Filología y Lingüística 39 (2): 143-171, 2013
ISSN: 0377-628X

EL *TREEBANK* DEL ESPAÑOL "IPROCOLDI": COMPONENTE ANOTADO DEL CORPUS CODIMEP-CR

Carla Victoria Jara Murillo

RESUMEN

En este artículo se describe el proceso que se siguió para crear el componente anotado con información lingüística (*treebank*) del Corpus de Mensajes Presidenciales Costarricenses (CODIMEP-CR), en el marco del proyecto No. 745-B1-244 Interfaz para el procesamiento de corpus lingüísticos digitales – IPROCOLDI. Ambos corpus se albergan en la interfaz IPROCOLDI (http://163.178.116.145/iprocoldi/).
Palabras clave: corpus anotado morfosintácticamente, *treebank* de español, lingüística de corpus, sintaxis española, PLN.

ABSTRACT

This paper describes the process followed in order to create a Spanish treebank in the framework of the research project No. 745-B1-244 Interfaz para el procesamiento de corpus lingüísticos digitales – IPROCOLDI (Interface for the processing of digital language corpora). The data for the treebank was extracted from the Corpus de Mensajes Presidenciales Costarricenses (CODIMEP-CR). The interface and the treebank are located at http://163.178.116.145/iprocoldi/.
Key words: treebanks, Spanish treebank, corpus linguistics, Spanish syntax, NLP.

1. Introducción

El proyecto IPROCOLDI nace como una inquietud por desarrollar en la Universidad de Costa Rica tanto la lingüística de corpus como herramientas computacionales en el marco del Procesamiento de Lenguaje Natural (PLN). En esa línea, los investigadores (Carla Victoria Jara Murillo, investigadora principal, y Antonio Leoni de León, investigador asociado) buscábamos desarrollar una interfaz que albergara conjuntos organizados de datos lingüísticos digitalizados (corpus) y algunas herramientas informáticas para la explotación de esos corpus.

Dra. Carla Victoria Jara Murillo. Profesora del Departamento de Lingüística de la Universidad de Costa Rica. Correo electrónico: carla.jara@ucr.ac.cr

Recepción: 23- 04- 2013
Aceptación: 14- 06- 2013

Otro de los objetivos principales era iniciar el desarrollo de un *treebank* del español a partir de un corpus del español escrito integrado por textos costarricenses. Para ello, se partió del CODIMEP-CR: Corpus de Mensajes Presidenciales Costarricenses (Jara Murillo 2011), un corpus elaborado como parte de un proyecto anterior en lingüística de corpus.

Para empezar a desarrollar nuestro *treebank*, partimos de la utilización de FIPS, una herramienta implementada en el Laboratoire d'Analyse et de Technologie du Langage (LATL), de la Universidad de Ginebra[1]. De acuerdo con Leoni de León, Schwab y Wehrli (2008: 37), FIPS "se inspira, fundamentalmente, del esquema teórico chomskyano (Chomsky, 1995, capítulo 1 con Howard Lasnik), con adaptaciones libres del modelo Minimalista (Chomsky, 2004), de Simpler Syntax (Culicover y Jackendoff, 2005) y de la Gramática *léxico-funcional* (Bresnan, 2001)" [2].

El primer aspecto que se abordó fue el tipo de información lingüística que se iba a anotar; en este sentido, decidimos incluir información léxica, morfológica y sintáctica. Para la anotación sintáctica se partió inicialmente del análisis obtenido del *parser* de FIPS, con base en el cual se postuló el formalismo IML (Iprocoldi Marked Language), que se detalla en la Tabla 1. En cuanto a la información léxica (lematización) y morfológica se decidió incorporar el etiquetado propuesto por EAGLES (Expert Advisory Group on Language Engineering Standards), por ser este un estándar cada vez más generalizado[3]. En el proceso de etiquetamiento morfológico (*tagging*) se tomó como base el *PoS tagger* de FreeLing (Open Source Suite of Language Analyzers), de la Universitat Politecnica de Catalunya (Padró and Stanilovsky 2012)[4].

El formalismo IML sigue básicamente el de FIPS y consta del siguiente etiquetado para la anotación sintáctica:

Tabla 1. Etiquetado de IML

Etiqueta IML	**Constituyente**
ST	Sintagma Temporal (nodo superior correspondiente a la oración)
SF	Sintagma Funcional (corresponde a SAs complementarios de sustantivos o verbos copulativos: *es <u>suficiente</u>, está <u>segura</u>, una institución <u>renovada</u>*)
SCo	Sintagma Conjuntivo
SC	Sintagma Complementizador
SD	Sintagma Determinante
SN	Sintagma Nominal
SA	Sintagma Adjetival
SV	Sintagma Verbal
SAd	Sintagma Adverbial
SP	Sintagma Preposicional

2. Definición y tipos de *treebanks*

De acuerdo con Schütze (1999), los *treebanks* son colecciones de ejemplos de análisis sintácticos de oraciones que sirven para construir analizadores estadísticos y distintos grupos

de investigación los han desarrollado en el marco del Procesamiento del Lenguaje Natural (PLN). En español, en comparación con lenguas como el inglés y el alemán, son relativamente pocos los *treebanks* que se han desarrollado y de ahí nuestro interés por impulsar este tipo de trabajo. De acuerdo con Navarro Colorado, "[...] los corpus en español no se han desarrollado para su utilización en PLN, y no se han creado métodos de anotación claros y consistentes, perfectamente enfocados a la explotación final del corpus en PLN" (2007: 21).

Según Sampson (2003), parece haber sido Geoffrey Leech quien acuñó el término *treebank* para referirse a un corpus anotado morfosintácticamente, en particular por el hecho de que una manera muy común de representar la estructura sintáctica es por medio de diagramas arbóreos. Sin embargo, en la actualidad un *treebank* no necesariamente se representa por medio de árboles (Nivre 2008). Las prácticas en este sentido se han diversificado, pero siempre el esquema de anotación estará determinado por la teoría sintáctica con base en la cual se realiza el análisis. Es común en la actualidad seguir un modelo lo más general posible o bien, modelos eclécticos, que toman aspectos de distintas teorías.

El *treebank* más conocido y que ha servido de base para otros proyectos en esta línea es el Penn Treebank. En los ejemplos del Penn Treebank, "trees are represented in a straightforward (Lisp) notation via bracketing. The grouping of words into phrases is fairly flat [...], but the major types of phrases recognized in contemporary syntax are fairly faithfully represented" (Schütze 1999: 412).

En cuanto a la representación, en la actualidad se opta o bien por la anotación de estructura sintagmática (como el Penn Treebank y el ICE-GB) o bien por la anotación de estructura de dependencias (ejemplos son el Prague Dependecy Treebank y el Quranic Arabic Dependency Treebank). Nosotros seguimos la representación sintagmática utilizando para ello los corchetes etiquetados, como muestra el siguiente ejemplo:

Hoy vemos los frutos de una obra de gobierno.[5]

[ST[SAd Hoy][SD] vemos [SV [SD los [SN frutos]][SP de [SD una [SN obra [SP de [SN gobierno]]]]]]]

Los archivos del *treebank* IPROCOLDI se encuentran en texto plano y están codificados en UTF-8. Este formato tiene las ventajas de la economía de recursos (archivos más pequeños y legibles) y la posibilidad de verlos y utilizarlos sin necesidad de herramientas informáticas especializadas. Cuando el análisis sintáctico se hace más complejo y detallado conviene utilizar otros formatos, por ejemplo el lenguaje de marcado XML, que permite mostrar la estructura jerárquica del análisis mediante indentaciones sucesivas y otros recursos estandarizados. Un sitio útil para revisar distintas propuestas de anotación es Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM)[6], de la Universidad de Leeds.

