METODOLOGIA PARA LA EXTRACCIÓN DE METADATOS SEMÁNTICOS DE TEXTOS EN ESPAÑOL UTILIZANDO PROCESAMIENTO DE LENGUAJE NATURAL:

SUBAPLICACIÓN PARA LA IDENTIFICACIÓN DE CONTEXTOS ESPACIALES Y TEMPORALES EN TEXTOS QUE DESCRIBAN INTERACCIONES ENTRE ACTORES

ERIKA TERESA DUQUE BEDOYA

UNIVERSIDAD EAFIT

DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS

2009

METODOLOGIA PARA LA EXTRACCIÓN DE METADATOS SEMÁNTICOS DE TEXTOS EN ESPAÑOL UTILIZANDO PROCESAMIENTO DE LENGUAJE NATURAL:

SUBAPLICACIÓN PARA LA IDENTIFICACIÓN DE CONTEXTOS ESPACIALES Y TEMPORALES EN TEXTOS QUE DESCRIBAN INTERACCIONES ENTRE ACTORES

ERIKA TERESA DUQUE BEDOYA

Trabajo de grado presentado como
requisito parcial para optar al título de
Magister en Ingeniería Informática

Asesor: PhD. JUAN GUILLERMO LALINDE PULIDO

MEDELLÍN
UNIVERSIDAD EAFIT
DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS
2009

Nota de aceptación

_____

_____

Presidente del jurado

_____

Jurado

_____

Jurado

_____

Medellín, 23 de noviembre de 2009

A mi gran amigo don Álvaro del Portillo,
quien ha sido el presidente
de mi tesis desde hace muchos años.

A mis padres, con su infinita paciencia
me enseñaron el amor al estudio y la sabiduría

A mis amigas de Arizá que siempre
confiaron en mí.

AGRADECIMIENTOS

Al profesor Juan Guillermo Lalinde Pulido, profesor e investigador del Departamento de Informática y Sistemas por su ayuda, paciencia y apoyo incondicional para la elaboración de este trabajo.

Al profesor y sacerdote Iván Darío Toro quien proporcionó de manera generosa los documentos históricos para elaborar el corpus de trabajo y tuvo la paciencia para ver los resultados de la investigación.

Al profesor Jorge Antonio Mejía, profesor del Instituto de Filosofía de la Universidad de Antioquia por sus valiosos consejos, por su infinita paciencia y por introducirme en los mundos del análisis textual.

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

GLOSARIO

ANÁFORA (*ANAPHORA*): hace referencia a una entidad que ha sido introducida previamente en el texto. Ejemplos de anáfora: "él la miró fijamente". El pronombre "él" es una anáfora.

AE: del acrónimo inglés *Analysis Engine.* Son estructuras de UIMA encargadas de analizar el texto ubicando automáticamente los metadatos que se encuentran en él y etiquetándolos.

ANOTADOR: son programas de JAVA que se utilizan en UIMA y que tienen elementos concretos de análisis para proceder con el etiquetamiento de una entidad determinada.

CAS CONSUMER: es el componente de UIMA que analiza los datos de las entidades de interés de acuerdo a los análisis del AE.

COLLECTION READER: es una interfaz del CPE de UIMA para acceder a los documentos que serán analizados.

CPE: del acrónimo inglés *Collection Processing Engine*. Es una aplicación de UIMA que permite analizar varios documentos al tiempo y visualizar las anotaciones o generar un archivo plano de texto que las contiene.

COBERTURA: medida de evaluación usada en los sistemas de recuperación de información y adaptada para los sistemas de extracción de la información. Define la proporción de documentos (o hechos en el caso de la EI) que son recuperados desde un conjunto de documentos relevantes.

CORPUS: Recurso lingüístico conformado por textos que están clasificados y ordenados de acuerdo a un criterio determinado. Pueden ser orales ó escritos.

DESCRIPTOR: es un recurso utilizado por UIMA que describe el documento analizado por el sistema en formato XML.

DETECCIÓN DE EVENTOS: Proceso por el cual se busca encontrar en los documentos de texto estructuras oracionales que expresen un evento y definir sus componentes.

DOCUMENT ANALYZER: Es una aplicación de UIMA que permite visualizar las anotaciones sobre un texto.

DOCUMENTO: Unidad de información digital con un contenido específico. Puede contener texto, audio, video o incluso otro tipo de medios

> **DE TEXTO:** Tipo de documento específico que solo contiene caracteres alfanuméricos en una secuencia específica.

EI: del acrónimo *Extracción de la información.* Ver *Extracción de la información.*

ENTIDAD: se refiere al sujeto en las oraciones de los textos. Las entidades son los elementos que realizan alguna acción y se asocian con otras entidades. Se puede hablar de las entidades también como actores.

ENTIDAD NOMBRADA: se refiere al nombre de una entidad dentro de un texto. Una *entidad* no es igual a un *nombre propio.* En inglés se le conoce como *Named Entity.*

ETIQUETA: es una marcación sobre un texto. Usualmente esta marcación se lleva a cabo utilizando XML.

EXTRACCIÓN DE INFORMACIÓN (EI): tecnología de las ciencias de la computación que busca obtener datos de los documentos relevantes que puedan ser consultados por un usuario. Utiliza un conjunto reducido de técnicas del lenguaje natural para hacer un análisis de los textos y "extraer" la información requerida.

GAZETTEERS*:* son diccionarios ó directorios geográficos que contienen información de una región o continente junto con estadísticas sociales y características físicas tales como montañas, ríos, carreteras, etc. Usualmente se utilizan de manera conjunta con mapas o atlas.

JAVA: es un lenguaje de programación orientado a objetos desarrollado por *Sun Microsystems*.

LEMA: son las entradas de un diccionario o enciclopedia.

METADATO: se trata de una información que describe otra información, son datos acerca de los datos. Son expresados en lenguajes de marcas como el XML y se usan normalmente

para caracterizar documentos. Facilitan en gran medida la labor de búsqueda de información.

DESCRIPTIVO: tipo de metadato cuyo contenido no puede obtenerse del texto asociado a él.

SEMÁNTICO: metadato cuyo contenido se encuentra especificado en el texto que relaciona de una manera explícita o implícita.

***PART OF SPEECH TAGGING (POS)*:** del inglés *Etiquetado de partes del discurso o de constituyentes del texto.*

PRECISIÓN: medida utilizada en la recuperación de la información y adaptada para la extracción de la información. Busca evaluar la proporción de documentos relevantes que son entregados al usuario entre un conjunto de documentos recuperados.

RECUPERACIÓN DE INFORMACIÓN (RI): tecnología de las ciencias de la computación que busca proporcionar al usuario documentos relevantes de acuerdo a su búsqueda. Utiliza técnicas estadísticas para el procesamiento de los documentos.

RELEVANCIA: propiedad de los documentos o datos extraídos de un texto. La relevancia se refiere a la manera en que se ajustan los datos recuperados a las necesidades de información del usuario.

RI: del acrónimo *Recuperación de la información.* Ver *Recuperación de la información.*

SGML: acrónimo de *Standard Generalized Markup Language* o Lenguaje de Marcado General Normalizado. SGML fue diseñado para permitir el intercambio de información entre distintas plataformas, soportes físicos, lógicos y diferentes sistemas de almacenamiento y presentación (bases de datos, edición electrónica, etc.), con independencia de su grado de complejidad.

***TDT*:** del acrónimo inglés *Topic Detection and Tracking.* Iniciativa de investigación por medio de la cual se busca desarrollar técnicas y algoritmos capaces de hacer detección y seguimiento de eventos en diferentes documentos mediante técnicas de recuperación y clasificación automática de información.

TEXTO: ver *documento de texto.*

UIMA: del acrónimo inglés *Unstructured Information Management Architecture.* Es una arquitectura de componentes de *software* para el desarrollo de aplicaciones que faciliten el análisis de información no estructurada (como los textos ó audio) y la integración con otras tecnologias. Fue desarrollada por IBM. Actualmente es un proyecto Apache.

URI: del acrónimo inglés *Uniform Resource Identifier*. Se trata de un conjunto de caracteres utilizados para identificar un recurso en Internet.

URL: del acrónimo inglés *Uniform Resource Locator.* Se trata de un conjunto de caracteres utilizados para determinar la ubicación de un recurso determinado en Internet.

XML: del acrónimo inglés *eXtensible Markup Lenguaje. Lenguaje de marcado extensible.* Lenguaje de marcado derivado del SGML que permite crear estructuras de datos asociadas o no a documentos de texto.

RESUMEN

En este trabajo describe el proceso por el cual se ha efectuado la extracción de información e identificación de eventos en un corpus construido para estos fines y compuesto por textos históricos pertenecientes a la Iglesia Católica en el s. XIX en Colombia entre los años 1869 y 1880, con un tamaño de 224 documentos. Este material pertenece a los archivos de la Arquidiócesis de Medellín y ha sido recopilado y suministrado por el padre Iván Darío Toro, Decano de la facultad de Filosofía y Teología de la Fundación Universitaria Luis Amigó y docente de la Escuela de Administración y Negocios de la Universidad EAFIT. Los procesos de extracción de información incluyeron la identificación automática de personajes, lugares y fechas por medio de la aplicación de algoritmos y heurísticas empleadas en las bibliotecas digitales. La identificación de eventos se llevó a cabo utilizando la combinatoria de las etiquetas extraídas previamente del corpus.

Palabras clave: DETECCIÓN DE EVENTOS / EXTRACCIÓN DE INFORMACIÓN / CORPUS HISTÓRICOS / RECUPERACIÓN DE INFORMACIÓN / LINGÜÍSTICA COMPUTACIONAL

INTRODUCCIÓN

Esta investigación surgió como parte del macroproyecto "Uso de Procesamiento de Lenguaje Natural para la Recuperación de Información en Documentos Históricos" liderado por los profesores Juan Guillermo Lalinde, investigador y docente del Departamento de Ingeniería de Sistemas de la Universidad EAFIT e Iván Darío Toro, profesor de la Universidad Luis Amigó y de la Universidad EAFIT. Fue propuesto con el fin de estudiar las metodologías relacionadas con las técnicas de extracción de información y lingüística computacional aplicadas en textos históricos, un tema que no ha sido explorado en este medio colombiano. El trabajo de investigación consistió en el desarrollo y la descripción de la metodología de trabajo utilizada para efectuar la extracción de información y análisis de eventos para los textos históricos de la Iglesia Católica en Colombia recopilados por el profesor Iván Darío Toro. Para efectuar este análisis textual fue preciso construir un corpus de estudio con los textos históricos y luego encontrar y manipular las herramientas informáticas que permitieran efectuar las tareas de extracción y de análisis de los datos mediante el desarrollo de algoritmos y heurísticas computacionales. De manera particular se hizo énfasis en la descripción de la metodología empleada por dos razones: la primera es que no resulta tan obvia a la hora de leer resultados en un *paper* científico ya que muchos detalles de análisis se omiten y en segundo lugar, porque estos temas han sido poco desarrollados en el medio colombiano y se espera que por medio de este trabajo otros investigadores puedan continuar con posteriores análisis textuales en el marco del macro proyecto.

Este trabajo cubrió varias etapas que se describen a continuación:

- El capítulo 1 es una breve introducción de los conceptos de lingüística computacional, extracción de información, lingüística de corpus y el estado de la cuestión en la detección de eventos y revisión de la literatura respectiva.
- El capítulo 2 trata sobre la construcción del corpus utilizado en el análisis textual, el cual contiene los documentos que proporcionó el historiador del macroproyecto y que

está constituido por cartas de la Iglesia católica en el s. XIX en Antioquia y Caldas. El corpus es la materia prima de cualquier análisis lingüístico.

- El capítulo 3 describe el proceso de caracterización y extracción de entidades del corpus por medio de heurísticas y algoritmos con el fin de obtener las anotaciones relacionadas con la detección de eventos, como son los nombres de personas, lugares y fechas.
- El capítulo 4 describe el proceso utilizado para la detección de eventos utilizando la información extraída.
- Conclusiones y trabajos futuros

De esta manera, el trabajo que se presenta en estas páginas ha sido el fruto de un proceso de maduración alrededor del manejo de las tecnologías relacionadas con la manipulación de textos históricos y los algoritmos y heurísticas que han permitido extraer información escrita en medio digital y que la mente humana puede reconocer y producir casi de manera automática como fruto de un proceso evolutivo del cerebro. De hecho, una de las lecciones aprendidas en este proceso ha sido el ejercicio de la humildad porque cuando se trabaja en temas de lingüística computacional se requiere tener un corazón grande para asombrarse ante la capacidad de inferencia y de aprendizaje que tiene el ser humano cuando lee ó escribe un texto y también se requiere humildad cuando se trata de emular este comportamiento en el terreno de la computación porque los resultados obtenidos no siempre son los que se esperarían en cuanto a la limitación que tienen los computadores para extraer datos y también debida a la complejidad propia de los procesos de implementación de la extracción de información. Por lo tanto, este trabajo es una descripción de cómo se utilizan los conceptos computacionales para alcanzar una mayor comprensión de fenómenos lingüísticos tales como la identificación de entidades y una aplicación derivada de esta tarea previa, que es el reconocimiento de eventos en textos escritos.

OBJETIVOS

Los siguientes son los objetivos de este trabajo de investigación:

Objetivo general del proyecto

- Desarrollar algoritmos y técnicas estadísticas basados en técnicas de procesamiento de lenguaje natural en español para la clasificación automática de personajes, lugares, fechas y relaciones en los mismos, que apoyen al usuario para la recuperación de contenidos en colecciones de documentos que han sido digitalizados (corpus).

Objetivos específicos

- Elaborar un corpus que permita la búsqueda, medición y análisis de la información del conjunto de documentos históricos suministrados (de la Iglesia Católica en su mayoría cartas y correspondencia), con base en los períodos de tiempo que presenten mayor acumulación de documentos y en la uniformidad estilística de la época.

- Definir algoritmos basados en procesamiento de lenguaje natural ó en las técnicas estadísticas que permitan identificar de manera automática personajes, lugares y fechas, para establecer relaciones entre ellos y proponerlas al usuario para la validación.

CAPÍTULO 1

MARCO CONCEPTUAL Y ESTADO DE LA CUESTIÓN PARA RECONOCIMIENTO DE ENTIDADES Y DETECCIÓN DE EVENTOS

"Quien pudiera como tú, a la vez quieto y en marcha,
cantar siempre el mismo verso, pero con distinta agua."
G. Diego. Romance del Duero

Palabras Clave. Recuperación de información, bibliotecas digitales, reconocimiento de entidades, corpus, detección de eventos.

Este capítulo describe los conceptos básicos de lingüística computacional, lingüística de corpus, extracción de información y el estado de la cuestión en detección de eventos.

1.1 MARCO CONCEPTUAL

1.1.1 Lingüística Computacional. Una de las revoluciones en el mundo documental ha sido la digitalización de documentos porque ha cambiado el tratamiento y la manipulación de la información y ha ampliado su horizonte de trabajo. Internet ha permitido acceder y compartir información en diferentes formatos a través del planeta y también ha contribuido a que muchos documentos que se encontraban en las bibliotecas estén disponibles para los investigadores sobre todo, el acceso a documentos históricos. Sin embargo el manejo documental también plantea retos. Lavid (2005) menciona 3 aspectos que han cambiado el tratamiento de la información con el advenimiento de la digitalización documental, a saber:

- Almacenamiento: la información ha cambiado su formato físico (impreso) por el formato digital. Los principales retos planteados en este aspecto son la adopción de

múltiples formatos de almacenamiento y su rápida caducidad en el tiempo, así como la conservación y preservación de la información.

- Reproducción de la información: antiguamente se transcribían las cartas para reproducir un mensaje. En el mundo actual a medida que la tecnología avanza, se presentan diferentes maneras de reproducir información y de enviarla de manera casi inmediata.
- Derechos de reproducción: el aspecto jurídico de derechos de autor y de reproducción de la información digital poco a poco se ha ido incorporando a las legislaciones de los países pero todavía es tema de estudio.

Esta revolución de la información digitalizada ha sido la puerta de entrada para la adopción de nuevas herramientas de trabajo en las llamadas **"Tecnologías del Lenguaje"** las cuales "*consisten en la aplicación del conocimiento sobre la lengua al desarrollo de sistemas informáticos capaces de reconocer, analizar, interpretar y generar lenguaje*" (Lavid, 2005). Este enriquecimiento de "saberes" en las ciencias humanas ha permitido explorar nuevos caminos de análisis y la adopción de metodologías de trabajo para el estudio del lenguaje puesto que el computador puede efectuar tareas necesarias y relativamente sencillas de implementar en los análisis lingüísticos, pero que son dispendiosas para un ser humano como por ejemplo el conteo de frecuencias o búsquedas de palabras y otras tareas[1]. Así mismo, el hecho de incluir el computador para el análisis documental ha permitido el advenimiento de nuevas tecnologías en cuanto a las búsquedas de información, su almacenamiento y preservación, la creación de bibliotecas digitales, la comprobación de teorías lingüísticas por medio del análisis de documentos representativos de una lengua, la utilización de tecnologías en el aprendizaje de nuevos idiomas, entre otras. El cambio en la metodología de trabajo utilizando el computador como herramienta también ha permitido la adopción de análisis cuantitativos que permitan la inclusión de modelos matemáticos y estadísticos para la descripción de fenómenos lingüísticos y también ha

[1] Una aplicación computacional que permite efectuar este tipo de análisis es CRATILO, desarrollado por el profesor Jorge Antonio Mejía del Instituto de Filosofía de la Universidad de Antioquia

permitido el enriquecimiento de los análisis cualitativos del lenguaje (Gelbukh y Sidorov, 2006). Como resultado se ha originado la **"Lingüística Computacional"** que combina los conocimientos de estadística y técnicas computacionales para el estudio del lenguaje. Otro concepto importante es el **Procesamiento del Lenguaje Natural (PLN),** que es una parte de la Inteligencia Artificial que investiga y formula mecanismos computacionales para facilitar la interacción entre el humano y la máquina estudiando el lenguaje humano (Carbonell, 1992). Aunque el objetivo de la lingüística computacional y del PLN son el mismo: estudiar el lenguaje humano y tratar de emularlo ó crear modelos del lenguaje humano por medio de prácticas computacionales, el PLN tiene un enfoque más técnico y orientado hacia la aplicación de herramientas computacionales en el lenguaje, mientras que la lingüística computacional tiene un enfoque orientado hacia el desarrollo de una teoría formal del lenguaje (Gelbukh y Sidorov, 2006).

Justamente el enriquecimiento de la metodología computacional en el análisis lingüístico propone una reflexión acerca del perfil que debería tener una persona que se dedica a la lingüística computacional aún cuando este tema merecería un análisis mucho más profundo y dedicado. Esta reflexión está orientada a analizar brevemente el cambio de paradigma que ofrece la complementariedad del perfil informático y el humanístico puesto que a medida que avanza la tecnología documental y el Internet, la información se encuentra más disponible y la adquisición de conocimientos informáticos es cada vez más "obligante". Una persona que se dedique al estudio de la lingüística computacional proviniendo del mundo "lingüístico" tiene el conocimiento teórico que le permite enfrentarse a un problema para abordarlo de manera específica pero también debería tener la suficiente apertura y flexibilidad mental para adaptarse a la utilización de herramientas computacionales que puedan brindarle elementos de análisis. Así mismo, debería tener cierto conocimiento formal desde la computación y la estadística para abordar los problemas siguiendo una metodología que permita proponer soluciones informáticas concretas para el análisis del fenómeno. De esta manera, es necesario tener conocimiento lingüístico pero es menester hacer uso del conocimiento informático. Es importante tener en cuenta que el hecho de estudiar una disciplina humanística es absolutamente compatible con otros conocimientos tales como la informática, enriqueciendo aun más el

perfil del humanista, que puede estar orientado hacia la aplicación de nuevas tecnologías en sus estudios y de hecho, esta es la tendencia. Así mismo, cuando se recorre el camino hacia la lingüística y se proviene del mundo "ingenieril", la complementariedad es interesante porque el estudio de un fenómeno humanístico enriquece el "bagaje" cultural y humanístico del ingeniero (que a veces desafortunadamente es muy pobre) y también requiere cierta "flexibilidad" mental para analizar los problemas desde otros puntos de vista. Además, el hecho de enfrentase al análisis de temas complejos como el estudio del lenguaje y sus problemas derivados se presenta como un reto para ser analizado y que tiene aplicación inmediata cuando se analizan las tecnologías de la Web y otros campos de aplicación (Jurasfky y Martin, 2000) que serán mencionados más adelante.

¿Por qué es importante estudiar Lingüística computacional? Es una pregunta que tiene un carácter pragmático, que cobra cada vez más importancia en el momento actual donde las tecnologías de Internet son más relevantes y que reúne varios frentes de trabajo:

- Facilitar la comunicación entre los humanos y las máquinas.
- Facilitar la comunicación entre personas que tienen lenguajes diferentes.
- Facilitar el análisis de tareas lingüísticas como el análisis de textos, construcción de diccionarios generales o terminológicos.
- Dominio de la Web.

Algunas de las aplicaciones de la lingüística computacional son:

- **Recuperación de información (RI) y Extracción de la información (EI):** estos campos han cobrado mucha relevancia porque facilitan las búsquedas de información en las bibliotecas digitales y en la Web. La recuperación de información estudia la manera de efectuar búsquedas de documentos relevantes de acuerdo a la consulta del usuario y la extracción de información hace una búsqueda de datos más selectos en el texto de acuerdo al interés del usuario. Es importante mencionar que el usuario siempre está interesado en efectuar búsquedas por conceptos (semántica) y los motores de búsqueda encuentran información de acuerdo a la grafía. En la sección 1.3 se exponen con mayor detenimiento tanto la RI como la EI puesto que son estudiados en este trabajo de investigación.

- **Interfaces hombre-máquina:** son sistemas construidos con el fin de facilitar la interacción entre un humano y la computadora tratando de adaptar las máquinas al lenguaje natural para evitar que el usuario manipule menús o lenguajes complicados.
- **Los sistemas pregunta-respuesta:** es una técnica que pretende dar respuesta corta y concisa a una pregunta formulada por el usuario sin que tenga que consultar los miles de documentos que puede devolver un buscador convencional. Utiliza métodos de recuperación de información y de extracción de información para encontrar los datos relevantes.
- **Traducción automática:** es uno de los temas que ha sido más estudiados en lingüística computacional desde sus inicios. Consiste en tratar de traducir documentos de un idioma a otro o entre varios a la vez. En un principio se consideró que esta tarea era fácil de resolver pero a medida que las investigaciones avanzaron se encontró que era necesario hacer un estudio más profundo de las lenguas, implicando el análisis gramatical y sintáctico para hacer traducciones entre ellas.
- **Enseñanza de las lenguas asistidas por computador:** es el diseño de programas educativos que permitan el aprendizaje de lenguas por medio del computador utilizando multimedia y reconocimiento de voz.
- **Herramientas para análisis de escritura:** son aplicativos que permiten hacer corrección ortográfica o gramatical de textos, conteo de palabras, frecuencias, concordancias y otras tareas que pueden ser dispendiosas pero que facilitan el análisis textual para los investigadores.
- **Bases de datos lexicográficas y terminológicas:** es la recopilación de datos lingüísticos de una lengua con el fin de construir diccionarios electrónicos que pueden ser generales o especializados (utilizando la terminología de un dominio concreto de conocimiento).
- **Análisis estilístico de textos:** es la identificación del origen de un texto basado en el estilo del autor. Estas aplicaciones son útiles para determinar el origen de documentos anónimos y para la identificación de plagios.

Esta investigación se ha ocupado de analizar información textual. En la próxima sección se expone un análisis más detallado de algunos conceptos sobre los tópicos relacionados con el análisis de corpus escritos, recuperación de información (RI), extracción de información (EI) que fueron estudiados a lo largo de este trabajo.

1.1.2 Aspectos metodológicos de diseño en la investigación lingüística. Puesto que esta investigación trata sobre lingüística computacional es importante considerar la metodología a seguir en este tipo de estudios. En general, de acuerdo a Lavid (2005) existen 4 aspectos a tener en cuenta para abordar un estudio en lingüística, a saber:

1. **El tamaño de la muestra:** hace referencia a la materia prima objeto de estudio. Es un factor determinante porque en la medida que la muestra sea más representativa es más extrapolable el resultado obtenido (Manning, 1999; Chantal, 2002). La muestra puede ser compuesta por corpus orales o corpus escritos.
2. **La disponibilidad de los datos en formato electrónico**: Es un aspecto crítico que es necesario tener en cuenta porque implica que los documentos a analizar se encuentren en texto electrónico y no simplemente digitalizados, además de que los procesos de conversión entre imagen de texto y el texto como tal son costosos en tiempo y en recursos. El documento digitalizado puede ser una foto o una imagen de escáner y este proceso es importante porque permite obtener la memoria del documento y preservarlo digitalmente, mas no es suficiente si se desea efectuar análisis computacionales porque las imágenes no permiten la manipulación directa del texto contenido en el documento. Es necesario efectuar otro pre-procesamiento para obtener el texto disponible. Usualmente se utilizan los OCR (*Optical Character Recognition*) o la digitación directa del texto.
3. **La adecuación de los programas existentes**: existen muchos programas de análisis textuales disponibles de acuerdo al tipo de estudio que se desee efectuar. Es primordial tener en cuenta primero el tipo de análisis a llevar a cabo, cuál será el corpus o texto de estudio y luego escoger la herramienta informática adecuada para proceder.

4. **Las necesidades de desarrollo de programas específicos**: cuando las aplicaciones existentes no son suficientemente flexibles para realizar los análisis requeridos se procede al desarrollo de programas informáticos. Para este trabajo de investigación se utilizó el software UIMA (Apache UIMA, 2009) que es una herramienta que permite construir herramientas para hacer análisis textual.

En este trabajo de investigación se ha seguido esta metodología, primero analizando la disponibilidad de los documentos analizados y construyendo el corpus de acuerdo con los criterios de representatividad requeridos (situación que se tratará en el capítulo 2) y luego estudiando las diferentes herramientas informáticas disponibles en el mercado para llevar a cabo el análisis y efectuando los desarrollos necesarios (capitulo 3 y 4).

1.1.3 Corpus y lingüística de corpus. Para trabajar en lingüística computacional no solo **es** indispensable tener el computador y conocer las técnicas estadísticas de análisis de datos sino que también es fundamental tener una fuente de análisis que proporcione los elementos para efectuar los estudios lingüísticos. De hecho**,** siempre se parte de la fuente lingüística para hacer los estudios de interés. Esa fuente de información es llamada "corpus" y tiene unas características propias que lo diferencian de un simple cúmulo de documentos: un corpus es una colección de documentos (textos **o** grabaciones) que ha sido colec**cionada** con fines específicos, que se considera representativa de la lengua que se quiere analizar (es decir que ofrece ejemplos significativos de la población de interés) y además es "balanceada" lo cual quiere decir que los criterios de **selección** del corpus se mantengan a lo largo del mismo de tal manera que sea posible acceder a una porción manteniendo sus características (Manning, 1999). Aunque el tamaño del corpus es un tema discutible porque en ocasiones es complicado saber su valor definitivo (sobre todo cuando se está construyendo), es recomendable que ofrezca las garantías de representatividad sobre todo para los aspectos relacionados con el análisis estadístico. Teniendo en cuenta la definición de corpus es posible también considerar las siguientes clasificaciones de recopilaciones de texto que citan Torruela y Llisterri (1999) para considerar qué es un corpus y que no es un corpus:

- Archivo o colección informática de textos: es una colección de textos informatizados que no tienen relación entre si.
- Biblioteca de textos electrónicos: es una colección de textos informatizados que tienen un formato estándar, siguiendo ciertas normas de contenido pero sin tener un criterio riguroso de selección.
- Corpus informatizado: es una recopilación de textos informáticos seleccionados de acuerdo a criterios lingüísticos, codificados de manera estándar y homogénea con el fin de reflejar el comportamiento de una o más lenguas.

De acuerdo a lo anterior, Internet no se considera un corpus porque es un "repositorio" de documentos que no están ordenados y ni obedecen a criterios uniformes. Sin embargo, es posible construir corpus formados por varias páginas web que obedezcan a criterios concretos de análisis, por ejemplo, el reconocimiento de personas en la web es un tema de investigación actual y su corpus de estudio puede conformarse por páginas de personajes colectadas de acuerdo con unos "lineamientos" que las hacen apropiadas para aplicar los métodos de análisis de reconocimiento de personas (Kalashnikov et al, 2008).

Al respecto de los tipos de corpus se tienen las siguientes clasificaciones de acuerdo a la naturaleza de los datos que manejan:

- Corpus orales: conformados por registros orales de un idioma y que son utilizados para reconocimiento del habla y enseñanza de lenguas.
- Corpus multimodales: incluyen voz e imágenes y son útiles para analizar el contexto de las formas lingüísticas.
- Corpus textuales: conformados por textos, son útiles para llevar a cabo estudios relacionados con terminología, fraseología, lexicografía. Un ejemplo es el corpus construido y utilizado en esta investigación.

Otro tipo de clasificación de corpus tiene que ver con el tipo de documentos que manejan. A continuación se presenta la propuesta por Alcántara (2007):

- **Corpus especializados:** se componen de documentos que han sido elegidos porque presentan características muy concretas, por ejemplo los corpus temáticos (medicina, biología, deportes) y los de registro (documentos del género periodístico).
- **Corpus generales:** son corpus abiertos que tratan temas generales. Ejemplos de estos tipos de corpus son *British National Corpus* (BTN) y el Corpus de Referencia del Español (CREA, http://corpus.rae.es/creanet.html).
- **Corpus comparables**: están compuestos por subcorpus que comparten todas las características básicas exceptuando el idioma. Se utilizan en estudios de lingüística comparada, traducción y enseñanza de las lenguas. Ejemplo: C-ORAL-ROM (http://lablita.dit.unifi.it/coralrom/).
- **Corpus paralelos**: están compuestos por subcorpus de textos idénticos pero en distintos idiomas. Pueden ser producidos simultáneamente. Ejemplo, los textos de legislación de la UE o de Canadá.
- **Corpus históricos**: son textos de distintos periodos históricos que se emplean para realizar estudios diacrónicos de la lengua. Un subconjunto de los corpus históricos son los Corpus Monitor que se utilizan en el estudio de los cambios idiomáticos en periodos de tiempo cortos. Ejemplo, el corpus CORDE de la Real Academia de la Lengua (http://corpus.rae.es/cordenet.html).

La utilización y construcción de corpus ha originado la llamada **"Lingüística de Corpus"** que es el estudio de la lengua mediante la recopilación y diseño de corpus. La lingüística de corpus es importante porque permite la construcción de elementos de estudio que luego pueden servir para la comprobación de teorías lingüísticas. Al respecto, Alcántara (2007) menciona cuatro aspectos metodológicos de la Lingüística de corpus:

- "[La lingüística de corpus] es una ciencia empírica ya que analiza los patrones reales de uso en textos naturales.
- Utiliza grandes colecciones de textos llamados corpus como base de sus análisis.
- Hace uso de las técnicas automáticas e interactivas a través de los ordenadores.
- Se basa en técnicas analíticas y cuantitativas".

1.1.4 Etiquetamiento de corpus. Un aspecto intrínseco del análisis de corpus es su **marcación** o **etiquetamiento,** porque finalmente los corpus son la materia prima para efectuar análisis del lenguaje pero los sistemas de etiquetamiento constituyen la herramienta para extraer las palabras de interés y los sistemas de etiquetado permiten escoger, almacenar y recuperar los datos del corpus. El lenguaje informático utilizado para efectuar las anotaciones o el etiquetamiento es el formato **XML** (*Extensible Markup Language*) que es un lenguaje de marcado desarrollado por el W3C (*Consorcio World Wide Web*) y que proviene a su vez del lenguaje SGML (*Standard Generalized Markup Language*). Las características que tiene XML con relación a su utilización como lenguaje de marcación y que lo han convertido en el estándar "*de facto*" para la anotación de recursos lingüísticos son:

- XML permite la libre configuración de las etiquetas sobre el corpus sin restricciones de marcación y además es compatible con el estándar UNICODE, lo que facilita la marcación para diferentes idiomas.
- XML permite conservar la estructura de los documentos, lo cual permite el intercambio de información entre aplicaciones informáticas, sobre todo cuando se dispone de varias herramientas lingüísticas que pueden ser aprovechadas.
- XML es independiente del medio, esto significa que es posible editar y mantener documentos y publicarlos automáticamente en diferentes medios, lo cual es una característica muy conveniente para los sitios Web.
- XML es un lenguaje abierto y gratis es decir que es posible utilizarlo sin restricciones de costo.

Alcántara (2007) ofrece una descripción muy completa acerca de las clases de herramientas informáticas utilizadas en el etiquetamiento:

- Asistente de etiquetado: son programas desarrollados para asistir el etiquetamiento de acuerdo al tipo de marcación que se quiera utilizar.
- Sistemas de etiquetado automático: realizan el etiquetado sin necesidad de la supervisión de un lingüista. Generalmente reciben un documento en un formato específico original y producen otro con los textos del primero pero con las etiquetas. El

desempeño de estos sistemas se mide de acuerdo a la precisión y la cobertura, parámetros que serán explicados en la sección siguiente. UIMA, que es la aplicación informática utilizada en este trabajo es un ejemplo de estos sistemas.

- Sistemas de etiquetado semi-automático: son sistemas en donde el lingüista ayuda a las aplicaciones a completar el etiquetado. La labor del humano consiste en revisar y corregir los resultados y desambiguarlos.

Existen algunas herramientas de etiquetamiento automático disponibles en Internet, algunas son GATE (**http://gate.ac.uk/index.html**) y UIMA (**http://incubator.apache.org/uima**) que fue la utilizada en este trabajo.

Cualquier herramienta informática utilizada en la anotación de los corpus debería cumplir con ciertos requerimientos mínimos de acuerdo con la recomendación de Leech (1993) citado por Abaitua (2000) y que se muestran a continuación:

1. "Facilitar la eliminación de las anotaciones de forma que sea posible recuperar la versión original de los textos.
2. Permitir la extracción de las anotaciones por sí mismas, de manera que puedan constituir una base de conocimientos autónoma, independiente del texto al que se deben.
3. Distribuir las normas en las que se basan las anotaciones para que los usuarios finales puedan interpretarlas sin dificultad
4. Indicar el procedimiento por el que se introdujeron las anotaciones en los textos y las personas responsables del proceso.
5. Alertar sobre la posibilidad de que el corpus anotado contenga errores. La anotación de un corpus es un acto de interpretación de estructuras y de contenidos y no es infalible.
6. Permitir la más amplia funcionalidad y reutilización del corpus acudiendo a propuestas con mayor aceptación y neutras en lo posible respecto a los formalismos o teorías.
7. Admitir la existencia de otras normas y estándares de anotación".

En el capítulo 3 se analizarán estas máximas de acuerdo con el trabajo llevado a cabo en el proceso de etiquetamiento.

Selección de las etiquetas del corpus. El tipo de etiquetas se escogen de acuerdo al tipo de análisis que se llevará a cabo, por ejemplo, un lingüista puede tener interés en anotar los componentes morfosintáctico del texto (que también es conocido como POS, "*part-of-speech-tagger*") o efectuar un análisis sintáctico ("*parsing*") o etiquetar las palabras que tengan un sentido pragmático de su interés. En caso de que se trate de un historiador o un genealogista, su interés puede radicar en el etiquetamiento de personas y lugares para lo cual se utilizará un etiquetamiento morfosintáctico y semántico. De acuerdo a lo anterior, pueden distinguirse varios tipos de etiquetamiento:

1. **Anotaciones sintácticas**: utiliza marcación estructural para marcar los párrafos, oraciones y frases.
2. **Lematización**: consiste en anotar los lemas presentes en el texto y es utilizado para análisis lexicales.
3. **Anotaciones morfosintácticas**: componentes morfosintácticos de la lengua.
4. **Anotaciones semánticas**: son las anotaciones de las relaciones semánticas entre elementos oracionales o clases semánticas de las palabras de un texto. Debido al auge de la web y las bibliotecas digitales, éste tipo de anotaciones ha cobrado importancia ya que permiten recuperar información a través de agentes de búsqueda especializados para el primer caso, o a través de técnicas heurísticas y recursos de conocimiento para el segundo.
5. **Anotaciones pragmáticas y discursivas**: son utilizados para el análisis de actos comunicativos del lenguaje.
6. **Anotaciones prosódicas**: utilizado para el etiquetamiento del lenguaje hablado.

En esta investigación se llevaron a cabo etiquetamientos sintácticos, de lemas y semánticos y que se analizan con mayor detalle en el capítulo 3.

1.1.5 Recuperación y extracción de información. El problema fundamental a la hora de trabajar con documentos de texto consiste en la necesidad de encontrar aquellos que se

ajusten a una necesidad de información específica y, luego encontrar la información requerida. Existen varios enfoques para lograr este objetivo: la recuperación de la información (RI) y la extracción de la información (EI).

Recuperación de información. La recuperación de información ha sido una de las tareas más estudiadas en lingüística computacional. El problema central es determinar cuáles documentos son relevantes para el usuario de acuerdo a su consulta. Los sistemas tradicionales de RI utilizan índices de palabras clave para indexar los documentos y la recuperación de información se efectúa de acuerdo a los modelos empleados para determinar la relevancia de los documentos o su "*ranking*"(Baeza-Yates y Ribeiro-Neto, 1999). A continuación se hará una breve descripción de los modelos empleados:

- **Modelo booleano:** se fundamenta en las consultas de información clásicas de las bibliotecas tradicionales mediante operadores lógicos ("AND", "OR", "NOT"). Los inconvenientes que presenta este modelo son en primer lugar la dificultad de ofrecer términos intermedios entre los documentos ya que su estrategia está basada en términos binarios y esto claramente afecta la recuperación de información pertinente y en segundo lugar, no siempre es posible traducir los requerimientos de lenguaje natural del usuario a una función booleana.
- **Modelo vectorial:** está basado en un concepto de medición de "similitud" de documentos por medio de la construcción de un vector de las palabras clave indexadas para cada documento. De esta manera es posible asignar pesos a las palabras clave y las consultas y, es posible recuperar documentos con cierto grado de similitud entre ellas y la recuperación es mucho mejor que la efectuada con el modelo anterior. Es el modelo más empleado en RI.
- **Modelo probabilístico:** este modelo intenta resolver el problema de la recuperación de información utilizando los conceptos de los sistemas probabilísticos: dado un conjunto ideal de respuestas a una consulta es posible recuperar el conjunto de documentos que más se aproxime a dicha respuesta. Por tanto, el problema de recuperación de información puede interpretarse como el proceso de especificar las propiedades del conjunto de respuestas ideal. Tal como afirma Baeza-Yates y Ribeiro-Neto (*op. cit*) las

propiedades se caracterizan por la semántica de las palabras clave indexadas. La ventaja de este método es que los documentos se clasifican por la probabilidad de encontrar la información relevante pero sus desventajas son la capacidad que debería tener el sistema para saber cuáles documentos son relevantes o no, y además no siempre se tiene en cuenta la frecuencia de las palabras presentes en el documento.

Puesto que la recuperación de información utiliza índices para efectuar sus operaciones, usualmente tiene subtareas de procesamiento o transformación de los documentos para dejar solo los términos de interés. Estas tareas son:

- Determinación del peso de los términos presentes en los documentos (medida de dispersión): puesto que se requiere recuperar los documentos de acuerdo con la consulta del usuario es preciso determinar los pesos de los términos presentes en los documentos. Existen dos factores importantes que definen la eficacia de la asignación de los pesos de los términos: el primero es la frecuencia de un término en un documento (*Term Frecuency*) y el segundo es la distribución de los términos a lo largo de la colección que se mide de acuerdo a la frecuencia inversa del documento (*Inverse Document Frecuency*). La combinación de los dos factores mide el peso asignado por término y con algunas variaciones también es posible medir los pesos de los términos por documento.
- Determinación de los términos relevantes que serán utilizados para la construcción de índices: usualmente la decisión de escogencia para una determinada palabra que será utilizada como término relevante está relacionada con la naturaleza sintáctica de la palabra (los nombres pueden contener mayor riqueza semántica que un adjetivo o un adverbio, etc.). Cuando se ha seleccionado los términos relevantes se procede con la tarea de determinar sus raíces o tarea de "*stemming*" que consiste en identificar cuáles palabras pertenecen a un mismo lema y eliminar parte de la palabra conservando solo su raíz. Luego se procede con la utilización de la lista de "*stopwords"* que implica la eliminación de aquellas palabras que tienen bajos valores de discriminación en cuanto a la recuperación de información porque son palabras muy frecuentes y no

suministran información útil para resolver el contenido del documento. Usualmente están compuestas por términos tales como pronombres, artículos, conjunciones, etc.

- Construcción de índices y estructuras de categorización de las palabras: ya obtenidos los términos relevantes por documento se procede a la indexación por medio de diferentes algoritmos tales como "*Soundex*", el cual es un algoritmo fonético utilizado para indexar palabras por su sonido y la construcción del índice invertido que muestra los términos junto al lugar de ocurrencia en el documento.

Debido a que estas operaciones de transformación de los documentos se efectúan con miras a la construcción de un índice para finalmente determinar la similitud entre documentos, la recuperación de información elimina mucha información que sería interesante tener en cuenta desde el punto de vista semántico para efectuar otros tipos de análisis textuales, pero de estas tareas se ocupan justamente los sistemas de extracción de información. Además, se requiere un componente semántico para que la recuperación de los documentos sea acorde con los requerimientos de información del usuario y un simple listado de términos o índice no es suficiente para recuperar los documentos pertinentes porque también se requiere el contexto de las palabras. Por esta razón los índices de recuperación de información suelen ser imprecisos (Baeza-Yates y Ribeiro-Neto, 1999).

Las medidas típicas de evaluación del desempeño en un sistema de recuperación de información son la cobertura y la precisión (Baeza-Yates y Ribeiro-Neto, 1999; Manning, 1999; Jurasfky y Martin, 2000):

- La **cobertura** (*recall*)**:** es el porcentaje de documentos relevantes que han sido recuperados sobre el total de documentos disponibles.

$$\text{Cobertura } (Recall) = \frac{\text{número de documentos relevantes}}{\text{Total de documentos}}$$

- La **precisión**: es el porcentaje de documentos recuperados que son pertinentes a la búsqueda del usuario.

$$\text{Precisión } (Precision) = \frac{\text{número de documentos relevantes por el usuario}}{\text{número de documentos relevantes del sistema}}$$

Los mejores ejemplos de sistemas de recuperación de información están dados por los buscadores de Internet: Google, Yahoo, Altavista, etc.

Extracción de información. El objetivo de la extracción de información no es sólo encontrar cuales documentos son relevantes para una consulta dada sino también encontrar datos determinados en el contenido de los documentos y suministrarlos al usuario de una manera organizada, esto es hacer una búsqueda de datos en el texto como tal. La clase de información que los sistemas de EI pueden extraer varía de acuerdo a las necesidades del usuario y el detalle que se requiera. Uno de los ejemplos clásicos de extracción de información es la identificación de entidades tales como nombres de personas o de organizaciones en los textos (también conocido como *Named Entity Recognition – NER*) ya que ofrecen particularidades lingüísticas interesantes tales como la desambiguación de nombres (por ejemplo, cómo distinguir si la palabra para el lugar geográfico "Cartagena" hace referencia a Colombia o a la localidad en España) y también la utilización de anáforas para reconocer si la referencia de un nombre propio en un documento es el mismo a lo largo del documento por medio de la referencia a su apellido o a un pronombre (Nist, 2005; Kalashnikov et al, 2008). Otra de las aplicaciones clásicas en la extracción de información es el reconocimiento de atributos para las entidades y la identificación de eventos (On et al, 2006).

Puesto que la extracción de información depende de la información semántica del texto se deben tener en cuenta las siguientes consideraciones:

- El reconocimiento de entidades utilizado en los sistemas de extracción de información depende fuertemente del idioma y del formato del texto; con base en estas características se definen unas "reglas de análisis" que son aplicadas al texto y que serán implementadas por el etiquetador, lo cual quiere decir que para proceder con el reconocimiento de una entidad de interés se analiza una porción del corpus (se hace un muestreo aleatorio), con base en ese análisis se programan las reglas para proceder con la detección de entidades y paulatinamente se van refinando a medida que el

corpus es recorrido. Estas reglas de análisis utilizan autómatas de estados finitos como las expresiones regulares y también es válido aprovechar los mecanismos lingüísticos del idioma concreto a analizar para hallar entidades y definir reglas, por ejemplo la utilización de preposiciones, conjunciones y otras categorías gramaticales (Jurasfky y Martin, 2000). Sin embargo, el estudio del lenguaje es complejo y usualmente el reconocimiento también se efectúa por medio de heurísticas que dependen también del formato del corpus. Esta dependencia del texto es importante porque las reglas de detección de entidades pueden ser diferentes para documentos notariales o para documentos de periódicos o cartas históricas. Algunos ejemplos de extracción de información se pueden ver en los trabajos de Mani y Wilson (2000), Mckay (2002), Muñoz et al (1998), Muller y Tannier (2004).

