar y un verbo léxico, es el verbo léxico el que queda anotado: (2a) *el motivo de haver variado el rumbo.* Si el evento viene expresado a través de una construcción perifrástica se

anota tanto el verbo finito como el no finito como eventos independientes, como por ejemplo: (3a) *bolvimos a ver;* Las perífrasis durativas son anotadas dejando marcado como evento sólo el verbo en gerundio: (4a) *vine navegando toda la noche en la mano*. En el corpus Modes TimeBank es muy común encontrar las construcciones *ir en vuelta* ("*iva enbuelta*" o "*iva en buelta*") o *dar fondo*. En la primera se observa el significado de *estar de regreso, volver. S*e anotará como un sólo evento donde quedará marcada la parte léxica como en la construcción (2a): (5a) *avistamos un navio grande que iva enbuelta del Oeste*. La segunda con el significado de fondear, quedará anotado como dos eventos independientes, al igual que las construcciones perifrásticas de (3a): (6a) *el 27 del mismo dio fondo 23 en este mismo puerto el paquevot correo el Patagon.*
Así mismo, las construcciones que contienen verbos con poca carga léxica o desemantizados (*tener, hacer*, *coger*, etc.) y se apoyan en un complemento nominal son anotados como eventos, quedando como eventos independientes. Son frecuentes estas construcciones en ModeS TimeBank: (7a) *salto una turbonada*; (7b) *al poner el sol hizo demarcación*; (7c) *entro calma.* Se llaman "verbos de apoyo" y se caracterizan por ser verbos no copulativos con escasa entidad semántica que van seguidos de un SN, pudiendo llevar artículo o no.

Sustantivos y adjetivos. En ModeS TimeBank se encuentran también "sustantivos eventivos", es decir, sustantivos considerados eventos los cuales designan un acontecimiento o suceso que contienen un límite temporal o denotan un tiempo. La mayoría de estos sustantivos son deverbales: (8a) *fueron causa de su arrivo a este puerto*; (8b) *salida de la Coruña a las 8 de la mañana del dia 21 de diciembre de 1768.*

Son considerados eventos los adjetivos o sustantivos que tengan una función predicativa como: (9a) *se quedo calma*; (9b) b*onancible amanecio el dia.* Es frecuente en ModeS TimeBank encontrar verbos que denotan fenómenos naturales *amanecer, anochecer,* etc. seguidos de un SN temporal y con predicativos antepuestos o pospuestos (9b).

Los verbos copulativos si van acompañados de complementos predicativos son considerados como eventos independientes: (10a) *A medio dia no observe el sol por estar cubierto de nubes.* En ModeS TimeBank sobre todo se encuentran predicados con verbos meteorológicos compuestos por un verbo copulativo seguidos de un nombre o adjetivo que hace referencia a un determinado fenómeno de la naturaleza (10a).

Preposiciones. En ModeS TimeBank también se observa el siguiente uso contenido en las guías: si el Sintagma Preposicional (SP) funciona como predicativo de un verbo, éste es anotado como un evento independiente. Estos SSPP son anotados dejando sólo marcada la preposición: (11a) *no vinieron los pliegos a bordo*; (11b) *viramos el ancla a pique.*

Son considerados eventos los SSVV cuyo núcleo verbal es *amanener, anochecer, oscurecer*, etc., Pueden ir seguidos de un SN temporal como *día, noche*, etc. También construcciones temporales de *al* + infinitivo, seguidas de un SN temporal como: (12a) *al romper el alva;* (12b) *amanecio el dia con orizontes con algunas nubes.*