Todas las instituciones que han emprendido la tarea de crear un *treebank*, señalan como las principales dificultades sus costos y el tiempo que debe dedicarse a su construcción; por ello cada vez se busca más avanzar en el diseño de herramientas automatizadas que ayuden en las labores de codificación. Así por ejemplo, los creadores del Proyecto 3LB, uno de los pocos que se han propuesto para la creación de un *treebank* en lengua española (además de catalán y euskera), señalan: "A pesar de que la construcción de un *treebank* es una tarea costosa, creemos que es una labor imprescindible para el desarrollo de aplicaciones reales en el área del Procesamiento del Lenguaje Natural (PLN) y como tal para el desarrollo de la sociedad de la información"[7].

3. Utilidad de los *treebanks*

El análisis y la explotación de corpus anotados es un componente fundamental de la llamada "lingüística de corpus", en la cual se utilizan los *treebanks* para estudiar diversos fenómenos léxicos, morfológicos y sintácticos. Por otra parte, en lingüística computacional sirven para desarrollar, probar o entrenar analizadores automáticos o semiautomáticos. Aplicados a corpus diacrónicos es posible estudiar más productivamente el cambio lingüístico. Además, los *treebanks* permiten evaluar la teoría sintáctica formal que en décadas pasadas se basaba en el método introspectivo tradicional del generativismo. La introspección es un método de evaluación de la gramática muy limitado, ya que se fundamenta solamente en la intuición del lingüista; en cambio, un *treebank* puede dar cuenta de la frecuencia de determinadas estructuras (qué tan comunes son en el uso real) y de su cobertura (qué fenómenos sintácticos nuevos emergen y para cubrir cuáles necesidades comunicativas).

Los *treebanks* creados mediante *parsers* (analizadores automáticos) sirven también para evaluar el grado de eficiencia de la herramienta misma; así, una vez que los resultados hayan sido revisados por humanos pueden proponerse nuevas reglas para mejorarlos. Uno de los objetivos de nuestro *treebank* es servir para la evaluación del *parser* utilizado observando los ámbitos gramaticales en los que el analizador automático dio errores. Así, la revisión manual del *treebank* por humanos va a redundar en la creación de las reglas sintácticas que no han sido incorporadas o resultaron defectuosas en la herramienta automática.

Desde la publicación del Penn Treebank, primer corpus anotado a gran escala, se han desarrollado *treebanks* para una cantidad relativamente apreciable de lenguas que aumenta conforme se desarrollan más herramientas y se destinan más fondos en los centros de investigación lingüística a construirlos. Así, se han desarrollado herramientas informáticas para la exploración y explotación de corpus lingüísticos, en particular herramientas de búsqueda (*search tools*). Las interfaces desarrolladas para estas tareas varían en su grado de sofisticación, desde los sistemas de búsqueda orientados a la programación computacional hasta los ambientes de exploración orientados a los lingüistas generales, con poca o ninguna formación en lingüística computacional o PLN.

4. *Treebanks* de lenguas del mundo

El artículo "treebank" en http://en.wikipedia.org/wiki/Treebank contiene direcciones web de *treebanks* para las siguientes lenguas: alemán, árabe, búlgaro, catalán, checo, chino, coreano, croata, danés, esloveno, español, estoniano, finés, francés moderno, francés histórico, griego moderno, griego antiguo, hebreo, hindi, holandés, húngaro, inglés moderno, inglés histórico, islandés, italiano, japonés, latín, noruego, persa, polaco, portugués moderno, portugués histórico, rumano, ruso, sueco, tailandés, turco, urdu y vietnamita[8].

Como es de esperar, la lengua más ampliamente representada es el inglés, que cuenta, además del ya clásico Penn Treebank, con los siguientes: Prague English Dependency Treebank; International Corpus of English (ICE); British Component of the International Corpus of English (ICE-GB); Diachronic Corpus of Present-Day Spoken English (DCPSE); Lancaster Parsed Corpus; Susanne Corpus / Christine Corpus / Lucy Corpus; Proyecto Verbmobil: Tübingen Treebank of Spoken English (TüBa-E/S), LinGO Redwoods Treebank; Proyecto AMALGAM: Multi-Treebank; Proyecto TalkBank: CHILDES Brown Eve Corpus with Dependency Annotation, SMULTRON: Parallel Treebank EN-DE-SV (inglés-alemán-

sueco). Además cuenta, para el inglés histórico, con el Penn Corpora of Historical English y el York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE). Las lenguas que le siguen al inglés son el alemán (Negr@ Corpus, TIGER Corpus, Tübingen Treebank of Spoken German (TüBa-D/S) y SMULTRON) y el italiano (Turin University Treebank (TUT), Venice Italian Treebank (VIT), Italian Syntactic-Semantic Treebank (ISST) y Siena University Treebank (SUT)). Para el español, se citan dos *treebanks*: UAM Spanish Tree Bank (de la Universidad Autónoma de Madrid) y CAST3LB (parte del proyecto 3LB, de varias universidades españolas); ambos se describen en el siguiente apartado.

5. Principales proyectos de corpus y *treebanks* del español

La recopilación bibliográfica que mantiene Joaquim Llisterri, de la Universitat Autònoma de Barcelona, en su sitio web "Corpus Linguistics and Written Language Resources - Bibliography" (Llisterri 2012) demuestra que el desarrollo de *treebanks* en español, si bien tuvo un inicio entusiasta en la década de 1990, ha mermado su avance en la actualidad. La bibliografía recopilada en ese sitio muestra que la investigación más reciente en este campo no supera el año 2002, a pesar de que el sitio está actualizado al 2012.

Según mi propia investigación sobre los proyectos en corpus y *treebanks* que se han realizado sobre el español (escrito), se concluye que los recursos disponibles por internet más completos en este campo son:

5.1. Base de Datos Sintácticos del Español Actual (BDS)

Se construyó con base en la parte contemporánea del Corpus ARTHUS: Archivo de Textos Hispánicos, de la Universidad de Santiago de Compostela. Es un proyecto del 2001 ubicado en el sitio http://www.bds.usc.es/, al parecer concluido, ya que su última actualización es de octubre de 2001.

5.2. Los corpus de la Real Academia Española (RAE)

El Corpus de Referencia del Español Actual (CREA) contiene en su parte escrita 5500 textos y unos 154 millones palabras, y en su parte oral, 9 millones de formas procedentes de trascripciones de la lengua hablada, con más de 1600 documentos. El Corpus Diacrónico del Español (CORDE) contiene 250 millones de palabras, de 1975 hacia atrás. Además se encuentra en proceso de construcción el Corpus del Español del Siglo XXI, en el que colaboran con la RAE las 21 asociaciones de la lengua española. El sitio de la RAE tiene una interfaz de búsqueda en la dirección http://corpus.rae.es/creanet.html. El banco de datos se describe así:

> A través de la aplicación de concordancias, los investigadores tienen a su disposición alrededor de 400 millones de formas de todos los períodos del español, tanto de España como de América, lo que constituye, sin duda, el recurso más importante del que se haya podido disponer jamás para el estudio de esta lengua.
>
> La nómina de autores y obras muestra las estadísticas generales del banco de datos, lo que permite valorar en todo su alcance los resultados obtenidos en la consulta a ambos corpus. Además, a través de esa aplicación específica, obtenemos los datos reales sobre los que se realiza cada consulta, bien sea de carácter general, bien combinando distintos criterios de selección, para obtener los datos estadísticos de la consulta realizada: número total de textos y su distribución geográfica, cronológica o temática. Se muestra también la referencia bibliográfica completa de cada uno de los textos, el criterio que ha servido para su clasificación cronológica en el banco de datos, la clasificación temática y el número de palabras de cada texto. (2012)[9]

5.3. Corpus del Español de Mark Davis (Brigham Young University)

Se encuentra en http://www.corpusdelespanol.org/; contiene textos en español desde el siglo XIII hasta el XX. Alcanza los 100 millones de palabras y es parte de un amplio proyecto que incluye varias lenguas. Presenta una de la interfaces de búsqueda más completas, en: http://www.corpusdelespanol.org/x.asp.