- Los estudios textuales se efectúan de acuerdo a los temas de interés y teniendo en cuenta sus características propias, lo que en otras palabras quiere decir que dependerá del "dominio" y de acuerdo a éste se escogerán las palabras que son importantes para etiquetar. Por ejemplo, para un biólogo las palabras de interés serán los nombres de las proteínas pero para una persona dedicada al estudio lingüístico puede ser la caracterización morfológica de las palabras o como en el caso concreto de ésta investigación que examina textos históricos, interesa analizar entidades tales como nombres de personas, nombres de lugares y las fechas. Sin embargo no siempre es posible obtener igual desempeño para el reconocimiento de todas las entidades de un texto porque como se mencionó anteriormente, dependen del idioma y de las reglas aplicadas. La figura siguiente muestra los indicadores actuales para el estado de la cuestión en extracción de información:

Items of Information	Percentile Reliability
Entities	90
Attributes	80
Facts	70
Events	60

Figura 1.

Indicadores del estado del arte en extracción de información (NIST, 2005)

- El desempeño de un sistema de extracción de información es evaluado usualmente por medio de los métodos de precisión y cobertura usados en la recuperación de información.

Otro de los campos de interés en la EI es la detección automática de eventos también conocida como "tareas de resolución de correferenciación" (*Coreference Resolution Tasks*) que surgió como una etapa posterior al reconocimiento de entidades. Estas tareas son de especial interés para los usuarios de las bibliotecas digitales y para el desarrollo de software para humanistas porque ofrecen aplicaciones interesantes en RI como son por ejemplo, la elaboración de resúmenes automáticos relacionados con la obtención de datos relacionados con la evolución cronológica de eventos (Kumaran y Allan, 2005), identificación de referencias de obras de arte (Davis et al, 2003) y obtener datos de personas en la web (Kalashnikov et al, 2008). Otros temas estudiados en la actualidad tienen que ver con la manera de detectar la evolución de los sucesos en el texto y la visualización de estas relaciones (Christel, 2006; Petras et al, 2006). En la sección siguiente se expondrá el estado del arte en el reconocimiento de entidades y la detección automática de eventos.

1.2 ESTADO DE LA CUESTIÓN EN RECONOCIMIENTO DE ENTIDADES Y DETECCIÓN DE EVENTOS

1.2.1 Reconocimiento de Entidades en Bibliotecas Digitales y Textos históricos. Uno de los primeros trabajos de reconocimiento de entidades para el idioma español fue el sistema *EXIT* (Muñoz et al, 1998) que encontraba entidades tales como nombres propios de personas y organizaciones para documentos notariales. Básicamente el sistema estaba compuesto por los siguientes partes: un etiquetador de categorías gramaticales, otro etiquetador para nombres, un módulo para reconocimiento de entidades y un analizador morfo-sintáctico. Para el reconocimiento de entidades utilizaron reglas heurísticas basadas en el contexto (la utilización de frases en mayúsculas, palabras claves, etc.) y diccionarios

de nombres, apellidos, lugares y empresas. El etiquetamiento de las categorías gramaticales se efectuó con el fin de detectar correlaciones entre los verbos y las entidades.

Otro proyecto interesante y pionero en la detección de entidades y eventos ha sido el proyecto Perseo (**www.perseus.tufts.edu**). Sus investigaciones en cuanto al reconocimiento de entidades, comenzaron reflexionando acerca de las necesidades que debería cubrir una biblioteca digital en cuanto a la información disponible, su organización y tendencias en las búsquedas de los usuarios (Crane et al, 2001) y para ello analizaron sus colecciones con el fin de preparar las marcaciones de metadatos para facilitar las búsquedas y desarrollar interfaces para visualizarlas. Posteriormente se enfocaron hacia el desarrollo de herramientas de búsqueda y heurísticas para vincular documentos teniendo en cuenta palabras claves y efectuar un proceso de etiquetado de nombres y lugares, además de la construcción de "listados de autoridades" basados en los registros de búsquedas, que debían facilitar el proceso de "desambiguación" de entidades (Crane y Wulfman, 2003). Estas actividades motivaron varios trabajos de reconocimiento de entidades para textos de las bibliotecas digitales y el reconocimiento de eventos en los mismos tales como Smith y Crane (2001) donde se muestran técnicas para desambiguación de entidades geográficas y los trabajos de Smith (2002a, 2002b) en donde se describen los métodos para detectar eventos en textos desestructurados y que se han convertido en trabajos clásicos en este tema.

El reconocimiento de entidades geográficas (lugares) tiene un especial interés para las bibliotecas digitales y fue de hecho, de las primeras actividades que se llevaron a cabo en el área de extracción de información, ya que se pensó en la posibilidad de georreferenciar lugares automáticamente en mapas y construir interfaces de usuario para visualizarlas. Las dificultades relacionadas con este tipo de entidades en cuanto a su desambiguación radican en que muchos lugares en América tienen iguales nombres en Europa y así mismo se pueden tener nombres iguales en un mismo país pero en diferentes regiones. Un primer trabajo relacionado con la tarea de identificar entidades geográficas fue desarrollado también en el marco del proyecto Perseo (Smith y Crane, 2001), en donde se identificaban las entidades temporales por medio de algoritmos de aprendizaje y heurísticas y se

efectuaba la desambiguación de nombres por medio del análisis del contexto (utilización de anáforas) y la utilización de varios elementos de conocimiento externos tales como "*gazetteers*" (diccionarios topográficos) y tesauros. Para efectuar la desambiguación localizaban los posibles nombres y de acuerdo a la frecuencia de aparición en el texto y su cercanía con otros topónimos (para ello definieron una ventana de aproximación) se definía una tabla para asignación de puntajes y el puntaje mas alto era el ganador. Este trabajo fue interesante porque permitió definir con rigurosidad varios criterios de recuperación de información para desambiguación de entidades espaciales. Otros trabajos sobre reconocimiento de entidades geográficas y la marcación de zonas geográficas de interés en mapas (Leidner et al, 2003) efectúan la desambiguación de entidades utilizando dos aproximaciones minimalistas a saber: la primera es llamada "referencia del discurso", y asume que si un lugar mencionado en el discurso se refiere a la misma localidad a lo largo del discurso, justamente esa palabra es asumida para ser usada con el mismo sentido a lo largo del discurso. La segunda heurística asume que, en casos donde hay más de un lugar mencionado en el texto, la región más cercana al lugar de interés es la que puede dar la interpretación para el conjunto (resolver ambigüedades por proximidad). Actualmente, las técnicas heurísticas para desambiguar lugares requieren la utilización de listados que proporcionen las coordenadas geográficas además de la extracción de información del texto (Leidner, 2007).

En cuanto al reconocimiento de nombres, este ha sido un tema que recientemente ha cobrado mayor importancia debido a que se ha tratado de encontrar mecanismos para el refinamiento de búsquedas en las bibliotecas digitales teniendo en cuenta que las tendencias de búsqueda de los usuarios de las bibliotecas digitales que provienen de las humanidades muestran un gran componente en la búsqueda de nombres propios (Buchanam et al, 2005) y porque también la identificación de nombres es un paso previo para la construcción de ontologías o de otros sistemas de interés como software para hallar genealogías, aparte del boom que ha provocado la Web y sus aplicaciones asociadas. Sin embargo, en las bibliotecas digitales y los textos históricos, los mecanismos utilizados para identificar entidades de personas son básicamente los mismos usados para detección de

lugares: la utilización de tesauros, diccionarios, listados de nombres y técnicas heurísticas basadas en los documentos y el idioma. Un trabajo interesante con relación a la identificación de nombres en bibliotecas digitales ha sido el análisis de los textos de periódico del s. XIX (Crane y Jones, 2006). Este trabajo es importante porque analizan las anotaciones de las entidades encontradas en textos históricos (periódicos del s. XIX) y describen las dificultades encontradas en el proceso, tales como la desambiguación de nombres geográficos y de personas. Otras propuestas de estudio con relación a la identificación de entidades en las bibliotecas digitales y textos históricos son, por ejemplo, la identificación de nombres con respecto a las conjunciones en el texto (Dale y Mazur, 2007) en donde se obtuvo un porcentaje de precisión y cobertura del 83% y 84% respectivamente o el estudio de reconocimiento de entidades en textos históricos utilizando "anidamientos" entre entidades con el fin de encontrar un nivel mínimo e inferir las reglas para encontrar otras entidades (Byrne, 2007), en donde se obtuvieron unas medidas de precisión y cobertura del 77% y 75%. Otros trabajos relacionados con extracción de entidades son, la utilización de algoritmos de agrupación para el análisis de referencias de un autor y sus citas bibliográficas (On et. al., 2006) o el reconocimiento de entidades para catálogos de obras de arte (Davis et al, 2003) en donde utilizan anáforas para el reconocimiento de entidades. Aunque el interés actual se concentra en la detección de búsquedas de personas en Internet utilizando por ejemplo, sistemas de correferencias entre páginas (Kalashnikov, 2008) todavía el análisis de entidades en bibliotecas digitales y en documentos históricos es un área por explorar.

El etiquetamiento de las "fechas" presentes en los documentos usualmente se lleva a cabo por medio de expresiones regulares y técnicas heurísticas que son relativamente sencillas de identificar aunque, como toda tarea de extracción de información, depende del contexto que no siempre es uniforme y presenta particularidades. Un buen trabajo acerca del reconocimiento de fechas en textos históricos utilizados en una biblioteca digital es el que muestra Mckay (2000) que describe un análisis de fechas para la biblioteca digital *Greenstone* y otro trabajo de interés es el de Mani y Wilson (2000) donde describen las heurísticas y el sistema desarrollado para identificación de fechas y expresiones de tiempo

en periódicos. El reconocimiento de las fechas también fue el inicio de otras tareas más complejas como son la detección de eventos y la identificación de relaciones temporales y por esta razón son simultáneos a la extracción automática de entidades. La identificación de eventos motivó la posibilidad de analizar nuevas metodologías para encontrar relaciones temporales presentes en el texto y ésta tarea no es fácil -sobre todo cuando se pretende describir su evolución- porque conlleva conocimiento lingüístico intrínseco y la utilización de fuentes de conocimiento. Así mismo, ha propuesto dos aspectos importantes de análisis: encontrar metodologías empleadas para detección de eventos como tal y proponer modelos de visualización apropiados y útiles para el usuario. Estas metodologías se expondrán en la próxima sección.

1.2.2 Reconocimiento de eventos. Desde la década de los noventas se ha tenido gran interés por la detección automática de sucesos en noticias como una tarea derivada de la recuperación de información. Esto dio origen a los sistemas TDT (*Topic Detection and Tracking*) cuyo objetivo es la identificación automática de eventos en flujos de noticias digitales. Este sistema utiliza clasificación automática de palabras y los temas son definidos como eventos específicos o "algo que está sucediendo" en un lugar en un "cierto tiempo" que usualmente es muy corto -la medida suele ser días- (Allan et al, 1998). Uno de los grandes problemas planteados en este campo consiste en determinar si un documento entrante está informando sobre un nuevo suceso o si se relaciona con otros ya recogidos por el sistema. En caso de que un documento no informe sobre un nuevo suceso el siguiente problema a resolver consiste en determinar a qué suceso pertenece. Otros problemas adicionales planteados son determinar las historias que conforman los distintos sucesos y descubrir las posibles relaciones entre ellos (Pons et al, 2001). Aunque los sistemas de TDT son empleados para detectar nuevas noticias su funcionamiento no es apropiado para hacer detección y seguimiento de eventos en textos históricos o en los textos manejados por una biblioteca digital debido a varias razones: en primer lugar puesto que el sistema TDT está enfocado a datos de noticias el tipo de etiquetamiento para entidades no es exhaustivo: los nombres propios no son manejados como entidades sino que son tratados como simples palabras porque lo que realmente tiene importancia para

este sistema es la detección de noticias nuevas o las relaciones que pueden ser encontradas entre ellas. Sin embargo las técnicas de TDT están ya incorporando a sus mecanismos de evaluación el reconocimiento de entidades tales como personas y lugares para asignarles peso en la evaluación de eventos y tener mayor pertinencia en cuanto a la clasificación por temas (Makkonen et al., 2002, citado por Kumaran y Allan, 2005) o a la clasificación en la detección y seguimiento de nombres en idiomas diferentes (Florian et al, 2004) pero todavía es tema de investigación. En segundo lugar, los sistemas TDT tampoco permiten hacer seguimiento para eventos en periodos muy largos de tiempo; usualmente el corpus que se analiza para sistemas TDT son noticias que tienen una redacción y estilo característicos (son concisas y con una redacción orientada hacia la descripción de hechos) y su "tiempo de vida medio" perdura a lo sumo una semana. Un texto histórico o un texto de una biblioteca digital es más discursivo y tiene otras características que hacen que el análisis sea diferente (Smith, 2002a, 2002b). Un primer aspecto de diferencia es que las bibliotecas digitales ofrecen una mayor riqueza en su contenido porque están compuestas por diferentes documentos que tienen formatos diversos y que no presentan una estructura documental estándar y esto hace que la estructura de metadatos en las bibliotecas digitales sea más compleja, no sucediendo así con los periódicos digitales que tienen estructuras que facilitan el etiquetamiento y la detección de eventos porque los títulos proporcionan información del suceso. En segundo lugar la información de las bibliotecas digitales o los textos históricos tampoco tienen las facilidades que presenta el "estilo periodístico" en cuanto a su redacción concisa. Los documentos históricos (como el utilizado por ejemplo en este trabajo de investigación) suelen ser totalmente discursivos. De acuerdo a Smith (*op. cit.*) los documentos históricos tienen 3 características esenciales: discursividad, digresión, y escalamiento. La discursividad tiene que ver con el estilo de los textos, es decir los textos históricos manejan un mayor discurso en sus palabras y no están divididos en unidades más discretas como si lo están las noticias. En segundo lugar las historias descritas en los documentos históricos tienden a ser digresivas (más dispersas) y a tratar diversos temas en sus escritos a diferencia de los textos de noticias que utilizan una redacción más precisa. Un documento histórico frecuentemente es mas disperso acerca de los eventos antes o después de un periodo de tiempo principal. Finalmente, los

documentos históricos tienen otra escala de tiempo más larga que las noticias y es necesario estudiar no solo la fecha de aparición de un suceso sino también emplear otros componentes lingüísticos que describan cambios temporales.

Teniendo en cuenta estos aspectos, para encontrar eventos en textos históricos ya sean suministrados por una biblioteca digital o por un historiador se debe tener una metodología de trabajo diferente, que debe considerar las características de los textos analizados porque los hallazgos de información dependen fuertemente del dominio, debe hacerse una tarea de identificación mas exhaustiva de entidades presentes en el texto y la aplicación de otros algoritmos o técnicas heurísticas para la identificación de eventos diferentes a los empleados en los sistemas TDT. Las técnicas empleadas para encontrar eventos para textos históricos se mencionan a continuación:

Smith (2002a) propuso una caracterización de las colocaciones encontradas en el corpus por medio de ventanas de asociación entre las diferentes entidades, el número de ocurrencias que contienen y verificando por medio de técnicas estadísticas la relación de asociación entre los lugares y las fechas. Este método mostró alto rendimiento en cuanto a la detección de eventos pero no alcanza a mostrar su evolución aunque es una buena aproximación del problema y en principio, no depende del idioma empleado porque el análisis fue estadístico y se efectuó sobre las entidades y las fechas sin considerar expresiones de tiempo más complejas.

Otros trabajos relacionados con la detección de eventos están relacionados con planteamientos de modelos de extracción de información temporal para la asignación automática de periodos de tiempo teniendo en cuenta el contexto y la “granularidad” de las palabras que implican procesos temporales. Como esta práctica es mucho más compleja porque se utilizan las palabras del idioma para asignación temporal y un análisis de los verbos, usualmente el método empleado es anotar a mano una porción de corpus y con base en esas anotaciones comenzar a analizar las expresiones temporales y el contexto para obtener las nuevas reglas de anotación. Mani y Wilson (2000) comenzaron a aplicar heurísticas y marcación automática de expresiones temporales en periódicos y obtuvieron una buena anotación en cuanto a la precisión y cobertura (valores cercanos a 96%) sin obtener aún relaciones temporales explícitas aunque encontraron varios modelos para

establecer las relaciones de los verbos con relación a la posición de las palabras indicadoras de expresiones de tiempo. Otro trabajo interesante fue el efectuado por Llidó et al (2001). Básicamente extraen fechas y se recupera el contexto en caso de encontrar patrones de tiempo ambiguos para determinar las fechas en las expresiones adyacentes a la palabra de interés, de esta manera si la fecha no incluye el año se buscaba la referencia mas próxima al año en el texto o si solo contenía el día se buscaba la fecha con mas información cercana y así con otros cuantos casos particulares ya que la mayoría de eventos necesitan fuentes de conocimiento más complejas para reconocerlos. En este modelo únicamente analizan fechas pero no otro tipo de entidades para hacer relaciones entre si. Analizaron 4 periódicos que contenían 1634 expresiones de tiempo y obtuvieron unos valores de recuperación de 96% en precisión y 95% en cobertura, aunque, para casos particulares. Otro trabajo interesante es el efectuado por Jones y Thompson (2003) que trata sobre la detección de eventos similares definiendo una metodología para detectar similitudes por medio de extracción de información en tripletas que contienen el nombre presente en la oración, el verbo y el tema del evento y se hacen comparaciones entre diferentes oraciones que contienen eventos para determinar su similitud. El problema surgió a la hora de la evaluación, que fue efectuada por medio de comparaciones entre humanos y el sistema. Aunque pudieron establecer similitudes entre dos oraciones no es posible hacerlo para tres y tampoco obtener una evaluación cuantitativa sino más bien cualitativa. Este método empleado es muy similar en cuanto a la organización de la información, al utilizado por Smith (2002a).

Al respecto de la construcción de corpus para estudiar relaciones temporales es importante mencionar el trabajo de Setzer y Gaizauskas (2002) porque se convertido en uno de los clásicos con relación al tema. Ellos plantearon la necesidad de construir un corpus piloto para hacer anotaciones de eventos, establecer relaciones entre ellos y utilizar teoría de grafos para su representación con el fin de mostrar información sobre todo de eventos que pueden llevar a la identificación de otros. Utilizaron heurísticas y el contexto para identificar expresiones de tiempo y efectuaron pruebas sobre corpus pequeños pero no hicieron pruebas de precisión y recuperación sobre los datos ya que debían refinar el

anotador y el reconocimiento de entidades que todavía tiene un gran componente humano. Realmente ahí radica una de las dificultades que se tienen en la EI para la medición de eventos y es justamente ¿cómo medirlos? ya que muchas de las técnicas que se utilizan deben ser comparadas con juicios emitidos por seres humanos. El problema de detectar eventos tiene la gran dificultad de que los anotadores humanos pueden expresar las relaciones equivalentes de eventos de diferentes maneras por ejemplo, pueden decir que un evento *e1* sucedió durante otro *e2* y que *e2* sucede antes de *e3*, dejando implícito que e1 también está antes de *e3*, mientras otros pueden listar las relaciones de manera explícita. Una opción puede ser preguntar por las relaciones entre todos los pares de eventos en un texto dado pero esto podría demandar muchos seres humanos haciendo este análisis, otra alternativa es utilizar reglas de inferencia para determinar eventos. (Muller y Tannier 2004, 2008) utilizan reglas de inferencia basadas las relaciones de Allen (Allen 1984, citado por los autores) para definir relaciones temporales y comparar las sentencias (con respecto a esas reglas) de las anotaciones humanas para obtener algún tipo de medida. Estos autores han planteado un modelo de inferencia para eventos utilizando la anotación de las fechas, el contexto de las relaciones temporales y así mismo han planteado un modelo de evaluación para la medición de los eventos ya que el modelo de precisión y cobertura no es suficiente. El modelo propuesto tiene 2 componentes: el primero se llama "finura" ("*finnesse*") que consiste en el promedio de los eventos que tienen información concordante con las anotaciones humanas o en otras palabras, que tanto se parece la información encontrada por el sistema a las anotaciones humanas. Luego proponen la "coherencia" que es una medida de cuán precisa es la información dada por el sistema (Muller y Tannier, 2004). Para efectuar la evaluación de la "evaluación" efectuaron varios experimentos utilizando el corpus *TimeBank* en inglés e hicieron varias comparaciones de desempeño con relación a la cobertura "estándar" y los modelos propuestos de finura y coherencia (Muller y Tannier, 2008). En este trabajo, encontraron que la información temporal es más global y que unos eventos son más importantes que otros y que la medida de la cobertura es inmune al tamaño del texto analizado, sin embargo estos experimentos fueron aplicados a textos de noticias y es necesario verificar si sigue igual comportamiento en otros tipos de textos y en mayor cantidad.

En general, los trabajos actuales sobre la identificación de eventos tienden a utilizar heurísticas relativas al esquema de etiquetamiento para las expresiones temporales del idioma, la determinación de reglas para inferir los eventos posibles en el texto con base en la gramática de los verbos y el contexto, la construcción de un corpus apropiado para la medición de eventos temporales y la utilización de mediciones apropiadas para determinar el desempeño del sistema aunque es todavía tema de investigación y usualmente lo que se utiliza son todavía las medidas de precisión y cobertura tradicionales.

1.2.3 Visualización de eventos. Otro aspecto sobre el análisis de entidades y la detección de eventos tiene que ver con los sistemas que se están desarrollando actualmente para facilitar el proceso de visualización de la evolución de los eventos llegando a lo que se conoce como "*TimeLines*", que son diagramas de tiempo en donde se puede mostrar la concentración de sucesos utilizando la información de las bibliotecas digitales. Algunos trabajos interesantes son por ejemplo, Petras *et al* (2006) en donde describen una infraestructura para análisis de metadatos llamada *Time Period Directory* la cual recupera los metadatos de fechas, lugares y personas tomados de los documentos de la biblioteca del Congreso de EE. UU y que son visualizados en mapas y en páginas web que describen el evento. Otro trabajo interesante con relación a la visualización es el de Christel (2006) que estudia una manera de visualizar rangos de fechas con respecto a los documentos analizados en una biblioteca digital mostrando un intervalo de tiempo en el eje X y las características de los documentos relacionados con dichas fechas en el eje Y. Sin embargo, la visualización todavía es tema de investigación porque depende en gran medida de los requerimientos de búsqueda del usuario y también es importante encontrar un equilibrio entre la información suministrada por los sistemas de RI y de EI.

CAPITULO 2

PROCESO DE CONSTRUCCIÓN DEL CORPUS UTILIZADO EN LA INVESTIGACIÓN (INICIOS EN LINGÜÍSTICA DE CORPUS)

"Ahora ven, escríbelo en una tablilla,
grábalo en un libro y que dure hasta el último día, para testimonio." Isaías 30:8

"He extendido esta carta más de lo usual,
ya que no tengo el tiempo necesario para hacerla breve." Blaise Pascal

Palabras Clave. corpus históricos, fraseología, lexicografía.

Este capítulo describe el proceso de construcción del corpus utilizado para esta investigación y sus características.

2.1 ELABORACIÓN DEL CORPUS

2.1.1 Descripción del corpus. Las razones por las cuales era necesario elaborar un corpus para este proyecto de investigación fueron dos: la primera razón fue la necesidad de tener el texto en un formato digital en donde serán aplicados los algoritmos y elementos de cálculo necesarios para las tareas de extracción de información y la segunda es la necesidad de conocer la estructura documental de los textos utilizados por los historiadores. El conocimiento de la estructura documental de los textos es necesario para proceder con el reconocimiento de entidades tal como se mencionó en el capítulo 1, ya que los sistemas de extracción de información utilizan los metadatos que dependerán del contenido de los documentos (Crane et al., 2001). El corpus que se ha elaborado está conformado por la documentación epistolar de la curia en Antioquia (Colombia) para el período comprendido entre 1869 a 1880. Los documentos han sido recopilados y suministrados por el historiador Pb. Dr. Iván Darío Toro de la Universidad Luís Amigó en sus investigaciones relacionadas con la historia de la Iglesia Católica en Colombia. El

corpus electrónico está conformado actualmente por 224 documentos y tiene un total de 114.085 palabras. Tomando el archivo que contiene mayor cantidad de documentos (CorpusSac1874), el promedio de palabras por documento es de 551 palabras, siendo el rango de 231 palabras mínimo y un máximo de 1684 palabras para los documentos contenidos.

2.1.2 Proceso de elaboración del corpus y sus dificultades. Una de las primeras dificultades que se encontró a la hora de elaborar el corpus y la tesis en general, fue encontrar poca bibliografía relacionada con el tema (prácticamente nula en Medellín) y el desconocimiento de la metodología de trabajo en la lingüística computacional. Sin embargo con el apoyo de la Universidad EAFIT fue posible efectuar una pasantía en el laboratorio de PLN de la UNED en Madrid (España) en diciembre-enero de 2005 por parte de la autora. En dicha experiencia el Profesor Julio Gonzalo hizo algunas aclaraciones y recomendaciones relacionadas con la metodología y la temática de la lingüística de corpus y fue enfático en afirmar que lo primero que debía tenerse en cuenta para iniciar el trabajo era justamente la elaboración del corpus. De regreso a Medellín se procedió a la elaboración del corpus y el primer paso lógico fue analizar los tipos de documentos que tenía el historiador y hacer la selección respectiva. Puesto que el Profesor Iván Darío Toro se encontraba dirigiendo un proyecto de digitalización de la Gaceta de Antioquia (el periódico de gobierno de Antioquia en el s. XIX) en el Archivo Histórico del Palacio Nacional, se analizaron -con su permiso- los documentos digitalizados para hacer unos ensayos con OCR y examinar la posibilidad de construir el corpus con ellos pero desafortunadamente el reconocimiento de OCR fue poco eficaz además de que era necesario ir al Archivo Histórico para obtener los documentos y se presentaron otras dificultades relacionadas con la manipulación de los archivos. Se procedió entonces a examinar otra documentación disponible por parte del historiador y éste proporcionó de manera generosa algunos de sus documentos recopilados: fotocopias de las cartas manuscritas del s. XIX de la curia. Estos documentos estaban previamente organizados por temas y estaban más disponibles para comenzar el trabajo de transcripción en medio electrónico. Entonces se decidió comenzar la elaboración del corpus con estos documentos.

Construcción del corpus: Para efectuar el análisis del corpus era necesario construirlo de manera digital, esto es, obtener el contenido en texto que pueda ser manipulado por las aplicaciones informáticas. Puesto que la mayoría de los documentos son manuscritos, se ha requerido un esfuerzo extra para su digitalización ya que el proceso de reconocimiento de patrones a través de OCR no fue eficaz. En su momento (entre el 2005 y 2006) se ensayaron varias herramientas de OCR para la conversión de los documentos manuscritos y no se tuvieron resultados satisfactorios porque el OCR no reconocía muy bien las letras y era poco práctico hacer la corrección ya que demoraba bastante tiempo. Adicionalmente los OCR presentaron otro inconveniente y es que muchas palabras del corpus están en desuso y es necesario que el OCR contenga palabras para español antiguo para reconocerlas. De este modo se llegó a la conclusión de que la utilización de OCR en el proceso de digitalización es un método impráctico para efectos de este proyecto ya que se consumía mucho más tiempo en la revisión de los textos que la transcripción directa de los mismos. Otra opción que se consideró fue analizar la posibilidad de que los estudiantes de Comunicación Social de la Universidad EAFIT colaborasen en el proceso de transcripción de textos, pero este procedimiento tampoco fue exitoso porque no se tenia suficiente disponibilidad por parte de los estudiantes y a nivel de construcción del corpus es conveniente que pocas personas lo elaboren ya que el porcentaje de errores es más pequeño. Entonces teniendo en cuenta estos inconvenientes, la autora se propuso transcribir el corpus con ayuda de otra persona. El hecho de conocer el corpus es una ventaja porque a medida que se iba transcribiendo fue posible identificar muchos patrones de reconocimiento de entidades que podían ser programados en el sistema y también tiene la ventaja de dar a conocer las particularidades lingüísticas del corpus.

En cuanto a la construcción del corpus como tal, uno de los aspectos a tener en cuenta en la transcripción de los documentos fue el tipo de editor de texto a emplear. En este tipo de trabajos no es deseable tener un editor de textos avanzado como el *Word* (herramienta y marca registrada de *Microsoft*) porque estos editores hacen corrección automática de errores y una de las condiciones que requiere el trabajo de transcripción textual es la fidelidad al texto original que puede contener palabras que puede ser consideradas por el

editor como errores de ortografía o errores gramaticales. Entonces se procedió a buscar otros editores de texto que evitaran estos inconvenientes. En principio se ensayaron los editores *WordPad* y *NoteBook* (herramientas y marcas registradas de *Microsoft*) pero tampoco fueron muy exitosos porque son herramientas muy limitadas en cuanto a que no tienen muchas herramientas de búsqueda y manipulación de los textos. Finalmente se decidió emprender la transcripción y elaboración del corpus con el programa *TextPad* en su versión de evaluación para idioma español (http://www.textpad.com). Otros editores de texto son EditPlus (http://www.editplus.com/), UltraEdit (http://www.ultraedit.com/) y ConText (http://www.contexteditor.org/). El TextPad tiene como ventaja la posibilidad de construir plantillas para insertar las etiquetas personalizadas requeridas en la elaboración del corpus como son por ejemplo, <DOC> y </DOC> que son etiquetas construidas para indicar cuándo empieza y termina un documento. Otras ventajas que ofrece el *TextPad* son la ausencia de corrección automática de errores y la facilidad de obtener diferentes tipos de búsquedas, como por ejemplo las búsquedas de palabras en varios documentos al tiempo generando un índice que muestra las ocurrencias de la palabra por documento incluyendo su contexto, lo cual es deseable para la construcción de diccionarios de arcaísmos y otros tipos de sub-productos que pueden ser usufructuados del corpus.

Otra dificultad que se presentó fue la estandarización de los errores y las palabras ilegibles presentes en los textos. Después de una búsqueda de textos relacionados con estos temas –que desafortunadamente son muy escasos en nuestro medio- se optó por utilizar los estándares de paleografía del Archivo Histórico de la Nación (Ladrón, 1996). Las etiquetas de paleografía fueron implementadas en *TexPad* como se mencionó anteriormente.

Y finalmente, otro escollo a superar fue el formato del corpus como tal, esto es ¿cómo es el formato documental de un corpus utilizado para análisis de EI? Para determinar las etiquetas y la forma del corpus, desde el principio de este trabajo se contó con el corpus en español utilizado para las pruebas TREC - *Text REtrieval Conference* (LDC, 2000) suministrado por el Profesor Juan Guillermo Lalinde. Las conferencias TREC son relativas a temas relacionados con recuperación de información y son patrocinadas por el Instituto Nacional de Normas y Tecnología de EE. UU (NIST) y su Departamento de Defensa.

Tienen como objetivo fomentar y apoyar la investigación en la recuperación de documentos a gran escala (http://trec.nist.gov/). El corpus contiene un conjunto de documentos utilizados para las tareas de evaluación en el TREC y consiste en aproximadamente 250 MB del textos del periódico mexicano El Norte y 200 MB de noticias de AFP (*Agence France Presse*) para el año de 1994 en idioma español. Los documentos de El Norte fueron usados para el TREC 3 y 4 y los documentos de la agencia AFP fueron utilizados para el TREC 5. Los 598 archivos de texto de AFP contiene 172.952 documentos y los 194 archivos de Infosel contiene 57.868 documentos y en promedio los archivos tienen un tamaño de 533KB. Este corpus tiene diferentes etiquetas construidas en lenguaje SGML (*Standard Generalized Markup Language*) y también incluyen los respectivos formatos DTD. Cada documento está etiquetado con las marcas <DOC></DOC> y contienen un único identificador etiquetado con las marcas <DOCNO></DOCNO>. Otros datos que contiene el encabezado son el número de palabras, idioma. Para separar párrafos utiliza las etiquetas y . En este trabajo, el corpus TREC ha servido como muestra para la construcción inicial del corpus histórico y también como corpus de contraste. La siguiente figura muestra una imagen del corpus de noticias:

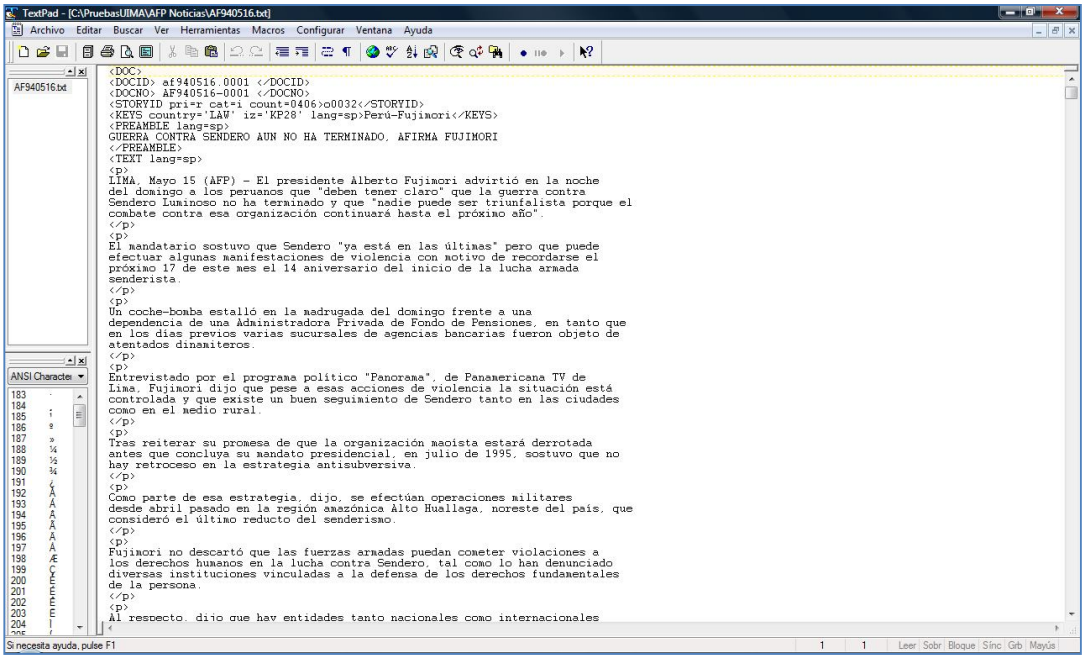
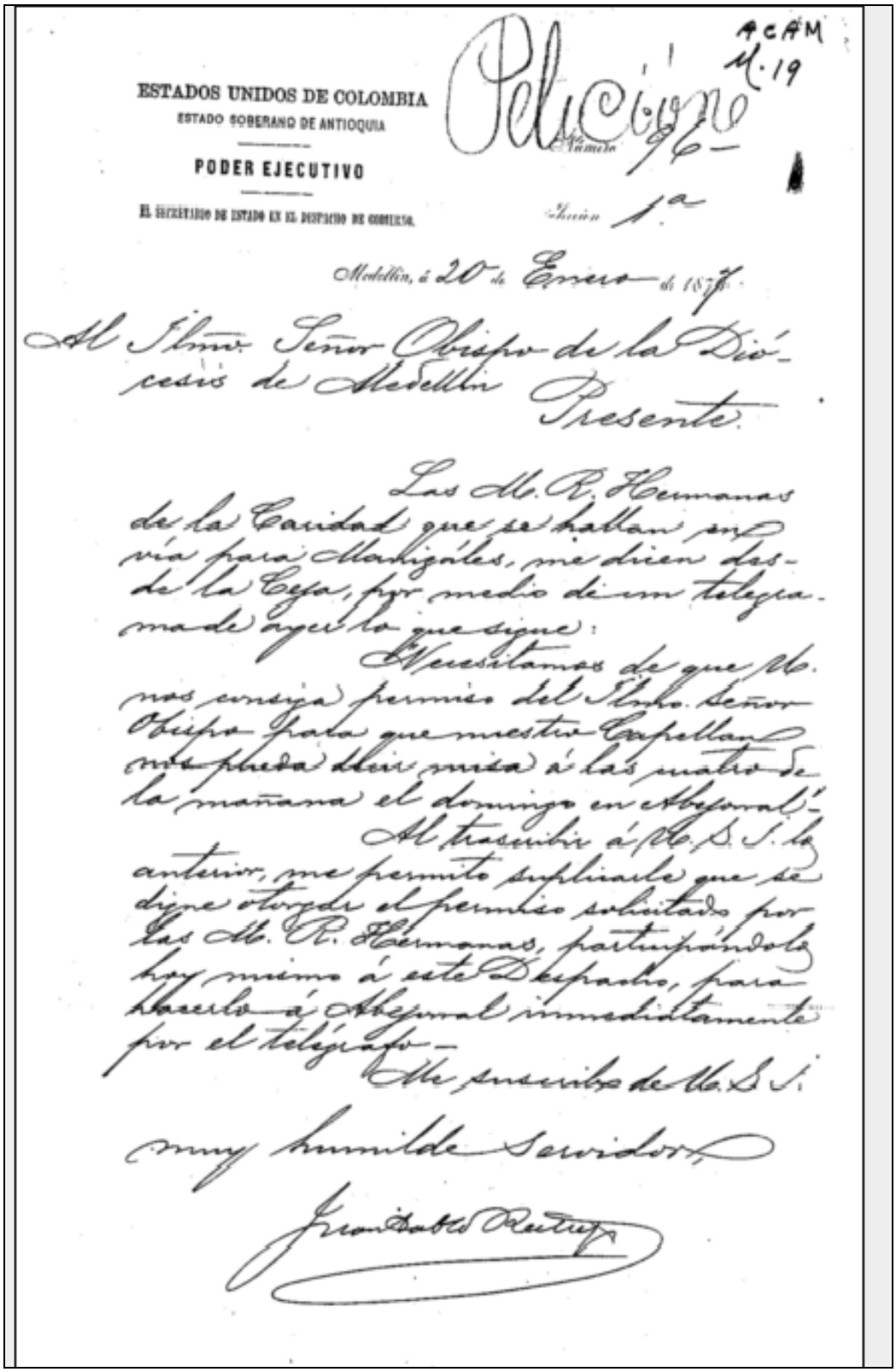


Figura 2.
Corpus de noticias AFP utilizado para el TREC (LDC, 2000)

Como se mencionó, el corpus utilizado para las conferencias TREC fue tomado en principio como modelo para la elaboración del corpus de sacerdotes (los documentos de obispos no se alcanzaron a incluir en este formato). El corpus histórico tuvo una configuración de etiquetas similar, solo que se pusieron las palabras clave del historiador, número de palabras por texto y la fecha estándar. Este ejercicio fue un poco arduo pero finalmente grato porque las etiquetas de los párrafos fueron útiles a la hora de analizar los eventos en los textos (tema que se aborda en el capítulo 4). De esta manera se etiquetaron 192 documentos. Posteriormente se construyó otra copia del corpus pero limpio, es decir solo conservando las etiquetas de inicio y fin del documento (<DOC> y </DOC>) y las de fechas para el corpus de cartas de obispos donde se incluyó la fecha del documento normalizada (<DATE>1874/01/13</DATE>). Este corpus limpio fue construido con el fin de facilitar el análisis en la identificación de los modelos de reconocimiento para los nombres de personas y lugares. En resumen: al principio del trabajo se construyó un corpus basado en el corpus de las conferencias TREC y que está conformado por 192 documentos. Luego en una etapa posterior del trabajo se construyó otro corpus que contenía los 192 documentos anteriores y otros nuevos pero que solo contenían las etiquetas de inicio y fin de documento y la fecha, es decir un corpus "limpio". Esta versión final del corpus tuvo un tamaño final de 224 documentos.

A continuación se muestra un ejemplo de los documentos manuscritos que conforman la colección del historiador y su respectiva transcripción a los corpus elaborados:



ACAM
M. 19

ESTADOS UNIDOS DE COLOMBIA
ESTADO SOBERANO DE ANTIOQUIA

PODER EJECUTIVO

EL SECRETARIO DE ESTADO EN EL DESPACHO DE GOBIERNO.

Peticiones
Número 1ª

Medellín, á 20 de Enero de 1877.

Al Ilmo. Señor Obispo de la Diócesis de Medellín
Presente.

Las M. R. Hermanas de la Caridad que se hallan en vía para Marinálas, me dicen desde la Ceja, por medio de un telegrama de ayer lo que sigue:

"Necesitamos de que Ms. nos consiga permiso del Ilmo. señor Obispo para que nuestro Capellan nos pueda decir misa á las nueve de la mañana el domingo en Abejoral."

Al trasmitir á Ms. S. S. lo anterior, me permito suplicarle que se digne otorgar el permiso solicitado por las M. R. Hermanas, participándolo hoy mismo á este Despacho, para hacerlo á Abejoral inmediatamente por el telégrafo.

Me suscribo de Ms. S. S.
muy humilde Servidor,

Francisco Restrepo

Figura 3.
Documento del corpus religioso, año 1877.

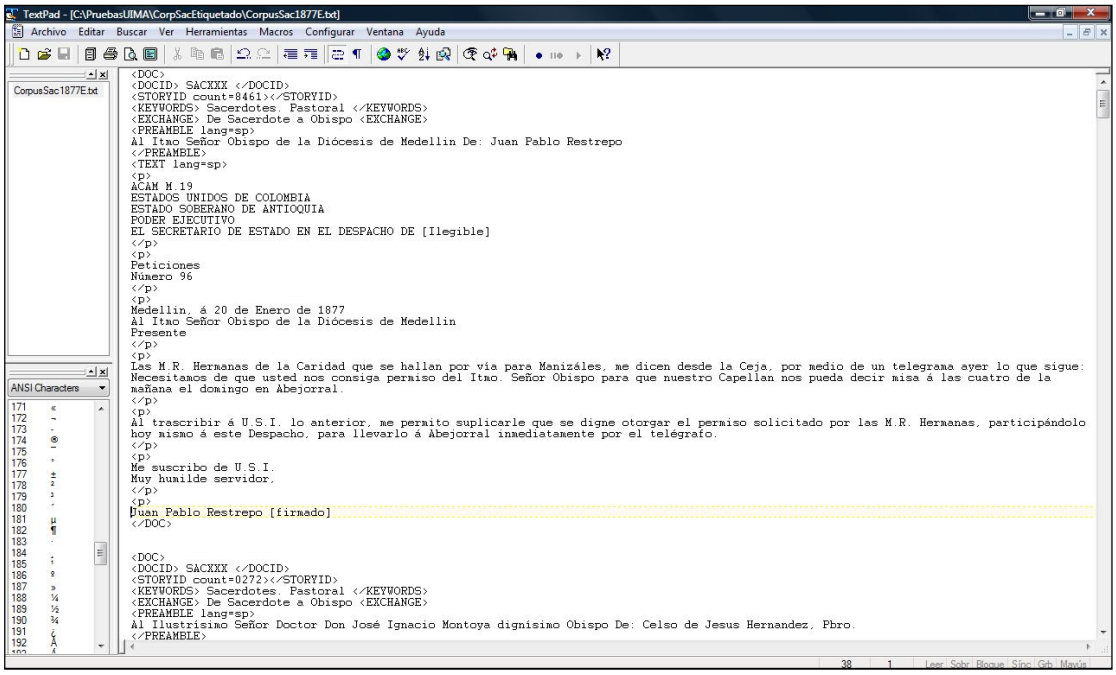


Figura 4.
El documento incorporado en el corpus de sacerdotes etiquetado de acuerdo al estándar del corpus de AFP, año 1877.

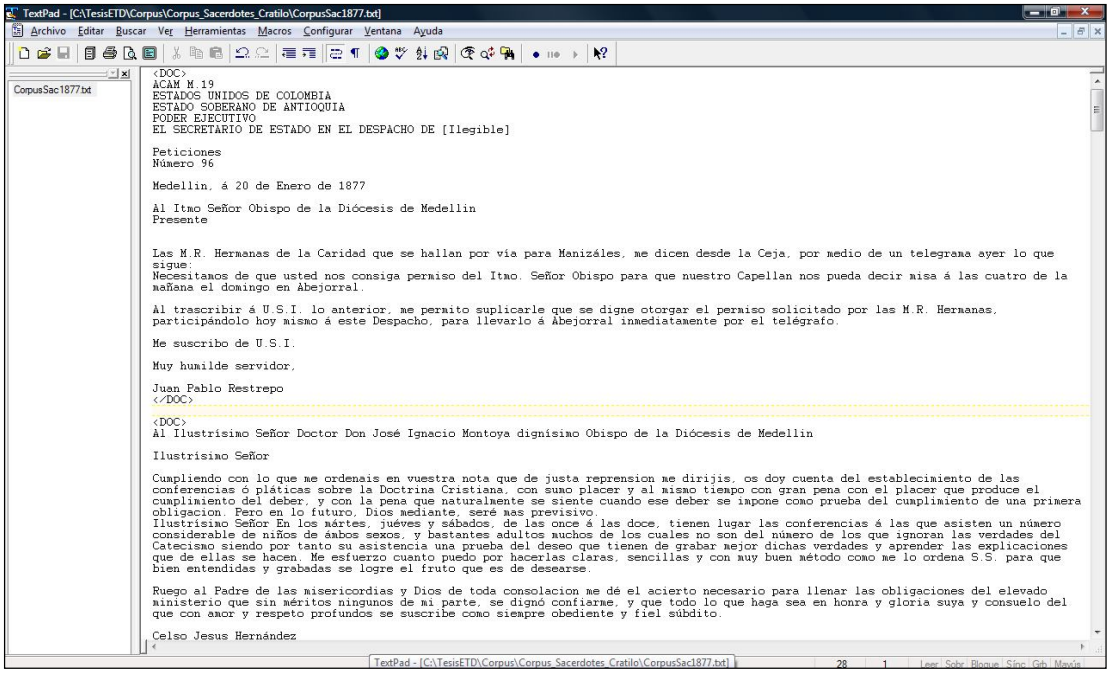


Figura 5.
El documento incorporado en el corpus de sacerdotes, año 1877.