Al realizar el análisis de los eventos y tiempos del corpus ModeS TimeBank, surgieron ciertas dudas sobre la consideración de expresiones relacionadas con las partes del día o fenómenos naturales, tales como: *amanecer, anochecer, poner el sol, romper el alba, obscurecer el día,* etc. ya que estos podían tener varias interpretaciones o significados. De los siguientes ejemplos extraídos del corpus se pueden encontrar una doble semántica. Por un lado, indica un tiempo (el tiempo en el que se produce el evento: *vi que el día amanecía*) pero por otro, indica que es algo que ocurre, sucede, por lo que también es un evento (*vi que amanceció*). Podría decir que es un tipo de *dot-object* (Pustejovsky, 95). Ejemplo: (13a) *anocheció este día con los orizontes aguazerados por todas partes*; (13b) *bonancible amaneció el dia con algunas nubes rojas y el tiempo de bastante mal semblante.* El tratamiento de las expresiones consideradas *dot-object* ha sido

la doble etiquetación, ya que posee la semántica eventiva y temporal al mismo tiempo, lo que lo trasforma en un objeto complejo, resultado de la unión de significados simples.

3.2.2 Anotación espacial

En ModeS TimeBank también se ha marcado la información espacial con el lenguaje de marcado SpatialML (Mani et al., 2008), un lenguaje de especificación para anotar y normalizar con coordenadas los diferentes tipos de expresiones espaciales. La etiqueta más representativa es PLACE que marca expresiones espaciales tanto absolutas como relativas. SpatialML está siendo revisado y ampliado, ya que se le está incorporando semántica del movimiento espacial (Pustejovsky et al., 2010). No se tiene constancia de ningún corpus lingüístico anotado como ambos lenguajes de marcado.

En el corpus ModeS TimeBank es común encontrar nombres propios como *Montevideo* o *Sierra de Maldonado*, sintagmas nominales como *este puerto* o *la referida torre* aunque también es frecuente encontrar coordenadas explícitas como *en la latitud 4 grados y 19 minutos Norte y longitud de 359 grados y 6 minutos*.

3.3 Estadísticas de ModeS TimeBank

Este apartado analiza la relevancia de los datos obtenidos en cuanto a la densidad y distribución de las etiquetas anotadas en el corpus, y presenta la calidad de la anotación en base al acuerdo entre los anotadores (*inter-annotation agreement*). El corpus ha sido anotado por dos anotadores con formación lingüística, uno de ellos con un entrenamiento previo en TimeML.

El acuerdo entre anotadores, en lo que concierne a las tres etiquetas anotadas en ModeS TimeBank, es de K_{cohen}=0,85 en relación a la etiqueta TIMEX3; Kc_{ohen}=0,98 respecto a PLACE y K_{cohen}=0,79 respecto a EVENT. La media es de K_{cohen}=0,87. El alto resultado de PLACE se debe a que pertenece a un conjunto cerrado y bien identificado de entidades, sin embargo en cuanto a los EVENT, la ambigüedad polisémica y de estructura gramatical supone una disminución en el acuerdo obtenido. Estas discrepancias serán tenidas en cuenta en trabajos futuros.

Densidad semántica. En la tabla 2 se observa la relación del número de palabras del corpus y el número de palabras que han sido etiquetadas con información semántica, y en la tabla 3 se observa la relación de etiquetas obtenidas a partir del recuento total del corpus.

Como se ve en la tabla 2, la entidad TIMEX3 está formada por 4850 palabras, sin embargo se identifican un total de 892 TIMEX3, ya que éstas pueden estar compuestas por más de una palabra.
La entidad PLACE está compuesta por 3560 palabras en un total de 477 unidades.

ModeS TimeBank Corpus	Núm. palabras	% sobre total palabras	% sobre total Etiquetado
TIMEX3	4850	18,94%	50,15%
EVENT	1261	4,92%	13,04%
PLACE	3560	13,90%	36,81%
Total etiquetadas	9671	37,76%	100,00%
Total palabras	25611	100,00%	----

Tabla 2: Relación de palabras en ModeS TB

	Etiquetas	Porcentaje
TIMEX3	892	33,91%
EVENT	1261	47,94%
PLACE	477	18,14%
Total	2630	100,00%

Tabla 3: Relación de etiquetas en ModeS TB

Si sumamos las palabras anotadas por cualquiera de los dos lenguajes nos da un total de palabras anotadas de 9671, esto es, se han anotado un 37,76% de vocablos del corpus, con un total de 2630 etiquetas provenientes de los lenguajes TimeML y SpatialML. El número de expresiones temporales del corpus es 18,94%, el número de eventos es 4,92%, y el número de lugares anotados en el corpus es 13,90%, por lo que se observa que la detección de este tipo de información resulta de gran interés en distintos niveles de análisis.