5.4. Corpuseye de VISL

El CORPUSEYE (Corpus Search Interface) fue creado por VISL (Visual Interactive Syntax Learning http://beta.visl.sdu.dk/) en la University of Southern Denmark. Contiene una interfaz para el español y extrae concordancias de los siguientes corpus anotados: ECI-ES2 (periódicos), Europarl-es (debates parlamentarios) y Wikipedia-es.

5.5. SFN - Spanish Framenet Corpus / Corpus del Español Actual (CEA)

Este corpus, ubicado en http://sfn.uab.es:9080/SFN/tools/cea/spanish, está lematizado y anotado morfológicamente, de acuerdo con un etiquetario propio. Tiene una interfaz de búsqueda basada en el IMS Open Corpus Workbench (CWB). La siguiente descripción del CEA está tomada de su sitio:

> Características: El Corpus del Español Actual (CEA) tiene 540 millones de palabras y está lematizado y etiquetado con información morfológica y/o categorial. El CEA está integrado por los siguientes textos: la parte española del corpus paralelo español-inglés Europarl: European Parliament Proceedings Parallel Corpus v. 6 (1996-2010), el módulo en lengua española del Wikicorpus v. 1.0, que contiene una parte importante de la Wikipedia (2006), y la sección en español del MultiUN: Multilingual UN Parallel Text 2000-2009, un corpus integrado por resoluciones de la Organización de las Naciones Unidas (ONU). (Subirats y Ortega 2012)

5.6. Corpus Textual Especializado Plurilingüe de IULA

Se ubica en http://www.iula.upf.edu/corpus/corpuses.htm y es desarrollado por el Instituto Universitario de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra. Contiene textos en catalán, castellano, inglés, francés y alemán que versan sobre sobre economía, derecho, medio ambiente, medicina e informática. Los textos se han marcado de acuerdo con el estándar SGML y sigue las directrices del Corpus Encoding Standard (CES) de EAGLES. Los textos están anotados con los etiquetarios morfosintácticos diseñados en el IULA. El proyecto es dirigido por M. Teresa Cabré y coordinado por Jordi Vivaldi.

5.7. El *SPANISH TREEBANK* de la Universidad Autónoma de Madrid

Desarrollado por el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid, se ubica en http://www.lllf.uam.es/ESP/Treebank.html[10], de donde se toma la siguiente descripción:

> Este proyecto se inició en diciembre de 1997, y para septiembre de 1999 el corpus contaba con 1.500 oraciones extraídas de periódicos (*El País Digital* y *Compra Maestra*) y anotadas sintácticamente. En este tiempo, se ha desarrollado la guía de anotación y herramientas para anotar y depurar. En la fase actual, se continúa la anotación manual con la ayuda de anotadores humanos y herramientas mejoradas. El objetivo de esta fase es conseguir 5.000 oraciones anotadas. También se han iniciado experimentos sobre el corpus. El trabajo futuro está orientado a la construcción semiautomática del corpus, basada en una gramática inferida del treebank.

5.8. El Corpus CESS-ECE y el Proyecto CAST3LB (ANCORA)

El Proyecto **CESS-ECE** (http://clic.ub.edu/cessece/index.php) tuvo como objetivo crear tres corpus: español (CESS-ESP), catalán (CESS-CAT) y euskera (CESS-EUS). CESS-ESP contiene 500.000 palabras y está anotado sintácticamente (con constituyentes y funciones) y semánticamente, con los sentidos nominales de WordNet. Utiliza dos herramientas de anotación: AGTK (Universidad de Pensilvania) y 3LB-SAT, una herramienta específica para la anotación semántica con sentidos de WordNet. Como resultado del proyecto se dispone de la guía de anotación sintáctica utilizada para el catalán y el español y el *treebank* AnCora. La relación entre estos proyectos se describe en Göhring:

> AnCora is a project at the Centre de Llenguatge i Computati_o (CLiC) from the University of Barcelona that develops in collaboration with TALP46 a \multilevel annotated corpora for Catalan and Spanish" (Taul_e et al. 2008). These AnCora-Ca and AnCora-Es corpora are built upon corpora from previous projects (3LB and CESS-ECE). Each corpus currently contains half a million words annotated at diferent linguistic levels. They are the largest multilayer annotated corpora freely available for these languages at the time of writing. The corpora include texts from various sources, mostly newspapers and news agencies; the Spanish corpus also contains a subset of the balanced LexEsp corpus. About 40% of each corpus stem from the Catalan and Spanish versions of El Periódico and collect the same news. (2009: 29)

El corpus CLiC-TALP para el español consta actualmente de 100.000 palabras anotadas manualmente a nivel morfosintáctico. El resto del corpus, hasta 5,5 millones de palabras, está anotado a nivel morfosintáctico de forma automática, con una tasa de error del orden de un 3%.

6. El Corpus Anotado de Español Costarricense (*TREEBANK* IPROCOLDI)

6.1. Aspectos preliminares

De acuerdo con Baker et al. (2006), la compilación de un corpus implica cinco etapas: el diseño del corpus, la planificación del sistema de almacenamiento, la obtención de permisos legales, la recopilación de los textos y finalmente la etapa de codificación. En esta última etapa suele distinguirse entre dos tipos de procesos: la inclusión de información acerca de los textos, por ejemplo mediante un encabezamiento tipo TEI, que especifica información o metadatos sobre la producción del texto (autor, fecha, datos de edición, participantes, etc.); a este proceso a veces se le identifica como **codificación** propiamente. El otro tipo de proceso es la inclusión de información lingüística, al que más específicamente se le denomina **anotación**. Un *treebank* es un segmento de un corpus que ha sido enriquecido con información lingüística suplementaria.

Para la construcción de un *treebank*, a su vez, se siguen varias etapas. Así por ejemplo, Göhring (2009: 34-5), en el proceso de construcción de la porción española del corpus paralelo SMULTRON[11], divide su trabajo en tres etapas: preprocesamiento, que incluye selección del corpus, del formato de anotación y de las herramientas de anotación; anotación; y posprocesamiento, que consiste en la revisión y mejoramiento de la anotación. Siguiendo de manera general ese mismo esquema, en los siguientes apartados se detallan los procesos que se siguieron en la construcción de nuestro *treebank*, el cual se alberga en la interfaz IPROCOLDI, en la dirección http://163.178.116.145/iprocoldi/.

El *TREEBANK* IPROCOLDI se creó en el marco del proyecto **IPROCOLDI: Interfaz para el procesamiento de corpus lingüísticos digitales**, financiado por la Universidad de Costa Rica. Se logró construir un repositorio de corpus digitales, el cual cuenta en este

momento con el corpus CODIMEP-CR (Corpus de Mensajes Presidenciales Costarricenses) y previstas para albergar corpus futuros.

6.2. Selección del corpus

Del CODIMEP-CR, corpus que se ubica en el sitio "Mensajes Presidenciales de Costa Rica" (https://sites.google.com/site/mensajepresidencialcr/) creado por la investigadora principal como parte de los objetivos de su proyecto "Preparación de un corpus digital de mensajes presidenciales costarricenses" (No. 745-A9-128, concluido), se trasladó a IPROCOLDI un conjunto de 244 documentos, de los 338 que componen el CODIMEP-CR. Los 94 documentos que se albergan en el sitio pero no se trasladaron corresponden a las transcripciones originales de los mensajes presidenciales del siglo XIX; es decir, de este siglo se trasladaron solamente las versiones modernizadas de esos mensajes. Así, el CORPUS CODIMEP-CR/IPROCOLDI consta de 244 archivos de texto plano codificado en UTF-8. De esta manera los documentos podrán ser procesados mediante herramientas de explotación de corpus, mientras que los que se albergan en el sitio Mensajes Presidenciales de Costa Rica se encuentran en formato PDF, por contar el sitio con un público meta mucho más general que IPROCOLDI.