2.1.3 Criterios de selección de los documentos del corpus. La selección de los documentos utilizados se realizó de acuerdo a la cantidad por año. Aunque el historiador tiene disponible en su colección física documentos comprendidos entre los años 1853 a 1988, se efectuó un conteo por año teniendo en cuenta la clasificación temática previamente asignada para determinar el periodo histórico en donde se encontraba la mayor cantidad de documentos. Esto es necesario para obtener una mayor cantidad de texto significativo de acuerdo a la época, teniendo en cuenta los criterios de representatividad y homogeneidad que debe presentar el texto para obtener realmente un corpus histórico apropiado para estudios lingüísticos (Tercedor, 1999; Torruela, 1999). Luego del conteo, se encontró que la mayor acumulación de documentos estaba en el periodo comprendido entre 1869 a 1880, tal como se muestra en el siguiente cuadro donde se muestra la distribución temática de documentos teniendo un total de 664 documentos físicos:

Año	Obispos	Sacerdotes	Política	Educación	Economía	Religiosidad	Geografía
1869	10	18	2	10	12	7	13
1870	12	9	4	2	2	1	1
1871	15	5	1	3	7	5	1
1872	10	16	1	8	10	9	0
1873	24	20	9	3	5	14	3
1874	12	30	12	6	4	11	2
1875	16	15	12	10	1	12	3
1876	2	14	14	15	0	3	2
1877	1	15	65	6	0	4	1
1878	4	14	31	3	4	3	0
1879	1	8	11	2	1	0	0
1880	4	10	1	2	0	5	0
Numero Documentos	**111**	**174**	**163**	**70**	**46**	**74**	**26**

Tabla 1.
Cuadro comparativo de la distribución de documentos disponibles para el corpus elaborado

Puesto que la mayor concentración de documentos estaba centrada en la temática “sacerdotes”, éstos fueron los primeros documentos que se comenzaron a transcribirse y posteriormente se continuó con los documentos de “santa sede” y “obispos” pero como los documentos de la santa sede eran tan pocos (8) se anotaron en el corpus de obispos. Con

respecto a la clasificación temática, a continuación se ofrece una corta descripción de cada una:

- Sacerdotes: pastorales, solicitudes al obispo para nombramiento de sacerdotes, resoluciones, etc.
- Santa Sede: documentos pontificios y nombramientos
- Obispos: pastorales, correspondencia entre obispos, resoluciones, etc.
- Política: asuntos políticos y las relaciones con la Iglesia.
- Educación: colegios, instituciones educativas, materias de seminarios.
- Economía: cuentas de la curia.
- Religiosidad: documentos descriptivos de manifestaciones religiosas.

Teniendo en cuenta la clasificación de los corpus y la naturaleza de los documentos que lo componen, este corpus es un corpus histórico que muestra el uso de la lengua española en Colombia durante el siglo XIX. Además tiene una particularidad y es que está compuesto por cartas de diferentes tipos de personas de la época y por lo tanto es representativo y puede ser utilizado para estudios lingüísticos más específicos tales como estudios pragmáticos, ortográficos, etc.

2.1.4 Organización del corpus elaborado. Los archivos del corpus se encuentran nombrados de acuerdo a su temática y de acuerdo al año. De esta manera se tienen 16 archivos de cartas de la temática Sacerdotes y 11 archivos de Obispos. Cada uno de estos archivos se encuentra ordenado cronológicamente. La siguiente tabla muestra la relación a nivel de tamaño y organización de los archivos que componen el corpus sin etiquetas (limpio) que fue utilizado para la caracterización de etiquetas semánticas:

Tipo de documento	Numero de archivos	Numero de documentos	Tamaño promedio (en bytes)	Tamaño total de archivo	Número de palabras	Tipo de archivo
Corpus sacerdotes	16	189	32,2 KB	506 KB	84483	Texto
Corpus obispos	11	35	16,3 KB	173 KB	29602	Texto
TOTALES	**27**	**224**		**679KB**	**114085**	

Tabla 2.
Cuadro comparativo de la relación de archivos del corpus a 29 de marzo de 2009.

Aunque en el momento se cuenta con 224 documentos transcritos, es posible aumentar aún más el tamaño del corpus porque el historiador cuenta con muchos más documentos disponibles pero sería necesario incluir recursos adicionales que faciliten esta labor. Además se consideró que el tamaño de la muestra era significativa considerando que son documentos históricos y de la misma época y por lo tanto servían de base para efectuar el trabajo de investigación.

2.1.5 Análisis del contenido de los documentos. El contenido de las cartas puede dividirse en tres tipos:

- Cartas de peticiones al obispo. Este tipo de correspondencia es interesante porque conforma un almacén de nombres y firmas de las personas corrientes de la época.
- Cartas informativas o aclaratorias acerca de algún hecho puntual.
- Cartas resolutivas o de temas de gobierno eclesiástico (Circulares, Pastorales y Oficios).

Otros documentos presentes en el corpus proceden de la Santa Sede, tales como Correspondencia del Papa o Nombramientos. En el corpus se incluyeron 8 documentos clasificados previamente por el historiador como "Santa Sede" correspondientes al periodo de interés (1869 a 1880) teniéndose 3 documentos pontificios y 5 nombramientos. Algunos de estos documentos utilizan imprenta (sobre todo los que tienen un carácter más imperativo, tales como comunicaciones de la Santa Sede o ciertas pastorales) y tienen como característica que incluyen textos en latín, los cuales se han ido recopilando para hacer un posterior análisis fraseológico.

2.1.6 Análisis de la estructura física de los documentos. Este corpus presenta un esquema uniforme en la estructura de los textos que lo conforman aunque se trata de diferentes tipos de documentos (cartas, oficios, circulares, pastorales, etc.). Las cartas presentan una estructura de fecha, nombre a quien se dirige, el texto, despedida y la firma pero tampoco es un formato uniforme aunque si ayuda para la identificación de fechas. Algunos ejemplos de la estructura de los encabezados del corpus se muestran a continuación:

Tipo de Documento	Encabezado
Oficio	Seminario Oficio Diócesis de Medellin i Antioquia Rectorado del Seminario Medellin, Mzo 8 de 1869 Al Illmo Sr. Obispo de la Diócesis
Documento notarial	ESTADOS UNIDOS DE COLOMBIA. Número 52 ESTADO SOBERANO DE ANTIOQUIA. PREFECTURA DEL DEPARTAMENTO DEL CENTRO. Medellin, 26 de agosto de 1869
Decretos Pontificios	DECRETO, MANDANDO PUBLICAR LAS LETRAS APOSTÓLICAS DE N.S.P. PIO IX DIRIJIDAS AL CLERO I PUEBLO DE ESTA DIÓCESIS, PARTICIPANDO EL NOMBRAMIENTO DE COADJUTOR DE ELLA, CON DERECHO A FUTURA SUCESION. Nos Valerio Antonio Jimenez, POR LA GRACIA DE DIOS I DE LA SANTA SEDE APOSTOLICA, OBISPO DE MEDELLIN I ANTIOQUIA…
Pastorales	ROMA ES DE LOS PAPAS, o CARTA PASTORAL QUE EL ILLMO, SEÑOR OBISPO DE Medellin i Antioquia DIRIJE A SUS DIOCESANOS en 8 de Febrero de 1871.
	Provincia Eclesiástica de Nueva Granada EN LA AMERICA MERIDIONAL DIOCESIS DE PANAMA Panamá, 21 de junio de 1871.
	Número ARQUIDIOCESIS DE SANTA FE DE BOGOTA Gobierno eclesiastico. Bogotá 18 de Diciembre de 1871 Ilustrísimo Señor Obispo de Medellin y Antioquia
Carta de feligreses a obispo	Medellin Fbro 19 de 1872 Al Illmo S. O. de la Diócesis D. Valerio A Jiménez. Ilmo Señor.

Tabla 3.
Algunos ejemplos de los formatos de las cartas del corpus

2.1.7 Tópicos futuros de investigación. Es posible ahondar en el estudio de la lingüística de corpus y efectuar otro tipo de trabajos relacionados con este tema. Algunos posibles temas para futuros trabajos son:

- Evaluación del cambio de los nombres y apellidos de personas y nombres de lugares: es un análisis de las variaciones que han sufrido algunos nombres encontrados en el corpus. Algunas variaciones son por ejemplo para lugares: Sanvicente, Donmatías, Piedrasblancas, Jirardota, Ansermaviejo.
- Evaluación de los cambios ortográficos de algunas ocurrencias de palabras
- Estudios pragmáticos de los discursos.
- Contraste del tipo de documentación utilizada con la actual.
- Análisis de las formas de tratamiento en cuanto a la sintaxis y la evolución del léxico, entre otras.

2.2 CARACTERIZACIÓN DEL CORPUS

El siguiente artículo relativo a las características lingüísticas del corpus fue enviado para el Congreso Internacional de la Asociación Española de Lingüística Aplicada (AESLA) efectuado del 3 al 5 de abril de 2008 en la ciudad de Almería (España):

Diseño y elaboración de un diccionario del léxico de un corpus de comunicación de 3 niveles (Obispos, sacerdotes y fieles) a partir de la documentación eclesiástica de archivos diocesanos en Colombia (1869 - 1880)

RESUMEN

Los corpus históricos constituyen una fuente de recursos lingüísticos útiles para el desarrollo de aplicaciones tales como la lexicografía, la terminología y la búsqueda translingüística de información. En este artículo presentamos y describimos el proceso de elaboración de un corpus histórico-religioso en Colombia entre los años de 1869 a 1880, el cual contiene los textos epistolares entre la Santa Sede, obispos, sacerdotes y fieles.

PALABRAS CLAVE

Corpus Históricos, fraseología, lexicografía, documentación eclesiástica, archivos diocesanos.

ABSTRACT

A historical corpus constitutes a source of linguistic resources. It is useful for the development of applications, i.e. lexicography and terminology, and for searching translinguistic information. In this article we discuss and describe the process of developing a corpus of historical and religious information in Colombia between 1869 and 1880, which contains the texts of letters between the Holy See, bishops, priests and faithful.

KEYWORDS

Historical Corpus, phraseology, lexicography, ecclesiastical documentation, diocesan files.

1. DESCRIPCIÓN DEL CORPUS

El corpus que actualmente estamos construyendo y que es la fuente de nuestro diccionario está conformado por la documentación epistolar de la curia en Antioquia (Colombia) para el período comprendido entre 1869 a 1880. Hemos escogido los documentos para este período histórico porque reunía la mayor concentración de documentos, lo cual favorece la representatividad y homogeneidad en el lenguaje de la época requeridos para lograr un estudio objetivo de sus características (Chantal 2002; Torruela y Llisterri 1999). Las cartas que conforman el corpus están ordenadas cronológicamente y se ha efectuado un análisis previo para agruparlas de acuerdo a conjuntos temáticos tales como la procedencia de las cartas (sacerdotes, obispos, santa sede) y temáticas tratadas (economía, política, culto, etc.). El corpus electrónico está conformado actualmente por 224 documentos y tiene un total de 114.085 palabras.

Puesto que se trata de documentos epistolares que se encuentran en su mayor parte manuscritos se ha requerido un esfuerzo extra para su digitalización, ya que el proceso de reconocimiento de patrones a través de OCR no siempre obtuvo resultados apropiados. Adicionalmente, en la transcripción de los documentos se ha utilizado los estándares de paleografía utilizados por el Archivo Histórico de la Nación (Ladrón 1996).

Los niveles de comunicación presentes en las cartas que conforman el corpus son:

- Correspondencia de sacerdotes al obispo: usualmente es la utilizada para resolver dudas con relación a temas eclesiásticos o comentar situaciones personales o de orden público en las parroquias. Algunas cartas tienen respuesta directa del obispo para solucionar los temas tratados.
- Correspondencia de los fieles a obispo: es utilizada por los vecinos y fieles para solicitar sacerdotes o elevar quejas.
- Correspondencia entre obispos: trata temas relacionados con asuntos de cooperación entre las diócesis.
- Correspondencia entre la Santa Sede y los obispos: temas relacionados con nombramientos, narración de sucesos entre Roma y la Nueva Granada.

1.1 CARACTERÍSTICAS DEL CORPUS

El uso del español utilizado en estas cartas es el característico del s. XIX en Colombia. Se observan diferencias con la morfología actual y en el uso ortográfico (ejemplos: aparece la "i" latina por la "y", la "s" por la "x", la "j" por la "g").

También se muestran varias formas ortográficas para una misma palabra puesto que el corpus está conformado por cartas de diferentes personas de distintas clases sociales. Sin embargo, el uso del español posee al mismo tiempo muchas semejanzas con el utilizado actualmente. La siguiente tabla muestra algunos fragmentos de cartas que ilustran las características citadas:

Fragmento	Procedencia Documento	Año
"La presente solicitud podria llevar las firmas de todos los habitantes de Salamina, con rarísimas excepciones, pero hemos creido que las que ella contiene son bastantes para probar la justicia que nos asiste"	Sacerdotes	1869
"Con el mayor respeto i consideracion manifiesto a S.S. Illma que desde el 20 de febrero procsimo pasado tomé posesion de este curato, i desde entonces estoi trabajando i ejerciendo mi ministerio como cura procurando cumplir fielmente con tan delicado cargo"	Sacerdotes	1869
"Es por esto dignísimo y Señor Obispo, que nos atrevemos, confiadas en vuestra lealtad, en dirijir nuestras suplicas, primeramente al Cielo y despues a voz dignisimo Pastor para que os digneis enviar a este pueblo, un Sacerdote Catolico, que enjugue las lagrimas que verten nuestros ojos al considerar tanta desgracia"	Sacerdotes	1870
"I ya que estais Illmo Sr. empuñando la bandera de la unidad, nosotros juramos esa bandera, al servicio de la cual ponemos nuestro prestijio i nuestra influencia i nuestro amor, elevando, como elevamos votos mui fervientes al Dios de las misericordias"	Obispos	1875

Tabla 1: fragmentos del corpus

1.2 ARCAÍSMOS

A medida que el trabajo de trascripción de los textos ha avanzado, se ha recopilado la información de los arcaísmos presentes en el corpus y las expresiones que son utilizadas por el gremio eclesiástico con el fin de elaborar diccionarios, teniendo en cuenta lo que propone Tercedor (1999) "La lexicografía y la terminología/terminografía, para que sean reales y fiables, deben basarse en córpora textuales representativos".

En la elaboración de los diccionarios se ha tenido en cuenta el contexto en el cual se utiliza la palabra ya que tiene un lugar importante la frase en donde se encuentra el término analizado (Orduña, 1999). Algunos ejemplos de arcaísmos encontrados se muestran a continuación:

Término original	Frase	Procedencia documento	Año
Egregias (Arcaísmo)	"movidos por vuestra Relijion i virtudes egregias"	Santa-Sede	1863
Fierro (Arcaísmo)	"Prevenimos i mandamos que no se ponga clavos de fierro ni ninguna clase de adornos en los altares que puedan dañar los estucados"	Obispos	1870
Fojas (Arcaísmo)	"el objeto sobre que ella versa i la foja del libro"	Obispos	1870
Fenecer (Arcaísmo)	"para examinar i fenecer las cuentas del Seminario"	Sacerdotes	1878
Letífero (Arcaísmo)	"dándoles letífero veneno en lugar de antídoto saludable"	Sacerdotes	1881

Tabla 2: ejemplos de las formas del diccionario

2. PRIMERA APROXIMACIÓN AL ANÁLISIS FRASEOLÓGICO

Tercedor (1999) en su escrito muestra que la fraseología presenta muchas variaciones en las cuales se encuentran expresiones fijas, proverbios y dichos. En efecto, en nuestro corpus encontramos dos tipos de conjuntos fraseológicos de interés para nuestro análisis: frases y dichos populares y tratamientos relativos a los saludos y despedidas.

2.1 FRASES Y DICHOS POPULARES

Debido al carácter epistolar del corpus se ha encontrado varios ejemplos de dichos populares, frases alusivas a las Sagradas Escrituras y algunas otras escritas en latín que provienen en su gran mayoría de las cartas de los sacerdotes a sus obispos. Estas frases tienen como objetivo ilustrar algún sentimiento concreto con relación a un tema en particular y reflejan en cierto modo, el lenguaje coloquial de la época y su utilización en el marco de la "confianza" con el destinatario que en este caso era el señor obispo o algún vicario. Algunas de estas frases son:

Procedencia	Frase	Año
Sacerdotes Dichos Populares	"De '*trecho en trecho'* fueron apareciendo en el mismo mes otras funciones de mas elevada alcurnia"	1875
Sacerdotes Dichos Populares	"No diga U.S.I. que '*ya las tórtolas estan tratando a las escopetas*'"	1877
Sacerdotes Dichos Populares	"Eso de '*ayúdate que Dios te ayudará'*, o de '*a Dios rogando y con el mazo dando'* ya era desconocido por aquí"	1878
Sacerdotes Dichos	"Espero en Dios, que el '*De profundis'* de ayer no tarde en convertirse en un jubiloso '*Sursum Corda'* mañana"	1878
Sacerdotes Dichos	"Como el hijo pródigo digo yó ahora, 'me levantaré é iré á la Casa de mi padre'"	1876
Sacerdotes Calificativos	"crecidos éstos hijos casi todos '*calaveras'"*	1874
Sacerdotes Calificativos	"Es verdad que hay un círculo de hombres sin religión, '*carituertos'* siempre para con los pobres clérigos"	1880

Tabla 3: ejemplos de fraseología del corpus

De especial interés son los *calificativos* ("calaveras", "carituertos") los cuales reflejan un tono despectivo y muestran un interés fraseológico por lo anómalo (Tercedor, 1999) y así mismo, algunas de las frases populares que todavía son utilizadas en Colombia ("Tórtolas tirando a las escopetas", "Ayúdate que Dios te ayudará" o "A Dios rogando y con el mazo dando").

2.2 SALUDOS Y DESPEDIDAS

En cuanto a los tratamientos sociales entre diversos niveles jerárquicos de la Iglesia, las expresiones utilizadas muestran respeto, autoridad y verticalidad. Tanto en el saludo como en la despedida de las cartas se dan: 22 maneras diferentes de saludar al obispo y más de 46 maneras diversas para despedirse. Algunos ejemplos se muestran a continuación:

A quien Se dirige	Saludo	Año
Santa Sede a obispo	"Illmo y Rmo Señor:"	1870
Obispo a obispo	"Mi muy respetado y venerado hermano:"	1874
Sacerdote a obispo	"Illmo sr. Obispo de la Diócesis."	1869
Fieles a obispo	"Ilustrisimo Señor Obispo de la Diocesis"	1869

Tabla 4: ejemplos de formas de tratamiento en el Saludo

A quien Se dirige	Despedida	Año
Santa Sede a obispo	"Dios guarte á V.S. Illma y Revda"	1870
Obispo a obispo	"Su atento hermano y amigo"	1876
Sacerdotes a obispo	"De Usía Ilustrisima Humilde subdito"	1877
Fieles a obispo	"Los que suscribimos pedimos a S.S.I. humildemente su bendicion"	1874

Tabla 5: ejemplos de formas de tratamiento en la despedida

La fraseología suele tener en cuenta los usos comunes de la lengua en el caso de los saludos y las despedidas que analizamos pero la verdadera riqueza fraseológica nace de la necesidad de expresar sentimientos de agradecimiento, bendición y respeto hacia una autoridad.

FUTURO DE ESTE TRABAJO

Se continuará en la elaboración del diccionario de términos religiosos y lingüísticos, con el fin de determinar una ontología computacional que permita efectuar análisis más detallados y automatizarlos en el tratamiento del corpus.

REFERENCIAS

Chantal M. 2002. "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento". Universidad de Málaga. [Documento de Internet disponible en http://elies.rediris.es/elies18/index.html]

Ladrón M. 1996. *Manual de Paleografía*. Santafé de Bogotá. Centro Editorial Javeriano CEJA.

Orduña. J. 1999. "La función definitoria de los ejemplos: a propósito del léxico filosófico del Diccionario de Autoridades". *Así son los Diccionarios*. Eds. Vila N., Calero M., Mateu R., Casanovas M, Orduña J. Lleida: 99-120.

Tercedor M. 1999. "La fraseología en el lenguaje biomédico: análisis desde las necesidades del traductor". Dpto de Traducción e Interpretación Universidad de Granada. [Documento de Internet disponible en http://elies.rediris.es/elies6/index.html#indice]

Torruela, J. y J. Llisterri. 1999. "Diseño de corpus textuales y orales". Departamento de Filología Española. Universidad Autónoma de Barcelona. [Documento de Internet disponible en http://liceu.uab.es/~joaquim/publicacions/Torruela_Llisterri_99.pdf]

CAPITULO 3

MODELO DE ETIQUETAMIENTO DE ENTIDADES

"La vista debe aprender de la razón" Johannes Kepler

"La cosa es tomar lo artificial con naturalidad." Mafalda (Quino)

Palabras Clave. Etiquetadores, reconocimiento de entidades, UIMA.

Este capítulo muestra la descripción de los modelos utilizados en el reconocimiento de entidades y su etiquetamiento de acuerdo a las características del corpus.

3.1 ESTRATEGIAS EMPLEADAS PARA EL ETIQUETAMIENTO

3.1.1 Herramienta utilizada para el etiquetamiento. Tal como se comentó en el capítulo 1 existen herramientas que permiten hacer el etiquetamiento de los corpus. Para ésta investigación la herramienta utilizada se llama UIMA - *Unstructured Information Management Architecture* – (Apache UIMA, 2009) que es una aplicación de libre uso disponible en Internet para implementar recursos lingüísticos utilizando los lenguajes de programación JAVA y C++, siendo compatible con el ambiente de programación Eclipse (Eclipse, 2009). Existen en el mercado otras herramientas para hacer análisis textual y se ensayaron algunas tales como GATE - *General Architecture for Text Engineering* (http://gate.ac.uk/) que es un anotador desarrollado por el grupo de PLN de la Universidad de Sheffield pero en su momento no hubo "compatibilidad" entre la autora y el programa. Otra herramienta empleada fue *WordSmith* (http://wordsmith.org/), que es una herramienta utilizada para hacer análisis frecuenciales de palabras, concordancias y estadísticas pero no fue utilizada porque tenía costo y su versión demo era muy limitada y finalmente *Cratilo*, que es una herramienta desarrollada por el profesor Jorge Antonio Mejía de la Universidad de Antioquia y que permite hacer análisis frecuencial y de

concordancias. La autora interactuó con ella al comienzo de la tesis para la elaboración de listados de palabras y para comprender conceptos claves para el desarrollo de la tesis. Finalmente se empleó UIMA debido a que es una herramienta libre que permite construir desarrollos propios en JAVA y además es compatible con ambientes de programación conocidos por la autora.

UIMA tiene una arquitectura basada en bloques llamados *Analysis Engines* (AE) que son los encargados de analizar el texto ubicando automáticamente los metadatos que se encuentran en él y etiquetándolos. Los AE están compuestos por uno o varios programas llamados *Anotadores,* que contienen los elementos concretos de análisis para proceder con el etiquetamiento de un texto. A su vez, los anotadores analizan un *artefacto* (que puede ser un texto o una grabación) y crean automáticamente metadatos sobre él. Un AE puede contener un solo anotador (Anotador Primitivo - *Primitive AE*) o también puede contener varios anotadores (Anotadores Agregados - *Aggregate AE*). Los anotadores producen sus análisis resultantes en forma de Estructuras Características (*Feature Structures*), las cuales son estructuras de datos que tienen un tipo y pueden contener un conjunto de atributos asociados. Con el fin de aclarar estos conceptos, en la figura 4 se muestra el texto "*Fred Center is the CEO of Center Micros. He is a graduate of State University and excels at golf*". Las anotaciones *Fred Center* y *He* son anotaciones del tipo *Persona* y la anotación *Center Micros* es del tipo *Organización*. (Apache UIMA, 2009).

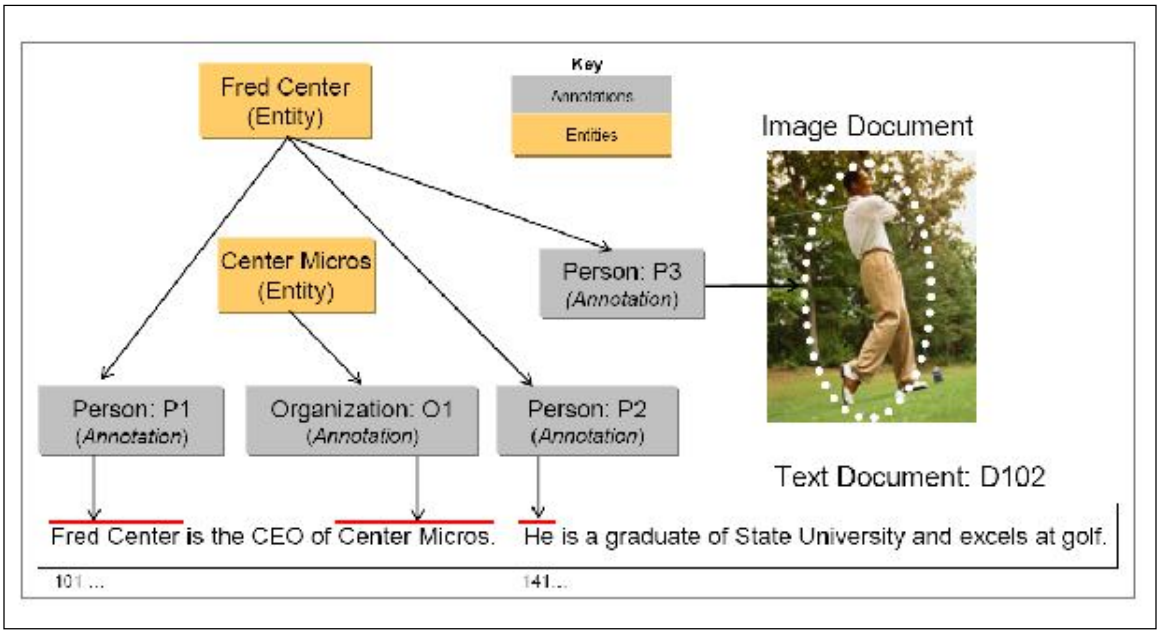


Figura 6.

Descripción del reconocimiento de entidades y etiquetamiento (Apache UIMA, 2009)

Para el computador un texto completo es una secuencia de caracteres que tienen asignados posiciones con respecto al inicio del texto. Una anotación tiene asociada una secuencia de caracteres llamada "*span*" y un span a su vez tiene una posición inicial y final con respecto al inicio del texto; en el ejemplo analizado, el *span* desde la posición 101 a la posición 111 contiene los caracteres "Fred Center" y es una anotación correspondiente al tipo "persona". La anotación es una manera de asignar metadatos semánticos en el texto, esto implica que una anotación tiene significado para el ser humano pero para el computador es un objeto que tiene unos atributos especiales. En este trabajo de investigación se desarrollaron varios anotadores utilizando la plataforma UIMA. La siguiente figura muestra una anotación en el corpus utilizando el anotador de personas desarrollado para esta investigación y empleando una herramienta de visualización de UIMA llamada "*CAS Visual Debugger*". En el lado derecho puede observarse resaltada la anotación perteneciente al tipo "Persona" y en el lado izquierdo se muestran sus atributos.

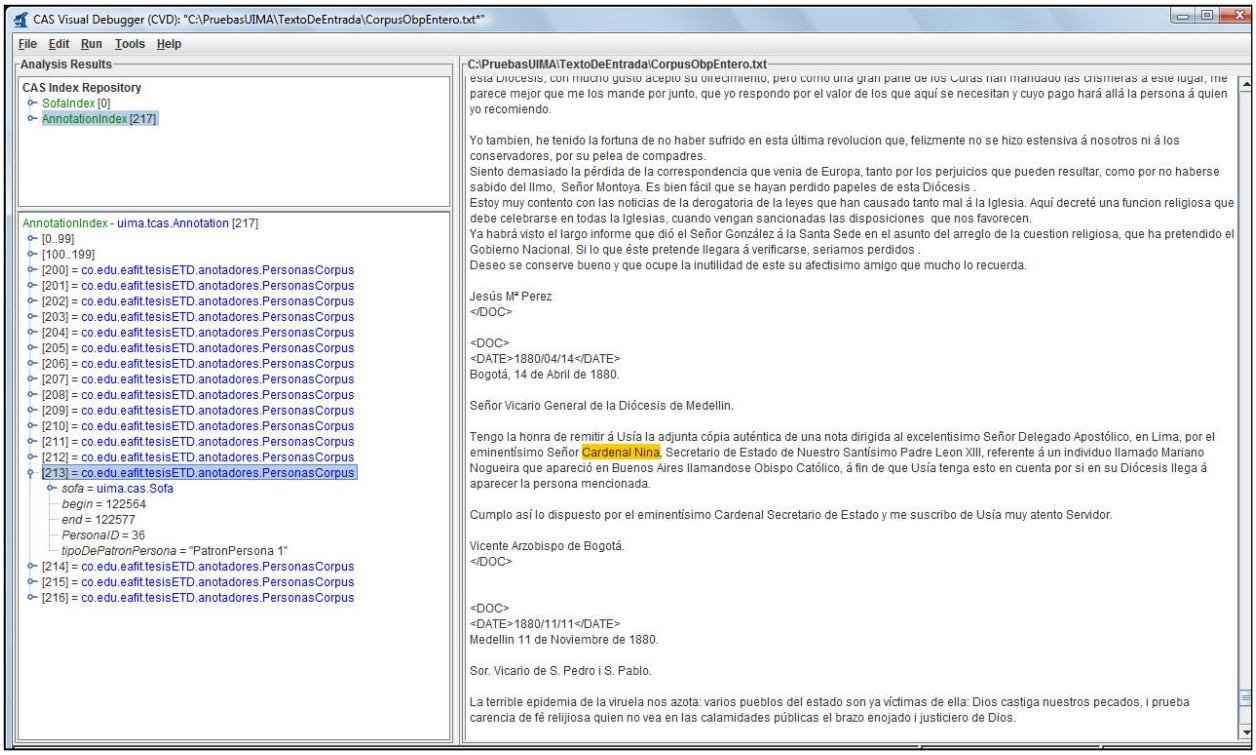


Figura 7.
Ejemplo de etiqueta de Persona en el Corpus utilizando la herramienta *CAS Visual Debugger* (Apache UIMA, 2009)

Las anotaciones se representan y comparten sus resultados en un bloque de UIMA llamado CAS (*Common Analysis Structure*) que contiene las anotaciones del documento analizado y permite asociarlas con el tipo de anotación y sus atributos para proceder a su análisis de acuerdo a los intereses del programador.

Descriptores de UIMA: Los AE requieren dos tipos de interfaces para su implementación:

- Descriptores: es la parte declarativa implementada en XML de los "tipos" que se quieren anotar en el texto. Los descriptores contienen información relativa a los metadatos del texto, el tipo al que va a pertenecer la anotación, el tipo de datos requerido para el CAS de entrada de datos y la clase de JAVA que implementa la anotación. Los descriptores tienen una interfaz que facilita introducir los datos de análisis. Esta interfaz tiene asociado su respectivo código XML.

- Código: es el código que implementa el algoritmo empleado para la anotación y que utiliza los datos definidos en el Descriptor. Puede ser implementado en JAVA o C++. Para el caso de este trabajo de investigación el código se implementó en JAVA.

Las siguientes imágenes ilustran los descriptores y el código en JAVA asociado a un anotador implementado en UIMA, en este caso el anotador LugaresCorpusDescriptor:

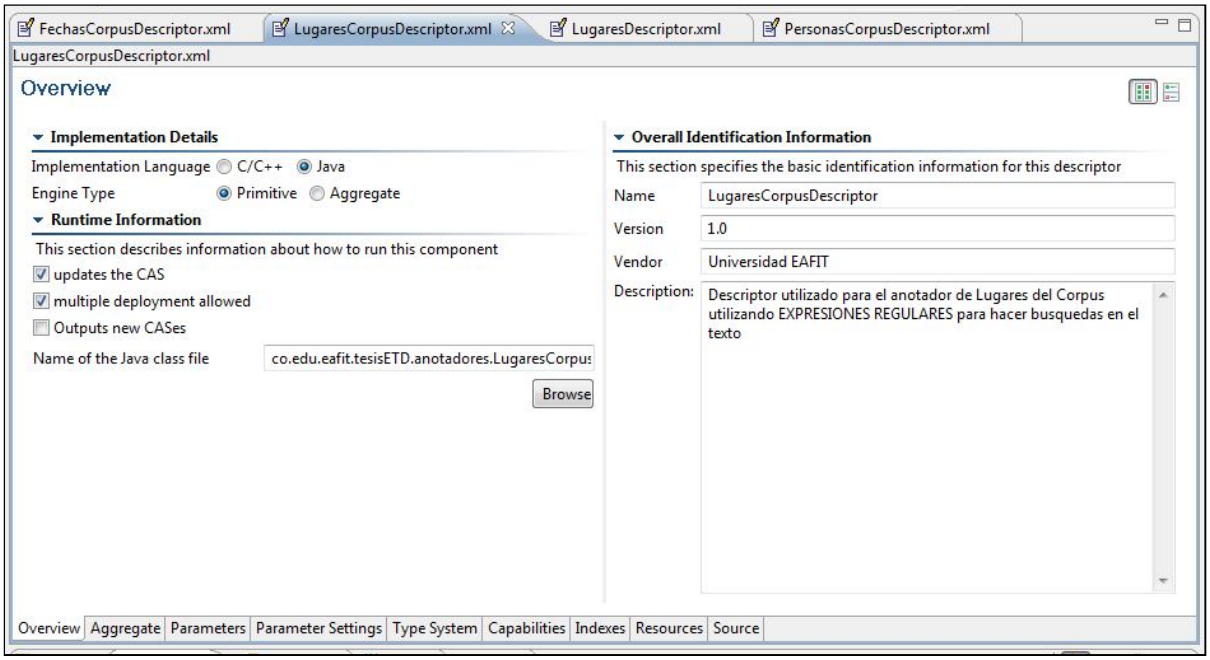


Figura 8.
Imagen del descriptor LugaresCorpusDescriptor

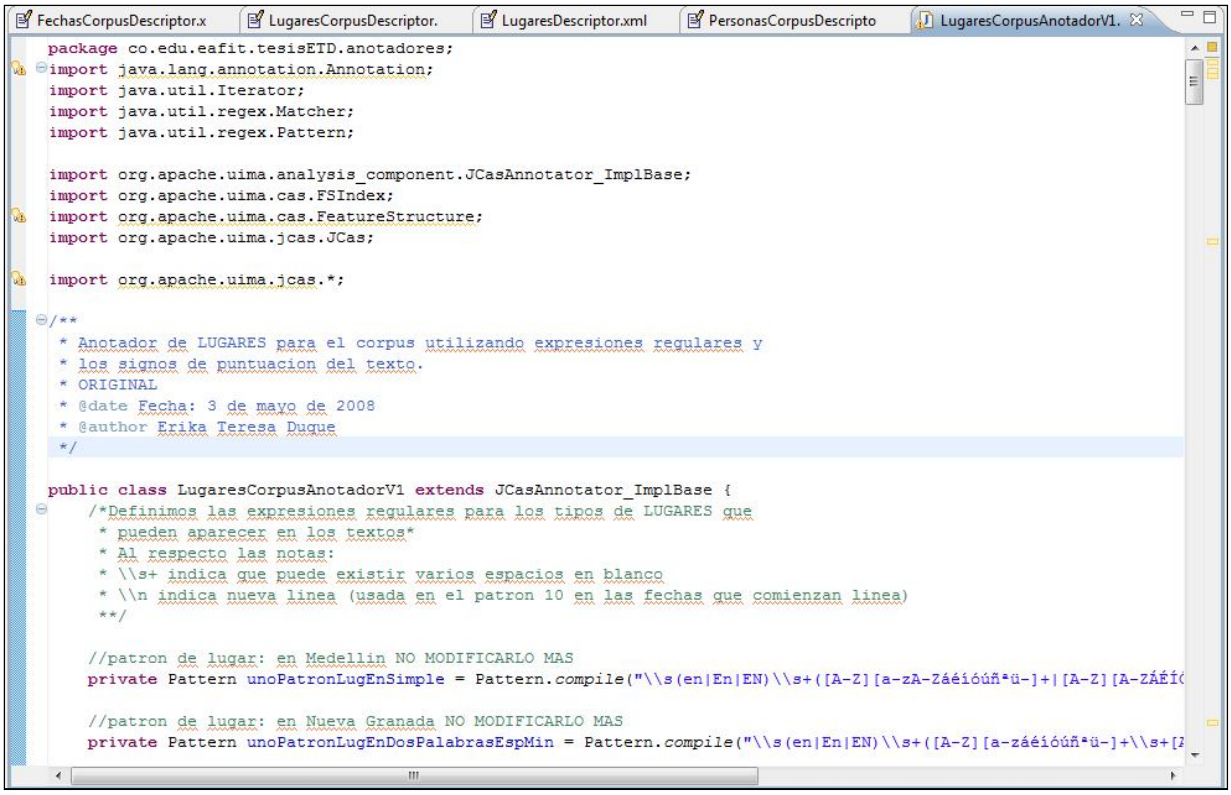


Figura 9.
Código en JAVA que soporta el descriptor LugaresCorpusDescriptor

Implementación de los anotadores en el trabajo de investigación: Los AE son construidos por el usuario de UIMA dependiendo de sus intereses de búsqueda. Puesto que en este trabajo de investigación analiza la extracción de información de personas, lugares, verbos y fechas, se utilizaron varios AE para reconocer y marcar estas entidades presentes en el corpus. Es importante aclarar que UIMA es solo una herramienta de anotación y que se debe programar de acuerdo a los intereses de extracción de información requeridos. Se implementaron 8 descriptores asociados a los tipos de entidades de interés y que se mencionan a continuación:

Descriptor del Anotador	Tipo de entidad a anotar	Tipo de anotador (agregado o primitivo)	Archivo de JAVA asociado	Comentarios
DescriptorDeEntidades	Todas	agregado	N / D	Es el descriptor que reúne todas las entidades a ser anotadas en el corpus
FechasCorpusDescriptor	Fechas	primitivo	FechasCorpusAnotador	Descriptor utilizado para identificar fechas utilizando patrones de expresiones regulares en el texto
AnotListadosFechasCorpus Descriptor	Fechas	primitivo	AnotListadosFechasCorpus	Anotador para aquellas fracciones de texto que denoten fechas y que no tengan un formato numérico
LugaresDescriptor	Lugares	agregado	N / D	Es el descriptor que reúne todas las anotaciones de lugares en el corpus
LugaresCorpusDescriptor	Lugares	primitivo	LugaresCorpusAnotadorV1	Descriptor utilizado para identificar lugares utilizando patrones de expresiones regulares en el texto
AnotListadosLugCorpusDes criptor	Fechas	primitivo	AnotListadosLugCorpus	Es un anotador de listado de lugares tales como países, ciudades, pueblos y barrios
DocCorpusAnotadorDescrip tor	Documento	primitivo	DocCorpusAnotador	Descriptor utilizado para anotar los documentos individuales del corpus.
PersonasCorpusDescriptor	Personas	primitivo	PersonasCorpusAnotador	Descriptor utilizado para identificar nombres de personas utilizando patrones de expresiones regulares en el texto

Tabla 4.
Listado de descriptores utilizados en el análisis de identificación de entidades

Visualización de las anotaciones: UIMA utiliza una aplicación llamada "*Document Analyser*" que consiste en una interfaz de usuario para escoger el listado de documentos a analizar y muestra en colores las diferentes marcaciones en el texto. UIMA siempre hace una copia del texto y en ese nuevo archivo hace las anotaciones en formato XML. De esta manera se cumple uno de los parámetros requeridos en las herramientas de análisis de corpus en donde se debe cumplir que en lo posible se conserve el texto original (Leech, 1993, citado por Abaitua, 2000). La siguiente imagen muestra los resultados de etiquetamiento para el corpus utilizando la herramienta mencionada de UIMA.

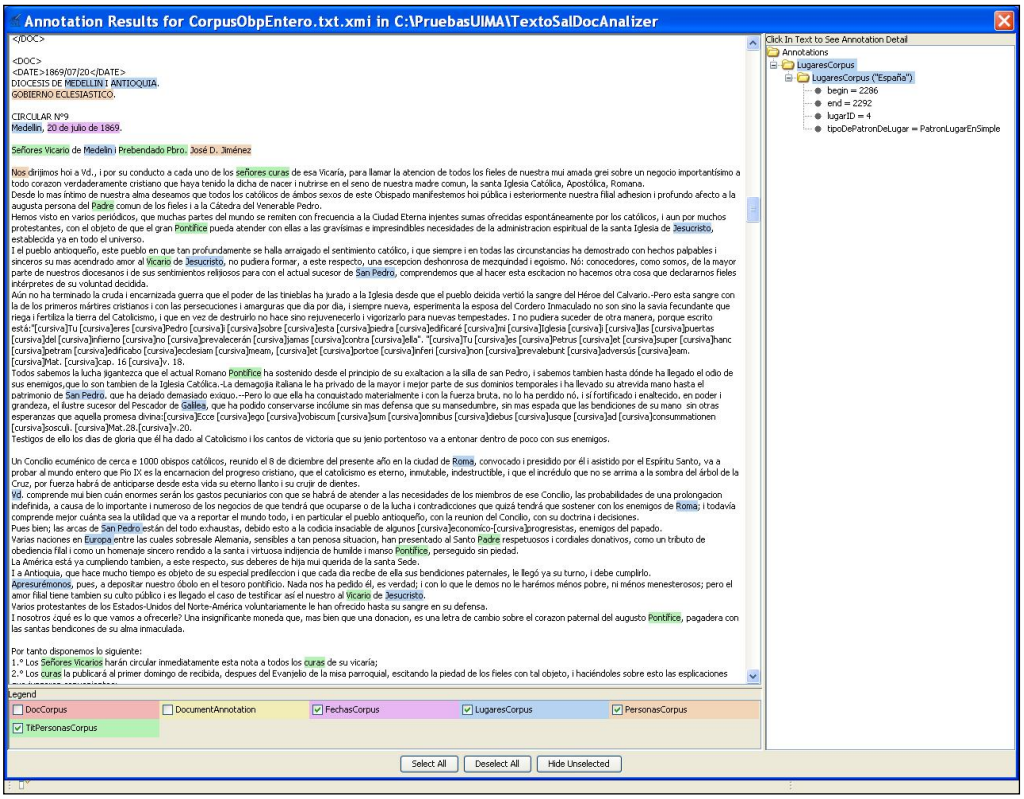


Figura 10.
Análisis del corpus utilizando la herramienta *Document Analyser* (Apache UIMA, 2009).

3.1.2 Descripción del método de etiquetamiento empleado. Como se mencionó en el capítulo 1, las técnicas para EI utilizan heurísticas y técnicas computacionales para la identificación de la información de interés. En esta investigación se utilizaron expresiones regulares combinadas con heurísticas para proceder con la identificación de los nombres pero antes de analizar los patrones empleados se explicará brevemente en qué consisten las expresiones regulares y como fueron empleadas para el reconocimiento de entidades y también se describen las otras heurísticas empleadas.

Expresiones regulares (ER). Las ER son un lenguaje formal para la caracterización de un conjunto de cadenas de texto (*strings*). Una cadena es una secuencia de caracteres alfanuméricos que comprende letras, números, signos de puntuación y otros caracteres tipográficos. Las ER son utilizadas en informática para aplicaciones relacionadas con validación de datos en formularios, construcción de compiladores para la validación sintáctica de programas y aplicaciones de extracción de información. Para el caso de la

extracción de información, las ER requieren un "modelo" que será empleado para efectuar las búsquedas en el corpus y que será escrito en un lenguaje de programación, por ejemplo PERL o JAVA. Muchos editores de texto permiten la utilización de patrones regulares para efectuar sus búsquedas.