La distribución de los diferentes tipos de expresiones de tiempo (etiqueta TIMEX3) y lugares geográficos (etiqueta PLACE) se puede observar en las tablas 4 y 5, respectivamente. En la tabla 5 se muestran

los valores de la entidad PLACE. Los valores COUNTRY (país) y CONTINENT (continente) hacen referencia a áreas políticas, RGN se utiliza para marcar áreas administrativas, WATER se utiliza para anotar lugares con características hidrográficas (ríos), MTN con características fisiográficas (montañas) y FAC hechos por el hombre o lugares artificiales (carreteras), PPLC, PPLA, PPL para lugares poblados (ciudades, capitales), LAT-LONG para latitudes y longitudes.

TIMEX3	Cantidad	Porcentaje
DATE	240	26,90%
TIME	433	48,54%
DURATION	218	24,44%
SET	1	0,11%
Total	892	100%

Tabla 4: Porcentaje de TIMEX3

PLACE	Cantidad	Porcentaje
FAC	17	3,56%
PPLC	9	1,88%
PPLA	11	2,31%
PPL	1	0,21%
RGN	32	6,71%
WATER	2	0,42%
CONTINENT	4	0,84%
COUNTRY	4	0,84%
MTN	5	1,05%
LAT_LONG	392	82,18%
Total	477	100%

Tabla 5: Porcentaje de PLACE

Se han comparado los resultados del corpus ModeS TimeBank con los de corpus TimeBank del inglés (Pustejovsky et al., 2006) y con los de Spanish TimeBank (Saurí, 2010) en cuanto al número de EVENT y TIMEX3 (Tabla 6), ya que son las etiquetas comunes en los tres corpus. Sin embargo, es necesario precisar que tanto el TimeBank como el Spanish TimeBank contienen otras etiquetas incorporadas en sus corpus.

A pesar de ello se hace evidente que la información con una temática temporal es tan numerosa que, ya sea en textos actuales o antiguos y en distintas lenguas, puede ser susceptible de distintas aplicaciones y áreas de estudio. Por otro lado, cabe mencionar, que la diferencia porcentual entre los corpus es tan significativa debido al dominio de ModeS TimeBank, ya que se refleja los acontecimientos con una ubicación minuciosa del espacio y del tiempo.

Corpus	**TimeBank**			**Spanish TimeBank**			**ModeS TimeBank**		
	Palabras	**Etiquetas**		**Palabras**	**Etiquetas**		**Palabras**	**Etiquetas**	
EVENT	-	7935	84,87%	-	12385	81,68%	1261	1261	58,56%
TIMEX3	-	1414	15,12%	-	2776	18,31%	4850	892	41,43%
Totales	61000	9349	100%	68000 (con puntuación)	15161	100%	25611	2153	100%

Tabla 6: Relación de distribución de las etiquetas temporales de los tres corpus

4. Consideraciones finales

En este trabajo se han descrito los distintos procedimientos seguidos para crear el corpus. Sin embargo, es necesario señalar que la ausencia de herramientas tanto para la etiquetación morfosintáctica como para la anotación semántica hace que la tarea sea costosa. A pesar de esto, la herramienta de anotación utilizada (BAT) ha facilitado el proceso de anotación manual debido a su flexibilidad y porque fue concebida dentro del marco de trabajo TimeML. Igualmente, el etiquetador morfosintáctico TreeTagger permite la incorporación de léxico nuevo marcado con lemas e información morfosintáctica básica, y posibilita la integración de recursos.