Decidimos utilizar el CODIMEP-CR para la creación de *treebank* por varias razones: en primer lugar disponíamos del corpus crudo (sin anotaciones) en formato PDF, de manera que el corpus completo se podía convertir de manera relativamente rápida a formato de texto. En segundo lugar, consideramos los siguientes aspectos lingüísticos:

a. El género discursivo del mensaje presidencial resulta apropiado para iniciar un *treebank* porque su diversidad tanto sintáctica como léxica es muy elevada, lo que permitía recopilar la variedad de estructuras que queríamos anotar.
b. El género además contiene diversas modalidades discursivas (descripción, narración, argumentación, etc.) y diversos modos oracionales (declarativo, imperativo, interrogativo, exclamativo, etc.).
c. Además, si bien es muy variado en todos estos aspectos lingüísticos, el registro es siempre muy formal, a menudo solemne, y sumamente cuidado en su forma, por lo cual igualmente servía nuestro propósito de mantener una homogeneidad diastrática y sobre todo diafásica.
d. Finalmente, queríamos utilizar textos costarricenses, aunque por el hecho de ser registro formal escrito, no presenta particularidades específicas del español de Costa Rica.

Otra de las razones es la conveniencia de que al tratarse de documentos públicos, no se requiere de trámites legales para su utilización y distribución a través de internet.

Por otra parte, se utilizaron los tres componentes del CODIMEP-CR: SIGLO XIX, SIGLO XX y SIGLO XXI y se mantuvo esta segmentación con miras a determinar, mediante futuros análisis, si hay alguna variación diacrónica de relevancia. De hecho, desde las etapas iniciales en la selección de las oraciones, lo primero que se hizo evidente fue la variación en el estilo sintáctico desde el siglo XIX hasta el XXI. En los documentos del siglo XIX, todas las oraciones tienden a ser extremadamente largas y, por ende, complejas, al punto de coincidir a menudo con párrafos completos. Como se verá en las siguientes secciones, este estilo sintáctico determinó el tamaño de los segmentos cronológicos del *treebank* identificados como SIGLOS.

6.2.1. Extracción del COREX (Corpus ORacional EXtendido) a partir del CODIMEP-CR

El COREX es un corpus compuesto por ca. 17 mil oraciones con una extensión máxima de 30 palabras. La Tabla 2 muestra el resumen de los datos:

Tabla 2. Composición del COREX – Corpus Oracional Extendido

CORPUS	**TAMAÑO**	**%**
CODIMEP-CR	7.83 MB (8.213.736 bytes) (1.288.293 palabras)	100
COREX	1.75 MB (1.836.458 bytes) (17.004 oraciones. Desglose: Siglo XIX: 682; Siglo XX: 14.151; Siglo XXI: 2.171)	22,3

La extracción de este primer corpus oracional se hizo manualmente con el fin de verificar que las oraciones cumplieran con los criterios que nos habíamos puesto inicialmente de que presentaran suficiente diversidad estructural como para ilustrar la mayor cantidad posible de tipos de oración en español (en registro formal escrito). El COREX completo significa una fuente de material invaluable para el futuro desarrollo de herramientas para el análisis sintáctico.

Los datos muestran que el COREX constituye el 22.3% del tamaño total del CODIMEP-CR. Al iniciar los análisis sintácticos con FIPS, pronto nos percatamos de que el porcentaje de error al someter oraciones tan largas al *parser* era de alrededor del 50%, por lo cual para reducir la longitud de las oraciones tomamos el criterio utilizado por los creadores del *Spanish Treebank* de la Universidad Autónoma de Madrid (Moreno, López y Sánchez 2003), de oraciones de un promedio de 15 palabras y lo readecuamos a un máximo de 15 palabras.

6.2.2. Extracción del COR (Corpus ORacional) a partir del COREX

En esta segunda extracción, se utilizó una expresión regular (^(\S+\s){1,14}\S+[.?!]$) para separar las oraciones de un máximo de 15 palabras de aquellas de mayor extensión. El COR así resultó constituido por un subconjunto del COREX que consta de 6.761 oraciones. El desglose de ambos conjuntos, en términos de siglos, es el siguiente:

Tabla 3. Composición del COR a partir del COREX

CORPUS	**TAMAÑO**	**DESGLOSE**
COREX	17.004 oraciones	Siglo XIX: 682 Siglo XX: 14.151 Siglo XXI: 2.171
COR	6.761 oraciones	Siglo XIX: 166 Siglo XX: 5.770 Siglo XXI: 825

Como señalé más arriba con respecto al estilo sintáctico del siglo XIX, puede apreciarse en la Tabla 3 que una vez aplicado el criterio de longitud de 15 palabras, el COREX de ese primer siglo, que contenía 682 oraciones, se redujo en el COR a 166; esto significa que casi un 75% del COREX fue separado del *treebank* porque las oraciones del siglo XIX tenían una longitud mayor a 15 palabras. En cambio, en la extracción del COR de los dos siglos siguientes, se separó una cantidad mucho menor de oraciones, alrededor del 66%.

6.2.3. Revisión manual del COR

El conjunto de oraciones del COR fue revisado con el fin de depurar el corpus de los errores que pudieran haber sido obviados en la extracción. Además, por medio de la herramienta de etiquetación morfológica de FreeLing[12], un paquete de herramientas de uso libre para el análisis lingüístico (Padró and Stanilovsky 2012), se etiquetaron todas las oraciones una por una, resultados a partir de los cuales se pudo determinar que algunas oraciones debían eliminarse del corpus. Un caso particular fue el de oraciones que aparecen en los documentos originales en mayúsculas. Con el fin de no intervenir de ninguna manera en la integridad de los textos, estas oraciones se integraron al COR como tales, dado que cumplían con el criterio de longitud. Sin embargo, luego del proceso de etiquetado se determinó que el etiquetador de FreeLing etiqueta estas oraciones en mayúscula (correspondientes a subtítulos en los textos originales) como nombres propios (NP), de la siguiente forma:

LA_PATRIA_ES_OBRA_EN_MARCHA:::::la_patria_es_obra_en_marcha:::::NP00000

UN_PUEBLO_SANO_ES_UN_PUEBLO_PRODUCTIVO:::::un_pueblo_sano_es_un_pueblo_productivo::::NP00000

Por lo tanto se procedió a eliminar estas oraciones. Junto con otros filtros menores que se aplicaron, se eliminaron en esta revisión un total de 93 oraciones, quedando constituido el COR para el *treebank* por 6.668 oraciones.

6.3. Anotación del *treebank*

Con el COR ya depurado se procedió al análisis sintáctico por medio del *parser* de FIPS y a su conversión al formalismo IML. Este procedimiento se realizó de manera semiautomática, ya que la herramienta devuelve la oración analizada en corchetes etiquetados, pero las oraciones deben introducirse en la herramienta manualmente. Realizado el análisis en su totalidad, se determinó que solamente un 15% de las oraciones dio algún error en el análisis automático por medio del *parser*. De acuerdo con esto, el *treebank* se estructuró en términos de dos subcomponentes: uno analizado y otro semianalizado, como se explica más adelante. El resumen de estos datos es el siguiente:

Tabla 4. Subcomponentes del *treebank* IPROCOLDI

COR	Oraciones analizadas	Oraciones semianalizadas	TOTAL
SIGLO XIX	144	25	169
SIGLO XX	4806	875	5681
SIGLO XXI	719	99	818
TOTAL	5669	999	6668
Porcentaje	85%	15%	100%

En la anotación de un corpus puede haber distintas “capas” (*layers*) de anotación. El denominado “esquema de anotación” (*annotation scheme*) debe explicar cuáles capas

se incluyen y los distintos procedimientos que se siguieron al anotar el corpus. En este sentido, Nivre (2008) hace una distinción entre corpus simplemente anotados y los *treebanks* propiamente. Un corpus anotado puede incluir una única capa de anotación al nivel de la palabra; la información que se anota será la clase de la palabra, su lema (lematización) y rasgos morfológicos de número, género, persona, tiempo, etc. Un *treebank* específicamente se refiere a que el corpus ha sido analizado sintácticamente, ya sea por medio de un analizador automático (*parser*), por humanos, o bien, por una combinación de ambos procedimientos, como suele ser la práctica actual. La siguiente tabla muestra los distintos tipos de anotación que puede presentar un corpus lingüístico anotado, en negrita las capas consideradas en nuestro *treebank* (tomada de Göhring 2009, quien se basa en Lemnitzer and Zinsmeister 2006).