Modelos empleados en las expresiones regulares. Como se mencionó anteriormente, las expresiones regulares son maneras de escribir "formas de palabras". A continuación se muestra un listado de formas de expresiones regulares utilizadas en lenguaje JAVA:

Modelo de construcción de la E.R.	Coincidencia	Observaciones
\ \s	Un espacio en blanco	
\ \S	Un caracter diferente a espacio en blanco	
\ \d	Un digito 0-9	Se muestran dos maneras diferentes de expresar la coincidencia para cada caso.
\ \p{Digit}		
\ \D	No es un dígito	
[^0-9]		
\ w	Cualquier carácter de palabra	
[a-zA-Z_0-9]		
\ p{Lower}	Un carácter alfabético en minúsculas	
[a-z]		
\ p{Upper}	Un carácter alfabético en mayúsculas	
[A-Z]		
[A-P] o [a-p]	Especifica el rango de letras mayúsculas entre A y P y a-p	Las expresiones regulares distinguen entre mayúsculas y minúsculas ("*case sensitive*")
[^abc]	Cualquier carácter menos a, b o c	

Tabla 5.

Ejemplos de construcción de expresiones regulares (JDK *Documentation*, 2009).

Las expresiones regulares también permiten efectuar operaciones lógicas como se muestra en la siguiente tabla:

Operaciones entre expresiones regulares		
XY	X seguido de Y (equivale a decir "X AND Y")	X puede ser cualquier expresión
X|Y	X o Y (equivale a decir "X OR Y")	
X*+	X cero o más veces	
X++	X una o más veces	
X{n}	X exactamente n veces	

Tabla 6.

Ejemplos de construcción de operaciones entre expresiones regulares (JDK *Documentation*, 2009).

Algunos ejemplos de expresiones regulares implementadas se muestran a continuación aunque se describirán con mayor detalle en las secciones siguientes:

Expresión regular	Ejemplos de palabras encontradas
\ \p{javaUpperCase}[a-zA-Záéíóúñªü°]+)\ \s+(\ \p{javaUpperCase}\ \p{Punct})	Carlos N.
(Pbro|Pbro.)+\ \s+((\ \p{javaUpperCase}[a-zA-Záéíóúñªü°-]+)	Pbro Ramírez
\ \s(en|En|EN)\ \s+([A-Z][a-záéíóúñªü-]+\ \s+[A-Z][a-záéíóúñªü-]+)	En Nueva Granada

Tabla 7.
Ejemplos de implementación de expresiones regulares.

Formalmente una expresión regular es lenguaje formal utilizado para la caracterización de un conjunto de cadenas de texto y pueden ser modeladas a través de autómatas finitos o máquinas de estados. Jurasfky y Martin (2000) explican con detalle el funcionamiento de las expresiones regulares y su relación con las máquinas de estado y sus aplicaciones en el procesamiento del lenguaje natural.

Metodología empleada para la elaboración de las expresiones regulares en esta investigación. Puesto que la idea primordial es extraer información acerca de las expresiones de tiempo, lugares y nombres de personas, las ER y las técnicas heurísticas se construyen con base en los modelos de las expresiones presentes en el texto siguiendo la siguiente metodología:

1. Determinar el tipo de entidad a ser encontrada en el corpus: esto es, primero tener una idea clara del tipo de dato a extraer.
2. Analizar una porción de corpus en donde se encuentra el tipo de entidad a reconocer y elaborar un listado previo del tipo de ocurrencias halladas en el corpus. Esta tarea se hizo de manera manual y no de manera automática (por ejemplo, escribiendo un programa para obtener un listado de las palabras) por dos razones: la primera de ellas fue que la elaboración del corpus contribuyó a la depuración, ya que a medida que se transcribían documentos se incluían las palabras de interés en el listado de las entidades. La segunda razón obedece a que la extracción de información depende del contexto, como se explicó en el capítulo

1, y muchas de las entidades de interés tienen nombres compuestos y se podía perder información y para este caso particular, el corpus era pequeño. Cuando se tienen grandes volúmenes de información es necesario recurrir a métodos automáticos para determinar los modelos de información a extraer.

3. Analizar el listado de las palabras por entidad, establecer similitudes y reunirlas en grupos para establecer patrones comunes entre ellas con el fin de configurar las expresiones regulares.
4. Implementar la expresión regular en JAVA. Los programas hechos en JAVA para identificar ER recorren el texto y lo compara con las ER propuestas en el código, cuando se encuentra un texto que coincide con el modelo de la ER se clasifica como una anotación.
5. Evaluar el sistema, esto es, determinar cuáles de las anotaciones realmente corresponden a los nombres de las entidades y cuales obedecen al modelo de la ER pero no son entidades (falsos positivos).

De manera particular el proceso fue más difícil para el reconocimiento de las entidades de personas debido a que el corpus es especialmente prolífico en ese aspecto puesto que incluía listados enteros de personas. Para este caso se escogió el archivo de texto más grande que corresponde a CorpusSac1874 que contiene 14.785 palabras y se elaboró una tabla donde se escogían los tipos de persona, se agrupaban y se les asignaba un modelo de expresión regular. A continuación se muestra una parte del cuadro elaborado para establecer los modelos de ER para los nombres de personas:

Modelos de personas para etiquetar, tomados como modelo documento de 1874			
Nombre	**Caracteristica**	**Largo**	**Tipo de patron**
Por Lilian Valencia Jesus	comienzo de linea y retorno de carro	27	tipoDePatronPersona = ERPersona 3B
Pro Silverio A Gomez	en el parrafo y seguido de minuscula	20	tipoDePatronPersona = ERPersona 3A
don Martin de Saldarriaga	comienzo de linea y retorno de carro	25	tipoDePatronPersona = ERPersona 5B
+ Carlos Obpo	comienzo de linea y retorno de carro	13	tipoDePatronPersona = ERPersona 6A
Juan Franº Mejia	comienzo de linea y retorno de carro	17	tipoDePatronPersona = ERPersona 6B
Pedro Pablo Salazar Euse	comienzo de linea y retorno de carro	26	tipoDePatronPersona = ERPersona 6C
Por Lilian Valencia Jesus	comienzo de linea y retorno de carro	25	tipoDePatronPersona = ERPersona 6C
Vicente Cevallos	comienzo de linea y retorno de carro	16	tipoDePatronPersona = ERPersona 7A
Martin E. Gaviria	comienzo de linea y retorno de carro	18	tipoDePatronPersona = ERPersona 7A
Pedro P Bernal	comienzo de linea y retorno de carro	14	tipoDePatronPersona = ERPersona 7C
Luis M. Martínez G.	comienzo de linea y retorno de carro	20	tipoDePatronPersona = ERPersona 7C
jóven Alejo Marulanda	en el parrafo y seguido de minuscula	22	tipoDePatronPersona = ERPersona 8B
Pablo F. Pineda Pbro.	comienzo de linea y retorno de carro	22	tipoDePatronPersona = ERPersona 8D
Jacobo J. H.	comienzo de linea y retorno de carro	13	tipoDePatronPersona = ERPersona 9A
Manuel d. J. Ocampo	comienzo de linea y retorno de carro	20	tipoDePatronPersona = ERPersona 9B
, Vicente A. Restrepo,	en el párrafo y con signos de puntuacion y retorno de carro	23	tipoDePatronPersona = ERPersona 9B
Marco A. Peláez J.	disparadora Católica y con signos de puntuación al final y retorno de carro	19	tipoDePatronPersona = ERPersona 9C

Tabla 8.
Detalle de la tabla de determinación de entidades de nombres de personas

Con base en las similitudes de expresiones regulares, se conformaban "familias de expresiones", de esta manera se tiene el tipo 9 de patrón que tiene a su vez expresiones similares agrupadas en tipología 9A, 9B, 9C y 9D. Esta subclasificación es importante porque facilita la asignación de modelos de ER en la clase de JAVA donde se implementan. El siguiente cuadro muestra parte de la clase de JAVA que implementa las familias de ER que modelan nombres de personas presentes en el corpus:

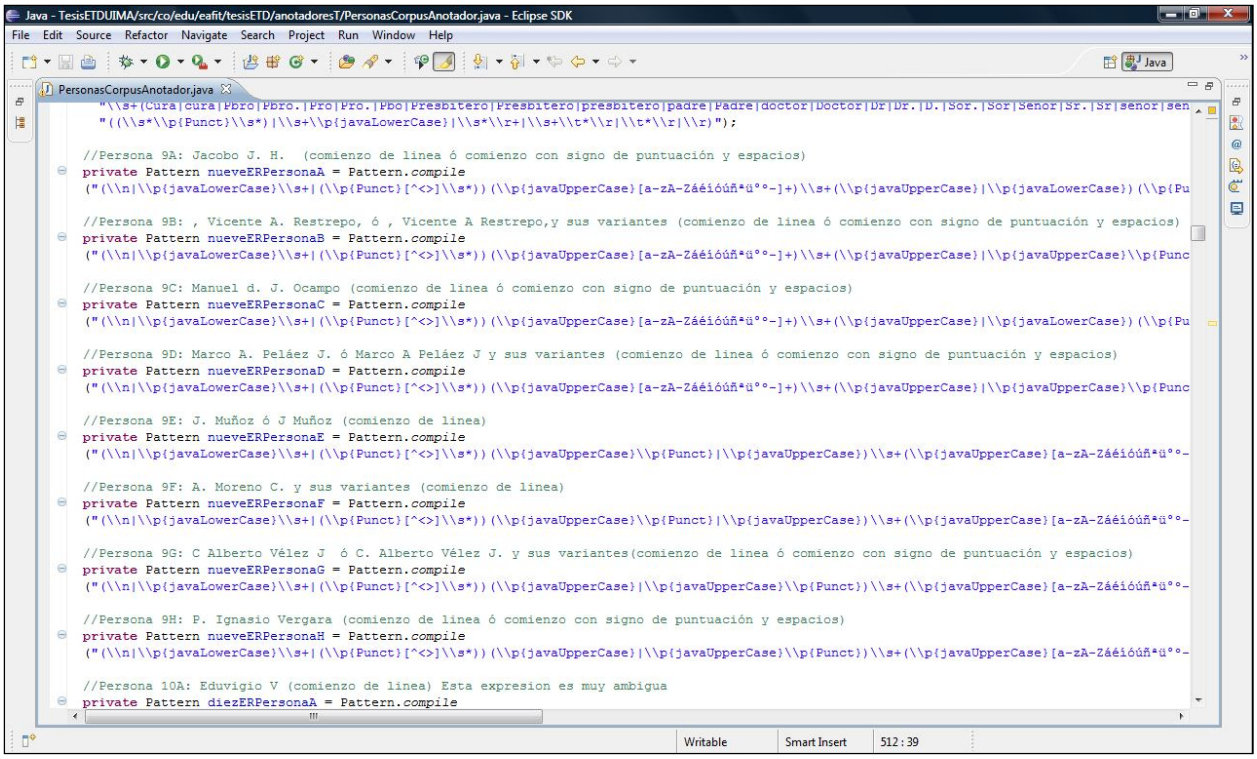


Figura 11.
Código de la clase de JAVA PersonasCorpusAnotador, donde se implementan las ER para detectar nombres de personas en el corpus

Las ER recorren un texto e identifican porciones que concuerdan con su tipo, pero como el trabajo de reconocimiento es semántico pueden generarse "errores " de etiquetamiento de dos tipos (Jurasfky y Martin, 2006): los "falsos positivos" ("*false positives*") que son frases incorrectamente encontradas por el patrón y los "falsos negativos" ("*false negatives"*) que son frases incorrectamente ignoradas (no halladas por el patrón).

Heurísticas empleadas elementos del contexto y utilización de fuentes de conocimiento externas. A continuación se describen las técnicas heurísticas empleadas para la extracción de información.

Heurística 1 Elementos del contexto. La EI y otras muchas aplicaciones de lingüística computacional requieren el uso de palabras del contexto para extraer de manera eficaz la información requerida y los componentes semánticos que permitan desambiguarla o por lo menos, identificar el tipo de significado que la información extraída tiene en el

contexto. En este trabajo de investigación se utilizaron "palabras disparadoras" de acuerdo a la notación empleada por Muñoz en su trabajo de extracción de información en documentos notariales (Muñoz, et al. 1998) o llamadas también palabras funcionales. Las palabras funcionales son palabras que rodean el nombre a extraer y que ayudan en su identificación como son, por ejemplo, los títulos de personas o la utilización de las preposiciones en el caso de la identificación de lugares. Estas palabras funcionales fueron utilizadas también en las ER para facilitar la identificación de nombres de personas y lugares y también fue necesario hacer un análisis previo del corpus para identificarlas y agruparlas. La siguiente figura ilustra el proceso empleado para las palabras funcionales para las ER de nombres de personas:

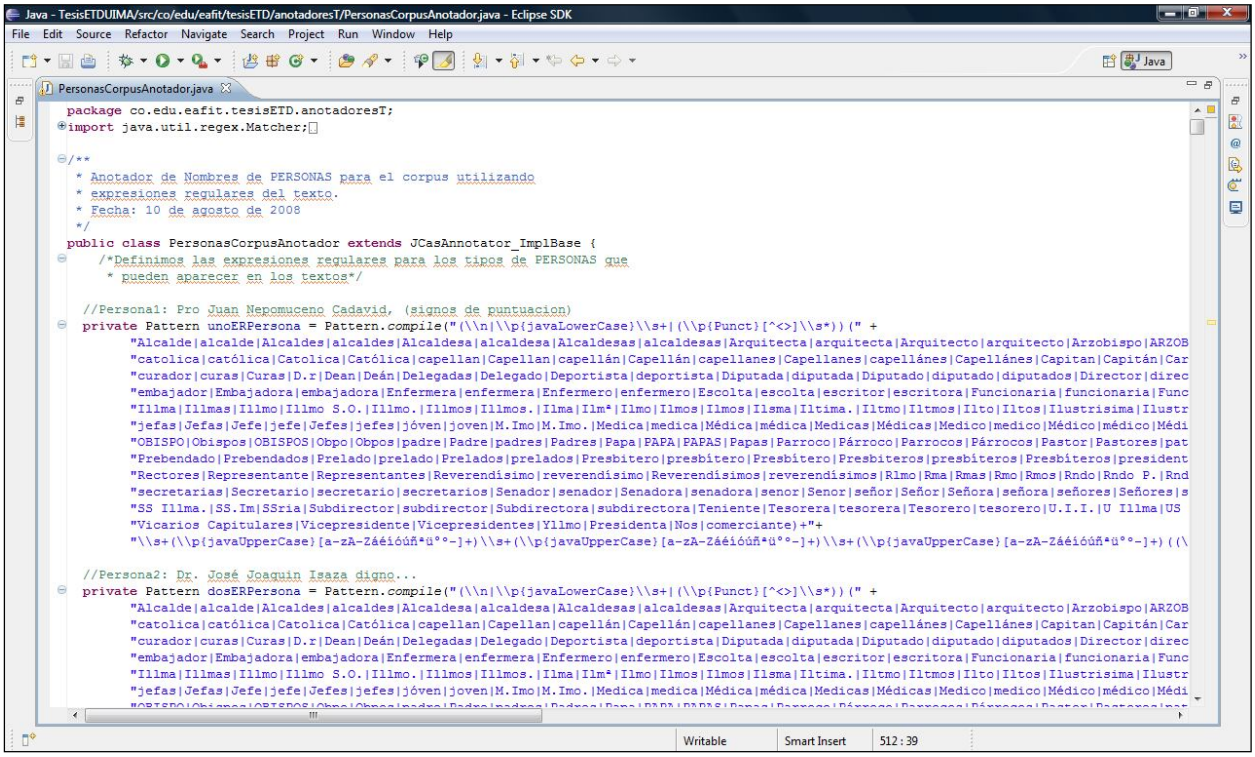


Figura 12.

Código de la clase de JAVA PersonasCorpusAnotador, donde se implementan las palabras funcionales como parte de las ER para detectar nombres de personas en el corpus

Heurística 2 Bases de conocimiento externas y listas de inclusión. En la extracción de información es importante también contar con fuentes de conocimiento externo que ayuden a refinar las búsquedas tal como se mencionó en el capítulo 1. En las bibliotecas digitales se utilizan accesos a diccionarios externos o a Internet. En este caso se implementaron listas de inclusión o listados de palabras clave que en este caso son lugares (capitales del mundo, pueblos de Antioquia y barrios de Medellín) y listados de palabras que pueden contener significado relacionado con fechas. Esta heurística fue implementada utilizando los descriptores de UIMA pero primero fue necesario hacer el desarrollo del código en java y unificarlo con el descriptor. El código de java compara las palabras de los listados con el texto y los anota si coinciden. Cuando ya se tiene programado el descriptor es posible ingresar más datos a los listados. La siguiente figura muestra uno de los descriptores empleados para la anotación de entidades utilizando ésta metodología:

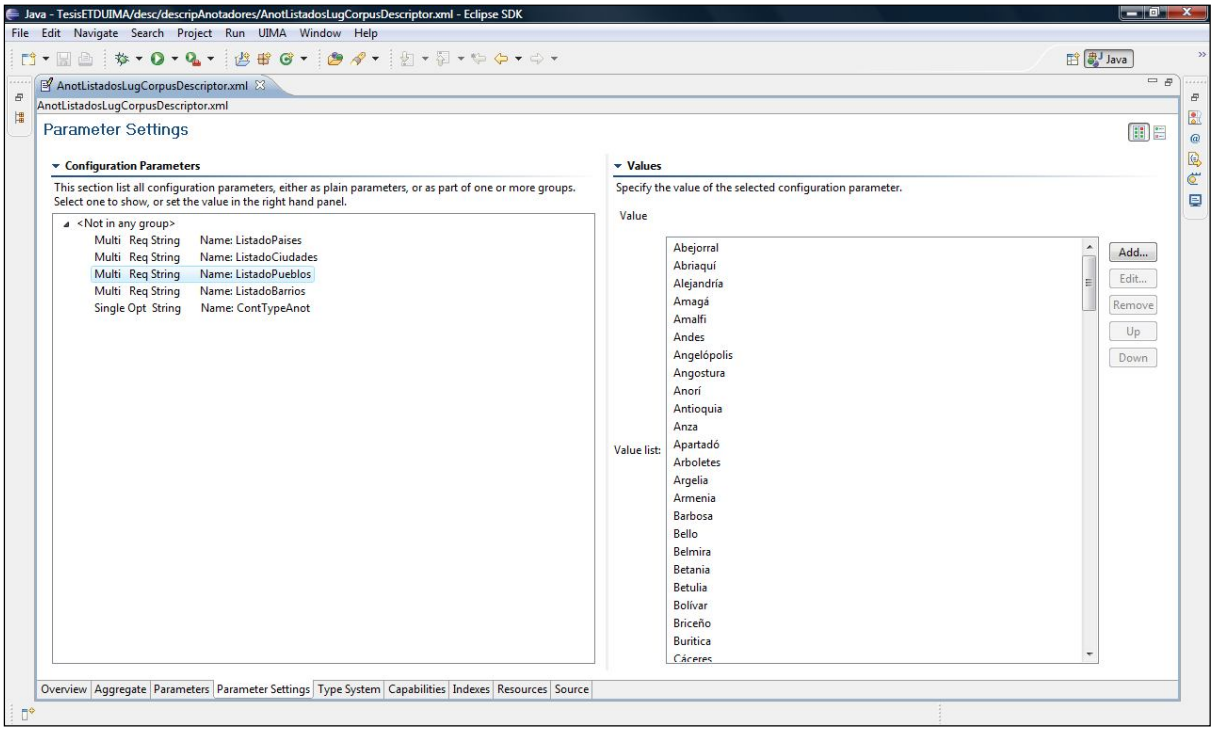


Figura 13.
Descriptor de lugares que utiliza listados para refinar la identificación de entidades de lugares

Es importante tener en cuenta que para el análisis de las entidades se utilizó el corpus empleado para este trabajo que fue descrito en el capítulo 2 pero para efectos de evaluación también se empleó el corpus de noticias utilizado para las pruebas TREC (LDC *Linguistic Data Consortium*, 2000) cuyas características ya fueron también mencionadas en dicho capitulo.

3.2 DEFINICIÓN Y CARACTERIZACIÓN DE LAS ENTIDADES PRESENTES EN EL CORPUS

3.2.1 Definición y análisis de características de las entidades. Aunque el objetivo de este trabajo es identificar anotaciones de nombres de personas, lugares y fechas, y establecer posibles relaciones entre ellas, una posible aplicación de este trabajo es diseñar un software que apoye a profesionales tales como los historiadores. Al efectuar un análisis de las actividades que estos profesionales llevan a cabo en su trabajo se encontró que requieren listados automáticos de personajes y lugares, preferiblemente discriminados por cada documento. Esta información es importante porque ofrece una visión de conjunto de los personajes que intervienen y en dónde se encontraban localizados en determinada situación histórica. Sin embargo, para elaborar estos listados de manera automática es necesario caracterizar las entidades de interés con el fin de encontrar reglas para la definición de los modelos tal como se explicó en la sección anterior. De acuerdo al corpus utilizado básicamente se tienen los siguientes tipos de "anotaciones" presentes en el texto:

- Documentos
- Fechas
- Nombres de lugares geográficos (topónimos)
- Nombres de personas (nombres propios)
- Verbos

Otro tipo de entidades presentes en el corpus son Inmuebles (seminarios, parroquias, iglesias) y organizaciones tales como Asociaciones (Diócesis, Junta Catedral, Capítulo Catedral, etc.), pero estas entidades no se tendrán en cuenta para el alcance actual de este trabajo y pueden ser objeto de otros análisis posteriores.

Es importante tener en cuenta que el enfoque utilizado para la anotación es semiautomático y que como se mencionó en el capítulo 1, la medida empleada para medir el desempeño de las anotaciones de un sistema de EI son la precisión (porcentaje de entidades identificadas por el sistema) y cobertura (porcentaje de identificaciones que corresponden a la entidad) y por tanto, se emplearán estas mediciones para medir el desempeño del sistema.

3.2.2 Etiquetas pertenecientes a los documentos del texto. Se etiquetaron los documentos que conforman el corpus y que están delimitados por las etiquetas <DOC> y </DOC> con el fin de determinar una ventana de medición para las anotaciones y los sucesos que puedan tenerse en cuenta en ellos. Esta ventana surge también con el propósito de hacer un índice de entidades halladas por documento con base en su rango de *span* y determinar el número de palabras por documento. La siguiente figura muestra el etiquetamiento por documento:

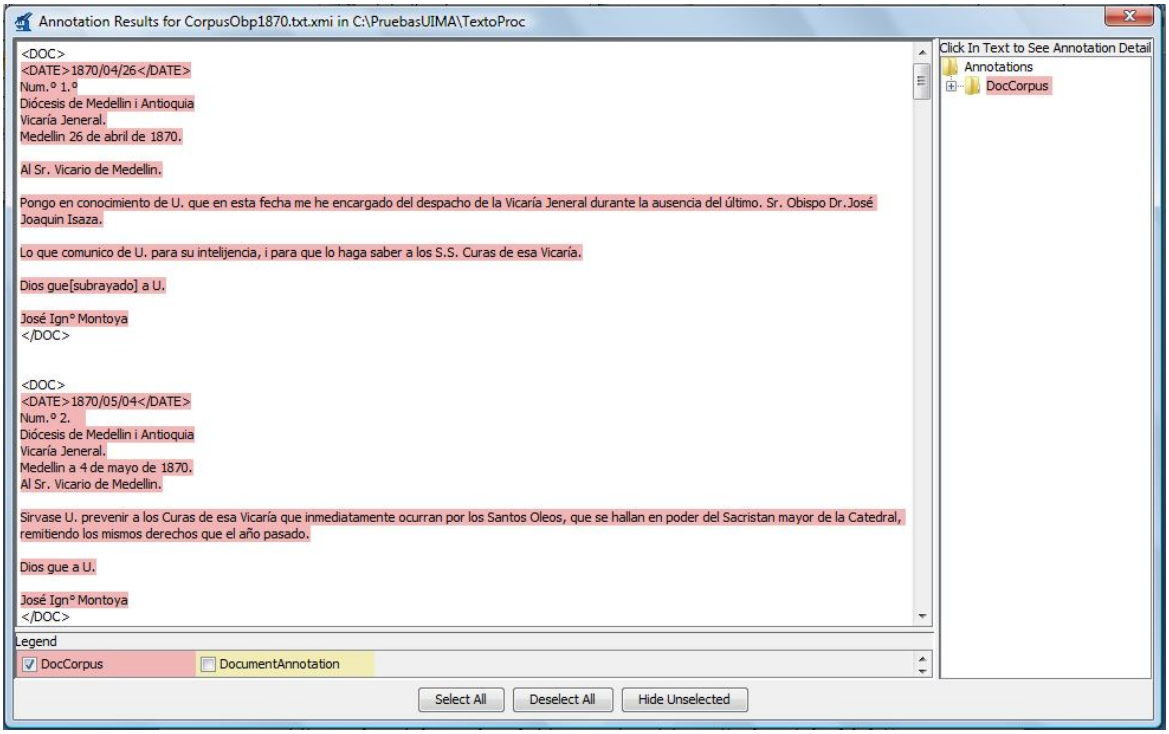


Figura 14.
Anotaciones correspondientes a los documentos que conforman el corpus.

3.2.3 Etiquetas para fechas. La mayoría de las fechas presentes en el corpus corresponden a las contenidas en el encabezado de las cartas y aunque presentan formatos diversos ha sido posible etiquetarlas por medio de ER porque contienen en su gran mayoría datos numéricos para el día, el mes y el año. También se encuentran algunas fechas que presentan otros formatos correspondientes a texto narrativo y requieren otro tratamiento para su etiquetamiento como por ejemplo, incluir listados de los números y sus posibles combinaciones. De igual manera, se encuentran otros tipos de fechas que deben ser resueltas por medio de anáforas y cuyo estudio sobrepasa el alcance de este trabajo. A continuación se muestran los modelos utilizados para el etiquetamiento.

Modelos de etiquetamiento utilizados

Expresiones regulares. Para el etiquetamiento de las fechas se utilizaron ER que describen las formas de fechas presentes en el texto. En un principio se pensó etiquetarlas mencionando la ciudad ya que como se trata de textos manuscritos, las fechas tienen el lugar de origen de la carta. Bajo este modelo se encontraron 34 maneras diferentes para etiquetar las fechas ya que influían dos factores:

- Los signos de puntuación entre el lugar y la fecha ya que pueden utilizarse espacio en blanco, coma o punto (Ej. Medellín, 17 de mayo de 1870 o Medellín 17 de mayo de 1870 o Medellín. 17 de mayo de 1870).
- La manera de describir el año (Ej. 1870 ó 1.810 ó 1,870)

Sin embargo, fue necesario restringir el modelo de etiquetamiento de las fechas prescindiendo de la ciudad porque la ciudad misma es otra entidad, otras fechas no lo incluían y es mucho mejor tener generalidad, de aquí la importancia de tener muy claro desde el comienzo del análisis los datos que se quieren analizar para evitar complicaciones futuras. Además se producían errores de ambigüedad (doble etiquetamiento) puesto que el modelo podía etiquetar dichas fechas incluyendo la ER completa (con la ciudad) y la ER simple (solo la fecha) ya que finalmente las ER son expresiones compuestas por el operador OR. El listado de ER definitivo incluyó 19 tipos de expresiones de fecha a etiquetar las cuales se muestran a continuación:

Nombre del modelo de fecha	Ejemplo de fecha etiquetada	Expresión Regular (ER) construida para el modelo de fecha
unoERFechaA	"Mayo 11"	(enero | Enero | ENERO | febrero | Febrero | FEBRERO | marzo | Marzo | MARZO | abril | Abril | ABRIL | mayo | Mayo | MAYO | junio | Junio | JUNIO | julio | Julio | JULIO | agosto | Agosto | AGOSTO | septiembre | Septiembre | SEPTIEMBRE | setiembre | Setiembre | SETIEMBRE | octubre | Octubre | OCTUBRE | noviembre | Noviembre | NOVIEMBRE | diciembre | Diciembre | DICIEMBRE | octubre | Octubre | OCTUBRE | noviembre | Noviembre | NOVIEMBRE)\ \s+\ \d{2}(\ \. | \ \s+)
unoERFechaB	"Agosto de 1874"	(Enero | ENERO | enero | Febrero | FEBRERO | febrero | Marzo | MARZO | marzo | Abril | ABRIL | abril | Mayo | MAYO | mayo | Junio | JUNIO | junio | julio | Julio | JULIO | agosto | Agosto | AGOSTO | septiembre | Septiembre | SEPTIEMBRE | setiembre | Setiembre | SETIEMBRE | octubre | Octubre | OCTUBRE | noviembre | Noviembre | NOVIEMBRE | diciembre | Diciembre | DICIEMBRE)\ \s+(de | De | DE)\ \s+\ \d{4}
unoERFechaC	- "Agosto 1 de 1874" - "Agosto 12 de 1874"	(Enero | ENERO | enero | Febrero | FEBRERO | febrero | Marzo | MARZO | marzo | Abril | ABRIL | abril | Mayo | MAYO | mayo | Junio | JUNIO | junio | julio | Julio | JULIO | agosto | Agosto | AGOSTO | septiembre | Septiembre | SEPTIEMBRE | setiembre | Setiembre | SETIEMBRE | octubre | Octubre | OCTUBRE | noviembre | Noviembre | NOVIEMBRE | diciembre | Diciembre | DICIEMBRE)\ \s+\ \d{1,2}+\ \s+(de | De | DE)\ \s+\ \d{4}
unoERFechaD	"dic.e 9/10 de 1874"	([a-zA-Z][a-z]+ | [a-zA-Z][a-z]+\ \.\ \w)\ \s+\ \d{1,2}+\ \s+(de | De | DE)\ \s+\ \d{4}
dosERFecha	- "Set. 1 de 1872" - "Set. 21 de 1872"	([a-zA-Z][a-z]+\ \.) \ \s+\ \d{1,2}\ \s+(de | De | DE) \ \s+\ \d{4}
tresERFecha	- "agosto 6 de 1.874" - "agosto 16 de 1.874"	(enero | Enero | ENERO | febrero | Febrero | FEBRERO | marzo | Marzo | MARZO | abril | Abril | ABRIL | mayo | Mayo | MAYO | junio | Junio | JUNIO | julio | Julio | JULIO | agosto | Agosto | AGOSTO | septiembre | Septiembre | SEPTIEMBRE | setiembre | Setiembre | SETIEMBRE | octubre | Octubre | OCTUBRE | noviembre | Noviembre | NOVIEMBRE | diciembre | Diciembre | DICIEMBRE | octubre | Octubre | OCTUBRE | noviembre | Noviembre | NOVIEMBRE)\ \s+\ \d{1,2}\ \s+(de | De | DE)\ \s+\ \d{1}\ \p{Punct}\ \d{3}
cuatroERFechaA	"4 de marzo"	\ \d{1,2}\ \s+(de | De | DE)\ \s+(enero | Enero | ENERO | febrero | Febrero | FEBRERO | marzo | Marzo | MARZO | abril | Abril | ABRIL | ,mayo | Mayo | MAYO | junio | Junio | JUNIO | julio | Julio | JULIO | agosto | Agosto | AGOSTO | septiembre | Septiembre | SEPTIEMBRE | setiembre | Setiembre | SETIEMBRE | octubre | Octubre | OCTUBRE | noviembre | Noviembre | NOVIEMBRE | diciembre | Diciembre | DICIEMBRE)
cuatroERFechaB	"4 marzo de 1880."	\ \d{1,2}\ \s+(enero | Enero | ENERO | febrero | Febrero | FEBRERO | marzo | Marzo | MARZO | abril | Abril | ABRIL | ,mayo | Mayo | MAYO | junio | Junio | JUNIO | julio | Julio | JULIO | agosto | Agosto | AGOSTO | septiembre | Septiembre | SEPTIEMBRE | setiembre | Setiembre | SETIEMBRE | octubre | Octubre | OCTUBRE | noviembre | Noviembre | NOVIEMBRE | diciembre | Diciembre | DICIEMBRE)\ \s+(de | De | DE)\ \s+(\ \d{4} | \ \d{4}\ \.)")
cuatroERFechaC	- "4 de febrero de 1874" - "12 de febrero de 1874"	\ \d{1,2}\ \s+(de | De | DE)\ \s+(Enero | ENERO | enero | Febrero | FEBRERO | febrero | Marzo | MARZO | marzo | Abril | ABRIL | abril | Mayo | MAYO | mayo | Junio | JUNIO | junio | Julio | JULIO | julio | Agosto | AGOSTO | agosto | Septiembre | SEPTIEMBRE | septiem

Nombre del modelo de fecha	Ejemplo de fecha etiquetada	Expresión Regular (ER) construida para el modelo de fecha
		bre|Setiembre|SETIEMBRE|setiembre|Octubre|OCTUBRE|octubre|Noviembre|NOVIEMBRE|noviembre|Diciembre|DICIEMBRE|diciembre)\ \s+(de|De|DE)\ \s+\ \d{4}
cincoERFecha	- "1 de Julio de 1,876" - "12 de Julio de 1,876"	\ \d{1,2}+\ \s+(de|De|DE)\ \s+[a-zA-Z][a-z]+\ \s+(de|De|DE)\ \s+\ \d\ \,\ \d{3}
seisERFecha	- "1 de Julio de 18,74" - "12 de Julio de 18,74"	\ \d{1,2}+\ \s+(de|De|DE)\ \s+[a-zA-Z][a-z]+\ \s+(de|De|DE)\ \s+\ \d{2}\ \,\ \d{2}
sieteERFechaA	- "1. de setiembre de 1868" - "10. de setiembre de 1868"	\ \d{1,2}\ \.+\ \s+(de|De|DE)\ \s+([a-zA-Z][a-z]+)\ \s+(de|De|DE)\ \s+\ \d{4}
sieteERFechaB	- "3. de noviembre 1876" - "12. de noviembre 1876"	\ \d{1,2}\ \.+\ \s+(de|De|DE)\ \s+([a-zA-Z][a-z]+)\ \s+\ \d{4}
ochoERFecha	- "1 de agosto 1874" - "12 de agosto 1874"	\ \d{1,2}\ \s+(de|De|DE)\ \s+([a-zA-Z][a-z]+)\ \s+\ \d{4}
nueveERFecha	- "1 de Julio de 1.876" - "12 de Julio de 1.876"	\ \d{1,2}\ \s+(de|De|DE)\ \s+([a-zA-Z][a-z]+)\ \s+(de|De|DE)\ \s+\ \d{1}\ \.\ \d{3}
diezERFecha	"marzo 1o de 1868"	([a-zA-Z][a-z]+) \ \s+(\ \d{1}[o])+\ \s+(de|De|DE)\ \s+\ \d{4}
onceERFecha	"agosto 1° de 1874"	([a-zA-Z][a-z]+) \ \s+\ \d{1}\ \°+\ \s+(de|De|DE)\ \s+\ \d{4}
doceERFecha	"1° de octubre de 1872"	\ \d\ \°+\ \s+([a-z][a-z]+)\ \s+([a-zA-Z][a-z]+)\ \s+(de|De|DE)\ \s+\ \d{4}
treceERFecha	- "1.872. Dicbre 2" - "1.872. Dicbre 27"	\ \d{1}\ \.\ \d{3}\ \.+\ \s+[A-Za-z][a-z]+ \ \s+\ \d{1,2}
catorceERFechaA	"1871 - 17 de octubre"	\ \d{4}\ \s+-\ \s+\ \d{1,2}+ \ \s+(de|De|DE)\ \s+[a-zA-Z][a-z]+
catorceERFechaB	"1871- 17 de octubre"	\ \d{4}-\ \s+\ \d{1,2}+\ \s+(de|De|DE)\ \s+[a-zA-Z][a-z]+
catorceERFechaC	"1871 -17 de octubre"	\ \d{4}\ \s+-\ \d{1,2}+\ \s+(de|De|DE)\ \s+[a-zA-Z][a-z]+
quinceERFecha	"1870. 24 de novbre"	\ \d{4}\ \.\ \s+\ \d{1,2}+\ \s+(de|De|DE)\ \s+[a-zA-Z][a-z]+

Tabla 9.

Modelos de fechas implementadas utilizando ER.

A continuación se muestra el resultado de etiquetar un texto del corpus utilizando los diferentes modelos de ER.

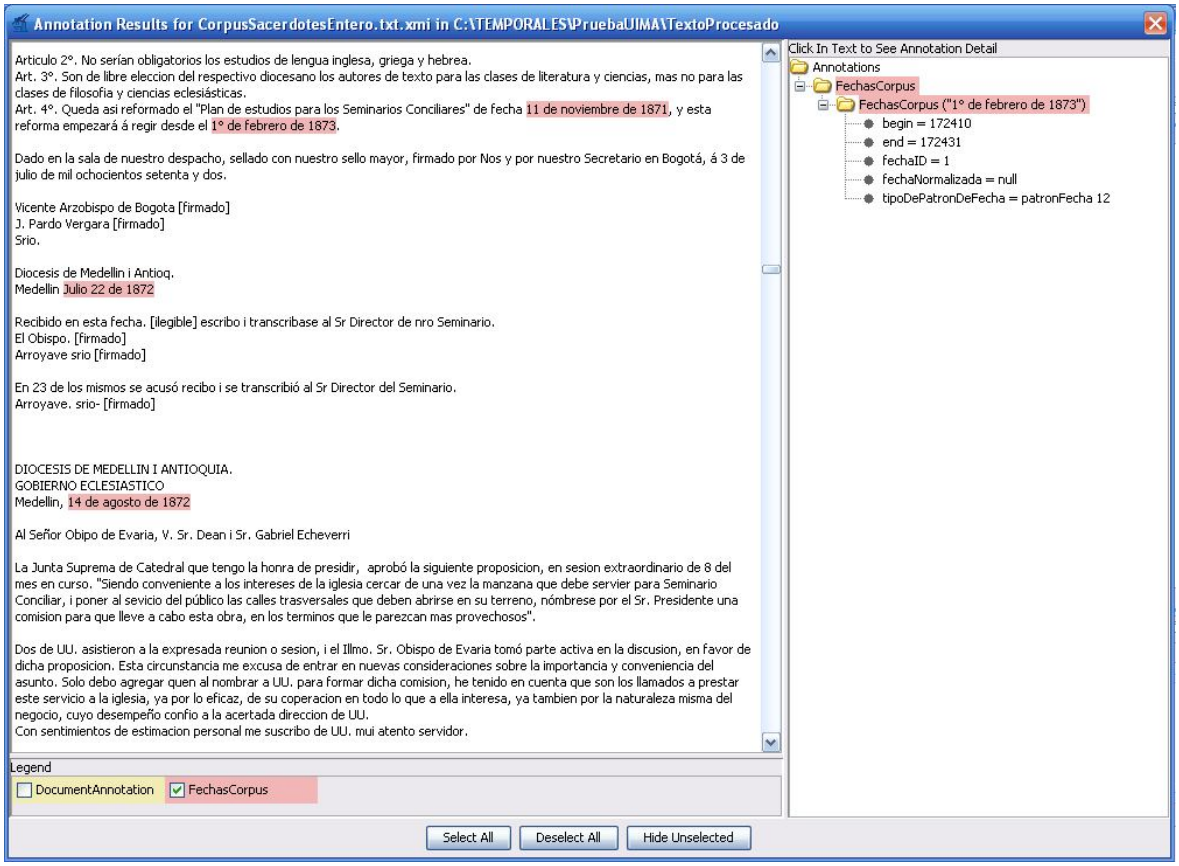


Figura 15.
Anotaciones correspondientes a modelos utilizando ER.

Listados que contienen palabras clave: puesto que el corpus también presenta otro tipo de fechas que utilizan palabras compuestas se elaboró un descriptor que contiene palabras con las que puede "armarse" ese tipo de fechas. En total se incluyeron 429 palabras discriminadas así: 108 tipos de palabras para meses y 293 ocurrencias y combinaciones para números (cientos, miles y otros). Algunas de las palabras empleadas en estos listados son:

Listados (parámetros)	Algunas palabras asociadas	Observaciones
ListadoMeses	Enero, febrero, marzo, abril… De enero, de mayo, Mayo de, enero de	
ListadoDias	Lunes, martes, miércoles, … Semana santa, cuaresma, pascua Días, noches, tardes, mañanas	Este listado fue eliminado de este análisis porque aunque son un elemento lingüístico útil para caracterizar la evolución de actividades en el tiempo, introducían mucha ambigüedad y su tratamiento requiere el análisis de anáforas, tema que puede dejarse como parte de un estudio más avanzado derivado de este trabajo.
ListadoCientos	Uno, dos, tres… veinte, veinticuatro, treinta, ochenta, doscientos	
ListadoMiles	Cien, trescientos, novecientos, mil	
ListadoNumerosEnLetras	Treinta y uno, treinta i uno, sexto, séptimo, veinte i dos	
ListadoNumerosDelMes	0, 1, 2,…31	

Tabla 10.
Listados de palabras incluidas en el análisis de fechas y ejemplos

Mediante estos modelos es posible etiquetar fechas que aparecen en el corpus tales como:

- *cinco de febrero de mil ochocientos setenta y cuatro*
- *primero de enero de mil ochocientos setenta y cuatro*
- *veintidos de Octubre del año de mil ochocientos sesenta i dos.*

La siguiente figura muestra las anotaciones utilizando este descriptor:

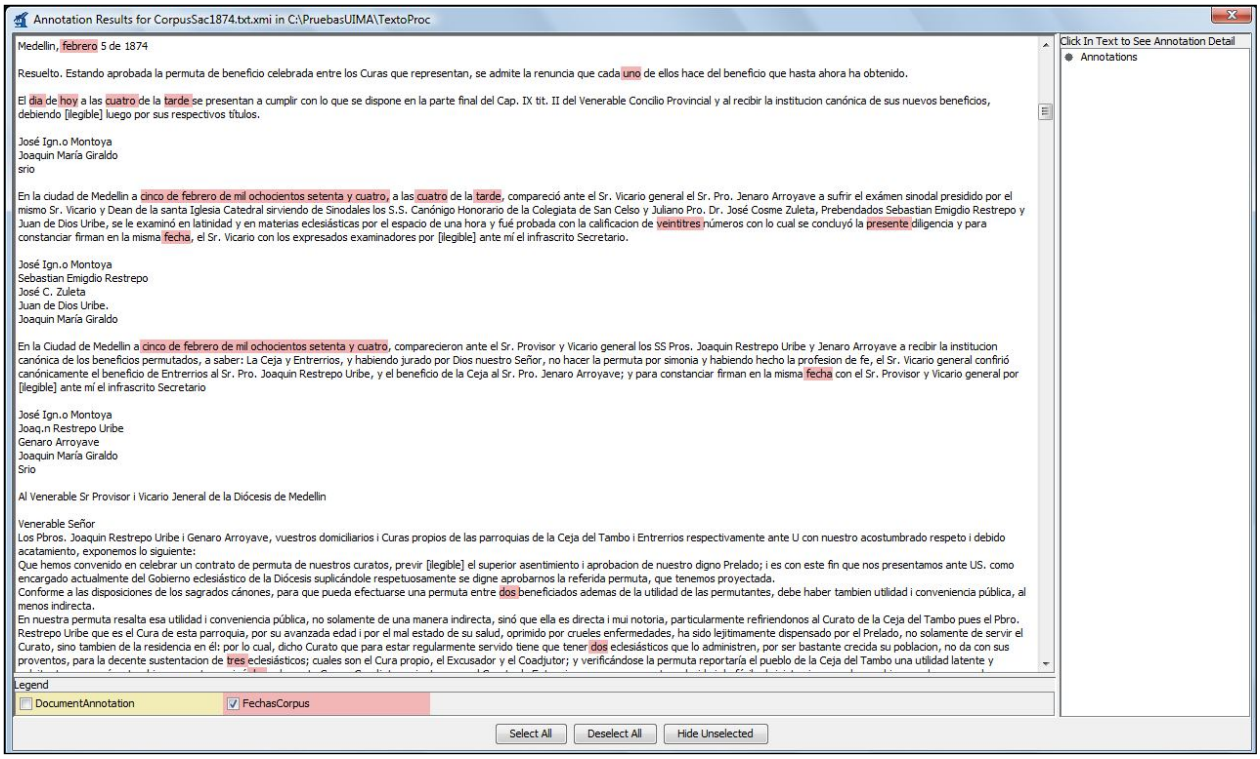


Figura 16.
Anotaciones de fechas que fueron implementadas con los listados

Sin embargo estos listados también introducen tres problemas de análisis:

1. Errores de asertividad en la precisión y cobertura: algunas de las palabras del listado no siempre etiquetan fechas completas aunque en el algoritmo empleado para el etiquetamiento "se forza" unir etiquetas que estuvieran contiguas y aquellas etiquetas que estén solas son eliminadas para disminuir en lo posible el impacto, por ejemplo de la frase "la suma de mil pesos" etiquetaba "mil".
2. Se etiquetan frases relacionadas con el manejo del tiempo cuyo grado de ambigüedad es mayor y por tanto requieren técnicas avanzadas para su análisis (algunas de ellas descritas en el capítulo 1) y que no están en el alcance de este trabajo. Desde este punto de vista se entiende porque el etiquetamiento de frases alusivas al tiempo motivó el análisis de eventos en textos lingüísticos. Algunas de estas frases son:
   - *medio día*
   - *cinco meses*

- *próximo agosto*
- *marzo próximo*
- *el año pasado*
- *el sábado en la noche*

3. En el texto también otro tipo de frases que son mucho más complejas de analizar y que requieren el estudio de anáforas para proceder con su desambiguación y no fueron susceptibles de etiquetamiento en este trabajo debido a su complejidad:
   - "*El martes que contamos 8 del que rije*…" (carta 14 de febrero de 1870)
   - "*Por el titulo espedido en 5 de febrero del corriente año…*" (carta abril 27 de 1870)

3.2.4 Etiquetas para nombres de lugares geográficos. La palabra "lugar" es demasiado amplia para ser empleada en un estudio de lingüística computacional ya que puede abarcar tanto lugares geográficos (pueblos, ciudades, departamentos, países) como lugares relacionados con inmuebles (edificios, seminarios, etc.). Para este proyecto analizaron los modelos de las ER necesarios para hallar lugares geográficos teniendo en cuenta que para identificar un "evento" los datos mínimos requeridos son la fecha y el lugar geográfico (Smith 2002a, 2002b) y por lo tanto se considerará como objeto de análisis posterior el etiquetamiento correspondiente a otras entidades tales como organizaciones.