En Sánchez-Marco et al. (2010) se menciona que el uso de recursos computacionales de corpus diacrónicos permite estudiar la evolución de la lengua y extraer patrones de comportamiento lingüísticos comparando grandes cantidades de datos. En nuestro caso, la comparación estadística del corpus con información

temporal ha permitido constatar que, debido a la alta densidad de información eventiva y temporal, el corpus creado puede ser muy útil de cara al desarrollo de sistemas y modelos de PLN. La creación de recursos lingüísticos, como el que aquí se ha presentado, abre tanto líneas de investigación teórica como el desarrollo de aplicaciones de diversa índole. Concretamente, ModeS TimeBank ha sido utilizado para iniciar el camino hacia la incorporación de textos en lenguaje natural en un SIG para su posterior visualización y análisis. En este sentido, el uso de lenguajes de marcado como TimeML o SpatialML ofrece amplias ventajas, debido entre otras cosas a su carácter estándar, a su sintaxis y su compatibilidad con distintos formalismos de representación (XML, bases de datos relacionales, etc.).

Para el futuro, nos planteamos la ampliación del corpus con textos del mismo periodo y dominio, así como también del mismo dominio y pero distintos periodos. De este modo, no sólo se da continuidad y un mayor alcance a la investigación, sino que se proporciona información muy valiosa para otras áreas del conocimiento fuera de la Lingüística, como la Historia y otras disciplinas dentro de las Ciencias Sociales.

Bibliografía

Guerrero Nieto, M., M.J. Rodríguez García, A. Urrutia, W. Siabato, M.A. Bernabé Poveda (2010) Incorporating TimeML into a GIS, CICLing 2010, proceedings IJCLA, vol.1 (1-2), 269–283.

Mani, I., J. Hitzeman, J. Richer, D. Harris, R. Quimby, B. Wellner (2008) SpatialML: Annotation scheme, corpora, and tools. Sixth International Language Resources & Evaluation (LREC'08).

Owens, J.B. (2008).What Historians Want from GIS? GIS Best Practices: Essays on Geography & GIS. Redlands, California: ESRI, 35-46.

Pustejovsky, J., J. Moszkowicz, M. Verhagen, (2010). The recognition and interpretation of motion in language. CICLing2010, 236-256.

Pustejovsky, J., M. Verhagen, R. Saurí, J. Littman, R. Gaizauskas, G. Katz, I. Mani, R. Knippen, A. Setzer (2006) TimeBank 1.2, Linguistic Data Consortium, Philadelphia.

Pustejovsky, J., R. Knippen, J. Littman, R. Saurí (2005) Temporal and Event Information in Natural language Text. Language Resources & Evaluation, 39(2-3): 123–164.

Pustejovsky, J. (1995). The generative lexicon. Cambridge, MA, MIT Press.

Sánchez-Marco, C., G. Boleda, J.M. Fontana, J. Domingo (2010) Annotation and Representation of a Diachronic Corpus of Spanish. LREC 2010, 2713-2718.

Saurí, R., Batiukova, O. Pustejovsky, J. (2009). Annotating Events in Spanish TimeML Annotation Guidelines. Barcelona Media Technical Report, BM 2009-01. (http://comunicacio.barcelonamedia.org/technical_reports/BM2009_01.pdf)

Saurí, R., Saquete, E., Pustejovsky, J. (2010). Annotating Time Expressions in Spanish. TimeML Annotation Guidelines. Barcelona Media Technical Report, BM 2010-02. (http://comunicacio.barcelonamedia.org/technical_reports/BM2010_02.pdf)

Saurí, R. (2010). Annotating Temporal Relations in Catalan and Spanish. TimeML Annotation Guidelines. Barcelona Media Technical Report, BM 2010-04. (http://comunicacio.barcelonamedia.org/technical_reports/BM2010_04.pdf)

Saurí, R. (2010). TempEval 2. Spanish Data Release. TempEval Release. Barcelona Media Technical Report, BM 2010-05. (http://comunicacio.barcelonamedia.org/technical_reports/BM2010_05.pdf)

Schmid, H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In International Conference on New Methods in Language Processing: 44–49.

Verhagen, M., R. Saurí, T. Caselli, J. Pustejovsky (2010) SemEval-2010 Task 13: TempEval-2. In 5th International Workshop on Semantic Evaluation, ACL 2010: 57–62.