Tabla 5. Posibles capas de anotación lingüística de un *treebank*. (Fuente: Göhring 2009: 6)

Nivel lingüístico	Anotación	Ejemplos
Morfosintaxis	**clase de palabra**	**etiquetas *PoS***
Morfología	**flexión** **lematización**	**etiquetas *PoS*** **lemas**
Sintaxis	**constituencia sintagmática** dependencia sintagmática orden de la oración	**categorías sintagmáticas** etiquetas de dependencias campos topológicos
Semántica	nombres de entidades (*named entities*, NE) sentidos de las palabras roles, *frames*	etiquetas *PoS* WordNet synset, etiquetas *PoS* etiquetas de roles, esquemas FrameNet
Pragmática	correferencia discurso	vínculos anafóricos cadenas de argumentos

En nuestro esquema, hemos incorporado dos capas o niveles de análisis: morfológico y sintáctico, y los distinguimos como: **nivel etiquetado** y **nivel (sintácticamente) analizado** respectivamente. El nivel etiquetado incluye lematización, clase de palabra y rasgos morfológicos mediante la incorporación de las etiquetas EAGLES para el español. El nivel analizado incluye la anotación de constituyentes por medio del formalismo IML, el cual se basa en el formalismo FIPS. Se detallan estos procesos en las secciones siguientes.

6.3.1. Tratamiento de las unidades lingüísticas (tokenization)

Las oraciones del COR, en total 6.668, se sometieron a la herramienta "PoS Tagging" de FreeLing. La herramienta se utilizó en línea: cada oración fue sometida al etiquetador y extraída de FreeLing para su incorporación al *treebank*. A manera de ejemplo, la primera oración del *treebank* se extrajo de la siguiente forma (léase verticalmente: vocablo, lema, etiqueta EAGLES):

Cuadro 1. Ejemplo de lematización y etiquetado morfológico

Hasta	***Pueblo***	***azote***	***endémicas***
hasta	***pueblo***	***azote***	***endémico***
SPS00	***NP00000***	***NCMS000***	***AQ0FP0***
el	***ha***	***cruel***	***o***
el	***haber***	***cruel***	***o***
DA0MS0	***VAIP3S0***	***AQ0CS0***	***CC***
día	***sufrido***	***de***	***epidémicas***
día	***sufrir***	***de***	***epidémico***
NCMS000	***VMP00SM***	***SPS00***	***AQ0FP0***
ningún	***el***	***enfermedades***	***.***
ninguno	***el***	***enfermedad***	***.***
DI0MS0	***DA0MS0***	***NCFP000***	***Fp***

Decidimos darle un formato horizontal no solo por conveniencia de espacio, sino por el diseño mismo del *treebank*, en triadas de archivos que contienen las oraciones numeradas correlativamente (ver apartado 6.4.). Para separar el lema y la etiqueta de cada vocablo se utilizó una convención que fuera clara y fácilmente identificable como símbolo de conexión: una secuencia de cuatro signos de dos puntos: *Hasta::::hasta::::SPS00.*

La principal dificultad con respecto a este proceso se dio con los signos de puntuación. En el siguiente apartado se indicarán los problemas específicos que se presentaron y cómo se resolvieron.

6.3.2. Anotación sintáctica de la estructura de constituyentes

Este proceso, paralelo al anterior, se llevó a cabo, igualmente, sometiendo las oraciones individuales al *parser* de FIPS y extrayendo de ahí la estructura analizada mediante corchetes etiquetados (*labelled bracketing*). Una vez con todas las oraciones analizadas, convertimos el etiquetado de FIPS al etiquetado IML. En este estadio también se llevó a cabo la clasificación de las oraciones en analizadas y semianalizadas, y se tomó la decisión de separar las semianalizadas como un subcomponente del corpus que permitiría estudiar con más claridad los casos en que el *parser* dio error y, eventualmente, hacer propuestas para su mejora y perfeccionamiento.

6.3.3. Concatenación de los niveles de anotación morfológica y de análisis sintáctico

Este proceso requirió que el investigador asociado, Dr. Antonio Leoni, diseñara un código para llevar a cabo esta tarea automáticamente y obtener así una estructura unificada de cada oración mostrando, además de la lematización, los dos niveles de análisis, el morfológico y el sintáctico. Estas estructuras se enlistan en los archivos del *treebank* que hemos identificado como **IML_ETIQUET**.

6.4. Estructura del *treebank*

De acuerdo con los procedimientos expuestos en secciones precedentes, en el diseño del *treebank* se incorporaron tres archivos correlativos por siglo:

a. Archivo de datos crudos: las oraciones sin anotación; se identifica como **COR**.
b. Archivo de datos analizados (el *treebank* propiamente); se identifica como **IML**.
c. Archivo de datos etiquetados y analizados; se identifica como **IML_ETIQUET**.

Se presenta la misma estructura para los dos subcomponentes: el analizado (**COR_analizado**) y el semianalizado (**COR_semianalizado**). En el Cuadro 2 se muestra la estructura completa del TREEBANK IPROCOLDI, que contiene un total de 6.668 oraciones:

Cuadro 2. Estructura del TREEBANK IPROCOLDI

```
+---Siglo_XIX: 144 oraciones analizadas y 25 semianalizadas. Total: 169.
| |   SIGLOXIX_IML_144.txt
| |   SIGLOXIX_IML_ETIQUET_144.txt
| |   SIGLOXIX_COR_analizado_144.txt
| \---COR_semianalizado
|       SIGLOXIX_IML_25.txt
|       SIGLOXIX_IML_ETIQUET_25.txt
|       SIGLOXIX_COR_semianalizado_25.txt
|
+---Siglo_XX: 4806 oraciones analizadas y 875 semianalizadas. Total: 5681.
| |       SIGLOXX_IML_4806.txt
| |       SIGLOXX_IML_ETIQUET_4806.txt
| |       SIGLOXX_COR_analizado_4806.txt
| \---COR_semianalizado
|       SIGLOXX_IML_875.txt
|       SIGLOXX_IML_ETIQUET_875.txt
|       SIGLOXX_COR_semianalizado_875.txt
|
+---Siglo_XXI: 719 oraciones analizadas y 99 semianalizadas. Total: 818.
| |   SIGLOXXI_IML_719.txt
| |   SIGLOXXI_IML_ETIQUET_719.txt
| |   SIGLOXXI_COR_analizado_719.txt
| \---COR_semianalizado
|       SIGLOXXI_IML_99.txt
|       SIGLOXXI_IML_ETIQUET_99.txt
|       SIGLOXXI_COR_semianalizado_99.txt
```

Este diseño permite una mayor flexibilidad en el uso del *treebank*, ya que si la investigación requiere solamente los datos sintácticos, puede utilizarse el archivo IML independientemente; mientras que si la investigación requiere datos lexemáticos, morfológicos y sintácticos se podrá utilizar el archivo que contiene la anotación completa.