Los lugares geográficos presentes en el corpus se encuentran en los encabezados de las cartas y en su texto. En principio se pensaría que esta posición facilitaría el modelo de etiquetamiento lo cual no es cierto ya que no siempre son fáciles de localizar lugares cuando no se tiene algún tipo de signo de puntuación de referencia. La mayoría de los lugares están conformados por una sola palabra que usualmente comienza por la letra mayúscula (Ej. Aranzazu) aunque también se encuentran poblaciones con dos palabras (Ej. Villa María) y excepcionalmente lugares con tres palabras (Ej. Ceja del Tambo) o cuatro palabras (Ej. Santa rosa de Cabal) presentando éstas últimas mayor nivel de complejidad.

Modelos de etiquetamiento utilizados

Modelamiento utilizando ER. Para hallar las ER relacionadas con los lugares geográficos presentes en el texto se hizo un análisis de sus características lingüísticas y se optó por

utilizar dos estrategias: la primera consistió en utilizar las palabras "claves" del idioma español para identificar un "lugar" en una oración como son las preposiciones (San Vicente, et al, 1991) y la segunda fue utilizar palabras "disparadoras" o las palabras del contexto. En la primera estrategia las palabras "claves" utilizadas fueron las siguientes preposiciones y conjunciones:

- Preposición "en": de acuerdo al DRAE (2003): "*Denota en qué lugar, tiempo o modo se realiza lo expresado por el verbo a que se refiere*", o de acuerdo al Diccionario del uso de la lengua de María Moliner (2003): "*Expresa el lugar dentro del cual está u ocurre la cosa de que se trata*". Ejemplo de uso de la preposición "en" el corpus: "*en Yolombó*".
- Preposición "de": de acuerdo al DRAE (2003) y al Diccionario de uso de la lengua (2003): "*Denota posesión o pertenencia*". Ejemplo de uso de la proposición en el corpus: "*de Bogotá*".
- Conjunciones "y" o "i", siendo ésta última una conjunción ya arcaica pero frecuente en el texto: aunque no son conjunciones que denoten lugar de manera explícita, se incluyeron en este análisis porque muchos lugares descritos en el corpus estaban acompañados por estas palabras, por ejemplo: "Obispo de Medellín y Antioquia" o "Obispo de Medellín i Antioquia". Sin embargo el uso de estas conjunciones también genera mucha ambigüedad y pueden también etiquetarse "falsos positivos" por la frecuencia que tienen estas palabras en el idioma.

En varios análisis previos se encontró que el nivel de asertividad para las ER que utilizaron la preposición "EN" es muy alto porque es una preposición utilizada para localizar y ubicar pero el nivel de asertividad para la preposición "DE" es mucho menor porque se presentaron muchos "falsos positivos". Es útil recordar que la preposición "de" tiene diversos usos en el idioma (se observan 23 usos distintos de acuerdo al DRAE, 2003) y debido a su alto nivel de ambigüedad fue necesario recurrir a las palabras del contexto para refinar la búsqueda, que fue la segunda estrategia utilizada. El criterio de escogencia de dichas palabras fueron los títulos de personas (Ej. "Cura de Angostura", "Prelado de Medellín", "Obispo de Antioquia") y también se elaboró un listado de palabras

observando porciones del corpus. Con el fin de dar mayor utilidad y generalidad a este trabajo, también se hizo el mismo ejercicio de búsqueda y análisis de palabras clave con varios archivos del corpus AFP de noticias en español (LDC, 2000) y se adicionaron al listado de palabras. Algunas de estas palabras se muestran a continuación.

Preposición "de"	Palabras claves presentes en el corpus histórico	Palabras claves presentes en varios artículos de noticias AFP
	administracion, administración, anexidades, beneficio, capilla, catedral, Catedral, catolicas, católicas, catolicos, católicos, cementerio, ciudad, coadjutor, coadjutoria, coadjutoría, curato, Curato, diario, distrito, distrito, Distrito, Distrito, distritos, Distritos, episcopal, escusador, fabrica, fábrica, fieles, Fieles, foráneo, foráneo, fracción, fracción, gobierno, Gobierno, habitante, habitantes, hijos, historia, hospital, Hospital, iglesia, Iglesia, interino, istmo, Istmo, limite, limites, municipal, municipio, Municipio, parroquia, Parroquia, parte, periodicos, periódicos, plaza, principal, propio, Pueblo, pueblo, puente, rei, relijiosas, republica, Republica, república, República, rey, santuario, seminario, Seminario, soberano, sociedad, Sociedad, templo, territorio, vecino, vecinos	Interior, cancillería, cancilleres, oriente, occidente, norte, sur, procedente, procedentes, población, estado, oposición, rebelión, salir, cerca, lejos, región, carnaval, localidad, localidades, gobierno, embajada, país, países, oriundo, natal, ciudad, banco, territorio, ciudadano, ciudadanos, torneo, kms, provincia, noroccidente, comercio, comerciantes, incursión, hidrográficas, provincia, departamento, militares, nuclear, afueras, afuera, historia, externa, desapareció, universidad, mediación, olímpicos, juegos, distrito, cárcel, mar.
Total Palabras:	**81**	**55**

Tabla 11. Palabras clave utilizadas en los corpus
(nota: algunas palabras aparecen sin tilde en el corpus)

Muchas de las palabras claves utilizadas para el corpus histórico fueron encontradas también en el corpus de noticias de AFP (LDC, 2000), tales como: distrito, gobierno, iglesia, provincia, estado, ciudad, presidente, etc. y realmente estos listados son complementarios ya que se trata del mismo idioma. Algunas palabras del corpus histórico no aparecen en el corpus de AFP porque son muy específicas (Ej. "Fracción", "coadjutor", "curato") o son arcaísmos (Ej. "anexidades", "relijiosos", "rei"). La siguiente tabla muestra los 24 tipos de patrones empleados para identificar lugares automáticamente en el texto utilizando ER:

Nombre del modelo de lugar	Ejemplo de lugar etiquetado	Expresión Regular (ER) construida para el modelo de lugar
unoERLugEnSimple	en Medellín	\\s(en|En|EN)\\s+([A-Z][a-zA-Záéíóúñªü-]+|[A-Z][AZÁÉÍÓÚÑª]+)(\\p{Punct}|(\\s\\p{javaLowerCase})|\\r\\n|\\r)
unoERLugEnDosPalabrasEspMin	en Nueva Granada	\\s(en|En|EN)\\s+([A-Z][a-záéíóúñªü-]+\\s+[A-Z][a-záéíóúñªü-]+)
unoERLugEnDosPalabrasEspMay	en NUEVA GRANADA	\\s(en|En|EN)\\s+([A-Z][A-ZÁÉÍÓÚÑª-]+\\s+[A-Z][A-ZÁÉÍÓÚÑª-]+)
dosERLugDeUnaPalabra	Cura de Medellin	\\s+(([a-z][a-záéíóúñªü-]+)|([A-Z][a-záéíóúñªü-]+))\\s+(de|De|DE)\\s+([A-Z][a-zA-Záéíóúñªü-]+)
dosERLugDeUnaPalabraMay	OBISPO DE ANTIOQUIA	\\s+[A-Z][A-ZÁÉÍÓÚÑª-]+\\s+(DE)\\s+([A-Z][A-ZÁÉÍÓÚÑüª-]+)
dosERLugDeDosPalabras	republica de Nueva Granada	\\s+[a-z][a-záéíóúñªü-]+\\s+(de|De|DE)\\s+[A-Z][a-zA-Záéíóúñªü-]+\\s+[A-Z][a-zA-Záéíóúñªü-]+
dosERLugDeDosPalabrasMay	OBISPO DE NUEVA GRANADA	\\s+[A-Z][A-ZÁÉÍÓÚÑüª-]+\\s+(DE)\\s+([A-Z][A-ZÁÉÍÓÚÑüª-]+\\s+[A-Z][A-ZÁÉÍÓÚÑüª-]+)
tresERLugarIMin	i/y Medellin	\\s+(i|I|y|Y)\\s+([A-Z][a-zA-Záéíóúñüª-]+|[A-Z][A-ZÁÉÍÓÚÑª-]+)
tresERLugarIMay	I/Y MEDELLIN	\\s+(i|I|y|Y)\\s+[A-Z][A-ZÁÉÍÓÚÑª-]+
diezERLugarEspecial1	Medellin, (comienzo de línea mas signo de puntuación)	\\r\\n+[A-ZÁÉÍÓÚ][a-zA-ZáéíóúñÁÉÍÓÚÑªü-]+
diezERLugarEspecial2	Medellin 18 de enero (comienzo de línea mas digito)de enero	\\r\\n+[A-ZÁÉÍÓÚ-][a-zA-ZáéíóúñÁÉÍÓÚÑªü]+ (\\s\\d{1,2})
diezERLugarEspecial3	Barbosa Febrero (comienzo de línea mas mes del año)	\\r\\n+[A-ZÁÉÍÓÚ-][a-zA-ZáéíóúñÁÉÍÓÚÑªü-]+\\s+(Enero|Febrero|Fbro|Marzo|Mzo|Abril|Mayo|Junio|Julio|Agosto|Septiembre|Setiembre|Set.e|Octubre|Noviembre|Nob.e|Diciembre|Dbre)
diezERLugarEspecial4	Barbosa febrero (comienzo de línea mas mes del año)	\\r\\n+[A-ZÁÉÍÓÚ-][a-zA-ZáéíóúñÁÉÍÓÚÑªü-]+\\s+(enero|en.o|fbro|febrero|marzo|mzo|abril|mayo|junio|julio|agosto|septiembre|setiembre|octubre|noviembre|diciembre)
diezERLugarEspecial5	San Carlos Abril (comienzo de línea mas mes del año)	\\r\\n[A-ZÁÉÍÓÚ-][a-zA-ZáéíóúñÁÉÍÓÚÑªü-]+\\s[A-ZÁÉÍÓÚ-][a-zA-ZáéíóúñÁÉÍÓÚÑªü-]+\\s+(Enero|Febrero|Marzo|Abril|Mayo|Junio|Julio|Agosto|Septiembre|Setiembre|Set.e|Set.|Octubre|Noviembre|Nob.e|Diciembre|Dbre)
diezERLugarEspecial6	San Carlos abril (comienzo de línea mas mes del año)	\\r\\n+[A-ZÁÉÍÓÚ-][a-zA-ZáéíóúñÁÉÍÓÚÑªü-]+\\s[A-ZÁÉÍÓÚ-][a-zA-ZáéíóúñÁÉÍÓÚÑªü-]+\\s+(enero|febrero|marzo|abril|mayo|junio|julio|agosto|septiembre|setiembre|set.e|octubre|noviembre|nob.e|diciembre|dbre)
diezERLugarEspecial7	Villa María 29 de enero	\\r\\n+[A-ZÁÉÍÓÚ-][a-zA-ZáéíóúñÁÉÍÓÚÑªü-]+\\s[A-ZÁÉÍÓÚ-][a-zA-ZáéíóúñÁÉÍÓÚÑªü-]+\\s+\\d{1,2}\\s+(de)\\s+(enero|fbro|febrero|marzo|mzo|abril|mayo|junio|julio|agosto|septiembre|setiembre|octubre|noviembre|diciembre|Enero|Febrero|Marzo|Abril|Mayo|Junio|Julio|Agosto|Septiembre|Setiembre|Set.e|Octubre|Noviembre|Nob.e|Diciembre|Dbre)
diezERLugarEspecial8	Dado en Medellin	\\r\\n+(Dado en)\\s+[A-Z][a-zA-Záéíóúñüª-]+

Nombre del modelo de lugar	Ejemplo de lugar etiquetado	Expresión Regular (ER) construida para el modelo de lugar
	á	
diezERLugarEspecial9	Gobierno Ecco Medellin	\ \r\ \n+(Gobierno|Gobno|Gbno|Gierno)\ \s+(eco|ecco|eclesiástico|ecco.|Ecco|Eco|Eclesiástico)(\ \s| \ \n)[A-Z][a-zA-Záéíóúñüª-]+
diezERLugarEspecial10	Gobierno Ecco. Medellin	\ \r\ \n+(Gobierno|Gobno|Gbno|Gierno)\ \s+(eco|ecco|ecco.|eclesiástico|Ecco|Eco|Eccio|Eclesiástico)\ \p{Punct}(\ \s| \ \n)[A-Z][a-zA-Záéíóúñüª-]+
diezERLugarEspecial11	Gbno Ecco. - Medellin, 3. de noviembre 1876	\ \r\ \n+(Gobierno|Gobno|Gbno|Gierno)\ \s+(eco|ecco|ecco.|eclesiástico|Ecco|Eco|Eccio|Eclesiástico)\ \p{Punct}\ \s+(-)\ \s+[A-Z][a-zA-Záéíóúñüª-]+
diezERLugarEspecial12	Vicaria Capitular Medellin Febrero 5 de 1875	\ \r\ \n(Vicaria Capitular)\ \s+[A-ZÁÉÍÓÚ-][a-zA-ZáéíóúñÁÉÍÓÚÑªü-]+\ \s+(enero|fbro|febrero|marzo|mzo|abril|mayo|junio|julio|agosto|septiembre|setiembre|octubre|noviembre|diciembre|Enero|Febrero|Marzo|Abril|Mayo|Junio|Julio|Agosto|Septiembre|Setiembre|Set.e|Octubre|Noviembre|Nob.e|Diciembre|Dbre)
diezERLugarEspecial13	en Rio de Janeiro	(\ \n| \ \s)(en|En|EN)\ \s+([A-Z][a-zA-ZáéíóúñªüZÁÉÍÓÚ-]+\ \s+(de|De|DE)\ \s+([A-Z][a-zA-Záéíóúñªüz ÁÉÍÓÚ-]+))(\ \r\ \n|(\ \s\ \p{javaLowerCase})| \ \ª| \ \p{Punct}|$)
diezPatronLugarEspecial14	San Carlos Abril (comienzo de línea mas mes del año)	\ \r\ \n[A-ZÁÉÍÓÚ-][A-ZÁÉÍÓÚÑªü-]+\ \s[A-ZÁÉÍÓÚ-][A-ZÁÉÍÓÚÑªü-]+\ \p{Punct}
diezERLugarEspecial15	de Medellín (incluye las palabras claves)	\ \s+(Administracion|Administración|administración|administracion|afuera|afueras|anexidades|Arquidiocesis|banco|beneficio|cancilleres|cancillería|capilla|Cardenal|carcel|cárcel|carnaval|Catedral|catedral|Catolicas|catolicas|católicas|Catolicos|catolicos|católicos|cementerio|cerca|ciudad|ciudad|ciudadano|ciudadanos|coadjutor|coadjutoria|coadjutoría|coadjuntor|comerciantes|comercio|Cura|cura|Curato|curato|Departamento|departamento|desapareció|diario|Diario|diario|Diocesis|diocesis|Diócesis|diócesis|distrito|Distrito|distrito|distrito|distrito|Distrito|Distritos|distritos|eclesiástica|embajada|episcopal|escusador|estado|externa|fabrica|fábrica|fieles|Fieles|foraneo|foráneo|fraccion|fracción|Gobierno|gobierno|gobierno|habitante|habitantes|hidrográficas|hijos|historia|historia|Hospital|hospital|Iglesia|iglesia|incursión|interino|Interior|interior|Istmo|istmo|juegos|kms|lejos|limite|limites|localidad|localidades|Mar|mar|militares|municipal|Municipio|municipio|natal|noroccidente|norte|nuclear|Obispo|Obpo|occidente|olimpicos|oposición|oriente|oriundo|pais|paises|Parroquia|parroquia|parte|periodicos|periódicos|plaza|población|Prelado|prelado|Presbitero|presbitero|Presidencia|presidencia|Presidente|presidente|principal|procedente|procedentes|propio|provincia|provincia|pueblo|Pueblo|puente|rebelión|region|rei|relijiosas|Republica|republica|República|república|Rey|rey|salir|santuario|seminario|Seminario|señoras|Señoras|señores|Señores|soberano|Sociedad|sociedad|sur|templo|territorio|territorio|torneo|universidad|vecino|vecinos|Vicaría|Vicario|vicario)\ \s+(de|De|DE)\ \s+([A-Z][a-zA-Záéíóúñªü]+)

Tabla 12.

Modelos de ER para identificar lugares

Listados de lugares: se implementaron listados de lugares para identificar los nombres que están conformados por varias palabras puesto que el modelo para implementarlas utilizando ER ofrece un mayor grado de ambigüedad. Ejemplos de lugares implementados son "Santa Fe de Bogotá" o "Ceja del Tambo". Los listados fueron discriminados así: listados de países (214 ocurrencias), listados de ciudades incluyendo las ciudades de Colombia y algunas más importantes del mundo (57), pueblos de Antioquia (130) y barrios de Medellín (287). En total se utilizaron 688 palabras. Los listados en su gran mayoría se sacaron de la Internet pero también se incluían palabras presentes en el corpus. La siguiente figura muestra diferentes tipos de lugares etiquetados:

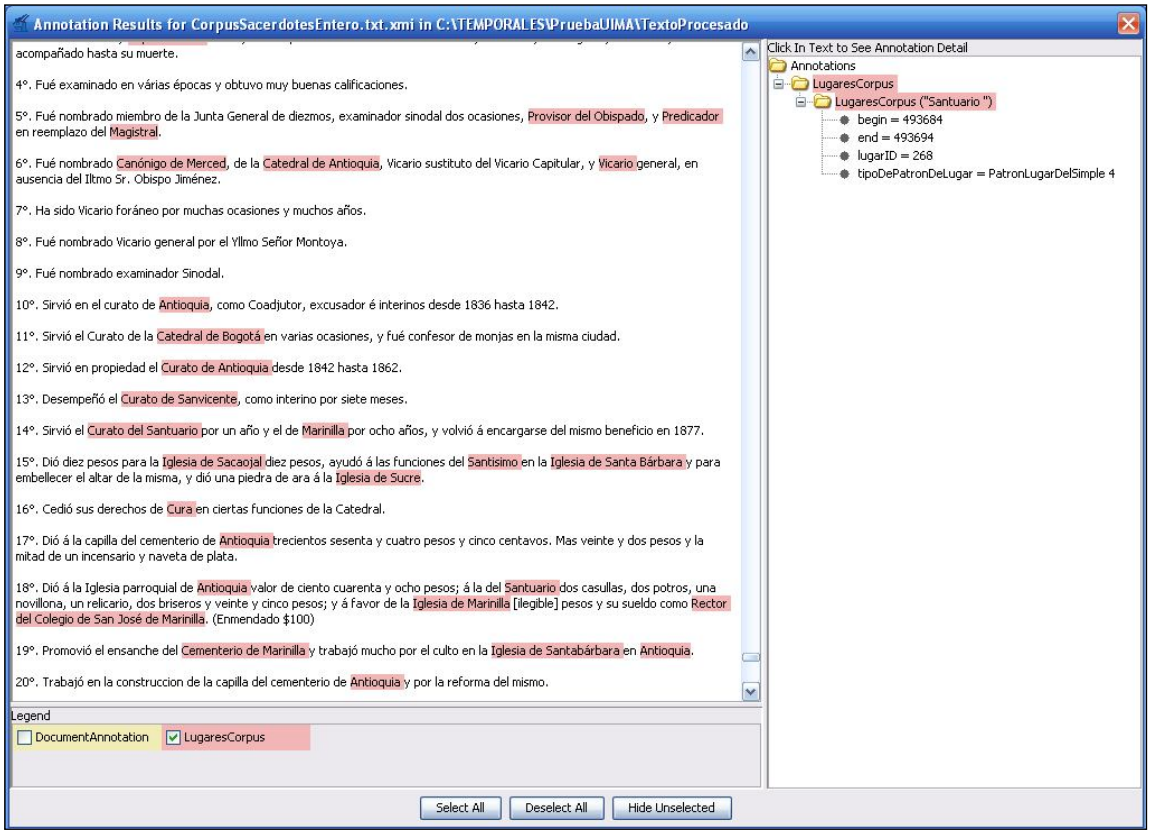


Figura 17. Etiquetas de lugares

3.2.5 Etiquetas para nombres de personas. De acuerdo a la DRAE (2003) un "nombre" se define como "*palabra que designa o identifica seres animados o inanimados*". En este proyecto se busca reconocer de manera automática nombres propios de personas incluidas en el texto

y no se tendrán en cuenta los nombres que puedan hallarse para objetos inanimados como por ejemplo, nombres de iglesias o de colegios (esta identificación de nombres puede incluirse en un análisis posterior). Por lo tanto, para efectos de reconocimiento de ER se definirá un nombre de persona como dos o más palabras que obedecen los siguientes criterios:

- Comienzan por una letra mayúscula y pueden ir precedidas por palabras "clave" para los nombres, tales como títulos.
- Varias palabras en mayúsculas que no sean inicio de frase.
- Varias palabras que comiencen con una letra mayúscula y que estén juntas en una frase o al comienzo de una línea.

La metodología utilizada para encontrar nombres de personas en el corpus es similar a la utilizada para hallar lugares: se analiza una porción del corpus para determinar las características lingüísticas y con base en ese análisis se configuran las ER y posteriormente se prueban los modelos con el resto del corpus.

El corpus histórico empleado para la caracterización es particularmente rico en nombres propios debido a su naturaleza epistolar y es común encontrar por ejemplo, los nombres de todo un vecindario entero solicitando un sacerdote. De esta manera se tienen varios tipos de combinaciones entre nombres y apellidos. A continuación se muestra una porción del corpus que ilustra esta consideración:

"Esperamos Ilustrísimo Señor, que en vuestra sabiduría, e inspirado por el amor al rebaño que apacentais, nos escuchareis con benevolencia dejando satisfechas nuestras ardientes i sinceras esperanzas.

Ilustrísimo Señor.
Salamina, 30 de junio de 1869

Lorenzo Escobar [firmado]
Juan de la C. Cevallos [ilegible] [firmado]
Luis Escovar [firmado]
Ramon C. [ilegible] [firmado]
Alfonso Robledo [firmado]
Manuel Isaza L [firmado]
Jose Ma Ospina Pineda [firmado]
Claudino Escobar [firmado]
Celestino García [firmado]
Dionisio J. Mejía. [firmado]
Narciso Londoño [firmado]
Francisco A. Escovar [firmado]
Mariano Ospina D. [firmado]
Juan N Trujillo A. [firmado]
Neru Alvarez [ilegible] [firmado]
Domingo Gallo [firmado]
Juan Mª Mejía [firmado]
Remipio Mª Mejia [firmado]
Ricardo Echeverri [firmado]

Nota. La presente solicitud podria llevar las firmas de todos los habitantes de Salamina, con rarísimas excepciones, pero hemos creido que las que ella contiene son bastantes para probar la justicia que nos asiste y la necesidad en que estamos de molestar vuestra atencion, para que os Digneis remediar los males que hoy nos afligen."

Figura 18.
Detalle del corpus histórico que muestra listados de personas

Puesto que el corpus se elaboró teniendo en cuenta las reglas paleográficas del Archivo General de la Nación (Ladrón, 1996), se observa en el texto anterior las etiquetas [ilegible] o [firmado] que también se tuvieron en cuenta en las expresiones regulares utilizadas y que fueron muy útiles a la hora de etiquetar los listados en donde los nombres

comenzaban al principio de la línea y de hecho, es necesaria ponerlas en el texto porque si se prescinden de ellas las expresiones regulares resultantes abarcan nombres consecutivos, lo cual implica que en el análisis de listados de personas es necesario incluir un dato adicional al sistema para hacer la distinción entre ellas. Para ilustrar este caso, se muestra la siguiente figura donde se tiene un listado de personas del corpus y cuando se observa una anotación, ésta abarca otros nombres de personas presentes en el texto:

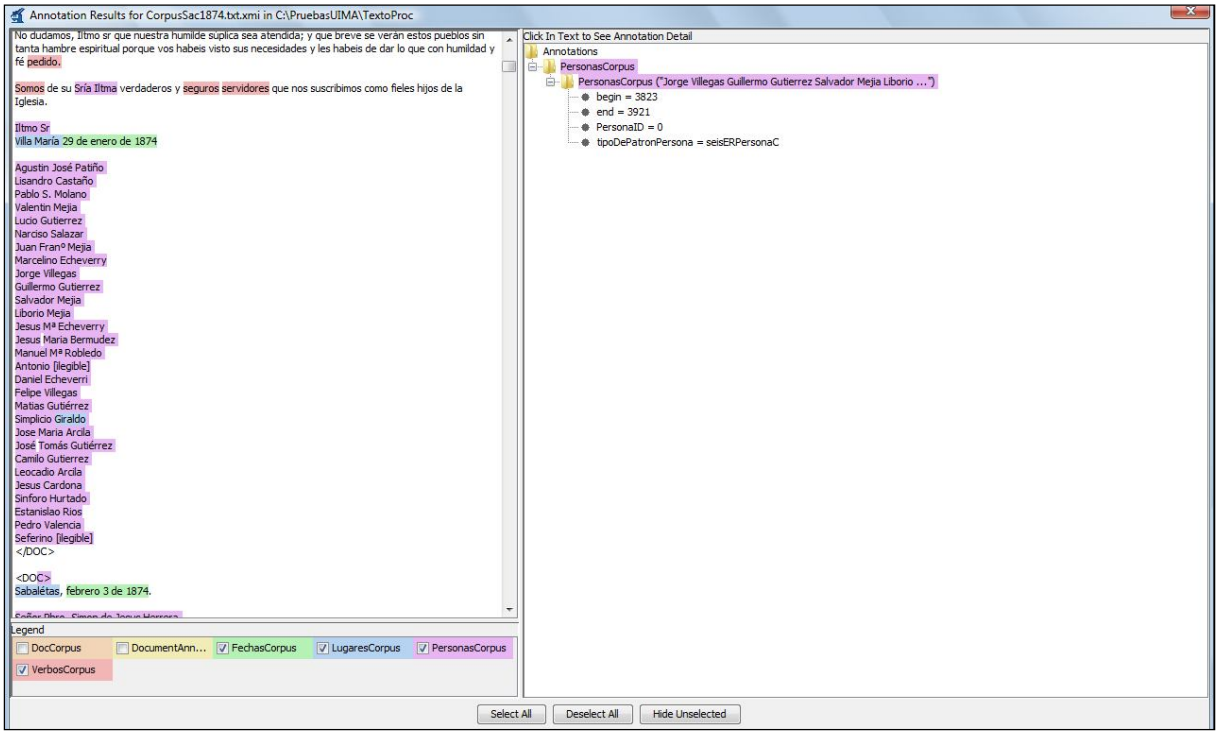


Figura 19.
Etiquetamiento incorrecto de nombres de personas presentes en un listado

Cuando se hace el mismo ejercicio en un listado que contenga una palabra que delimite donde termina y comienza otro nombre propio el resultado es mucho más preciso, como se muestra en esta porción de corpus:

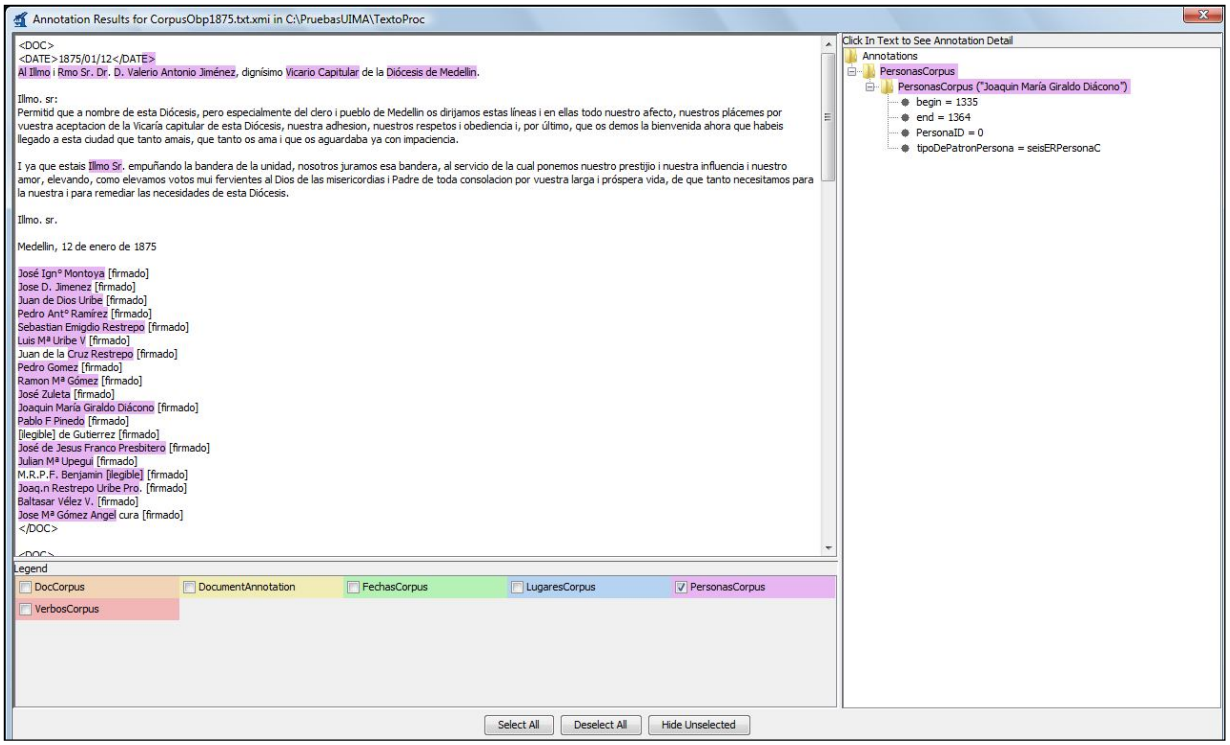


Figura 20.
Etiquetamiento corregido para nombres de personas presentes en un listado

El siguiente cuadro muestra los diferentes tipos de modelos de ER empleados para etiquetar nombres en el corpus:

Nombre del modelo persona	Ejemplo de persona etiquetado	Expresión Regular (ER) construida para el modelo de persona
unoERPersona	Pro Juan Nepomuceno Cadavid,	(Alcalde | alcalde | Alcaldes | alcaldes | Alcaldesa | alcaldesa | Alcaldesas | alcaldesas | Arquitecta | arquitecta | Arquitecto | arquitecto | Arzobispo | ARZOBISPO | Arzobispos | ARZOBISPOS | aviador | aviadora | canonigo | Canonigo | canónigo | Canónigo | canonigos | canónigos | Cantante | cantante | catolica | católica | Catolica | Católica | capellan | Capellan | capellán | Capellán | capellanes | Capellanes | capellánes | Capellánes | Capitan | Capitán | Cardenal | Cardenales | clerigo | Clerigo | clérigo | clerigos | clérigos | Clérigos | clero | Clero | coadjutor | Coadjutor | coadjutores | Coadjutores | Coronel | coronel | cura | Cura | Curador | curador | curas | Curas | D.r | Dean | Deán | Delegadas | Delegado | Deportista | deportista | Diputada | diputada | Diputado | diputado | diputados | Director | director | Directora | directora | Dn | Dn. | Doctor | doctor | Doctora | doctora | Doctoras | doctoras | Doctores | Doctores | Don | Doña | doña | don | Dr | Dr. | Embajador | embajador | Embajadora | embajadora | Enfermera | enfermera | Enfermero | enfermero | Escolta | escolta | escritor | escritora | Funcionaria | funcionaria | Funcionario | funcionario | General | general | Generales | generales | Guardaespalda | guardaespalda | Guardaespaldas | guardaespaldas | Ilmo | Illma | Illmas | Illmo | IllmoS.O. | Illmo. | Illmos | Illmos. | Ilma | Ilmª | Ilmo | Ilm

Nombre del modelo persona	Ejemplo de persona etiquetado	Expresión Regular (ER) construida para el modelo de persona
		os | Ilmos | Ilsma | Iltima. | Iltmo | Iltmos | Ilto | Iltos | Ilustrisima | Ilustrísima | Ilustrisimo | Ilustrísimo | Ilustrisimos | Ilustrísimos | Imo | Imos | Ingeniera | ingeniera | Ingeniero | ingeniero | Itmo | Jefa | jefa | jefas | Jefas | Jefe | jefe | Jefes | jefes | jóven | joven | M.Imo | M.Imo. | Medica | medica | Médica | médica | Medicas | Médicas | Medico | medico | Médico | médico | Médicos | médicos | Ministra | Ministras | Ministro | ministro | Ministros | ministros | monseñor | Monseñor | N.S.P. | Notario | notario | Notarios | notarios | Obispo | OBISPO | Obispos | OBISPOS | Obpo | Obpos | padre | Padre | padres | Padres | Papa | PAPA | PAPAS | Papas | Parroco | Párroco | Parrocos | Párrocos | Pastor | Pastores | patinador | Patinador | patinadora | Patinadora | Pbo | Pbro | Pbro. | Pbros | Pbros. | policia | policía | Pontifice | pontifice | Pontífice | pontífice | Prebendado | Prebendados | Prelado | prelado | Prelados | prelados | Presbitero | presbítero | Presbítero | Presbiteros | presbíteros | Presbíteros | presidenta | Presidente | presidente | Presidentes | presidentes | Pro | Pro. | Profesor | profesor | Profesora | profesora | Provisor | Provisores | Por | Rector | Rectores | Representante | Representantes | Reverendísimo | reverendísimo | Reverendísimos | reverendísimos | RImo | Rma | Rmas | Rmo | Rmos | Rndo | Rndo P. | Rndos | RR II | S.Sria | S.R. | S.S. | S.S.IIlma | S.S.I | S.S.ILLMA | sacerdote | Sacerdote | sacerdotes | Sacerdotes | Secretaria | secretaria | secretarias | Secretario | secretario | secretarios | Senador | senador | Senadora | senadora | senor | Senor | señor | Señor | Señora | señora | señores | Señores | señoria | Señoria | señoría | Señoría | Sñ. | Sor | Sor. | sr | Sr | sr. | Sr. | sra | Sra | Sra. | Sres | sres. | Sres. | Sría | srio | srio. | srios | srios. | Srr | Srr. | SSIIlma. | SS.Im | SSria | Subdirector | subdirector | Subdirectora | subdirectora | Teniente | Tesorera | tesorera | Tesorero | tesorero | U.I.I. | U IIlma | US Ilma | US. | US. IIlma | US.I | US. IIlma | Usía | Usía Ima | vicario | Vicario | Vicario Capitular | Vicario capitular | vicarios | Vicarios | Vicarios Capitulares | Vicepresidente | Vicepresidentes | YIlmo | Presidenta | Nos | comerciante)+\ \s+(\ \p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-Záéíóúñªüºº]+)
tresERPersonaA	Pro Dr Juan N. Cadavid o Pro Dr Juan N Cadavid	(LISTADO DE PALABRAS CLAVES ANTERIORES)\ \s+ (\ \p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+) \ \s+(\ \p{javaUpperCase}| \ \p{javaUpperCase}\ \p{Punct})\ \s+ (\ \p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)
tresERPersonaB	Dr. Rudecindo Ma. Correa ó Dr. Valerio Ant. Jimenez	(LISTADO DE PALABRAS CLAVES ANTERIORES)\ \s+ (\ \p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\ \s+(\ \p{javaUpperCase} [a-zA-Záéíóúñªüºº-]{1,2}\ \p{Punct})\ \s+ (\ \p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)
tresERPersonaC	presbítero José M. Gómez Anjel o presbítero José M Gómez Anjel	(LISTADO DE PALABRAS CLAVES ANTERIORES)\ \s+ (\ \p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+) \ \s+(\ \p{javaUpperCase}| \ \p{javaUpperCase}\ \p{Punct})\ \s+ (\ \p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)

Nombre del modelo persona	Ejemplo de persona etiquetado	Expresión Regular (ER) construida para el modelo de persona
cuatroERPersona	Pbro José María Gomez Angel	(LISTADO DE PALABRAS CLAVES ANTERIORES)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-záéíóúñªü°-]+)
cincoERPersonaA	Pbro. Simon de Jesus Herrera	(LISTADO DE PALABRAS CLAVES ANTERIORES)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(de|De|DE)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
cincoERPersonaB	Subdirectora Petronila de Macías	(LISTADO DE PALABRAS CLAVES ANTERIORES)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+) \\s+(de|De|DE)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
cincoERPersonaC	Secretaria Juana Arango de V. (que puede estar en el principio de una linea o antes de un título	(LISTADO DE PALABRAS CLAVES ANTERIORES)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA- Záéíóúñªü°º-]+)\\s+(de)\\s+(\\p{javaUpperCase}{1,2}|\\p{javaUpperCase}{1,2}\\p{Punct})
cincoERPersonaD	Pbro. doctor José Ramos Duran de C.	(LISTADO DE PALABRAS CLAVES ANTERIORES)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA- Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(de|De|DE)\\s+(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct })
seisERPersonaA	Juan Ventura	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+) \\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
seisERPersonaB	Juan Ventura Suluaga	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA- Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
seisERPersonaC	Pedro Pablo Salazar Euse	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase} [a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
sieteERPersonaA	Francisco J. Cardona o Tomas J Bernal	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+) \\s+(\\p{javaUpperCase}\\p{Punct}|\\p{javaUpperCase})\\s+(\\p{javaUpp erCase}[a-zA-Záéíóúñªü°º-]+)
sieteERPersonaB	Nicolás Londoño Z. ó Francisco Muñoz M	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}\\p{Punct}|\\p{javaUpperCase})\\s*
sieteERPersonaC	Juan N Trujillo A o Juan N Trujillo A. o Juan N. Trujillo A o Juan N. Trujillo A.	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}\\p{Punct}|\\p{javaUpperCase})\\s+(\\p{javaUpperCas e}[a-zA-Záéíóúñªü°º-]+)\\s*(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})
ochoERPersonaA	señor Koch o Pbro Ramirez	(Cura|cura|Pbro|Pbro.|Pro|Pro.|Pbo|Presbitero|Presbítero|presbítero|padr e|Padre|doctor|Doctor|Dr|Dr.|D.|Sor.|Sor|Señor|Sr.|Sr| señor|senor|Canónigo|Canonigo|secretario|srio)+\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
ochoERPersonaB	señor Gómez Anjel o	(Cura|cura|Pbro|Pbro.|Pro|Pro.|Pbo|Presbitero|Presbítero|presbítero|padr

Nombre del modelo persona	Ejemplo de persona etiquetado	Expresión Regular (ER) construida para el modelo de persona
	senor Moritz Koch	e|Padre|doctor|Doctor|Dr|Dr.|D.|Sor.|Sor|Señor|Sr.|Sr|señor|senor|Canónigo|Canonigo|secretario|srio|jóven|joven|Jóven)+\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)
ochoERPersonaC	Pro Silverio A. Gomez o Pro Silverio A Gomez y sus variantes	(Cura|cura|Pbro|Pbro.|Pro|Pro.|Pbo|Presbitero|Presbítero|presbítero|padre|Padre|doctor|Doctor|Dr|Dr.|D.|Sor.|Sor|Señor|Sr.|Sr|señor|senor|Canónigo|Canonigo|secretario|srio)+ \\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)
ochoERPersonaE	Pablo F. Pineda Pbro.	(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (Cura|cura|Pbro|Pbro.|Pro|Pro.|Pbo|Presbitero|Presbítero|presbítero|padre|Padre|doctor|Doctor|Dr|Dr.|D.|Sor.|Sor|Señor|Sr.|Sr|señor|senor|Canónigo|Canonigo|secretario|srio)
nueveERPersonaA	Jacobo J. H.	(\\n|\\p{javaLowerCase}\\s+|(\\p{Punct}[^<>]\\s*))(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (\\p{javaUpperCase}|\\p{javaLowerCase})(\\p{Punct})\\s+ (\\p{javaUpperCase}|\\p{javaLowerCase})((\\s*\\p{Punct}\\s*)|\\s+\\p{javaLowerCase}|\\s*\\r+|\\s+\\t*\\r|\\t*\\r|\\r)
nueveERPersonaB	, Vicente A. Restrepo, o , Vicente A Restrepo, y sus variantes	(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)
nueveERPersonaC	Manuel d. J. Ocampo	(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (\\p{javaUpperCase}|\\p{javaLowerCase})(\\p{Punct})\\s+(\\p{javaUpperCase}|\\p{javaLowerCase})((\\s*\\p{Punct}\\s*))\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)
nueveERPersonaD	Marco A. Peláez J. o Marco A Peláez J y sus variantes	(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})
nueveERPersonaE	J. Muñoz o J Muñoz	(\\p{javaUpperCase}\\p{Punct}|\\p{javaUpperCase})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)
nueveERPersonaF	A. Moreno C. y sus variantes	(\\p{javaUpperCase}\\p{Punct}|\\p{javaUpperCase})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+(\\p{javaUpperCase} \\p{Punct}|\\p{javaUpperCase})
nueveERPersonaG	C Alberto Vélez J o C. Alberto Vélez J. y sus variantes	(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})
nueveERPersonaH	P. Ignasio Vergara	(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})
diezERPersonaA	Eduvigio V	(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+ (\\p{javaUpperCase}\\p{Punct}|\\p{javaUpperCase})
diezERPersonaB	Ramon Ma. Lónjas	(\\p{javaUpperCase}[a-zA-Záéíóúñªüºº-]+)\\s+(\\p{javaUpperCase}[a-zA-

Nombre del modelo persona	Ejemplo de persona etiquetado	Expresión Regular (ER) construida para el modelo de persona
		Záéíóúñªü°º-]+\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
diezERPersonaC	Jesus Ma. Garcia G o Jose Ma. Hernández P.	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})
diezERPersonaD	Jose Mª Angel H o Jose Mª Angel H.	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+((\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-])|(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]\\p{Punct}))\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})
onceERPersonaA	C de Monroy o Roman de Hoyoz	(\\p{javaUpperCase}|(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+))\\s+(de|De|DE)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
onceERPersonaB	Joaquin de Posada A. o Joaquin de Posada A	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(de|De|DE)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})
onceERPersonaC	J. María de Vivez o J María de Vivez	(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(de|De|DE)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
onceERPersonaD	Juan de J Orozco o Juan de J. Orozco	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(de|De|DE)\\s+(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
onceERPersonaE	Juan de Jesus Echeverri o Amigo de Juan Arenas	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(de|De|DE)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°ºº-]+)
onceERPersonaF	Nacianceno Hernandez Arango de Lorenzo	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°ºº-]+)\\s+(de|De|DE)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
onceERPersonaG	,Pastora Vásquez de Villa. o ,Inés Posada de V. o ,Inés Posada de V	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(de|De|DE)\\s+(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct}|(\\p{javaUpperCase}[a-zA-Záéíóúñªü°ºº-]+))
doceERPersonaA	José del Cármen López	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(del|Del|DEL)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°ºº-]+)
doceERPersonaB	Josefa Alvarez del Pino	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(del|Del|DEL)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°ºº-]+)
treceERPersonaA	Juan de la C. T. y sus variantes	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(de|De|DE)\\s+(la|La|LA)\\s+(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})
treceERPersonaB	Juan de la C. Tovon	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(de|De|DE)\\s+(la|La|LA)\\s+(\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
catorceERPersonaA	Pablo Antonio Balovizar i Alvarez	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)

Nombre del modelo persona	Ejemplo de persona etiquetado	Expresión Regular (ER) construida para el modelo de persona
		\\s+(i|I|y|I)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
quinceERPersonaA	José Ign.o Montoya	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+ \\p{Punct}{1}[aeion])\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
quinceERPersonaB	Joaq.n Restrepo Uribe	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+\\p{Punct}[aeion]) \\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
diezYSeisERPersona	i Genaro Arroyave y a José Joaquin	([iíaáoó])\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+((\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)| (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+))
diezYSieteERPersona	Faustina Estrada (hija)	(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA- Záéíóúñªü°º-]+)\\s+(\\p{Punct}(Madre|madre|Hija|hija) \\p{Punct})
diezYOchoERPersonaA	Antonio [ilegible] o Antonio Maria [ilegible]	((\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)|(\\p{javaUpperCase}[a-zA- Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase} [a-zA-Záéíóúñªü°º-]+))\\s+(\\p{Punct} (ilegible|Ilegible)\\p{Punct})
diezYOchoERPersonaB	Gerardo [ilegible] Arias o Gerardo Antonio [ilegible] Arias	((\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)|(\\p{javaUpperCase}[a-zA- Záéíóúñªü°º-]+)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+))\\s+(\\p{Punct}(ilegible|Ilegible) \\p{Punct})\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
diezYOchoERPersonaC	[ilegible] Ramirez Muñoz	(\\p{Punct}(ilegible|Ilegible)\\p{Punct})\\s+(\\p{javaUpperCase}[a-zA- Záéíóúñªü°º-]+)\\s*(\\p{javaUpperCase} [a-zA-Záéíóúñªü°º-]+)
diezYNueveERPersonaA	Por Roque Franco Jesus Ramirez	(Por|por|POR)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)
diezYNueveERPersonaB	Por Roque Franco Jesus Ramirez B o Por Roque Franco Jesus Ramirez B	(Por|por|POR)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})
diezYNueveERPersonaC	Por Antonio Arias Cardona Jesus Ramirez B o Por Antonio Arias Cardona Jesus Ramirez B.	(Por|por|POR)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})
diezYNueveERPersonaD	Por Manuel Redondo Juan P. Correa o Por Manuel Redondo Juan P Correa	(Por|por|POR)\\s+(\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}[a-zA-Záéíóúñªü°º-]+)\\s+ (\\p{javaUpperCase}|\\p{javaUpperCase}\\p{Punct})\\s+(\\p{javaUpperCas

Nombre del modelo persona	Ejemplo de persona etiquetado	Expresión Regular (ER) construida para el modelo de persona
		e}[a-zA-ZáéíóúñªüºÑ-]+)
diezYNueveERPersonaE	Por Juan Jose Dias Jose Ma. Ramirez o Por Juan Jose Dias Jose Ma Ramirez	(Por|por|POR)\ \s+(\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ ((\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]{1,2} \ \p{Punct})|(\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]{1,2})) \ \s+(\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)
veinteERPersonaA	a ruego de Efecino [ilegible]	(a ruego de|á ruego de|A ruego de|Á ruego de)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{Punct}(ilegible|Ilegible)\ \p{Punct})
veinteERPersonaB	a ruego de Juan Jose Dias	(a ruego de|á ruego de|A ruego de|Á ruego de)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)
veinteERPersonaC	a ruego de Juan Jose Dias Rodriguez	(a ruego de|á ruego de|A ruego de|Á ruego de)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)
veinteERPersonaD	a ruego de Juan J. Dias	(a ruego de|á ruego de|A ruego de|Á ruego de)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}|\ \p{javaUpperCase}\ \p{Punct})\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)
veinteERPersonaE	a ruego de Juan J. Dias Rodriguez	(a ruego de|á ruego de|A ruego de|Á ruego de)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}|\ \p{javaUpperCase}\ \p{Punct})\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)\ \s+ (\ \p{javaUpperCase}[a-zA-ZáéíóúñªüºÑ-]+)

Tabla 13.
Modelos de ER para identificar nombres de personas

En total se construyeron 64 modelos de ER que describen los nombres de las personas en el corpus y, a continuación se muestran las siguientes observaciones metodológicas en cuanto a su proceso de construcción y caracterización:

Familias de nombres: para efectuar la caracterización de la mayor cantidad de nombres de personas y proceder a su modelaje por medio de ER se escogieron varias porciones donde aparecían listados de personas y se recogieron en una hoja en Excel donde se agrupaban y clasificaban de acuerdo a las características que ofrecían los nombres, por esta razón se

tienen los modelos divididos en varias categorías dadas por números que a su vez, se subdividen en letras del alfabeto con el fin de agruparlos por "familias de nombres" y encontrar el modelo de ER más general o particular según el caso. En total se configuraron 20 familias de nombres. A continuación se muestran ejemplos de familias de nombres:

Tipo de "familia de nombres de personas"	Tipo de persona	Nombre de persona	Modelo de ER
Personas 6	Persona 6A	Juan Ventura	seisERPersonaA
	Persona 6B	Juan Ventura Suluaga	seisERPersonaB
	Persona 6C	Pedro Pablo Salazar Euse	seisERPersonaC
Personas 9	Persona 9A	Jacobo J. H.	nueveERPersonaA
	Persona 9B	Vicente A. Restrepo o Vicente A Restrepo	nueveERPersonaB
	Persona 9C	Manuel d. J. Ocampo	nueveERPersonaC
	Persona 9D	Marco A. Peláez J. o Marco A Peláez J	nueveERPersonaD
	Persona 9E	J. Muñoz o J Muñoz	nueveERPersonaE
	Persona 9F	A. Moreno C. y sus variantes	nueveERPersonaF
	Persona 9G	C Alberto Vélez J o C. Alberto Vélez J.	nueveERPersonaG
	Persona 9H	P. Ignasio Vergara	nueveERPersonaH

Tabla 14

Ejemplos de familias de nombres empleados para modelar las ER

Caracterización de las ER: las familias de nombres de personas comprendidas entre uno a cinco utilizan palabras clave para la identificación de los nombres. Estas palabras clave han sido recopiladas del corpus de sacerdotes y el corpus de noticias de AFP (LDC, 2000) y corresponden a títulos de personas o títulos de profesiones o algunos lugares geográficos. Las restantes familias de ER fueron construidas teniendo en cuenta el contexto y las tipologías de los nombres. Existen modelos de nombres que incluyen letras o signos especiales, por ejemplo: "Juan Franº Mejía" o "Jose Mª Ángel". Las ER incluyen estos signos con el fin de detectarlos en el texto y por lo tanto son reconocibles. En caso tal de que se requiera incluir otro símbolo es necesario extenderlas.