Las oraciones de cada archivo se encuentran numeradas correlativamente, es decir, se identifican con el mismo número en cada archivo. Así, la primera oración del *treebank*, correspondiente a la oración 1 del archivo COR del siglo XIX, es la siguiente:

> XIX-1. Hasta el día ningún Pueblo ha sufrido el azote cruel de enfermedades endémicas o epidémicas.[13]

En el archivo IML se encuentra en la misma línea 1 el análisis sintáctico:

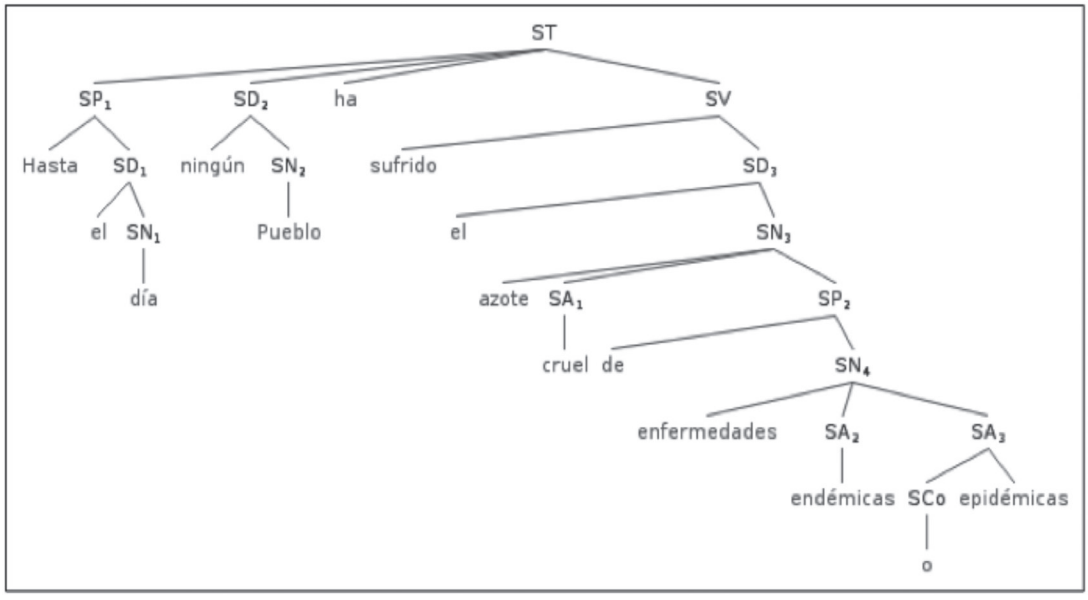
> XIX-1. [ST[SP Hasta [SD el [SN día]]][SD ningún [SN Pueblo]] ha [SV sufrido [SD el [SN azote [SA cruel][SP de [SN enfermedades [SA endémicas][SA[SCo o] epidémicas]]]]]]

Finalmente en el archivo IML_ETIQUET se encuentra en la línea 1 el análisis completo de la misma oración en el siguiente formato (cada vocablo aparece seguido de su lema y luego de la correspondiente etiqueta morfológica):

XIX-1. [ST[SP Hasta::::hasta::::SPS00 [SD el::::el::::DA0MS0 [SN día::::día::::NCMS000]]] [SD ningún::::ninguno::::DI0MS0 [SN Pueblo::::pueblo::::NP00000]] ha::::haber::::VAIP3S0 [SV sufrido::::sufrir::::VMP00SM [SD el::::el::::DA0MS0 [SN azote::::azote::::NCMS000 [SA cruel::::cruel::::AQ0CS0][SP de::::de::::SPS00 [SN enfermedades::::enfermedad::::NCFP000 [SA endémicas::::endémico::::AQ0FP0][SA[SCo o::::o::::CC] epidémicas::::epidémico::::AQ0FP0]]]]]]]

Para la visualización de la estructura sintáctica, hemos utilizado una herramienta de uso libre, phpSyntaxTree[14], que permite convertir el análisis a un diagrama de árbol, como puede verse en el siguiente ejemplo correspondiente a la misma oración *supra*:

Cuadro 3. Diagrama arbóreo de la oración XIX-1 del *treebank*



6.5. Aspectos relacionados con la puntuación

Al someter las oraciones al *parser*, se observó que la puntuación determina la segmentación de las oraciones para el análisis. Así por ejemplo, en el COR aparecían oraciones como la siguiente:

Ustedes me conocen, saben que así he sido siempre: soy fiel a mis convicciones.[15]

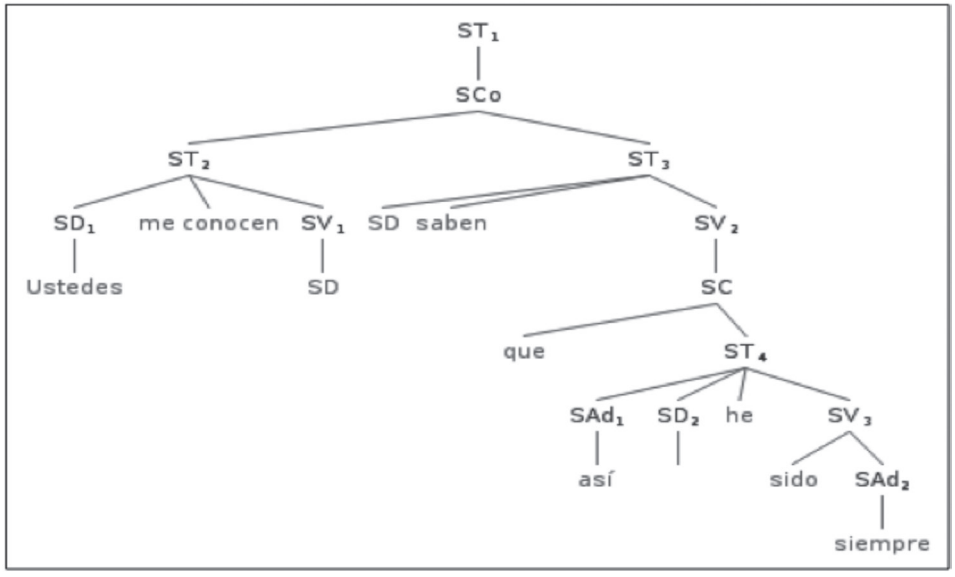
Mientras que el signo de coma queda subsumido en el análisis de las dos primeras oraciones, el signo de dos puntos induce al *parser* a separar la última oración asignándole un análisis independiente. De acuerdo con esto, se optó por mantener en un mismo análisis las oraciones yuxtapuestas por medio de coma, y se independizaron las oraciones yuxtapuestas por dos puntos y por punto y coma. Así, la oración anterior se separó en las dos oraciones XXI-214 y XXI-215. En el *treebank* aparecen del siguiente modo:

XXI-214. [ST [SCo[ST[SD Ustedes] me conocen [SV [SD]]][ST[SD] saben [SV [SC que [ST[SAd así][SD] he [SV sido [SAd siempre]]]]]]]]

XXI-215. [ST[SD] soy [SV [SF[SD] [SA fiel [SP a [SD mis [SN convicciones]]]]]]]

El diagramador de árbol muestra la estructura de XXI-214 como un sintagma conjuntivo (SCo) compuesto por dos sintagmas temporales:

Cuadro 4. Diagrama arbóreo de la oración XXI-214



En el caso de oraciones separadas por punto y coma, se siguió el mismo procedimiento. Encontramos el siguiente ejemplo del siglo XX:

Es cierto; pero no siempre habrá de valernos nuestra buena estrella.[16]

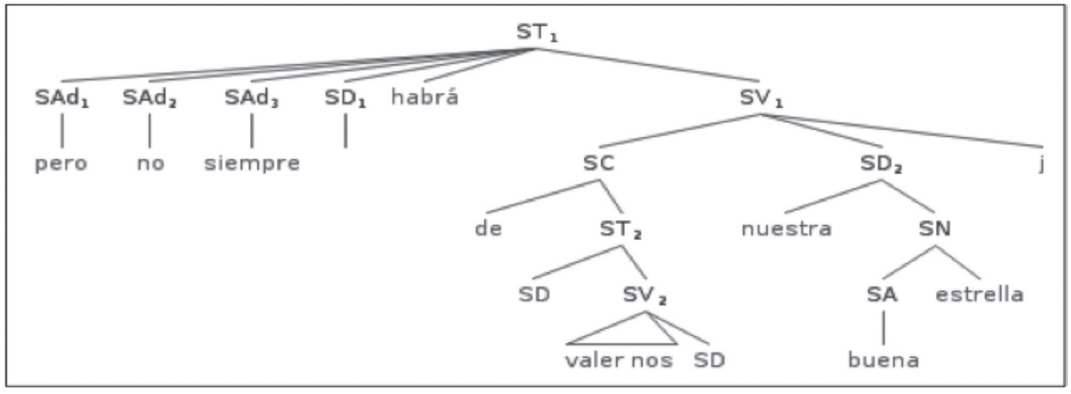
que se analiza de la siguiente forma:

XX-520. [ST[SD] Es [SV [SF[SD] [SA cierto]]]]

XX-521. [ST[SAd pero][SAd no][SAd siempre][SD] habrá [SV [SC de [ST[SD] [SV valer nos [SD]]]] [SD nuestra [SN[SA buena] estrella]]j]]

En este caso, el signo de puntuación separa las oraciones, pero además la segunda oración está coordinada con la anterior por medio de la conjunción adversativa *pero*. El diagramador muestra este elemento como un sintagma adverbial (SAv):

Cuadro 5. Diagrama arbóreo de la oración XX-521



Precisamente este ejemplo nos sirve para ilustrar la ventaja de contar con dos archivos separados: el análisis sintáctico (IML) y el análisis enriquecido con información morfológica (IML_ETIQUET). Mientras que en el diagrama anterior los constituyentes *pero no siempre* aparecen indistintamente como SAds, en el análisis etiquetado tenemos las distinciones más finas sobre la clase de las palabras:

XX-521. [ST[SAd **pero:::::pero:::::CC**][SAd **no:::::no:::::RN**][SAd **siempre:::::siempre:::::RG**] [SD] habrá:::::haber:::::VMIF3S0 [SV [SC de:::::de:::::SPS00 [ST[SD] [SV valer:::::valer:::::VMN0000 nos:::::nos:::::PP1CP000 [SD]]]][SD nuestra:::::nuestro:::::DP1FSP [SN[SA buena:::::bueno:::::AQ0FS0] estrella:::::estrella:::::NCFS000]]j]]

De esta manera podemos elegir el archivo que conviene de acuerdo con el detalle informativo que requerimos.

El siguiente ejemplo –oraciones separadas por el signo de dos puntos– permite hacer una observación importante: no siempre las dos oraciones resultaron en análisis completos. Cuando el análisis automático de una de ellas dio error, esta quedó ubicada en el subcomponente semianalizado, como sucede con la siguiente oración:

Es este un grave cargo: dichosamente es también un cargo absurdo.[17]

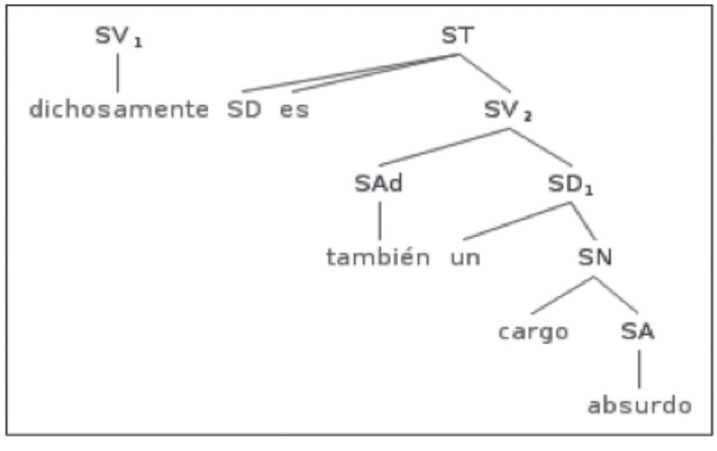
Los análisis fueron los siguientes (se indica el subcomponente semianalizado mediante una **S** después del número romano que indica el siglo):

XX-81. [ST[SD] Es [SV [SF[SD] [SA este]][SD un [SN[SA grave] cargo]]i]]

XXS-24. [SV dichosamente] [ST[SD] es [SV [SAd también][SD un [SN cargo [SA absurdo]]]]]

Si diagramamos la estructura de la oración semianalizada, podemos ver que el problema que tuvo el *parser* fue que no logró adscribir el constituyente adverbial *dichosamente* al ST.

Cuadro 6. Diagrama arbóreo de la oración XXS-24



Este tipo de errores que comete el *parser* al realizar el análisis permitirá recomendar nuevas reglas para el mejoramiento del análisis sintáctico automatizado, siendo este uno de nuestros principales objetivos al desarrollar el *treebank*.

En la mayoría de los casos, sin embargo, las oraciones yuxtapuestas y relacionadas por medio de punto y coma o dos puntos, aparecen en el componente analizado como oraciones sucesivas, como en el siguiente ejemplo:

Nuestro lema debe ser: En el bien social está el de cada uno de nosotros.[18]

que se analiza de la siguiente manera:

XX-334. [ST[SD Nuestro [SN lema]]i debe [SV [ST[SD] [SV ser]]]]

XX-335. [ST[SP En [SD el [SN bien [SA social]]]][SD] está [SV [SD el [SP de [SD cada [SD uno [SP de [SD nosotros]]]]]i]]

La última observación que debo hacer con respecto al tratamiento de la puntuación en relación con el *parsing* es que cuando el constituyente que sigue al signo de puntuación no es un sintagma temporal (ST) o no tiene la estructura explícita de un ST (por ejemplo, si el verbo está elidido), los sintagmas constituyentes no fueron separados. Ejemplos:

También fue menor la introducción de carne extranjera: 90,542 kilos, contra 141.072 en 1908.[19]

XX-123. [ST[SAd También][SD] fue [SV [SF[SD] [SA menor]][SD la [SN introducción [SP de [SN carne [SA extranjera]]]]i]] [SD 90,542 [SN kilos][SP contra [SD 141.072 [SP en [SD 1908]]]]]

Los nombres han cambiado; la mala suerte no.[20]

XX-636. [ST[SD Los [SN nombres]] han [SV cambiado]] [SD la [SN[SA mala] suerte [SN no]]]

7. Algunos aspectos de la anotación que requieren revisión y depuración

En las siguientes secciones se presentan ejemplos de algunos aspectos en los cuales deberá concentrarse la revisión manual y la corrección humana del *treebank*. No estamos todavía en condición de presentar un inventario exhaustivo de los problemas ni de proponer reglas específicas para resolverlos, sino simplemente de ejemplificar algunos de ellos a partir de una primera inspección general del corpus.