Ambigüedades en los nombres: Existen varias familias de nombres que presentan un grado de ambigüedad alto como son por ejemplo aquellos nombres compuestos por un titulo y un apellido (Ej. "Pbro. Cadavid") y existe otra característica interesante

relacionada con las personas que firmaron por otras, como por ejemplo "Por Manuel Redondo Juan Pablo Correa", donde es muy difícil separar los dos nombres propios.

Para los corpus analizados no se presentaron las dificultades que muestran otros tipos de estudios similares a esta investigación pero hechos para el idioma inglés tales como el trabajo de Crane y Jones (2006) donde mostraban que el grado de ambigüedad de su sistema era alto porque muchos lugares geográficos se llaman igual que los nombres de personas. Para el caso de este trabajo de investigación los nombres de personas están diferenciados de los nombres geográficos y no se presentaron ambigüedades en este aspecto.

3.2.6 Etiquetamiento de los verbos que indican acciones en el texto. La manera en la cual se describen las acciones en el idioma español es por medio de los verbos. Un verbo, de acuerdo al Diccionario de Uso del Español (2000) se define como "Palabra con que se expresan las acciones y estados de los seres, y los sucesos". Puesto que a los historiadores les interesa analizar los fenómenos que se presentan en el texto y que demuestran las acciones, era importante desarrollar una metodología para proceder con el etiquetamiento de los verbos. Teniendo en cuenta que en informática siempre se debe buscar la manera más fácil de hacer las cosas, en un principio se intentó etiquetar los verbos utilizando los sufijos porque se basan en la terminación de los verbos y, por lo tanto, contienen una menor cantidad de palabras para comparar. Algunos de los sufijos utilizados para la identificación de verbos en el idioma español son:

- Infinitivo: AR (Ej. caminar), ER (Ej. temer), IR (Ej. venir)
- Formas no personales como son los Gerundios (Ej. caminando, temiendo, riendo),
- Participios de pasado (Ej. caminado, temido, reído).

La idea en un principio fue identificar las palabras que tenían estas terminaciones y anotarlas como verbos. Para su implementación se utilizó un descriptor que contenía los sufijos discriminados por el tipo de tiempo verbal empleado. Luego la clase de Java asociada al anotador efectuaba la identificación de las palabras en el texto y procedía a su

marcación como verbo. La siguiente figura muestra el descriptor implementado para sufijos:

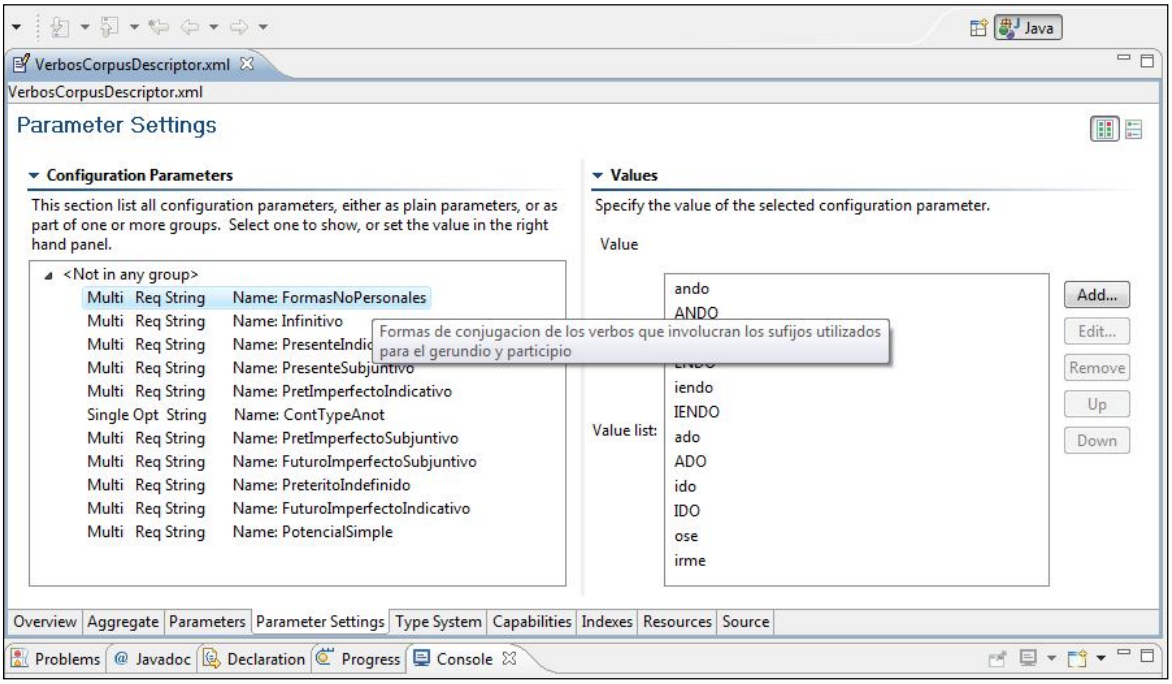


Figura 21
Listados implementados con los sufijos para etiquetar verbos

El problema que se encontró bajo este modelo era que el anotador era demasiado ambiguo por lo tanto se presentaban demasiados “falsos positivos” y la anotación de los verbos por tanto no era muy exitosa (muchos adjetivos del idioma terminan en “ado”, “ido”). Además, era necesario incluir los signos de puntuación en el análisis y el modelo se complicaba. Teniendo en cuenta estos aspectos se decidió implementar la detección de los verbos utilizando la raíz en lugar de los sufijos aunque ello implicaba obtener todos los verbos del idioma. Para ello se recopiló un listado de verbos regulares e irregulares del español en Internet (Wikcionario, 2008) y se hizo un anotador con ellos pero fue preciso hacer un preprocesamiento de estas palabras antes de incluirlas en el anotador como tal. En primer lugar, se obtuvo el listado de los verbos y luego se procedió a extraer las terminaciones de los verbos que estaban en infinitivo (terminaciones AR, ER e IR) y dejar solo su raíz con el fin de aportar mayor generalidad a la búsqueda porque usualmente los verbos utilizados en el idioma están conjugados y la raíz permanece invariable. De esta

manera el anotador tiene listados de verbos con sus raíces y se procede a anotar la palabra cuando encuentra coincidencia y pone como fin de la palabra el primer espacio en blanco que encuentra luego de la coincidencia para completar la palabra anotada. El siguiente cuadro muestra un detalle del pre-procesamiento que se hizo a los verbos.

Verbo Infinitivo	Terminación	Largo (longitud)	Raíz
abastecer	er	9	abastec
ablandecer	er	10	ablandec
ablentar	ar	8	ablent
abluir	ir	6	ablu
abnegar	ar	7	abneg
aborrecer	er	9	aborrec
abstraer	er	8	abstra
acaecer	er	7	acaec
acentuar	ar	8	acent
acertar	ar	7	acert
aclarecer	er	9	aclarec

Tabla 15
Detalle del preprocesamiento de verbos para español

Aunque en el anotador se encuentran 1550 verbos listados, ésta cantidad varía mucho porque a medida que se analizan los textos es posible complementar y alimentar esta base de conocimiento con otras palabras puesto que los verbos implementados pertenecen al español actual y debido a que el corpus maneja arcaísmos existen verbos poco frecuentes o que presentan variaciones ortográficas de la época y es necesario incluir esas formas y variantes de los mismos en los listados para que puedan ser etiquetados en el texto. De esta manera es necesario por ejemplo, incluir variantes para la palabra “dirigir”: “dirij” y “dirig” o para la palabra “expresar” es necesario incluir “espres” o “expres” o para la palabra “examinar” es necesario anotar “examin” o “esamin”, etc.

También surge un nuevo problema con este modelo y es la anotación de los verbos irregulares como “haber”, “ir” o “ser”. Para este tipo de casos particulares es necesario incluir expresamente en los listados las ocurrencias de esos casos. Esto indica que la anotación de los verbos es una tarea difícil desde el punto de vista lingüístico porque es necesario alimentar continuamente la base de conocimiento para lograr la mayor cobertura

de anotaciones posibles. La siguiente figura muestra algunas de las ocurrencias incluidas en los listados del descriptor de verbos:

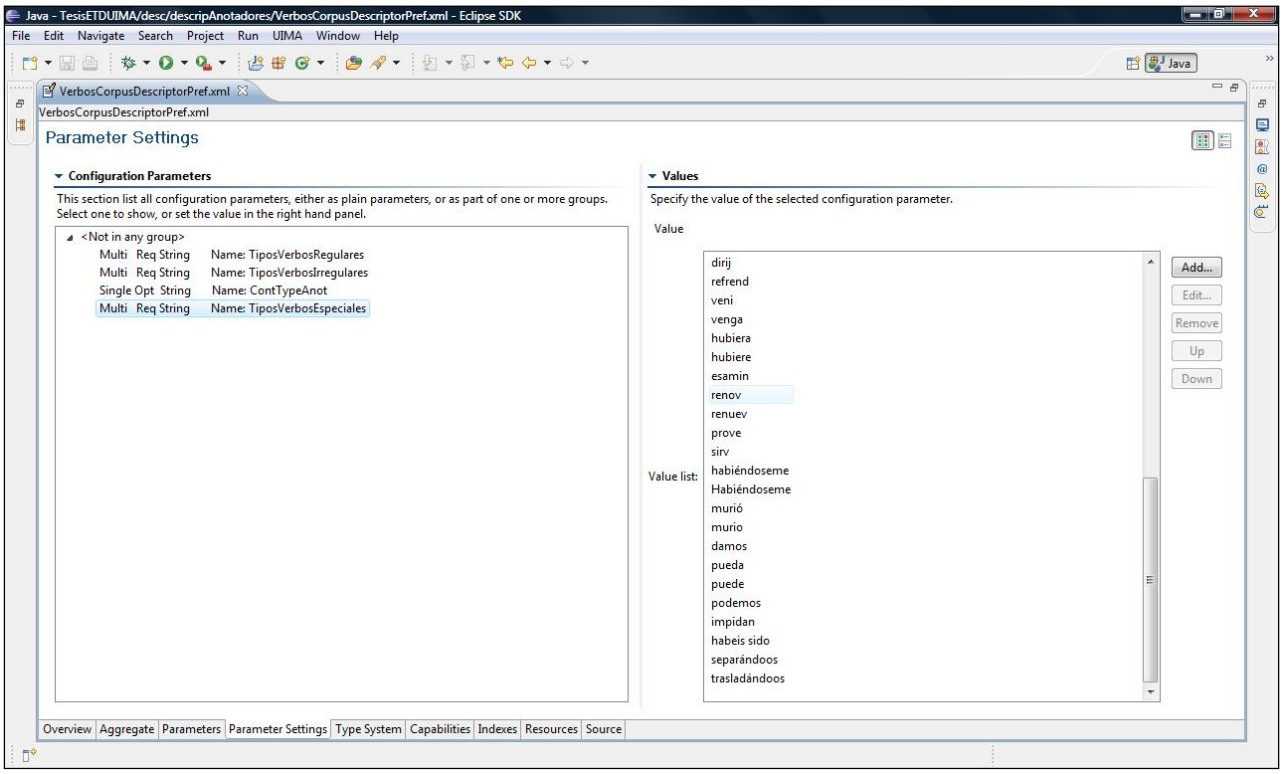


Figura 22.
Listados implementados con las raíces para etiquetar los verbos

La siguiente figura muestra parte de las anotaciones del corpus de sacerdotes:

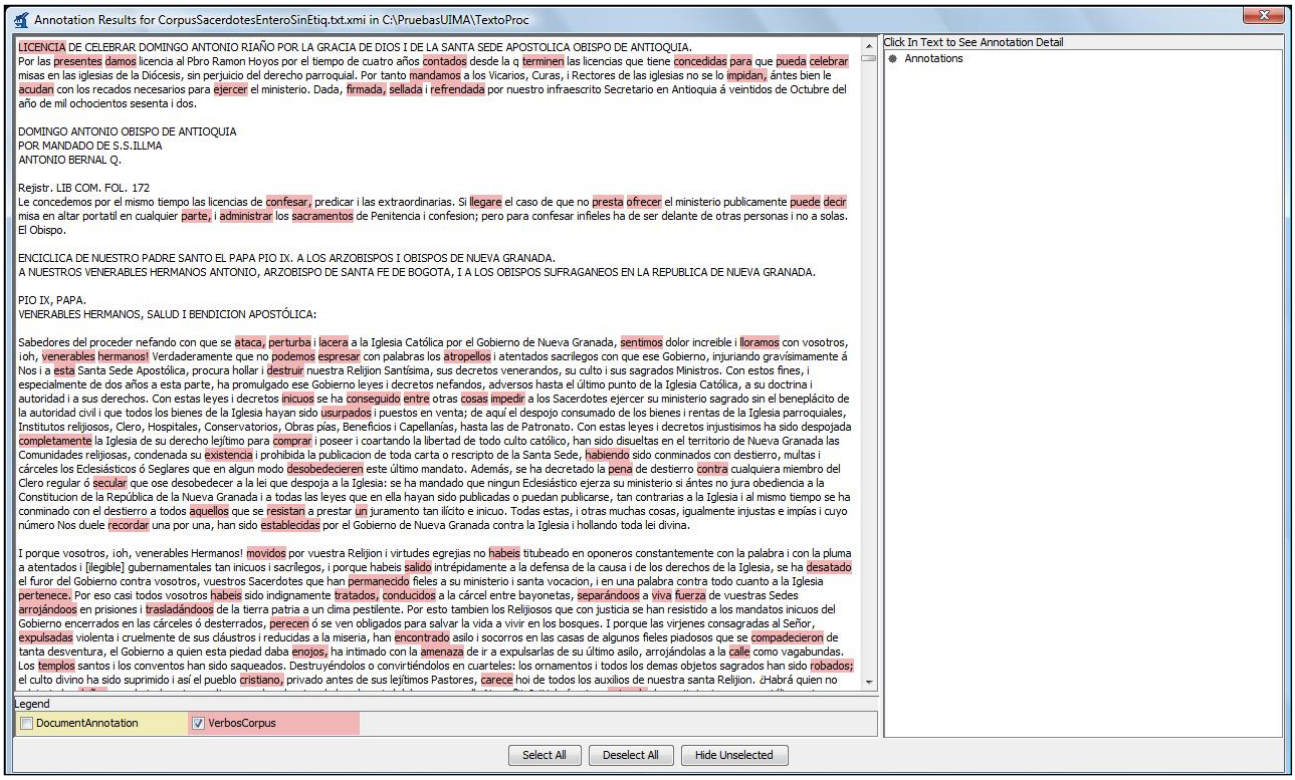


Figura 23.
Anotaciones de los verbos en el corpus histórico

3.3 EVALUACION DE LAS ETIQUETAS

3.3.1 Herramientas utilizadas en la evaluación. La herramienta empleada para analizar las etiquetas fue el *Collection Processing Engine* (CPE) que es una aplicación de UIMA que permite analizar varios documentos al tiempo y visualizar las anotaciones o generar un archivo plano de texto que las contiene (depende de lo que el programador desee efectuar). A diferencia de la herramienta *Document Analyser*, el CPE permite hacer análisis de varios documentos simultáneamente y constituye en sí un motor de búsqueda de metadatos pero es el usuario del sistema quien decide qué tipo de datos son recolectados y como se pueden visualizar por medio de la programación en JAVA de los archivos que componen el CPE especifico para la aplicación. El CPE está compuesto por varios subsistemas a saber:

- *Collection Reader*: es una interfaz para acceder a los documentos que serán analizados. Se compone de dos archivos: un respectivo descriptor y un archivo de JAVA llamado *FileSystemCollectionReader*.
- *Analysis Engine* (AE): es el componente que contiene los descriptores de los metadatos a analizar. En este caso, se elaboró un descriptor llamado *DescriptorDeEntidades* que recoge todos los datos de las entidades de nombres de personas, lugares, fechas y verbos para analizarlos en un documento de manera secuencial (en lenguaje UIMA es un *Descriptor Agregado* porque recoge en un solo AE varios descriptores). Si sólo se quisiera analizar una determinada entidad se escogería el descriptor de esa entidad.
- *CAS Consumer*: es el componente que analiza los datos de las entidades de interés de acuerdo a los análisis del AE. Es programable en JAVA de acuerdo a los requerimientos del usuario y también tiene asociado un descriptor. Para el análisis de cobertura y precisión de las entidades se programó un archivo que muestra el listado de entidades por documento.

La siguiente figura muestra un esquema de funcionamiento interno del CPE de UIMA:

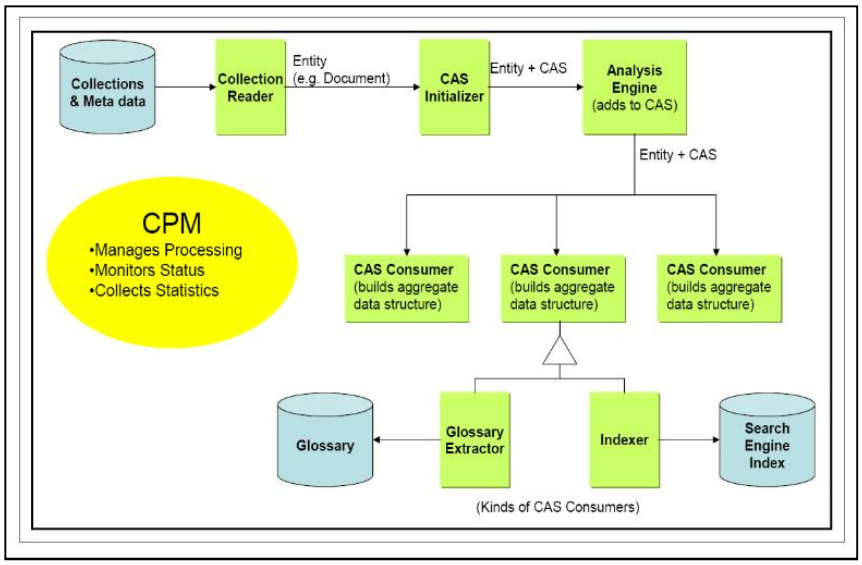


Figura 24.
Proceso interno de funcionamiento del CPE de UIMA (UIMA, 2009)

La siguiente figura muestra la interfaz del CPE empleada para el análisis de entidades:

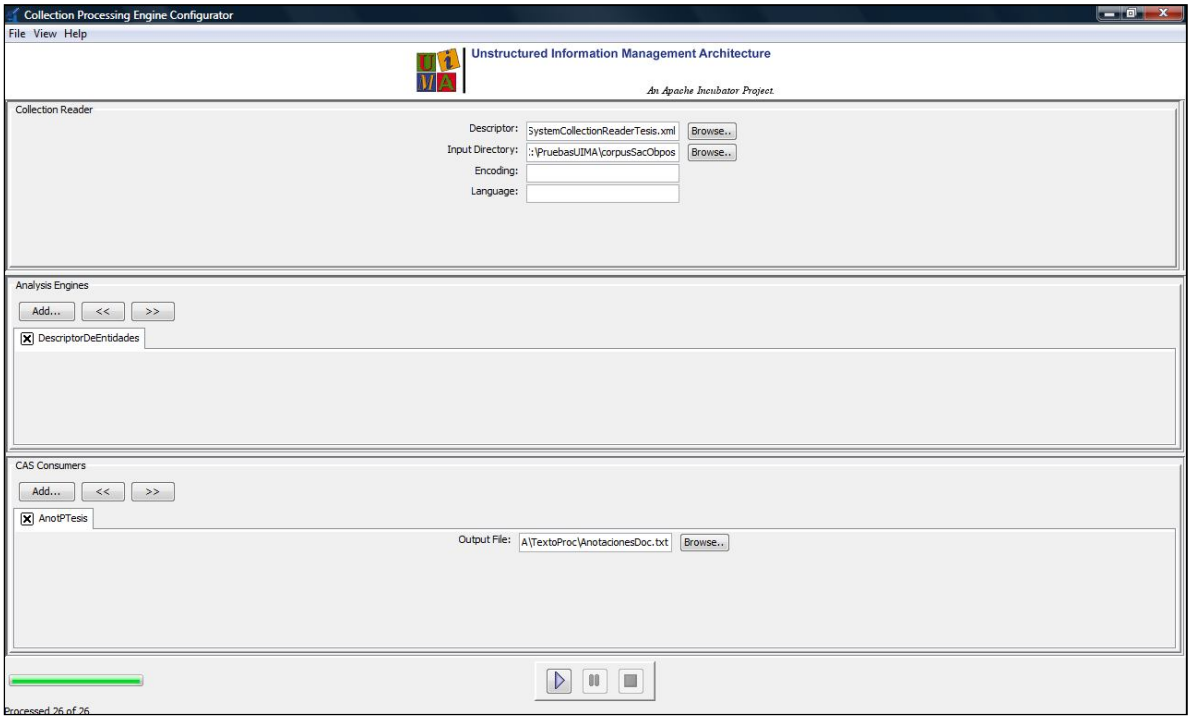


Figura 25.
Interfaz *Collection Processing* (UIMA, 2009)

Cuando el CPE termina de hacer un análisis de documentos muestra una ventana donde se observa el rendimiento del sistema durante el análisis como se observa en la siguiente figura:

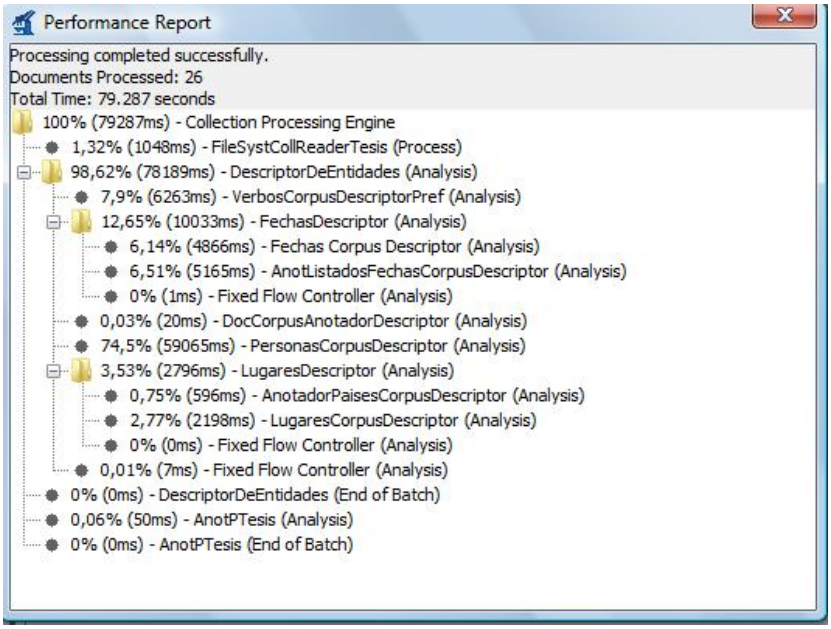


Figura 26.
Ejemplo de resultados del análisis de documentos del CPE.

3.3.2 Metodología para la medición del etiquetamiento del sistema. La manera de medir el rendimiento en cuanto al etiquetamiento es calcular los porcentajes de cobertura y la precisión tal como se explicó en el capítulo 1. La **cobertura** se midió tomando el número total de etiquetas sobre las etiquetas evaluadas para la entidad de interés y la **precisión** se midió tomando el número total de etiquetas de la entidad analizada sobre las etiquetas que son correctas. Para obtener los datos se hace un análisis de los documentos que conforman el corpus mediante el CP y éste produce un archivo de texto que contiene los datos de las entidades por documento. Este archivo luego es procesado para identificar si la palabra marcada pertenece o no a la categoría analizada (desambiguación) y luego se calcula el total de aciertos / desaciertos para obtener la cobertura y la precisión. Estas mediciones se hicieron manualmente porque son metadatos semánticos y el ser humano es quien debe decidir si una anotación pertenece a una categoría o no.

3.3.3 Análisis de resultados para el etiquetamiento. Para la evaluación del sistema se analizaron dos corpus:

- El corpus histórico compuesto por 26 documentos que contienen 221 cartas históricas de los sacerdotes y obispos, con un tamaño de 724KB.
- Un documento tomado aleatoriamente del corpus utilizado para el TREC (LDC, 2000) que sirve de corpus de contraste, el cual contiene 297 noticias de periódicos pertenecientes a mayo de 1994 y que tiene un tamaño de 576KB. Aunque solo se tomó un documento de los que conforman este corpus de noticias, el CPE puede etiquetar cantidades más grandes de texto sin problema, por ejemplo, se hizo una prueba de etiquetamiento para 10MB de noticias de AFP en donde se generó un archivo de etiquetas de un tamaño de 5MB.

Rendimiento del sistema de durante el análisis. A continuación se muestran las tablas de datos resultantes del análisis utilizando el CPE para los corpus histórico y de contraste (son los datos que el sistema muestra de acuerdo a la figura 25) y que muestran el rendimiento del sistema durante el análisis para cada uno:

Datos de rendimiento del CPE para el corpus histórico

Tipo de corpus	Corpus histórico
Numero de archivos que conforman el corpus	26 archivos
Numero de documentos	221
Tiempo total de análisis de etiquetas	78.544 segundos
Porcentaje de tiempo de análisis para *Collection Reader*	0.26% (202 ms)
Porcentaje de tiempo de análisis del *DescriptorDeEntidades*	99,02% (77771 ms)
Porcentaje de tiempo de análisis en el proceso del *Cas Customer*	0.73% (570 ms)
TOTAL PORCENTAJE DE TIEMPO DE ANALISIS DEL SISTEMA	**100%**
Porcentaje de tiempo de análisis para etiquetas de fechas	12,75% (10015 ms)
Porcentaje de tiempo de análisis para etiquetas de documentos	0,03%(24 ms)
Porcentaje de tiempo de análisis para etiquetas de lugares	3,75% (2944 ms)
Porcentaje de tiempo de análisis para etiquetas de nombres de personas	74,33% (58378 ms)
Porcentaje de tiempo de análisis para etiquetas de verbos	8.14% (6393ms)
TOTAL PORCENTAJE DE TIEMPO DE ANALISIS PARA ETIQUETAS (DescriptorDeEntidades)	**99%**

Tabla 16.
Medidas de rendimiento en etiquetado para el corpus histórico

Se observa que el 99% del tiempo de análisis se consume en la identificación de las etiquetas (módulo DescriptorDeEntidades) y el 1% restante se consume en el procesamiento de datos de entrada y salida. El etiquetado de nombres de personas consume el 74.33% de tiempo con relación a los procesos de etiquetamiento de lugares, fechas y verbos.

Datos de rendimiento para el corpus de contraste archivo AFP940527

Tipo de corpus	Corpus de contraste
Numero de archivos que conforman el corpus	1 archivo
Numero de documentos	297
Tiempo total de análisis de etiquetas	69.927 segundos
Porcentaje de tiempo de análisis para *Collection Reader*	0.1% (73 ms)
Porcentaje de tiempo de análisis del *DescriptorDeEntidades*	92,9% (64959 ms)
Porcentaje de tiempo de análisis en el proceso del *Cas Customer*	7% (4894 ms)
TOTAL PORCENTAJE DE TIEMPO DE ANALISIS DEL SISTEMA	**100%**
Porcentaje de tiempo de análisis para etiquetas de fechas	11,1% (7760 ms)
Porcentaje de tiempo de análisis para etiquetas de documentos	0,01%(5 ms)
Porcentaje de tiempo de análisis para etiquetas de lugares	3,65% (2549 ms)
Porcentaje de tiempo de análisis para etiquetas de nombres de personas	73,23% (51207 ms)
Porcentaje de tiempo de análisis para etiquetas de verbos	4.92% (3438 ms)
TOTAL PORCENTAJE DE TIEMPO DE ANALISIS PARA ETIQUETAS (DescriptorDeEntidades)	**92.9%**

Tabla 17.
Medidas de rendimiento en etiquetado para el corpus de noticias

De manera similar al corpus histórico, se observa que en el corpus de noticias AFP940527 el 92,9% del tiempo de análisis se consume nuevamente en la identificación de las etiquetas (módulo DescriptorDeEntidades) y el 7,09% restante se consume en el procesamiento de datos de entrada y salida. El etiquetado de nombres de personas también consume gran parte del porcentaje de entidades con el 73.23% de tiempo con relación a los procesos de etiquetamiento de lugares, fechas y verbos.

En general los datos de etiquetamiento son muy similares a pesar de que se tratan de corpus diferentes, sin embargo llama la atención el hecho de que el corpus histórico tiene un mayor porcentaje de verbos etiquetados (8.14%) con relación al corpus de noticias (4.92%). Esto puede ser un indicio de que el corpus histórico es más discursivo que el corpus de noticias.

Análisis de la cobertura y precisión del sistema de extracción de información. Tal como se mencionó en el capítulo 1, las medidas de evaluación de los sistemas de extracción de información se efectúan por medio del cálculo de la cobertura y la precisión. La cobertura y la precisión del sistema de extracción de información se calcularon de la siguiente manera siguiendo la metodología propuesta por Jurasfky y Martin (2000):

- La **cobertura (*recall*)**: es el porcentaje dado por el número de etiquetas correspondientes a la entidad analizada sobre el total de etiquetas del sistema.

$$\text{Cobertura } (Recall) = \frac{\text{número de etiquetas de la entidad analizada}}{\text{Total de etiquetas evaluadas}}$$

- La **precisión (*precision*)**: es el porcentaje dado por el número de etiquetas de la entidad analizada que son acertadas (es decir que sí corresponden a la entidad) sobre el número de etiquetas de la entidad analizada.

$$\text{Precisión } (Precision) = \frac{\text{número de etiquetas acertadas}}{\text{número de etiquetas de la entidad analizada}}$$

A continuación se muestran las tablas resultantes del análisis de estos dos parámetros tanto para el corpus histórico como para el corpus de contraste:

Corpus histórico (26 documentos que contienen 221 cartas)						
Total etiquetas: 7196						
Tipo de entidad	**Evaluadas**	**Correctas**	**Incorrectas**	**Cobertura**	**Precisión**	**Principal causa de Error**
Fechas	598	453	145	9%	76%	Expresiones de tiempo que no son fechas.
Nombres de Lugares	2164	1381	783	33%	64%	Falsos positivos tomados como lugares (ej. "Sacerdotes", "Pastor", "Señor").
Nombres de personas	3756	2788	968	58%	74%	Falsos positivos de sustantivos que no son nombres propios de personas (ej. "Sociedad Católica", "Seminario de Antioquia").
TOTALES	**6518**	**4622**	**1896**		**71%**	

Tabla 18.
Medidas de precisión y cobertura para el corpus histórico

Corpus de contraste (1 documento que contiene 297 noticias)						
Total etiquetas:						
Tipo de entidad	**Evaluadas**	**Correctas**	**Incorrectas**	**Cobertura**	**Precisión**	**Error principal**
Fechas	676	413	263	10%	61%	Expresiones de tiempo que no son fechas.
Nombres de Lugares	2375	1963	412	37%	83%	Falsos positivos de nombres de personas o de instituciones (ej. "Naciones Unidas", "OLP")
Nombres de personas	3412	1741	1671	53%	51%	Sustantivos que no son nombres propios de personas (ej. "Vuelta de Suiza", "Tercer Mundo").
TOTALES	**6463**	**4117**	**2346**		**64%**	

Tabla 19.
Medidas de precisión y cobertura para el corpus de contraste

Al respecto de la evaluación se tienen los siguientes comentarios:

Evaluación de las fechas: los mayores errores encontrados en el etiquetamiento de las fechas tienen que ver con aquellas expresiones de tiempo que no son consideradas fechas y que el sistema detecta (son el 62% con relación al total de errores), por ejemplo "diez meses", "próximo agosto", "dos años", "quince o veinte días", "próximo pasado mes", etc. Este inconveniente aparece para los dos corpus. Estas expresiones temporales requieren un análisis de anáforas para su desambiguación. La presencia de estos errores indica que el sistema muestra datos que permitirían continuar con un futuro análisis de eventos utilizando causalidades que es tema de investigación actual. Con respecto a las fechas etiquetadas correctamente el sistema detecta expresiones tales como "23 del mes de noviembre del presente año", "uno de diciembre de mil ochocientos setenta", "doce del presente mes", "veintidós de Febrero de mil ochocientos setenta i uno", "seis de agosto de mil ochocientos setenta y tres", "3 de julio de mil ochocientos setenta", "dos de junio del año del Señor, mil ochocientos sesenta y tres" y las expresiones que involucran esquemas mas numéricos como los mostrados en los modelos implementados por las ER.

Evaluación de los lugares: En el momento de la desambiguación, la mayor dificultad se presentó en el momento de determinar si una etiqueta correspondía a un lugar o no (sobre todo en el corpus de AFP) y para verificar fue preciso consultar Internet, es decir que el sistema permite una mayor cobertura si se alimenta las bases de conocimiento. La

desambiguación de los lugares en el corpus histórico no fue compleja porque se conocía el corpus y por tanto la identificación no fue complicada. También es importante analizar que no se presentaron los errores que se presentan en otros sistemas similares debido a homónimos como ocurre para otros corpus en inglés.

Problemas de Ambigüedad para los nombres de personas: la principal dificultad en la desambiguación de los nombres de las personas fue que muchas etiquetas correspondían también a nombres de instituciones, es decir que en el momento de desambiguar se tomaron como desacertadas algunos nombres de instituciones y fue necesario entonces redefinir el concepto de "nombre". En el análisis se tomaron como nombres de personas los títulos como son por ejemplo "Señor Presbítero Ocampo" o "U. Illma". Puesto que se incluyeron las etiquetas de [Firmado] en el corpus de histórico, los nombres de las personas presentes en los listados no tuvieron inconvenientes en ser etiquetados y reconocidos como tal. También en el corpus de AFP se etiquetaron como nombres de personas algunos países que no aparecieron en las etiquetas de lugares.

Evaluación de los verbos: La idea con los verbos es tratar de determinar las acciones en el párrafo. Sin embargo, es importante aclarar que todavía faltan muchas palabras que pueden ser incluidas en el listado de verbos es decir que el modelo es susceptible de mejoras, lo cual implica que es posible enriquecer la base de conocimiento del sistema. Con relación a los verbos etiquetados, en el corpus histórico se etiquetaron 4601 formas en total como verbos, de los cuales se tienen 1054 lemas diferentes que equivalen a un 23% sobre el total de palabras etiquetadas. Para el corpus de contraste se etiquetaron en total 517 formas como verbos de las cuales se tienen 497 lemas diferentes que equivalen a un 96%. Estos resultados dan una idea de lo discursivos que son los textos, esto es, el corpus histórico es mucho más disperso y narrativo que el corpus de contraste compuesto por noticias, que es mucho más concreto y resumido en su contenido. Este resultado es de esperarse porque los corpus son diferentes y tienen temas, géneros épocas distintas. Es posible hacer un análisis más extenso sobre el número de ocurrencia y las frecuencias de

las palabras para tener una idea de lo que puede tratar el texto pero estos estudios no están en el alcance de este trabajo y puede tratarse como un tema de análisis posterior.

En general, los valores de precisión del sistema se encuentran en los valores normales para este tipo de trabajos, por ejemplo, Crane y Jones (2006) en su trabajo de análisis de entidades para el periódico *Richmond Times Dispatch*, que es un periódico del s. XIX muestran unos valores de precisión similares para nombres de personas (76.56%), lugares (97.42%) pero el valor de fechas si está por encima del encontrado en el sistema de esta investigación (96.46%). En general es posible hacer mejoras en los valores de las fechas y los lugares para mejorar la cobertura de etiquetamiento.

CAPITULO 4

MODELO DE DETECCIÓN DE EVENTOS

"La humildad tiene dos polos: lo verdadero y lo bello." Víctor Hugo

"Yo no aclamo haber controlado los eventos, sino que confieso plenamente que los eventos me han controlado a mí." Abraham Lincoln

Palabras Clave. Detección de eventos.

Este capítulo describe el proceso de detección de eventos para los diferentes corpus de análisis utilizados en esta investigación. La metodología empleada para la detección de eventos en este trabajo está basada en la que empleó Smith (2000a, 2000b) debido a que es el acercamiento más simple y que además brinda posibilidades de otros análisis futuros.

4.1 METODOLOGIA Y MODELO DE RECONOCIMIENTO DE EVENTOS EMPLEADOS

4.1.1 Definición de evento. Tal como se explicó en el capítulo 1, los estudios relacionados con la identificación y definición de lo que es un evento han tenido su origen en los análisis de información de los sistemas TDT y en la manera en la cual es posible anotar relaciones temporales de temas en los textos. Los sistemas TDT utilizan clasificación automática de temas en un flujo permanente de noticias y los que van apareciendo son definidos como eventos o "algo no trivial que ocurre cerca de un lugar en un cierto tiempo" (Yang et al, 1998). Es justamente en el "cambio en el tiempo" donde radica la particular dificultad para identificarlos porque no basta con etiquetar fechas que tienen un cierto modelo estándar como por ejemplo, las fechas presentes en el corpus que fueron identificadas por ER para este trabajo, sino que es necesario analizar también otros tipos de etiquetas que denotan cambios temporales y que requieren de análisis de anáforas y

heurísticas especiales para determinar cuánto es el intervalo de tiempo que se describe en el texto. Además existen otros problemas adicionales relacionados con la continuidad de un evento a lo largo del texto y que requiere la identificación de una "línea del tiempo" o (durabilidad) para determinar su evolución y otro temas son también las posibles relaciones entre varios eventos entre sí que pueden no ser triviales y que el ser humano "infiere" pero a la hora de llevarlo en la implementación utilizando el computador es difícil hacerlo. Por estas dificultades que no son fáciles de solucionar y que aún son tema de investigación, se optó por determinar que, para este trabajo de investigación la hipótesis de trabajo es que un evento será detectado cuando aparezcan de manera simultánea etiquetas correspondientes a fecha, lugar y nombres de personas en una ventana de aproximación textual elegida para efectuar el análisis (ésta ventana se refiere al tamaño de texto a analizar que puede ser un documento del corpus o un párrafo o frases, la elección dependerá del tipo de análisis que se desee efectuar), de acuerdo con Setzer y Gaizauskas (2002) en donde, a la hora de identificar un evento es necesario tener en cuenta dos aspectos importantes: especificar el objetivo temporal de la representación que deseamos obtener de un texto y luego identificar la información auxiliar que debe ser extraída porque es útil para llegar a la meta de la representación temporal. Teniendo en cuenta esto, también se consideraron en el análisis las etiquetas de los verbos que indican información adicional pero que también pueden incluir cierto ruido a la hora de la identificación y por tanto no se tuvieron en cuenta para este análisis. El sistema implementado para detectar eventos para este trabajo muestra por documento o por párrafo las anotaciones de fechas, lugares, personas y el texto de interés que las contiene. El sistema implementado bajo la plataforma de UIMA arroja un archivo de texto plano que contiene los datos y el usuario final luego hace una lectura y evaluación de la pertinencia del evento localizado de acuerdo a su interés.

4.1.2 Heurísticas empleadas para la detección de eventos. Para la implementación de la detección de eventos se tuvieron en cuenta varios aspectos prácticos que surgieron en la medida que se hacían ensayos y que tienen que ver con la naturaleza de los textos

analizados, puesto que los procesos de extracción de información tienen mucha dependencia de dominio. Estos aspectos se mencionan a continuación:

a. **Definir una ventana de aproximación documental para el análisis de eventos.** Este es un aspecto importante porque permite determinar cuál es la ventana de visión del evento y hace referencia al intervalo de análisis de texto escogido que puede ser un documento (por ejemplo una carta entera del corpus histórico) o párrafos o frases. En este sistema se definieron dos ventanas de aproximación: documentos y párrafos y se definieron unos marcadores para delimitar el documento y el párrafo con el fin de facilitar su identificación.
b. **Definir las etiquetas que definen un evento y la información a mostrar.** De acuerdo a la definición de un evento, además de la aparición de una o varias etiquetas de fechas y lugares también es necesario incluir el texto donde fueron encontradas con el fin de contextualizar el hecho como tal.
c. **Definir como mostrar los datos.** Debido a que el sistema está pensado para que un profesional de las ciencias humanas pueda efectuar algún tipo de análisis posterior teniendo en cuenta la información suministrada y tratando en lo posible de facilitarle su manipulación, se optó por entregar un archivo de texto plano de modo que sea fácil manipularlo luego en un procesador de texto o en otras aplicaciones informáticas tales como Excel mediante el uso de macros o otro tipo de aplicaciones. Además, es importante mostrar los datos de manera ordenada con el fin de identificar primero los eventos (los párrafos que contienen 4 metadatos) y luego de manera descendiente los otros datos que contengan entre 3, 2 y 1 etiquetas y se ha elaborado por tanto una macro en Excel para facilitar este ordenamiento.

4.1.3 Herramientas utilizadas para el análisis. Para el análisis de detección de eventos se utilizó nuevamente la herramienta CPE de UIMA pero se construyeron nuevos descriptores para el análisis en el AE y el *Cas Customer,* es decir que el CPE puede ser programado para diferentes actividades dependiendo de los diferentes tipos de análisis textuales. Para el caso de eventos, los descriptores y archivos de código en JAVA implementados se muestran en la siguiente tabla:

Componente de CPE	Nombre del descriptor	Archivos asociados con el descriptor	Observaciones
AE	EventosParrafoDescriptor (descriptor agregado)	DocumentoDescriptor VentanaTextoDescriptor FechasDescriptor PersonasCorpusDescriptor LugaresDescriptor	Contiene los descriptores elaborados para anotar entidades por párrafo de fechas, personas y lugares que fueron explicados en el capítulo 3. Cada descriptor contiene su respectivo archivo de JAVA que lo implementa.
Cas Consumer	CasConsumerEventosDescrip	AnotEventosFechaTesis AnotEventosSoloParrafoTesis AnotEventosSoloMetadatosTesis	Contienen el código en JAVA para mostrar las etiquetas y el texto asociado con ellas. Son archivos implementados en JAVA.

Tabla 20.
Listados de componentes del CPE y los descriptores y archivos asociados

En el *Cas Customer,* que es el programa que contiene el código que permite escribir los resultados del análisis en un archivo de texto plano, se implementaron tres tipos de archivos de JAVA que soportan datos distintos para diferentes análisis:

Archivo de JAVA	Observaciones
AnotEventosSoloParrafoTesis	Es un archivo que permite generar en un archivo plano los párrafos analizados (solo el texto).
AnotEventosSoloMetadatosTesis	Es un archivo que permite generar en un archivo plano los metadatos (las etiquetas) que fueron analizadas por párrafo.
AnotEventosFechaTesis	Es un archivo que permite generar en un archivo plano tanto los metadatos (las etiquetas) como el párrafo analizado. Finalmente este fue el archivo que se implementó para la detección de eventos.

Tabla 21.
Descripción de diferentes archivos implementados en JAVA para el análisis de eventos

4.2 IMPLEMENTACIÓN DEL SISTEMA PARA DETECCIÓN DE EVENTOS

Partiendo de la definición de evento, se procedió a implementar el sistema por medio de las heurísticas mencionadas en el apartado anterior y que se describirán a continuación:

4.2.1 Definición de las ventanas de aproximación. La implementación de estas ventanas de aproximación se llevó a cabo por medio de la construcción de anotadores que marcan el texto de acuerdo a algún rasgo común que permita identificar documentos enteros o párrafos. Se implementaron dos tipos de anotadores: uno para etiquetar documentos

llamado *DocCorpusAnotador* y otro para marcar párrafos del texto llamado *VentanaTextoDescriptor*. El anotador de documentos fue el primero en ser implementado porque los documentos del corpus histórico y el corpus de noticias incluyen las etiquetas <DOC> y </DOC> y las anotaciones quedaban delimitadas de acuerdo a las posiciones donde comienzan y terminan estas etiquetas como se muestra en la figura:

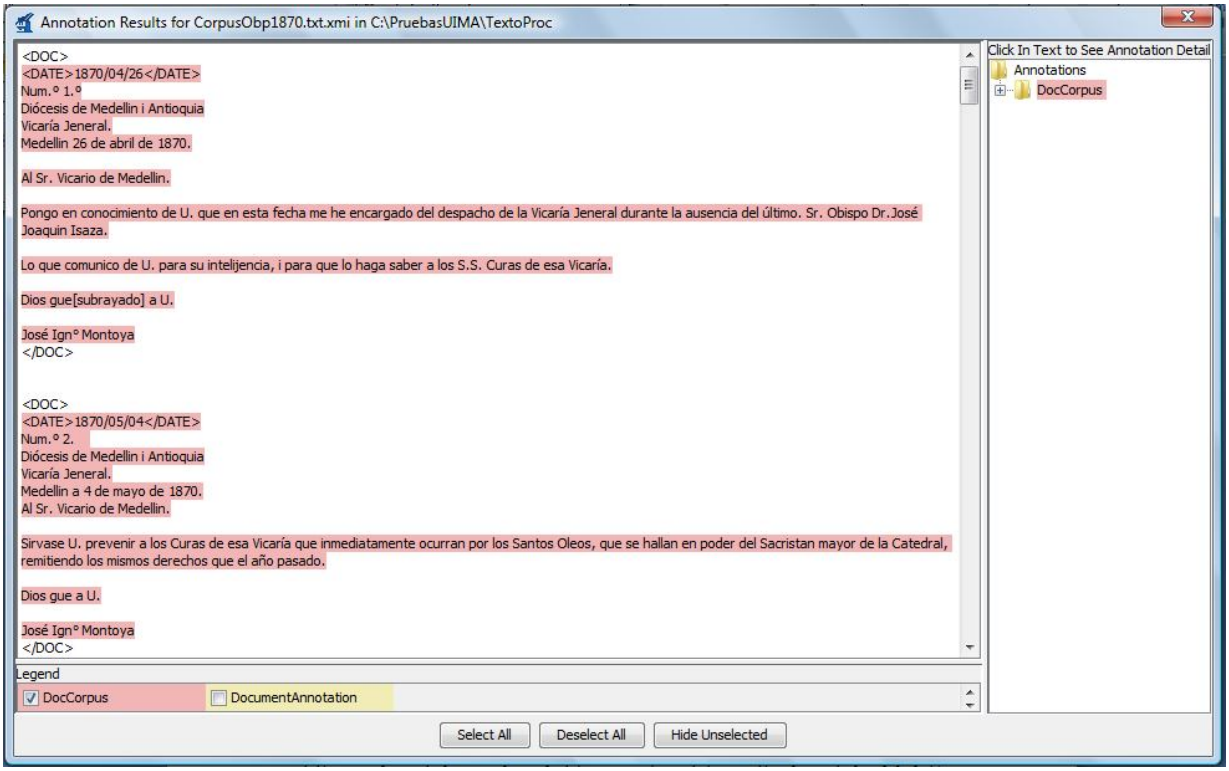


Figura 27.
Ejemplo descriptor de Documentos

Sin embargo a la hora de detectar eventos sucedió que la ventana puede ser demasiado grande y el sistema muestra información muy global que también puede ser buena para otros tipos de análisis pero para los eventos puede refinarse aun más. Un ejemplo se muestra en la siguiente figura donde hay una porción del texto de salida para el análisis del corpus histórico del año 1880 y se muestran los datos para el documento 2:

```
<++++NUEVO DOCUMENTO++++>
Localizacion del Documento (URI):C:\PruebasUIMA\TextoDeEntrada\CorpusSac1880E.txt

co.edu.eafit.tesisETD.anotadoresT.DocCorpus CONTADOR 1 Empieza EN = 5 Termina EN = 2772
FECHAS;12 de enero de 1880;279;298;1
FECHAS;unos cuatro ó;1398;1411;1
FECHAS; cuatro meses ;1929;1943;1
LUGARES;Pasto;94;99;1
LUGARES;  Hatoviejo;266;277;1
LUGARES;Medellin;366;374;1
LUGARES;Itagüí;529;535;1
LUGARES;Hatoviejo ;1157;1167;1
LUGARES;Itagüí ;1919;1926;1
LUGARES;Medellin;2687;2695;1
PERSONA;Al Vicario General De: Pbro. Baltasar Botero;187;231;1
PERSONA;Al V. s;311;318;1
PERSONA;Vicario Gral;321;333;1
PERSONA;D. Sebastian Emigdio Restrepo  Medellin;335;374;1
PERSONA;V. Prelado;387;397;1
PERSONA;Véalo U.;706;714;1
PERSONA;I. El;714;719;1
PERSONA;Pbro. Hincapié;1379;1393;1
PERSONA;Pbro. Hincapié;1886;1900;1
PERSONA;I. Lo;1997;2002;1
PERSONA;El Pbro. Hincapié;2031;2048;1
PERSONA;Si U.;2224;2229;1
PERSONA;Pbro. Hincapié;2256;2270;1
PERSONA;Baltasar Vélez V. ;2576;2594;1
```

Figura 28.
Ejemplo de resultados del análisis de etiquetas por documento

Pensando en afinar un poco la búsqueda se implementó un descriptor por párrafos. Como toda actividad de EI, era necesario definir primero qué es un párrafo para iniciar la caracterización y delimitarlo: un párrafo para este caso de estudio es un texto que comienza en una línea con una letra mayúscula y termina con un punto seguido por un retorno de carro. Sin embargo en el corpus histórico analizado los textos que lo conforman fueron escritos originalmente por diferentes personas y por tanto ocurren una serie de casuísticas que dificultan la caracterización porque no siempre los párrafos terminan con punto o no siempre comienzan por letra mayúscula. Esta dificultad muestra una vez más la dependencia que tienen los sistemas de EI con relación a los textos analizados para su implementación. Para solucionar el problema de la delimitación del párrafo se acudió felizmente a uno de los primeros ensayos que se hicieron en la construcción del corpus histórico que incluían como modelo el corpus de noticias de AFP. Este corpus tiene un encabezado que incluye metadatos referentes a la cantidad de palabras del texto, la fecha y otros, pero cada párrafo está delimitado por las etiquetas y lo cual hace de este

método muy dependiente del texto y del etiquetado porque es un proceso difícil encontrar un error mientras el sistema procesa los datos, pero es efectivo en cuento a la delimitación requerida. A continuación se muestra una parte del corpus de AFP y parte del corpus histórico construido con esta misma estructura para ilustrar la delimitación de los párrafos:

```
<DOC>
<DOCID> af940522.0007 </DOCID>
<DOCNO> AF940522-0007 </DOCNO>
<STORYID pri=u cat=i count=0115>o0027</STORYID>
<KEYS country='LAW' iz='WQ36' lang=sp>Ruanda-combates</KEYS>
<PREAMBLE lang=sp> URGENTE &bell;&bell;&bell; </PREAMBLE>
<TEXT lang=sp>
<p>
EL FPR SE APODERO DEL AEROPUERTO DE KIGALI
</p>
<p>
NAIROBI, Mayo 22 (AFP) - Los rebeldes del Frente Patriótico Ruandés (FPR) se apoderaron este domingo del aeropuerto internacional de Kigali, confirmó un vocero de la Misión de las Naciones Unidas para la asistencia en Ruanda (MINUAR), Abdul Kabia.
</p>
<p>
Desde hacía varios días el FPR se disputaba el aeropuerto con las fuerzas gubernamentales, que disponen cerca de las pistas de un importante cuartel, Camp Kanombe. Dentro del aeropuerto habían instalado baterías de defensa antiaérea.
</p>
<p>
Camp Kanombe sigue en poder de las fuerzas gubernamentales pero está cercado por el FRP, añadió la misma fuente.
</p>
<p>
AFP
</p>
</TEXT>
<TRAILER lang=sp>
<DATE>0944 GMT 94/05/22</DATE></TRAILER>
</DOC>
```

Figura 29.

Ejemplo de la estructura del corpus de contraste AFP Noticias (LDC, 2000)

```
<DOC>
<DOCID> SAC036 </DOCID>
<STORYID count=0155></STORYID>
<KEYWORDS> Sacerdotes. Moralidad </KEYWORDS>
<EXCHANGE> De Sacerdote a Obispo <EXCHANGE>
<PREAMBLE lang=sp>
Ilustrisimo Señor Doctor Valerio Antonio Jimenez De: Valerio Martinez
</PREAMBLE>
<TEXT lang=sp>
<p>
Antioquia en.o 4 de 1871
</p>
<p>
Ilustrisimo Señor Doctor Valerio Antonio Jimenez
Medellin
</p>
<p>
Ilustrisimo Señor.
</p>
<p>
Cumpliendo con lo ordenado por UIIlma" con fha" 20 de agosto del año ppdo" he trascrito, al Señor Pro Diego
Leal el oficio que UIIlma" me remitió. Con fha" de ayer, estubimos 3., dicho Señor Pro me dió contestacion al
presitado oficio, citandome para que en asocio de él, hicieramos la revision de los ornamentos que existen en la
Iglesia mayor de esta ciudad. Hoy hemos practicado la expresada revicion tanto en los ornamentos, como en el
inventario que existe en poder del Señor Pro Leal, y resulta que no hay ninguno de los que pertenecieron  al
Iltmo Señor. Riaño; pues, los que hai actualmente en ella son los costeados por el Venerable Capitulo en años
anteriores.
</p>
<p>
Le incluyo el oficio, contestacion del Señor Pro Leal para lo que estime conveniente.
</p>
<p>
Soy de Usia Iltma" atento servidor y Capellan"
</p>
<p>
Valerio Martinez [firmado]
</p>
</TEXT>
<PLACE> Medellin </PLACE>
<TRAILER lang=sp>
<DATE>1871/01/04</DATE></TRAILER>
</DOC>
```

Figura 30.
Ejemplo de la estructura del corpus histórico etiquetado

Teniendo este esquema de trabajo fue mucho más sencillo implementar el anotador para párrafos aprovechando la presencia de las etiquetas pero esta caracterización tiene el inconveniente de depender de la forma del texto porque se requiere la presencia de la etiquetas extras para delimitar los párrafos, aunque también puede programarse para incluir las casuísticas propias de los corpus utilizando ER, lo cual puede servir para

efectuar un análisis más general que permita analizar textos sin etiquetar pero su implementación es mucho más compleja. Se implementó entonces un anotador llamado *VentanaTextoDescriptor* tal como se observa en la siguiente figura:

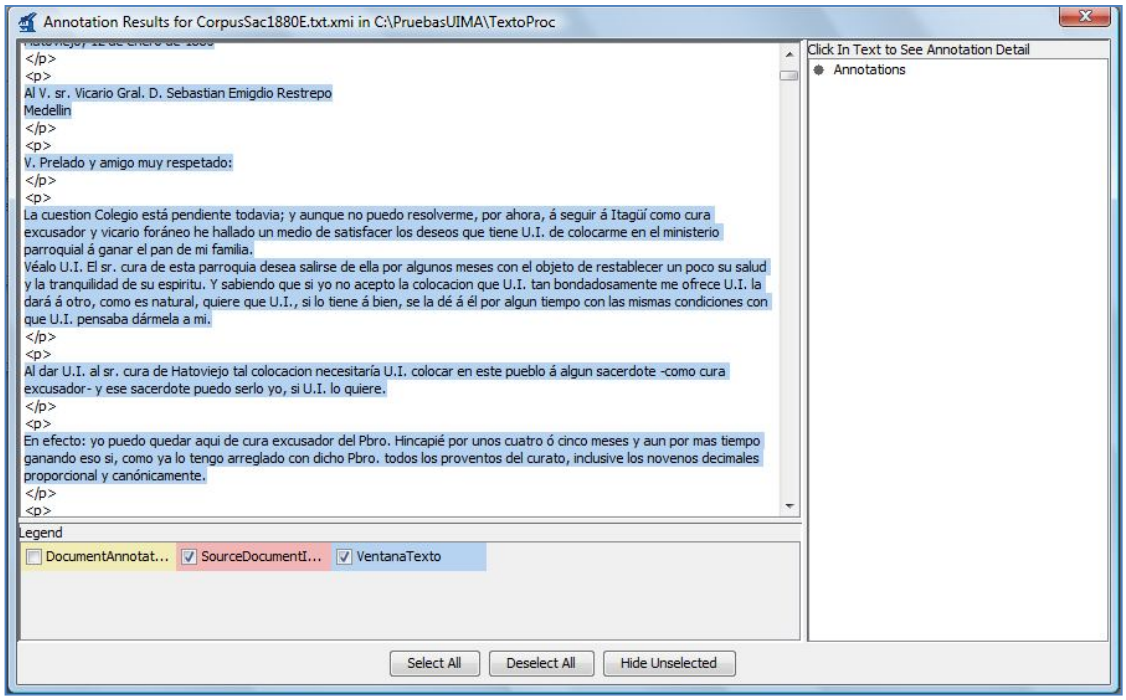


Figura 31.
Ejemplo descriptor para los párrafos

4.2.2 Metodología para el análisis de eventos. El proceso fue el siguiente: teniendo ya los anotadores por párrafo se procedió a programar el código para extraer los metadatos semánticos (es decir las anotaciones de Fecha, Lugar y Persona) utilizando el CPE y para ello programando el código respectivo de JAVA del *Cas Customer*. Para efectuar esta programación era preciso definir como se analizarían los documentos presentes en cada uno de los archivos porque el análisis de eventos en este proceso implica una participación humana fuerte y por cuestiones de facilidad en el análisis, se determinó procesar cada archivo por aparte para facilitar la revisión detallada. De esta manera se programó el código para generar un archivo plano con los metadatos y el texto por párrafo de cada documento presente en cada archivo del corpus (en este punto de detección de eventos es necesario aclarar que son diferentes los metadatos a las anotaciones: los metadatos son los

tipos de etiquetas implementadas y las anotaciones ya son los objetos o instancias concretas). Para que el CPE procesara el archivo de interés, éste se ponía en una carpeta de entrada de datos (llamada *TextoDeEntrada*) y el archivo resultante del análisis se nombraba de acuerdo al año del texto, por ejemplo AnotEventosFecha1871E.txt. A continuación se muestra una porción del archivo plano resultante de un análisis con el CPE implementado para detectar eventos en el corpus de sacerdotes de 1873:

<+++++NUEVO DOCUMENTO++++>
Localizacion del Documento (URI):C:\PruebasUIMA\TextoDeEntrada\CorpusSac1873E.txt
PARRAFO|1|286|341
PERSONAS|Al Illmo Sor Obispo de Evaria Provisor Vicario Jeneral|287|340
Al Illmo Sor Obispo de Evaria Provisor Vicario Jeneral
PARRAFO|2|352|382
FECHAS|6 de marzo de 1873|362|380
Medellin 6 de marzo de 1873.
PARRAFO|3|393|760
LUGARES|Marcelina |422|432
LUGARES|Mercédes|434|442
PERSONAS|Illmo Señor: Los|394|411
Illmo Señor:
Los deudos de Marcelina i Mercédes. Ospina quieren unas exequias para el 11 de los corrientes, trigésimo dia del [ilegible] su muerte, i que se celebre en el Monasterio del Cármen. En tal virtud, pido a U.S.I respetuosamente se sirva concederme licencia para presidir i presenciar en aquella iglesia en dichas exequias.
Soi de U.S.I su afmo servidor.
PARRAFO|4|771|808
PERSONAS|José Mª Gómez Angel|772|791
José Mª Gómez Angel [firmado]
Cura
PARRAFO|5|819|869
FECHAS| marzo 6 de 1873|851|867
LUGARES| Medellin|841|851
Gobierno eclesiástico
Medellin marzo 6 de 1873.
PARRAFO|6|880|1229
FECHAS| dia 11 del presente mes. |1023|1049
PERSONAS|Señoras Marcelina|947|964
PERSONAS|Mercedes Ospina|967|982
PERSONAS|Sor Cura|1101|1109
PERSONAS|Madre Priora|1148|1160
Concedemos por nuestra parte permiso para que las exequias de las Señoras Marcelina y Mercedes Ospina, se celebren en la Iglesia del Cármen el dia 11 del presente mes. Siendo la Iglesia del Cármen una Iglesia exenta, el Sor Cura se entenderá con el Capellan i con la Madre Priora para que le permitan funcionar en ella, en las exequias expresadas.

Figura 32.
Porción del archivo resultante del análisis para el corpus histórico 1873

El archivo resultante contiene eventos pero no están ordenados y contiene varios campos que se explican a continuación:

- **Encabezado del archivo:** está compuesto por una URI que identifica el archivo que se está analizando y su localización.

- **Metadatos asociados al párrafo:** los metadatos por párrafo básicamente son 4:
    - PÁRRAFO: muestra un número que tiene el contador del párrafo en el documento empezando por el primero que aparece.
    - FECHAS: muestra la fecha etiquetada y los números de inicio y fin del *span*.
    - LUGAR: muestra el lugar etiquetado y los números de inicio y fin del *span*.
    - PERSONA: muestra el nombre etiquetado y los números de inicio y fin del *span*.

    Nota: de acuerdo a la definición de evento, éste ocurre cuando se encuentran en un mismo párrafo metadatos de FECHA, LUGAR y nombres de PERSONAS. Entonces los archivos resultantes contienen eventos y también otras combinaciones resultantes entre los metadatos que pueden ser sub eventos.

- **Texto del párrafo:** se muestra el texto del párrafo analizado donde se encuentran los metadatos anteriores para un análisis humano posterior.

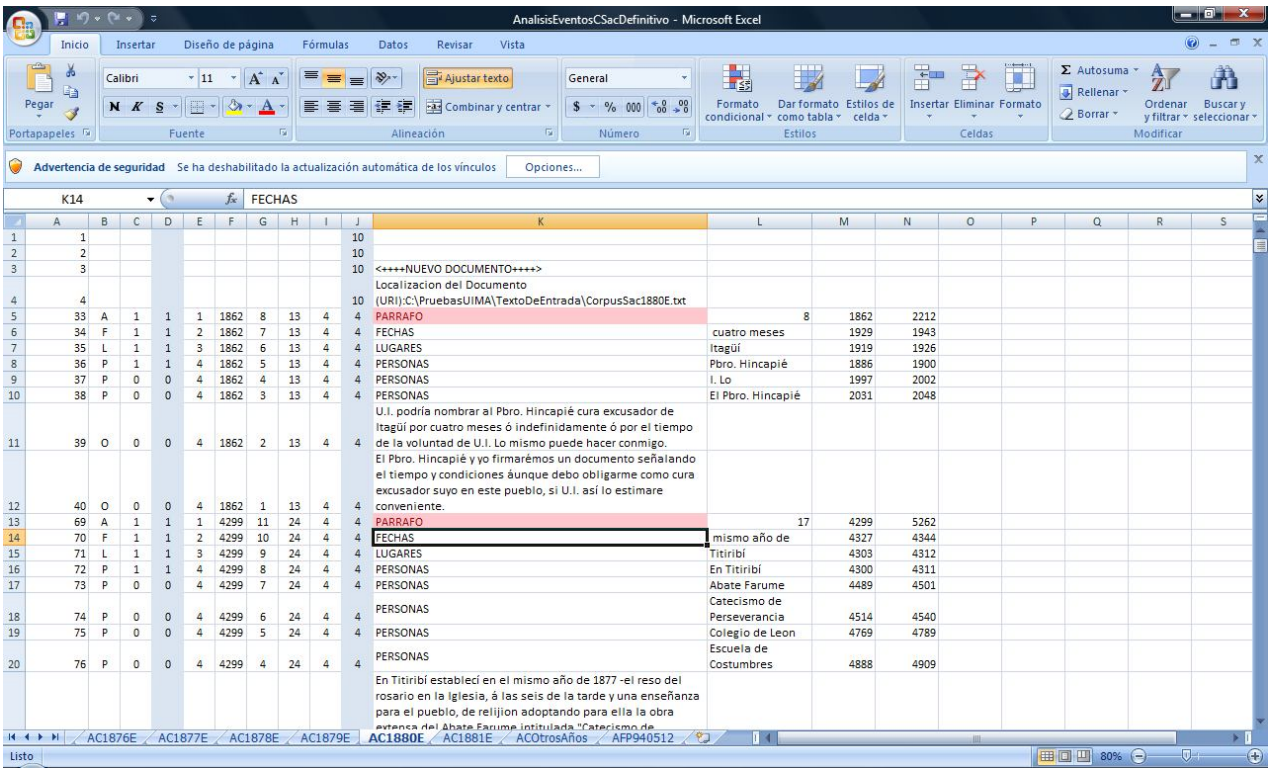
El archivo utiliza el signo "|" para los separadores por columna con el fin de facilitar la importación a Excel y efectuar el procesamiento de los eventos. Este procesamiento se llevó a cabo en varias hojas de cálculo (una por archivo) que contienen las fórmulas para el ordenamiento de los eventos obtenidos, esto es, se ordenan de acuerdo a los 4 metadatos de interés (PÁRRAFO, FECHA, LUGAR, PERSONAS y el texto del párrafo) y de manera descendente para mostrar las otras posibles combinaciones de datos o sub eventos. Dos ejemplos de eventos encontrados se muestran a continuación:

<++++NUEVO DOCUMENTO++++>			
Localización del Documento (URI):C:\PruebasUIMA\TextoDeEntrada\CorpusSac1876E.txt			
	Número del párrafo	**Comienzo**	**Fin**
PÁRRAFO	1	264	649
FECHAS	nueve de enero de mil ochocientos setenta y seis e	331	381
FECHAS	dos años	498	508
LUGARES	Medellin	313	322
PERSONAS	Vicaria Capitular de Medellin	292	321
PERSONAS	Dean Pbro	389	398
PERSONAS	D José Ignacio Montoya	400	422
PERSONAS Rector del Seminario Conciliar		433	463
En el Despacho de la Vicaria Capitular de Medellin a diez y nueve de enero de mil ochocientos setenta y seis el V. sr Dean Pbro. D José Ignacio Montoya -nombrado Rector del Seminario Conciliar de la Diócesis- por un período de dos años prestó el juramento de cumplir bien y fielmente los deberes de su encargo. Y para constancia firma con el Prelado y el infrascrito Secretario			
PARRAFO	9	1728	1775
FECHAS	Enero 28 de 1876	1756	1773
LUGARES	Medellin	1746	1756
PERSONAS	Vicaria Capitular Medellin Enero	1729	1762
Vicaria Capitular Medellin Enero 28 de 1876.			

Figura 33.

Ejemplos de eventos encontrados para el corpus histórico 1876

Una imagen del archivo en Excel utilizado para el ordenamiento se muestra a continuación:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	1									10				
2	2									10				
3	3									10	<++++NUEVO DOCUMENTO++++>			
4	4									10	Localizacion del Documento (URI):C:\PruebasUIMA\TextoDeEntrada\CorpusSac1880E.txt			
5	33	A	1	1	1	1862	8	13	4	4	PARRAFO	8	1862	2212
6	34	F	1	1	2	1862	7	13	4	4	FECHAS	cuatro meses	1929	1943
7	35	L	1	1	3	1862	6	13	4	4	LUGARES	Itagüí	1919	1926
8	36	P	1	1	4	1862	5	13	4	4	PERSONAS	Pbro. Hincapié	1886	1900
9	37	P	0	0	4	1862	4	13	4	4	PERSONAS	I. Lo	1997	2002
10	38	P	0	0	4	1862	3	13	4	4	PERSONAS	El Pbro. Hincapié	2031	2048
11	39	O	0	0	4	1862	2	13	4	4	U.I. podría nombrar al Pbro. Hincapié cura excusador de Itagüí por cuatro meses ó indefinidamente ó por el tiempo de la voluntad de U.I. Lo mismo puede hacer conmigo.			
12	40	O	0	0	4	1862	1	13	4	4	El Pbro. Hincapié y yo firmarémos un documento señalando el tiempo y condiciones áunque debo obligarme como cura excusador suyo en este pueblo, si U.I. así lo estimare conveniente.			
13	69	A	1	1	1	4299	11	24	4	4	PARRAFO	17	4299	5262
14	70	F	1	1	2	4299	10	24	4	4	FECHAS	mismo año de	4327	4344
15	71	L	1	1	3	4299	9	24	4	4	LUGARES	Titiribí	4303	4312
16	72	P	1	1	4	4299	8	24	4	4	PERSONAS	En Titiribí	4300	4311
17	73	P	0	0	4	4299	7	24	4	4	PERSONAS	Abate Farume	4489	4501
18	74	P	0	0	4	4299	6	24	4	4	PERSONAS	Catecismo de Perseverancia	4514	4540
19	75	P	0	0	4	4299	5	24	4	4	PERSONAS	Colegio de Leon	4769	4789
20	76	P	0	0	4	4299	4	24	4	4	PERSONAS	Escuela de Costumbres	4888	4909
											En Titiribí establecí en el mismo año de 1877 -el reso del rosario en la Iglesia, á las seis de la tarde y una enseñanza para el pueblo, de relijion adoptando para ella la obra extensa del Abate Farume intitulada "Catecismo de			

Figura 34.
Archivo en Excel utilizado para efectuar el ordenamiento de datos para la detección de eventos

Este proceso de ordenamiento en Excel es aparentemente sencillo pero fue un poco complicado de implementar porque era preciso definir muy bien los campos y analizar los retornos de carro para evitar espacios entre los párrafos que pudieran dificultar el análisis, así que era necesario efectuar varios ensayos de programación tanto en Excel como en el archivo de JAVA del CPE.

4.3 ANALISIS DE RESULTADOS PARA LA DETECCIÓN DE EVENTOS

4.3.1 Análisis cuantitativo de eventos. De acuerdo al ordenamiento resultante del archivo en Excel se elaboró una tabla con el número de veces que aparecen las combinaciones de etiquetas por párrafo para los dos corpus analizados (histórico y AFP). Las convenciones utilizadas para las etiquetas son:

A = etiqueta de párrafo, F = etiqueta de fecha, L = etiqueta de lugar, P = etiqueta de nombre de persona.

Archivos	Número de Documentos analizados	COMBINATORIA DE LAS ETIQUETAS							
		Total Párrafos	Número de eventos AFLP	Porcentaje eventos (AFLP) con relación al total de párrafos	Número de sub-eventos ALP	Porcentaje eventos (ALP) con relación al total de párrafos	Número de sub-eventos AFL	Número de sub-eventos AFP	Etiquetas que Aparecen 1 o 2 veces
1868	5	37	3	8%	11	30%	0	0	23
1869	17	110	21	19%	32	29%	4	0	53
1870	11	87	20	23%	19	22%	1	4	43
1871	9	141	12	9%	55	39%	1	2	71
1872	16	132	31	23%	35	27%	2	3	61
1873	18	139	14	10%	41	29%	3	4	77
1874	29	289	36	12%	62	21%	3	4	184
1875	22	129	19	15%	31	24%	5	5	69
1876	15	87	11	13%	28	32%	2	5	41
1877	14	105	9	9%	24	23%	5	3	64
1878	13	113	14	12%	25	22%	4	3	67
1879	8	62	2	3%	15	24%	2	4	39
1880	10	151	18	12%	37	25%	4	3	89
1881	2	23	1	4%	6	26%	0	0	16
Otros Años	2	12	3	25%	8	67%	0	0	1
TOTALES	**191**	**1617**	**214**		**429**		**36**	**40**	**898**
PORCENTAJES DISTRIBUCION DE EVENTOS				**13%**		**27%**	**2%**	**2%**	**56%**

Tabla 22.
Frecuencia de aparición de etiquetas para el corpus histórico etiquetado

Archivos	Número de Documentos analizados	COMBINATORIA DE LAS ETIQUETAS							
		Total Párrafos	Número de eventos AFLP	Porcentaje eventos (AFLP) con relación al párrafo	Número de sub-eventos ALP	Porcentaje eventos (ALP) con relación al párrafo	Número de sub-eventos AFL	Número de sub-eventos AFP	Etiquetas que Aparecen 1 ó 2 veces
AFP940512	129	1289	171	13%	276	21%	25	37	781
PORCENTAJES DISTRIBUCION DE EVENTOS				**13%**		**21%**	**2%**	**3%**	**61%**

Tabla 23.
Frecuencia de aparición de etiquetas para el corpus de contraste AFP Noticias (LDC, 2000)

Tal y como se observa, los eventos determinados por la presencia de 4 etiquetas representan un 13% con respecto al total de etiquetas tanto para el corpus histórico etiquetado como para el corpus de noticias. Este resultado es curioso porque se trata de corpus totalmente diferentes. Los sub eventos de Lugar y Persona representan también valores globales similares siendo 27% para el corpus histórico y 21% para el corpus de noticias, así que la hipótesis que se tenía desde el principio de una mayor detección de eventos utilizando las 4 etiquetas no es válida y la mayor cantidad de eventos se encuentran en la detección de etiquetas de lugares y personas en un mismo párrafo. Esto en cierto modo significó una sorpresa para la autora porque en la bibliografía examinada no se encontró una descripción detallada del comportamiento en las etiquetas semánticas y tampoco se esperaba este resultado sabiendo que los corpus son diferentes. Pero en cierto modo había de esperarse porque en general las fechas o las frases relativas al tiempo se escriben con menor frecuencia. Esto demuestra una vez más que estos temas relativos al análisis lingüístico textual dependen de las personas y de la manera en la cual piensan para escribir un texto, independientemente del comportamiento que se esperaría utilizando un método cuantitativo. Cabe recordar que la hipótesis de eventos que se maneja en este punto está determinada por la presencia de etiquetas y sus correlaciones pero no tiene todavía un análisis semántico, el cual se analizará en la próxima sección. También es interesante añadir que la mayoría de los párrafos se concentran en la categoría donde aparecen 1 ó 2 etiquetas que pueden ser por ejemplo, un lugar o solo el párrafo.

4.3.2 Análisis semántico de los eventos. Para efectuar el análisis semántico de los eventos (es decir de los párrafos que tenían las 4 etiquetas presentes) fue preciso hacer una lectura de cada uno para determinar si efectivamente correspondían a un evento, y realmente es así como se evalúan los eventos: por medio de juicios humanos. A continuación se muestra el análisis para el corpus histórico y parte del análisis para el corpus de noticias en las siguientes tablas:

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos (AFLP)	De que se tratan los eventos hallados	Si es un evento	No es un evento
1868	5	37	3	Encabezado de carta		1
				Licencias para Presbítero	1	
				Llamado para presentarse a concurso por parroquia	1	
1869	17	110	21	La gobernación pide el edificio del seminario	1	
				Inicio de carta		1
				Permiso al obispo para ordenarse habiendo participado en la guerra	1	
				Inicio de carta		1
				Inicio de carta		1
				Inicio de carta		1
				Toma de posesión de curato	1	
				Inicio de carta		1
				Posesión sacerdotes junta conciliar Medellín	1	
				Respuesta del obispo por la solicitud del edificio seminario	1	
				Toma de posesión de cargo sacerdote	1	
				Nombrar coadjutor para Fredonia	1	
				continuación carta coadjutor		1
				Jubilación padre Emigdio Restrepo	1	
				Solicitud de cambio de Cura Salamina	1	
				Inicio de carta		1
				Solicitud de cambio de Cura en Sanvicente	1	
				Inicio de carta		1
				Defensa sacerdote Lino Garro	1	
				Inicio de carta		1
				Nombrar sustituto de sacerdote	1	
1870	11	87	20	Queja cura de Jirardota	1	
				Nombramiento obispo de Evaria	1	
				Respuesta de la queja de Jirardota	1	
				Inicio de carta		1
				Nombramiento junta catedral	1	
				Consulta sobre linderos parroquiales en Sardina y canoas	1	
				Continuación consulta linderos	1	
				Respuesta del obispo sobre consulta	1	
				Inicio de carta		1
				Compra solares para construcción catedral	1	
				Compra solares para construcción catedral	1	
				Inicio de carta		1

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos (AFLP)	De que se tratan los eventos hallados	Si es un evento	No es un evento
				Resolución para construcción seminario Medellín	1	
				Respuesta del obispo para construcción seminario Medellín	1	
				Orden de destinación de limosnas para el Colegio Pio Latino de Roma	1	
				Solicitud de cura en Yolombó	1	
				Inicio de carta		1
				Inicio de carta		1
				Pago construcción catedral	1	
				Contabilidad de la diócesis	1	
1871	9	141	12	Pastoral Roma es de los Papas	1	
				Continuación pastoral Roma es de los papas	1	
				Continuación pastoral Roma es de los papas	1	
				Continuación pastoral Roma es de los papas	1	
				Inicio de carta	1	
				Citación del padre Gómez a comparecer ante autoridades	1	
				Inicio de carta	1	
				Nombramiento nuevo obispo de panamá	1	
				Inicio de carta		1
				Inicio de carta		1
				Plan de estudios seminaristas	1	
				Perdida de ornamentos monseñor Riaño	1	
1872	16	132	31	Firmas solicitud de cura	1	
				Delimitación parroquias Aránzazu y Filadelfia	1	
				Contabilidad del seminario	1	
				Inicio de carta		1
				Inicio de carta		1
				Acusaciones sacerdote de Carolina	1	
				Inicio de carta		1
				Decisión sobre acusación de cura por el obispo	1	
				Inicio de carta		1
				Decisión del obispo sobre acusación de cura	1	
				Permiso especial del obispo para residir en concordia	1	
				Inicio de carta		1
				Decisión del obispo sobre acusación de cura	1	
				Inicio de carta		1
				Inicio de carta		1
				Plan de estudios seminaristas	1	

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos (AFLP)	De que se tratan los eventos hallados	Si es un evento	No es un evento
				Plan de estudios seminaristas	1	
				Plan de estudios seminaristas	1	
				Plan de estudios seminaristas	1	
				Plan de estudios seminaristas	1	
				Inicio de carta	1	
				Inicio de carta	1	
				Construcción catedral Medellín	1	
				Inicio de carta		1
				Acusaciones sacerdote de Angostura	1	
				Principio de auto de la curia		1
				Fin de carta		1
				Inicio de carta		1
				Inicio de carta		1
				Inicio de carta		1
				Inicio de carta		1
1873	18	139	15	Juicio padre Leoncio Villa	1	
				Resolución juicio padre Leoncio Villa	1	
				Resolución juicio padre Leoncio Villa	1	
				Inicio de carta		1
				Prorrogación de licencias eclesiásticas	1	
				Inicio de carta	1	
				Respuesta acusaciones padre Leoncio Villa	1	
				Resolución diezmos señor Canvas	1	
				Resolución diezmos señor Canvas	1	
				Quejas cura de Salamina	1	
				Quejas cura de Salamina	1	
				Quejas cura de Salamina	1	
				Solicitud curas para Manizales	1	
				Quejas cura de Salamina	1	
				Resultados colecta seminario Medellín	1	
1874	29	289	36	Solicitud de sacerdotes para Villa María	1	
				Cambio de cura en Ansermaviejo	1	
				Resultados examen padre Jenaro Arroyave	1	
				Cambio de curatos entre la Ceja y Entrerrios	1	
				Inicio de carta		1
				División de curato en Medellín	1	
				Descripción de la situación económica del obispado	1	

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos (AFLP)	De que se tratan los eventos hallados	Si es un evento	No es un evento
				Solicitud de sacerdote	1	
				Inicio de carta		1
				Solicitud de dos sacerdotes para Guática y Arrayanal	1	
				Inicio de carta	1	
				Solicitud de reemplazo curato de Vahos por sarampión	1	
				Resultado juicio padre Bernardo Ocampo	1	
				Retiro de oposición a curato del padre Leoncio Villa	1	
				Inicio de carta		1
				Respuesta retiro oposición padre Leoncio Villa	1	
				Inicio de carta		1
				Inicio de carta		1
				Inicio de carta		1
				Inicio de carta		1
				Informe de desempeño del cura Silverio Gómez en Aguadas	1	
				Informe de desempeño del cura Silverio Gómez en Aguadas	1	
				Inicio de carta		1
				Vicisitudes de desplazamiento entre San Rafael y Marinilla	1	
				Respuesta de oposición padre Vicente Ceballos	1	
				Informe financiero Seminario	1	
				Informe financiero Seminario	1	
				Informe financiero Seminario	1	
				Informe financiero Seminario	1	
				Informe financiero Seminario	1	
				Informe financiero Seminario	1	
				fin de carta		1
				Solicitud prorroga becario del seminario Simón Urrea	1	
				Hoja de vida y actividades sacerdote	1	
				Fin de carta		1
				Hoja de vida y actividades sacerdote	1	
1875	22	129	19	Fin de carta		1
				Fin de carta		1
				Fin de carta		1
				Solicitud de pago de sermones	1	
				Juramento posesión del vicerrector seminario	1	
				Juramento posesión profesor seminario	1	

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos (AFLP)	De que se tratan los eventos hallados	Si es un evento	No es un evento
				Juramento posesión profesor seminario	1	
				Juramento posesión profesor seminario	1	
				Juramento posesión profesor seminario	1	
				Felicitaciones de nombramiento de vicario capitular	1	
				Dimisorias para el cura José Dolores Jiménez	1	
				Juramento posesión profesor seminario	1	
				Relato celebraciones varias en la ceja	1	
				Relato celebraciones varias en la ceja	1	
				Relato celebraciones varias en la ceja	1	
				Inicio de carta		1
				Dimisorias para el cura Tomas Molina	1	
				Traslado restos del padre Joaquín Naranjo a Manizales	1	
				Proceso juicio padre Leoncio Villa	1	
1876	15	87	11	Juramento posesión Rector del seminario	1	
				fin de carta		1
				Inicio de carta		1
				Resolución del obispo con relación a coadjutores de parroquia	1	
				Nombramiento de obispo para José Ignacio Montoya	1	
				Descripción de situación pueblos de oriente con respecto a situación económica y violencia	1	
				fin de carta		1
				Suspensión de orden de captura sacerdote	1	
				Envío de autos de visita a San Mateo al obispo	1	
				fin de carta		1
				fin de carta		1
1877	14	105	9	Inicio de carta		1
				Resolución de permuta de curatos	1	
				Pastoral sobre comportamiento del clero ante la situación de guerra	1	
				Protesta de sacerdote por la pastoral sobre comportamiento del clero	1	
				Protesta de sacerdote por la pastoral sobre comportamiento del clero	1	
				Reapertura de iglesias y culto en el cauca	1	
				fin de carta		1
				Traslado de sacerdote para Medellín	1	
				Traslado de sacerdote para Medellín	1	

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos (AFLP)	De que se tratan los eventos hallados	Si es un evento	No es un evento
1878	13	113	14	Inicio de carta		1
				Inicio de carta		1
				Defensa de acusación	1	
				Inicio de carta		1
				Inicio de carta		1
				Inicio de carta		1
				Recomendación sacerdote de Sanjerónimo	1	
				Relato muerte de sacerdote enfermo	1	
				Acusación hacia 2 matrimonios por no asistir a la ceremonia	1	
				fin de carta		1
				Inicio de carta		1
				Inicio de carta		1
				Relato ejercicios espirituales	1	
				Amonestación paternal al cura Epitacio Quiros	1	
1879	8	62	2	Inicio de carta		1
				Informe proceso de cura de Neira	1	
1880	10	151	18	Nombramiento cura excusador Itagüí	1	
				Relato de ejercicios piadosos en Titiribí	1	
				Solicitud de dimisorias	1	
				Respuesta del obispo sobre las dimisorias	1	
				Aprobación de las dimisorias	1	
				Inicio de carta	1	
				Inicio de carta	1	
				Respuesta a solicitud de sacerdote para Valparaíso	1	
				Asignación de renta para el padre Pedro Gómez	1	
				Resolución de exámenes padre Emigdio Restrepo	1	
				Donaciones diferentes iglesias	1	
				Aparte de hoja de vida sacerdote	1	
				Aparte de hoja de vida sacerdote	1	
				fin de carta		1
				Inicio de carta	1	
				Relato de cura excusador de Itagüí	1	
				Relato de cura excusador de Itagüí	1	
				Inicio de carta	1	
1881	2	23	1	Inicio de carta		1
Otros años	2	12	3	Licencias para presbítero Ramón Hoyos	1	

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos (AFLP)	De que se tratan los eventos hallados	Si es un evento	No es un evento
				Pronunciamiento de la santa sede contra la violencia a la iglesia colombiana	1	
				Pronunciamiento de la santa sede contra la violencia a la iglesia colombiana	1	
TOTALES	**191**	**1617**	**215**		**150**	**65**

Tabla 24.
Análisis semántico detallado para eventos en corpus histórico

Muchos de los desaciertos en eventos para este corpus ocurrieron porque los inicios y finales de cartas cumplen con el requisito de contener las 4 etiquetas y también se presentaron falsos positivos en algunos párrafos que no contenían fechas sino cantidades de dinero que son detectados como fechas. También es importante añadir la identificación de los eventos fue sencilla porque se conocía de antemano el corpus (esa es justamente una de las ventajas de haberlo construido) pero esta identificación no es trivial. Una vez más sale a relucir la dependencia del texto a analizar en este tipo de estudios.

Con relación al corpus de AFP, puesto que es más extenso que el corpus histórico y se trata de un solo archivo que contiene varios documentos, a continuación se muestra una parte del análisis en la siguiente tabla:

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos	De que se tratan los eventos hallados	SI son eventos	NO son eventos
AFP940512	129	1289	171	Colombia concede ciudadanía a hijo de ex espías soviéticos	1	
				Aldrich Amesd, estadounidense condenado a cadena perpetua por espionaje	1	
				Una mujer radicada en Argentina dice haber sido testigo de un atentado cometido en Roma hace 50 años	1	
				ONU acepta solicitud de México para asistencia electoral	1	
				Eliminados tenistas argentinos del torneo Coral Springs	1	
				Un sobrino del ex Presidente chileno Augusto detenido en Santiago por proceso de estafa	1	
				Las exportaciones brasileras de café sumaron 4.236.720 sacos (60 kilos), por 348.351.000 dólares,	1	
				La policía chilena detuvo a un centenar de estudiantes, que el jueves se apoderaron del edificio central de la Universidad Tecnológica Metropolitana	1	
				Los jóvenes exigen la renuncia de la directora de la escuela de Trabajo	1	

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos	De que se tratan los eventos hallados	SI son eventos	NO son eventos
				María Boccaci de Ciani, de 81 años, radicada en la villa turística de Bariloche, donde vive Priebke desde hace 46 años, estimó que el ex capitán nazi, cuya extradición solicitó Italia, "hoy no merece castigo".	1	
				Continuación de la noticia del nazi	1	
				Continuación de la noticia del nazi	1	
				La Organización de Estados Americanos (OEA) debe jugar un papel decisivo para resolver el caso haitiano,	1	
				Catorce ciudadanos chinos que habían ingresado a territorio boliviano procedentes de Sao Paulo, Brasil, con pasaportes falsos	1	
				Continuación de la noticia de los chinos	1	
				Unos cuarenta jóvenes ultraderechistas armados con cuchillos y porras daban la caza a los extranjeros en la ciudad de Magdeburgo	1	
				Festival de cine de Cannes	1	
				Programación de vuelos aeropuerto de Orly	1	
				Manifestación de campesinos de Rio de Janeiro	1	
				Continuación manifestación	1	
				El ciclista español Miguel Induráin realizará su último ensayo antes de participar en vueltas ciclísticas	1	
				Continuación noticia ciclista	1	
				Solicitud de ayuda al pueblo palestino de OMS	1	
				Continuación ayuda al pueblo palestino de OMS	1	
				Acusación de acoso sexual contra Bill Clinton	1	
				Juego de partido de alto riesgo en Chile	1	
				Continuación juego de partido de alto riesgo en Chile	1	
				Recibimiento del Papa en el hospital Gemelli a otro obispo	1	
				Gira internacional del candidato Luiz Ignacio Lula Da Silva	1	
				Continuación noticia gira internacional del candidato Luiz Ignacio Lula Da Silva	1	
				Análisis de levantamiento del embargo internacional de armas en EEUU	1	
				Venta de compañía de navegación en Rio de Janeiro	1	
				Continuación venta de compañía de navegación en Rio de Janeiro	1	
				Solicitud de levantamiento de prohibición de entrada de periodistas a Jericó	1	
				Continuación periodistas en Jericó	1	
				Continuación periodistas en Jericó	1	
				Aumento de embarques de soya en Brasil	1	
				Extradición a Bolivia de ex presidente	1	
				Continuación Bolivia ex presidente	1	
				La muerte del “Trottoirs de Buenos Aires”, trece años de tango en Paris	1	

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos	De que se tratan los eventos hallados	SI son eventos	NO son eventos
				Continuación muerte del Trottoirs	1	
				Desaparición de escritor autor del principito	1	
				Continuación escritor autor del principito	1	
				Lanzamiento de álbum de grupo de rock alternativo francés	1	
				Gira de cantante de salsa venezolano en París	1	
				Continuación noticia gira de cantante de salsa	1	
				Destrucción de fortificaciones en Yemen	1	
				Consecuencias de negado de visa de Maradona en Japón	1	
				Continuación noticia de negado de visa de Maradona en Japón	1	

Tabla 25.
Análisis semántico detallado para eventos en corpus de contraste AFP Noticias (LDC, 2000)

Para éste corpus la identificación del evento no fue muy complicada porque el estilo periodístico conciso hace que la información sea muy precisa y, por tanto, en muchas ocasiones solo era necesario leer el primer renglón del párrafo para identificarlo.

Una de las conclusiones de este análisis es que se analizaron dos tipos de texto diferentes y la eficiencia del sistema variará dependiendo del tipo de texto.

Continuando con el análisis semántico de eventos, con base en los resultados anteriores se elaboraron dos tablas de resultados globales que se muestran a continuación y donde se muestra la "precisión semántica", lo que significa si correspondía a un evento o no.

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos (AFLP)	SI son eventos	NO son eventos	Precisión Semántica
1868	5	37	3	2	1	66.7%
1869	17	110	21	12	9	57,1%
1870	11	87	20	15	5	75%
1871	9	141	12	10	2	83,3%
1872	16	132	31	17	14	54,8%
1873	18	139	15	14	1	93,3%
1874	29	289	36	26	10	72,2%
1875	22	129	19	15	4	78,9%
1876	15	87	11	6	5	54,5%
1877	14	105	9	7	2	77,8%
1878	13	113	14	6	8	42,9%
1879	8	62	2	1	1	50%
1880	10	151	18	16	2	88,9%
1881	2	23	1	0	1	0%
Otros años	2	12	3	3	0	100%
TOTALES	**191**	**1617**	**215**	**150**	**65**	
PORCENTAJE				**70%**	**30%**	

Tabla 26.
Análisis semántico para eventos en el corpus histórico

Archivos	Documentos analizados	Párrafos analizados	Numero Eventos (AFLP)	SI son eventos	NO son eventos	Precisión semántica
AFP940512	129	1289	171	170	1	
TOTALES	**129**	**1289**	**171**	**170**	**1**	
PORCENTAJE				**99%**	**1%**	**99%**

Tabla 27.
Análisis semántico para eventos para el corpus de contraste AFP Noticias (LDC, 2000)

Los porcentajes de asertividad globales con respecto al número total de eventos hallados en eventos fueron 70% para el corpus histórico y 99% para el corpus de noticias.

Aunque la identificación de los eventos en los textos se hizo sin dificultad para los dos corpus, los problemas de identificación del evento surgen cuando se disminuye el número de las etiquetas presentes en el párrafo porque en primer lugar se pierde información necesaria para entender el contexto y se utilizan otros componentes lingüísticos más

complejos que exigen mayor uso de inferencia humana pero que son poco triviales para el proceso computacional, por ejemplo a continuación se muestran varios párrafos aleatorios que contienen 2 y 1 etiquetas para los dos corpus, donde se concentraba la mayor cantidad de porcentaje con respecto a los párrafos analizados:

Corpus	Numero de etiquetas	Tipo de etiquetas	Párrafo
Histórico	2	Persona y Párrafo	No se vaya a enojar conmigo, porque entonces si acabaré yo de enloquecerme. A veces deseo yo perder el juicio para que algunos me hagan el honor de decir, aunque ya loco yo no necesite ni acepte honores: "Vean UU.: Baltasar si tenia juicio puesto que lo perdió".
Histórico	2	Lugares y Párrafo	Acompaño a U.I. una comunicacion del cura de Filadelfia y la resolucion que al pie de ella he dictado en conciencia
Histórico	1	Párrafo	En los campos hay algunas escuelas, y se enseña en ellas la doctrina. Pero hay que confesar que las escuelas son muy pocas y que se necesitan grandes esfuerzos para llenar esta faltas.
AFP	2	Lugares y Párrafo	Claro que tuvimos que ver con la rebelión de Chiapas, porque con la reflexión cristiana empujamos a los indios a recuperar su dignidad, a ser conscientes de sus deberes pero también de sus derechos, dijo.
AFP	2	Persona y Párrafo	Al parecer, el Papa habría manifestado su interés en conversar personalmente con monseñor Ruiz, aunque su estado de salud no le permitiría seguramente un encuentro demasiado prolongado.
AFP	2	Fechas y Párrafo	En el marco del programa de la privatización de estatales, iniciada en octubre de 1991, ya fueron transferidas al sector privado 24 de las 66 empresas públicas incluídas en la primera fase de ese programa, cuyas ventas totalizaron 7.861 millones de dólares.
AFP	1	Párrafo	Hoy no vamos a lanzar ninguna pedrada, afirmó uno de los adolescentes que subieron a los techos de autos y camiones con banderas palestinas para ver pasar a los policías.
AFP	1	Párrafo	Los evacuados señalaron que la mayoría de los comercios están cerrados y que se forman largas colas en las gasolineras abiertas.

Tabla 28.
Texto de los párrafos para 2 y 1 etiquetas en el corpus histórico y de contraste AFP Noticias (LDC, 2000)

4.3.3 Evaluación del sistema de detección de eventos. Una de las principales dificultades en la detección de eventos es justamente como evaluarlos porque requiere definir muy bien que se entiende por “evento” y luego es necesario acudir a juicios humanos para determinar si lo hallado realmente corresponde o no corresponde y justamente ahí radica parte de la

dificultad porque es un proceso costoso en tiempo y en talento humano porque es necesario entrenar a las personas que efectuarían el análisis de acuerdo a la definición de evento propuesta, además de que la desambiguación automática todavía es un problema no resuelto y también el juicio evaluado por humanos tiene sus riesgos en cuanto a su imparcialidad. Para este trabajo por ejemplo, solo se evaluaron los eventos que fueron determinados por la presencia de 4 etiquetas y que corresponden al 13% de las mediciones efectuadas para los dos corpus, pero para evaluar si los párrafos que contienen 3 etiquetas contienen eventos o no es preciso hacer éste mismo ejercicio que requiere componente humano y de hecho, en la evaluación de eventos se utilizan corpus más pequeños (véase los trabajos comentados en la reseña del capítulo 1) y aquí se utilizaron dos corpus diferentes que contienen varios documentos, por lo tanto se deja como tema de análisis posterior la evaluación de los otros párrafos con las otras combinaciones menores a 4.

Con relación a cómo medir el desempeño del sistema, es también otra difícil pregunta porque todavía es tema de investigación. Las medidas de precisión y cobertura se utilizan para RI y EI pero es difícil ponderarlos en la medición de eventos porque implica saber que tanta cobertura tuvo una noticia en los párrafos que la componen o que tan preciso es un párrafo de otro y esto es claramente un problema semántico que aún está sin solucionar. Sin embargo, para este trabajo se utilizó la precisión de los eventos en cuanto a su asertividad con respecto de las correferencias entre etiquetas y de esta manera fue posible "medir" que la precisión semántica global fue de 70% y 99% para los párrafos que contenían las cuatro etiquetas para corpus histórico y de noticias respectivamente, pero como se mencionó, las medidas de evaluación semánticas para eventos son tema de estudio actual y se encuentran fuera del alcance de este trabajo.

Adicionalmente las medidas de correlación entre etiquetas permiten visibilizar algunas características textuales que indican otras maneras de abordar estos problemas semánticos por ejemplo, en las tablas 22 y 23 donde se analizan los porcentajes de etiquetas se observa que más del 50% de los párrafos tienen una o ninguna etiqueta (eliminando la etiqueta de párrafo), lo cual indica que la mayor parte de la información presente en los dos corpus es

mas discursiva y por tanto deben utilizarse otros mecanismos como por ejemplo, la utilización de etiquetas lingüísticas presentes en los textos que indiquen mayor información relativa a la composición de las frases.

5. CONCLUSIONES

A continuación se presentan las siguientes conclusiones de acuerdo a los temas tratados:

Conclusiones relacionadas con los aspectos metodológicos

- Metodología en lingüística computacional: para efectuar cualquier tipo de labor en lingüística computacional es indispensable tener claro desde el principio dos aspectos: el objeto de estudio, el objetivo del análisis y el corpus que se analizará. El éxito de la investigación dependerá de qué tan claros y definidos se encuentren tanto los objetivos como el objeto de estudio. Además es necesario tener una dosis de humildad para enfrentar estos temas y "*continuar la marcha y no desfallecer en el camino*" porque los computadores y las técnicas utilizadas no funcionan siempre con la misma efectividad que tendría un ser humano en muchas tareas. Aun falta mucho camino por recorrer en inteligencia artificial para entender muchos procesos que el ser humano puede hacer de manera automática y que la naturaleza ha refinado por miles de años.
- La metodología descrita a lo largo de trabajo brinda elementos de comprensión de la lingüística computacional para otros futuros investigadores que deseen abordar el estudio de estos temas y también es posible aplicarla para el análisis de extracción de información en los sistemas de bibliotecas digitales y estudios lingüísticos en general.
- Construcción de corpus: para iniciar cualquier estudio lingüístico computacional es indispensable obtener el corpus donde se analizarán los datos y de hecho es su materia prima. La elección de los documentos de análisis y el formato digital utilizado son factores importantes a la hora de su construcción porque permiten definir tanto los temas de estudio como el tratamiento que deberá tenerse a la hora de su manipulación computacional. Así mismo existen metodologías para la construcción de corpus Treebank, pero el proceso como tal puede ser costoso en tiempo y en talento humano sobre todo por el esfuerzo y cuidado que debe tenerse a la hora de su implementación, por lo tanto debe sopesarse si para efectuar un estudio se construye el corpus ó se consigue uno especifico, sin embargo esta elección dependerá de los objetivos de

análisis del investigador y de las herramientas con que se cuenta para determinar si las que existen pueden ser refinadas ó si es necesario desarrollar otras herramientas de estudio.

- Etiquetamiento de corpus: para tener éxito en una labor de etiquetamiento de corpus es indispensable tener claridad sobre el objetivo que se persigue en el análisis lingüístico, definirlo muy bien para caracterizar las etiquetas en el texto y encontrar una herramienta que permita efectuar los etiquetamientos con facilidad de acuerdo a las habilidades del investigador.
- Detección de eventos: es un tema de interés actual pero es necesaria la adquisición de herramientas matemáticas y lingüísticas avanzadas para su estudio.

Conclusiones relacionadas con los aspectos temáticos:

- En este proyecto de investigación se construyó un corpus para análisis lingüístico, se implementó un sistema de etiquetamiento que incluía los nombres de personas, lugares y fechas por medio de una herramienta computacional, se analizó una metodología de trabajo para detectar eventos en textos históricos basadas en otros trabajos similares y se describió la metodología empleada en todo el proceso. Esto muestra que se llevó a cabo el objetivo general planteado en el proyecto de investigación.
- El análisis de textos históricos y la construcción de un corpus para su estudio lingüístico permite estudiar datos lingüísticos valiosos. En este trabajo fue posible caracterizar algunos de los aspectos lingüísticos del s. XIX en Colombia por medio de la digitalización de los documentos de la Iglesia Católica pero además la construcción del corpus permitirá realizar otros estudios posteriores por parte de otros investigadores. Esta conclusión muestra que se logró llevar a cabo el objetivo específico 1 planteado en el proyecto de investigación.
- La caracterización del etiquetamiento de los nombres de lugares y personas depende del tipo de corpus que se tenga, esto es, el corpus histórico construido es muy rico en nombres de personas y su caracterización ha permitido que el sistema pueda analizar otros textos, lo cual indica que muchos de los resultados dependerán de la escogencia

del corpus de trabajo y también permitirá efectuar otro tipo de análisis de PLN que requerían de este trabajo previo tales como la aplicación de algoritmos de aprendizaje y otras estrategias de inteligencia artificial.

- La caracterización de entidades desarrollada presentó valores de cobertura y precisión en los rangos típicos para este tipo de sistemas. Los valores de precisión para fechas, lugares y nombres de personas fueron muy similares para el corpus histórico y el corpus de contraste, obteniendo valores de 76% y 61% para fechas, 64% y 83% para lugares (donde se observa que para el corpus de contraste hubo mayor asertividad debido a que los nombres de lugares son más universales) y 71% y 51% para los nombres de personas. Los falsos positivos que se presentaron durante la caracterización de nombres para lugares y personas fueron entidades que no correspondían a la entidad analizada, por ejemplo la mayor cantidad de falsos positivos en personas correspondieron a nombres de lugares. Sin embargo es posible refinar aun mas los modelos de identificación y reconocimiento de nombres de lugares y fechas ya teniendo clara la metodología de trabajo.
- Aunque los verbos fueron etiquetados y se obtuvieron buenos índices de cobertura es necesario llevar a cabo un análisis más profundo sobre sus tipologías lingüísticas con el fin de lograr mayor resultado en el análisis para la detección de eventos y otros tipos de análisis del léxico de los textos utilizados.
- Aunque es un tema de investigación actual, se planteó una posible metodología para la detección de eventos en textos pero es necesario un estudio más amplio del tema porque, aunque la hipótesis de trabajo sobre detección de eventos que consistía en demostrar que la presencia de las etiquetas de nombres propios, lugares y fechas en un mismo párrafo eran un indicio de un evento no se cumplió, todavía no se descarta y este resultado negativo ocurrió debido a limitaciones metodológicas. Todavía es posible indagar al respecto empleando recursos lingüísticos mas avanzados.
- Considerando que se definieron como eventos la ocurrencia de las etiquetas de nombres de personas, lugares y fechas en un mismo párrafo, el porcentaje de eventos puros en los dos corpus analizados correspondieron al 13% de los párrafos analizados, un porcentaje pequeño en comparación con los subeventos donde aparecían etiquetas

de lugares y personas y que correspondió al 27% para el corpus histórico y 21% para el corpus de contraste. Los subeventos donde se encontraron etiquetas de fecha y lugar ó fecha y persona corresponden a una minoría en el porcentaje total (2% para ambos corpus). De igual manera, se comprobó que la cantidad de párrafos que contienen 1 ó 2 etiquetas a lo sumo, corresponden a valores de 56% para el corpus histórico y el 61% para el corpus de contraste. Estas medidas están basadas en la combinatoria de las etiquetas y evidencia la dificultad existente para la evaluación de eventos.

- En cuanto a la evaluación semántica de los eventos hallados, el análisis mostró diferencias entre los dos corpus analizados. Para el corpus histórico la asertividad de los eventos fue 70%, siendo el 30% restante correspondiente a párrafos que eran inicio de carta y no correspondían al texto de la carta como tal. El corpus de contraste si tuvo una asertividad semántica de eventos del 99%. En este sentido, el análisis evidencia la diferencia de estilo y de narrativa que presentan ambos textos.
- Puesto que en la investigación se definieron los eventos y se desarrolló un sistema para detectarlos y ordenarlos para su posterior análisis y validación por parte del usuario, ésta y las anteriores conclusiones muestran que se logró llevar a cabo el objetivo específico 2 planteado en el proyecto de investigación.

6. TRABAJOS FUTUROS

En estos temas de lingüística de corpus, lingüística computacional y detección de eventos existen muchos temas de estudio que pueden abordarse. Para el caso concreto de esta investigación se proponen algunos temas de estudio que pueden ser llevados a cabo en trabajos de pregrado ó doctorado:

- Etiquetamiento de corpus: es posible profundizar en la caracterización de otros criterios de etiquetamiento en los corpus como son por ejemplo, etiquetamientos de tipo pragmático (sobre todo para el corpus histórico que es bien prolífico en estos temas discursivos) y los etiquetamientos lingüísticos para determinar por ejemplo, los tipos de errores ortográficos más comunes y las variaciones de los nombres de personas y lugares y apellidos, así como etiquetar las monedas y las palabras utilizadas en la terminología católica. Además es posible refinar aun más los modelos de etiquetamiento para lugares y fechas.
- Otro trabajo es comparar el sistema de etiquetamiento empleado con otras herramientas disponibles en el medio y medir el desempeño del sistema.
- Detección de eventos: es posible refinar los sub-eventos que contienen 3 y menos etiquetas y caracterizarlo con mayor precisión para obtener datos adicionales que conlleven a la elaboración de una metodología de detección de eventos más profunda pero este es un tema que puede ser tratado como trabajo para un posible doctorado.
- Detección de redes sociales: es otro tema que puede abordarse ya teniendo en cuenta la metodología descrita en este trabajo y el corpus digitalizado: utilizar técnicas de análisis de redes sociales para determinar quien influye sobre quién y otras características sociales más sutiles. También es un tema para un posible trabajo de doctorado.

ANEXO A PSEUDOCÓDIGOS

1. DESCRIPCION DEL ALGORITMO EMPLEADO PARA ELIMINAR ETIQUETAS REPETIDAS

Fue necesario implementar un algoritmo para eliminar etiquetas repetidas en el sistema porque los programas en java analizan las ER una a una cuando recorren el texto y ocurre con frecuencia que una misma entidad sea reconocida por varios modelos al tiempo, lo cual era de esperarse y además constituye una situación problemática porque si una misma entidad está etiquetada varias veces no es posible establecer los datos de conteo verdaderos a la hora de efectuar el análisis. La siguiente grafica ilustra esta situación de múltiple etiquetamiento para una entidad:

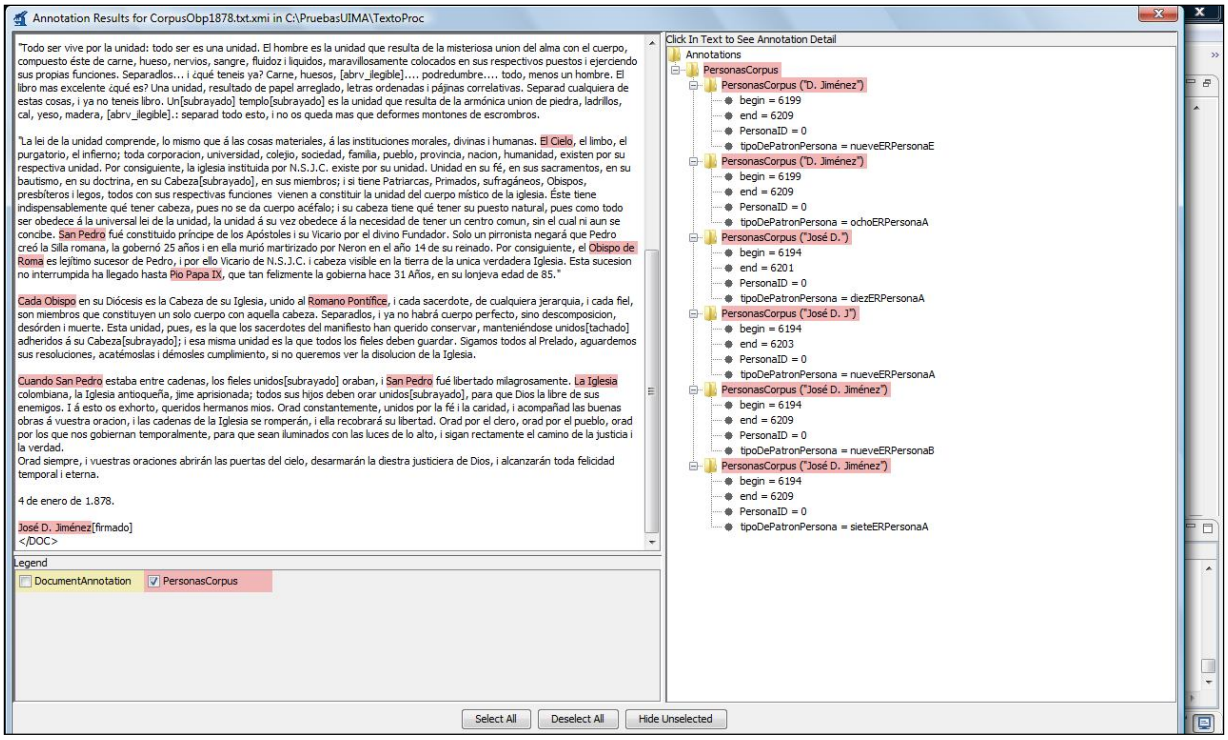


Figura 35. Anotaciones de personas mostrando repetición en las ER para el corpus histórico

Como se observa en el ejemplo, para la entidad "José D. Jiménez" existen cinco ER que la identifican como acertada incluyendo expresiones tales como "José D." o "D. Jiménez" o

"José D. Jiménez" siendo ésta la ER deseable. El ser humano a la hora de identificar las expresiones identifica automáticamente cuál de ellas es la correcta pero el computador debe distinguir cual es la ER deseada. El criterio adoptado para solucionar este problema fue escoger la etiqueta más grande para una entidad y eliminar las etiquetas contenidas en ella. Para implementarlo es necesario hacer una iteración sobre el tipo de entidad de interés (por ejemplo si es PersonasCorpus) y hacer sucesivas comparaciones entre los números de inicio y final de las etiquetas para determinar cuál es la etiqueta más grande y cuales etiquetas están contenidas y proceder a eliminar las etiquetas sobrantes. A continuación se muestra el pseudocódigo de esta implementación:

```
Posiciones de inicio y fin de la anotación anterior a analizar
oldStart = inicio de la anotación de personas anterior
oldEnd = fin de la anotación de personas anterior
while (iPers.hasNext()) {
        pAnotNew = (PersonasCorpus)iPers.next();
        newStart = inicio de la anotacion de personas (pAnotNew)
        newEnd = inicio de la anotacion de personas (pAnotNew)

        newStart >= oldStart
        if (newStart < oldEnd) {
                if (newStart == oldStart) {
                                Borrar etiqueta pAnotNew

                } else if (newEnd <= oldEnd) {
                                Borrar etiqueta pAnotNew
                        }
else {
                Renombrar el inicio y el fin de la etiqueta anterior
                pAnotOld(newEnd);
                oldEnd = newEnd;
                Borrar etiqueta pAnotNew
        }

} else {
                Renombrar el inicio y el fin de la etiqueta anterior
                oldStart = newStart;
                oldEnd = newEnd;
                pAnotOld = pAnotNew;
}
```

Figura 36. Pseudocódigo para eliminar etiquetas repetidas

Luego de aplicar el algoritmo queda una sola anotación válida tal como se muestra en la figura para el mismo ejemplo mostrado:

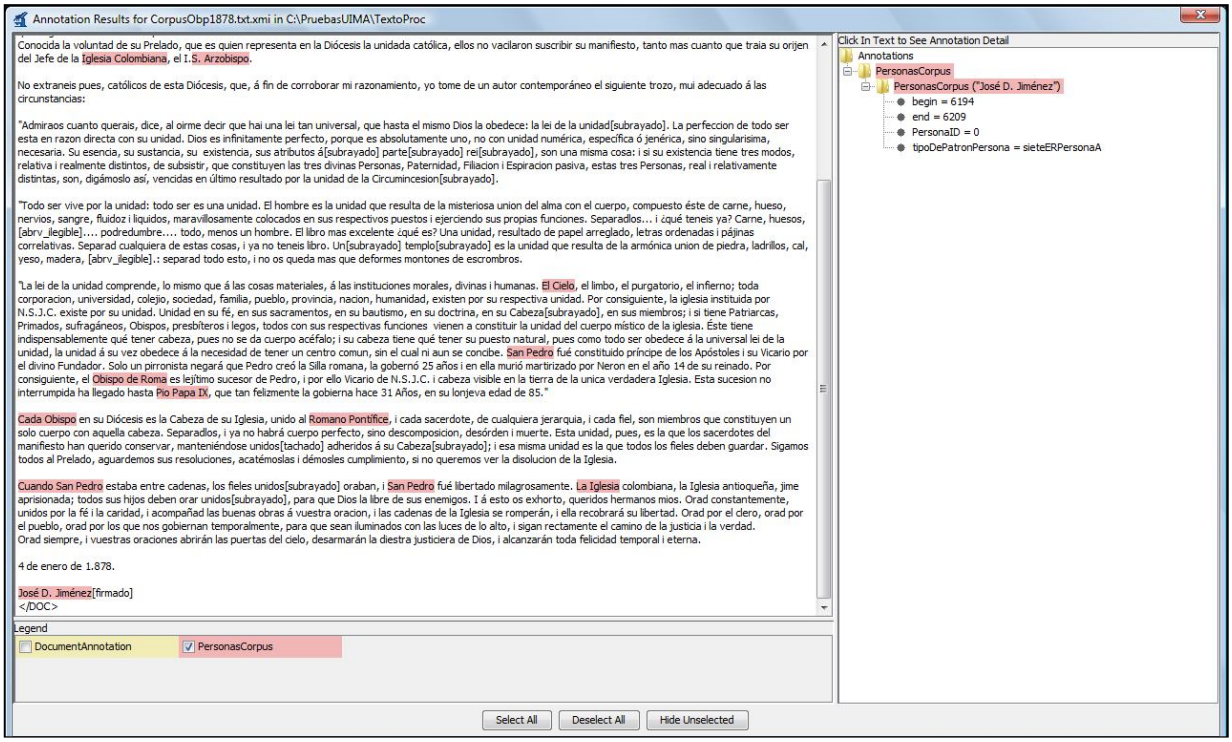


Figura 37. Anotación de persona para el corpus histórico luego de aplicar el algoritmo descrito

2. DESCRIPCION DE LOS ALGORITMOS EMPLEADOS PARA ETIQUETAR LISTADOS DE PALABRAS

Fue necesario implementar un algoritmo para recorrer los listados de palabras y compararlas con el texto analizado y si la encontraba, etiquetarla. Esto se aplicó para fechas, nombres de lugares predeterminados y los verbos.

El siguiente pseudocódigo muestra como construir un iterador del tipo de anotación (Fecha, Lugar, Persona, Verbo) que se va a buscar en el texto. Esto es necesario porque las anotaciones se guardan en un índice general del UIMA y es necesario recuperarlas definiendo otro índice especifico de acuerdo al tipo de anotación.

```
Obtener un iterador sobre las anotaciones contenidas en el tipo de Anotacion definida:

FSIterator it = aCAS.getAnnotationIndex(TipoDeAnotacion).iterator();

    Hacer un loop sobre el iterador:
    while (it.isValid()) {
        Obtener el iterador
        AnnotationFS anotacion = (AnnotationFS)it.get();

        Obtener el texto de la anotacion encontrada
        String coveredText = anotacion.getCoveredText();

        Obtener la anotacion de la posicion donde comienza la iteracion
        anotBegin = anotacion.getBegin();

        Ir a la clase para recorrer el texto
         anotarRango(aCAS, coveredText, anotBegin);

        Avanzar sobre el iterador:
        it.moveToNext(); }
```

Figura 38. Pseudocódigo para análisis de listado de palabras

Teniendo luego el tipo de anotación, es necesario comparar el tipo de palabras con el texto y anotar cuando se identifique una de ellas. El listado de palabras se guarda en un arreglo de *Strings* y cada arreglo se va comparando con el texto. Este es el pseudocódigo del método anotarRango:

```
Hacer un For para recorrer el arreglo que contiene el listado de
palabras:

    for (i = 0; i < aVerbo.size()) {

      La anotación comienza cuando el texto coincide con un espacio en blanco concatenada
      con el inicio de la palabra del listado:
      start = aText.indexOf(" ".concat(aVerbo.get(i)))+1;

      La anotación termina cuando el texto coincide con un espacio en blanco a partir de
      la posición del inicio:
      end = aText.indexOf(" ", start);

      Luego se procede a crear la anotación de acuerdo al tipo de anotación de la API de UIMA
      createAnnotation(aCAS, start, end); }
```

Figura 39. Pseudocódigo método anotarRango

3. DESCRIPCION DE LOS ALGORITMOS EMPLEADOS PARA ETIQUETAR ENTIDADES UTILIZANDO EXPRESIONES REGULARES

La otra heurística empleada para la anotación consistió en definir las ER para identificar las palabras de interés. Primero se definieron las ER que fueron mostradas en el capítulo 3 y se les nombraba. El segundo paso consistía en comparar las ER con el texto y en caso de coincidencia, efectuar la anotación, tal como se muestra en el siguiente pseudocódigo:

```
/ /buscar fecha con el modelo de laER (en este caso, unoERPersona)

   Definir un buscador de la ER que recorra el texto:
   Matcher matcher = unoERPersona.matcher(docText);

   Mientras matcher sea encontrado:   while (matcher.find()) {

   AnotadorPersona pAnot = new AnotadorComodin(aJCas);

    Definir el comienzo de la anotacion:
    pAnot.setBegin(matcher.start());

   Definir el final de la anotacion:
    pAnot.setEnd(matcher.end());

   Agregar la anotacion a los indices generals de UIMA:
   pAnot.addToIndexes();   }
```

Figura 40. Pseudocódigo detección de ER

BIBLIOGRAFÍA

1. Abaitua, J., (2000). "Tratamiento de corpora bilingües", en Seminario La ingeniería lingüística en la sociedad de la información. Fundación Duques de Soria, Julio 2000. Disponible en internet en (http://paginaspersonales.deusto.es/abaitua/konzeptu/ta/soria00.pdf)

2. Alcántara Plá, M. (2007). "Introducción al análisis de estructuras lingüísticas en corpus". Ediciones UAM (Universidad Autónoma de Madrid).

3. Allan, J., Papka, R., Lavrenko, V., (1998). "*On-line new event detection and tracking*", en *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pag. 37–45.

4. Apache UIMA (2009), UIMA (*Unstructured Information Management Architecture,* en http://incubator.apache.org/uima. Fecha de consulta 31 de marzo de 2009.

5. Baeza-Yates, R., Ribeiro-Neto, B. (1999). "*Modern Information Retrieval*". Ed. Addison-Wesley.

6. Byrne, K. (2007), "*Nested Named Entity Recognition in Historical Archive Text*", en International Conference on Semantic Computing, 2007. ICSC 2007, Septiembre, pag. 17-19.

7. Buchanam, G., Cunningham, S.J,. Blandford, J., Rimmer, J., Warwick, C. (2005). "*Information seeking by humanities scholars*", en ECDL05: *Proceedings of the 9th European Conference on research and advanced technology for Digital Libraries*, *Lecture notes in Computer Science*, pag. 218- 229.

8. Carbonell, J. (1992). "El procesamiento del lenguaje natural, tecnología en transición", Congreso de la Lengua Española, Sevilla. [Documento de Internet disponible en http://cvc.cervantes.es/obref/congresos/sevilla/tecnologias/ponenc_carbonell.htm] Fecha de consulta 25 de noviembre de 2009.

9. Chantal M. (2002), "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento", Universidad de Málaga. [Documento de Internet disponible en http://elies.rediris.es/elies18/index.html] Fecha de consulta 31 de marzo de 2008.

10. Christel, M. G., (2006), *"Windowing Time in Digital Libraries"*, en International Conference on Digital Libraries JCDL'06, June 11–15, Chapel Hill, NC, USA.

11. Crane, G., Smith D. A., Wulfman, C. (2001), *"Building a hypertextual digital library in the humanities: a case study on London"*, en *Proceedings of the First* ACM+IEEE JCDL'01, June 24-28, Roanoke.

12. Crane, G., Jones, A., (2006), *"The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection"*, en International Conference on Digital Libraries JCDL'06, June 11–15, Chapel Hill, NC, EEUU.

13. Crane, G., Wulfman, C. (2003), *"Towards a Cultural Heritage Digital Library"*, en *Proceedings of the 2003 Joint Conference on Digital Libraries* (JCDL'03), pag. 75- 86, Houston, Texas USA.

14. Deitel, P. y Harvey M. (2004), "Como programar en Java", 5° edición, Ed. Pearson Educación, México.

15. Davis, P., Elkson, D., Klavans, J. (2003), "*Methods for Precise Named Entity Matching in Digital Collections*", en *Proceedings of the 2003 Joint Conference on Digital Libraries* (JCDL'03), pág. 125 – 127, Houston. USA.

16. Dale, R. y Mazur, P. (2007), "*Handling Conjunctions in Named Entities*", en CICLing 2007, A. Gelbukh (editor), pág. 131–142, Springer-Verlag Berlin Heidelberg.

17. DRAE (2003), "Diccionario de la lengua española", CD-ROM. Real Academia Española, 22° edición. Madrid.

18. Eclipse (2009), Eclipse, en http://www.eclipse.org/ Fecha de consulta 31 de marzo de 2009.

19. Florian, R., Hassan, H., Ittycheriah, A., Jing, H., Kambhatla, N., Luo, X., Nicolov, N., Roukos, S. (2004), "*A Statistical Model for Multilingual Entity Detection and Tracking*", en *Proceedings of the seventh conference on Natural language learning at* HLT-NAACL 2004. Pag. 1-8. Boston, Massachusetts, EEUU.

20. Gelbukh, A y Sidorov, G. (2006). "Procesamiento automático del español con enfoque en recursos léxicos grandes". Ciudad de México, Instituto Politécnico Nacional, pp. 240.

21. JAVA (2009), JAVA *Sun Microsystems,* en http://www.sun.com/ Fecha de consulta 31 de marzo de 2009.

22. *JDK Documentation* (2009), JAVA *Sun Microsystems,* en http://download.java.net/jdk7/docs/ Fecha de consulta 16 de abril de 2009.

23. Jiménez, M. (2001), *"Reconocimiento y generación de entidades semánticas"*, en *Proceedings of the MT Summit VIII*, Enero.

24. Jones, D. y Thompson, C. (2003) "*Identifying Events using Similarity and Context*", en Proceedings *of the seventh conference on Natural language learning at* HLT-NAACL 2003, Volumen 4, pág 135 – 141, Edmonton, Canada.

25. Jurasfky D. y Martin J. (2000) *"Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition".* Ed. Prentice-Hall.

26. Kalashnikov, D.V., Zhaoqi C., Mehrotra, S., Nuray-Turan, R. (2008) "*Disambiguation Algorithm for People Search on the Web*", en *IEEE Transactions on Knowledge and Data Engineering,* Vol. 20, pág. 1550-1565.

27. Kumaran, G., Allen, J. "*Using Names and Topics for New Event Detection*" (2005). *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. ACM. Vancouver, Canada

28. LDC Linguistic Data Consortium (2000), TREC Spanish Corpus, CD-ROM, Universidad de Pensilvania.

29. Ladrón M., (1996), *Manual de Paleografía*. Santafé de Bogotá. Centro Editorial Javeriano CEJA.

30. Lavid, J. (2005), "*Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*". Ed. Cátedra (Colección de Lingüística). Madrid.

31. Llidó, D., Berlanga, R., Aramburu, M.J. (2001), "*Extracting Temporal References to Assign Document Event-Time Periods*". *Lecture Notes In Computer Science*, LNCS 2113, pág. 62–71, 2001. Springer-Verlag Berlin Heidelberg.

32. Leidner, J.L., Sinclair, G., Webber, B., (2003), *"Grounding spatial named entities for information extraction and question answering*" en A. Kornai y B. Sundheim editors, HLT/NAACL'03 *Workshop: Analysis of Geographic References, Association for Computational Linguistics* 2003, pág. 31-38, Edmonton, Alberta, Canadá. 2003. Disponible en Internet (http://acl.ldc.upenn.edu/W/W03/W03-0105.pdf)

33. Leidner. J.L. (2004a), *"Towards a reference corpus for automatic toponym resolution evaluation"* en *Workshop on Geographic Information Retrieval held at SIGIR*, Sheffield, UK.

34. Leidner. J.L. (2004b), "*Toponym Resolution in Text: "Which Sheffield is it*?", en Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR 2004), Sheffield, UK.

35. Leidner, J.L. (2007), "*Toponym Resolution in Text (Annotation, evaluation and Applications of Spatial Grounding)"* en ACM SIGIR Forum, Vol. 41 No. 2 December 2007. Amsterdam, Holanda.

36. Kumaran, G., Allan, J.,(2005) "*Using Names and Topics for New Event Detection"* en *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pág. 121–128. Vancouver, British Columbia, Canada,

37. Manning, C., Schutze, H. (1999), *"Foundations of Statistical Natural Language Processing"*. 6° Edición. MIT Press.

38. Mani, I., Wilson, G. (2000), "*Robust temporal processing of news*", en Proceedings of the 38th *Annual Meeting on Association for Computational Linguistics* (ACL), pag. 69 – 76, Hong Kong.

39. Mckay, D., (2000), "*Mining dates from historical documents"*, en Technical report, Department of Computer Science, University of Waikato, 2000. Disponible en Internet http://www.cosc.canterbury.ac.nz/open/NZCSPG/papers/dmm9@cs.waikato.ac.nz.pdf. Fecha de consulta 4 septiembre de 2008.

40. Moliner, M. (2001), "*Diccionario de uso Español*", CD-ROM, 2° edición. Madrid.

41. Muller P., Tannier X,. (2004), "*Annotating and measuring temporal relations in texts*", en *Proceedings of the 20th international conference on Computational Linguistics*, Genova, Suiza.

42. Muñoz, R., Montoyo, A., Llopis, F., Suárez, A. (1998), "*Reconocimiento de entidades en el sistema EXIT"*, en Procesamiento del Lenguaje Natural 23, 47-53.

43. Murray, R. L. (2005), *"Toward a metadata standard for digitized historical newspapers"*, en International Conference on Digital Libraries Proceedings of the 5th ACM/IEEE-CS JCDL'05, Junio 7-11, Denver, EEUU.

44. NIST (2005), "*What is Information Extraction?*" En http://www.itl.nist.gov/iaui/894.02/related_projects/muc. Fecha de consulta 1 de marzo de 2009.

45. Orduña, J., (1999), "*La función definitoria de los ejemplos: a propósito del léxico filosófico del Diccionario de Autoridades*", en Así son los Diccionarios. Eds. Vila N., Calero M., Mateu R., Casanovas M, Orduña J. pág. 99-120, Lleida, España.

46. On, B.-W., Elmacioglu, E., Lee, D. (2006), *"An effective approach to entity resolution problem using quasi-clique and its application to digital libraries"* en Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL'06. June, pag. 11-15.

47. Petras, V. Larson, R. R. Buckland, M. (2006), *"Time Period Directories: A Metadata Infrastructure for Placing Events in Temporal and Geographic Context"*, en *International Conference on Digital Libraries* JCDL '06 June, pag. 11–15, Chapel Hill, EE.UU.

48. Pons, A., Berlanga, R. Llidó, D., (2001), *"Técnicas de agrupamiento semántico-temporal para la identificación de sucesos en bases de noticias digitales"*, en IX Conferencia de la Asociación Española para la Inteligencia Artificial, pág. 603-612, Gijón, España.

49. Ponce, I., Zárate, J., Olivares, J. (2006), "*Recuperación de información de páginas web mediante una ontología que es poblada usando clasificación automática de textos*", en: IEEE *Looking Forward*, IEEE *Computer Society*, Vol. 13.

50. Project Gutemberg (2009), en http://www.gutenberg.org/wiki/Main_Page Fecha de consulta 28 de julio de 2009.

51. Rydberg-Cox, J.A., Mahoney, A., Crane, G., (2001), "*Document Quality Indicators and Corpus Editions*", en *International Conference on Digital Libraries* JCDL '01, June, pag. 24–15, Roanoke, Virginia, USA.

52. San Vicente, E., Taibo, C., Beltrán, M., Merino, A., Ballús P. (1991), "*Lengua y Literatura*", Áreas consultor didáctico. Ed. Planeta, Madrid.

53. Schilder, F., Katz, G., Pustejovsky, J., (2007), *"Annotating, Extracting and Reasoning About Time and Events*". F. Schilder et al. (Eds.), Reasoning about Time and Events, LNAI 4795, pag. 1–6. Springer-Verlag Berlin Heidelberg 2007

54. Setzer, A. y Gaizauskas, R. (2002), *"On the importance of annotating temporal event-event relations in text"*, en LREC 2002 Workshop on Annotation Standards for Temporal Information in Natural Language, Las Palmas, Gran Canaria, España.

55. Smith, D.A. y Crane, G. (2001), *"Disambiguating geographic names in a historical digital library"*, en Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL, Darmstadt, Germany, September 4-9, pág. 127–136.

56. Smith, D.A., (2002a), "*Detecting and Browsing Events in Unstructured text*", en Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pág. 73-80, Tampere, Finland.

57. Smith, D.A., (2002b), *"Detecting Events with Date and Place Information in Unstructured Text*", en Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries JCDL'02, July 2002, pág. 191-196, Portland, USA.

58. Smith, D.A. y Mann, G. S. (2003), *"Bootstrapping toponym classifiers"*, en HLT-NAACL 2003 Workshop: Analysis of Geographic References, Eds. A. Kornai and B. Sundheim, pág. 45–49.

59. Tannier X. Muller P., (2008), "*Annotating and measuring temporal relations in texts*", en Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Mayo 2008, Marrakech, Morocco.

60. TEI (2009), TEI: *Text Encoding Iniciative*, en http://www.tei-c.org/index.xml_ Fecha de consulta 10 de febrero de 2009.

61. Tercedor, M. (1999), "*La fraseología en el lenguaje biomédico: análisis desde las necesidades del traductor*", en Departamento de Traducción e Interpretación Universidad de Granada. [Documento de Internet disponible en http://elies.rediris.es/elies6/index.html#indice]

62. Torruela, J. y Llisterri, J, (1999), "*Diseño de corpus textuales y orales*", en Departamento de Filología Española. Universidad Autónoma de Barcelona. [Documento de Internet disponible en http://liceu.uab.es/~joaquim/publicacions/Torruela_Llisterri_99.pdf]

63. Wikcionario (2008), Categorías de Verbos Regulares en http://es.wiktionary.org/wiki/Categor%C3%ADa:ES:Verbos_regulares. Fecha de consulta 2 de diciembre de 2008.