7.1. Problemas de lematización

El *treebank* requiere una revisión manual detallada para resolver los errores de lematización que produjo FreeLing. Por ejemplo, en la oración 54 del siglo XIX, *Y creo más*[21], la forma verbal *creo* fue lematizada con el infinitivo *crear*, en lugar del correcto *creer.*

XIX-54. [ST[SCo Y::::y::::CC][SD] **creo::::crear**::::VMIP1S0 [SV [SAd más::::más::::RG]]]

7.2. Errores del *parsing* automático

*7.2.1. El clítico **se***

La descripción y el análisis de la partícula **se** ha representado uno de los más difíciles retos para los analizadores automáticos del español. En su propio *treebank* del español, Moreno et al. (2003) la tratan en la mayoría de los casos como pronombre, excepto cuando funciona como marcador de intransitividad o de impersonalidad. De acuerdo con esto, distinguen cinco tipos de anotación:

a. *se* (PRON) por *le*, sustitución que obedece a razones fonéticas. Ejemplo: *Se lo dio*.
b. *se* (PRON) en construcción reflexiva o recíproca, que cambia de acuerdo con persona y número: *me* (SG P1), *te* (SG P2), *se* (SG P3), *nos* (PL P1), *os* (PL P2), *se* (PL P3). Al

ser semántica la diferencia entre el reflexivo y el recíproco, ambos se anotan igual. Ejemplo: *Yo me lavo.*

c. *se* pronominal o intrínseco: cuando es parte del verbo, aparecen unidos en el lexicón (*enterarse*). Se anota como palabra compuesta, sea que preceda (separado) o siga (unido) al verbo. Cuando precede se trata como pronombre con sus propios rasgos y sus marcas de discontinuidad y referencia, para mantener su autonomía como pronombre que forma parte del verbo. Cuando sigue al verbo se deja igual y se añade el rasgo #CLITIC. Este caso representa un grave problema para el *parser* de FIPS, como se verá más adelante.

d. *se* como marcador de intransitividad: incluye las tres construcciones llamadas "pasiva con *se*", "*se* inacusativo" y "*se* medio", que tienen en común la pérdida de un argumento. Se distinguen semánticamente pero como solo interesan para la anotación los criterios sintácticos, se anotan igual. Así, este *se* no se trata como pronombre, sino como un marcador de construcciones intransitivas. Ejemplo: *El libro se rompió*. Se anota mediante la etiqueta categorial SE-MARK y el rasgo INTRANSITIVE.

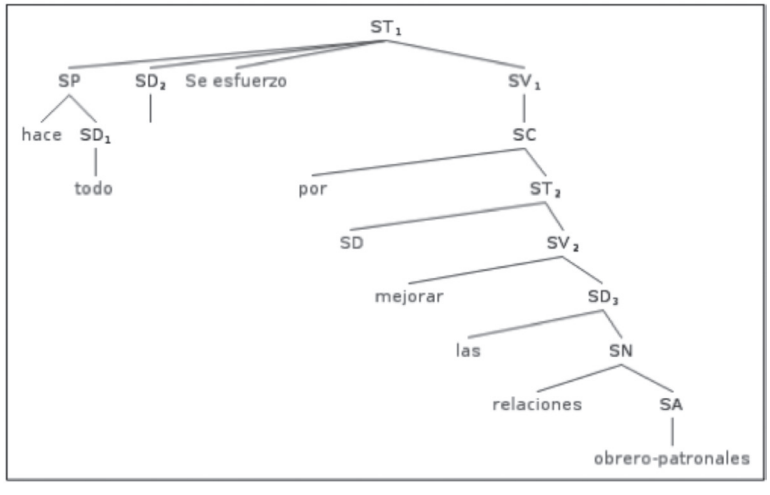
e. *se* impersonal: estas construcciones se caracterizan por no tener sujeto sintáctico; *se* es un pronombre que marca impersonalidad: *Se vende piso* (Alguien vende un piso). Se anota mediante la etiqueta categorial SE-MARK y el rasgo IMPERSONAL (Moreno et al. 2003: 155-6).

En nuestro propio análisis, *se* presenta varios problemas según la construcción en que aparece. Cuando aparece en posición proclítica, y por ende como palabra independiente, el *parser* lo traslada a algún otro constituyente, en lugar de conservarlo como elemento constitutivo del sintagma temporal, regla que sigue FIPS con respecto a otros clíticos. Los dos ejemplos siguientes corresponden el primero (XX-2123) al componente analizado y el segundo (XXS-690) al semianalizado:

XX-2123. *Se hace todo esfuerzo por mejorar las relaciones obrero-patronales*.[22]

[ST[SP hace [SD todo]][SD] Se esfuerzo [SV [SC por [ST[SD] [SV mejorar [SD las [SN relaciones [SA obrero-patronales]]]]]]]

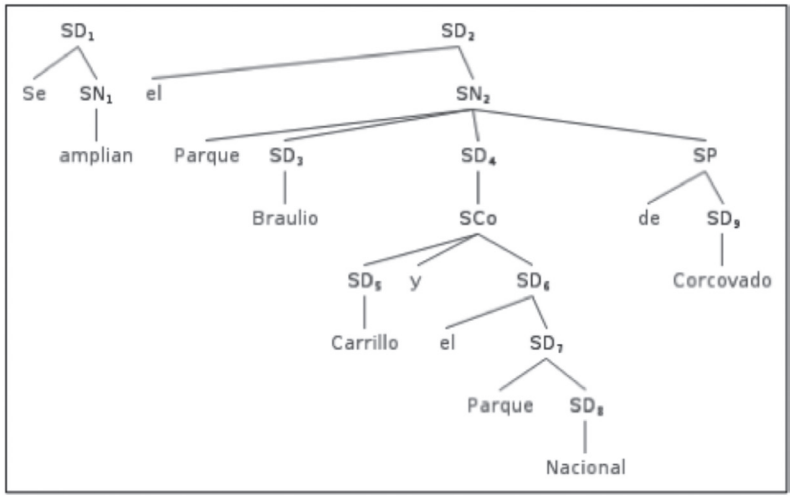
Cuadro 7. Diagrama arbóreo de XX-2123



XXS-690. *Se amplían el Parque Braulio Carrillo y el Parque Nacional de Corcovado*.[23]

[SD Se [SN amplían]] [SD el [SN Parque [SD Braulio][SD [SCo[SD Carrillo] y [SD el [SD Parque [SD Nacional]]]]][SP de [SD Corcovado]]]]

Cuadro 8. Diagrama arbóreo de XXS-690

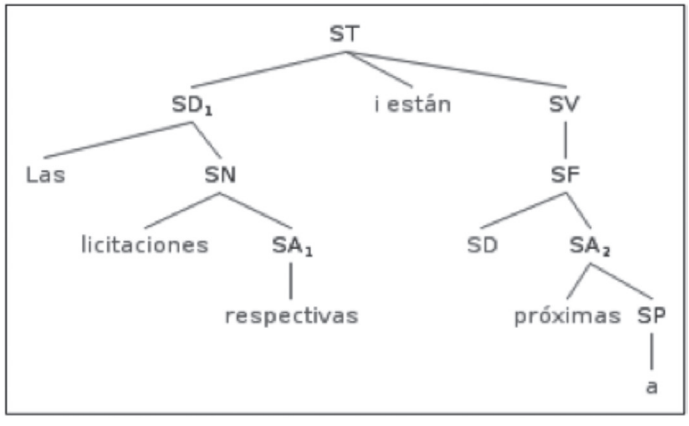


Como puede verse en este último ejemplo (semianalizado), el *parser* ni siquiera reconoce el verbo como núcleo de un ST. Sin embargo, la situación más grave se da en el caso del *se* pronominal enclítico, ya que FIPS elimina completamente la palabra (el verbo en infinitivo y terminado en *-se*) en el análisis resultante. Así por ejemplo, la oración: *Las licitaciones respectivas están próximas a publicarse*, se indica como semianalizada en el *parser*, aunque el árbol se diagrama correctamente:

XXS-438. *Las licitaciones respectivas están próximas a publicarse*.[24]

[ST[SD Las [SN licitaciones [SA respectivas]]]i están [SV [SF[SD] [SA próximas [SP a]]]]]

Cuadro 9. Diagrama arbóreo de XXS-438

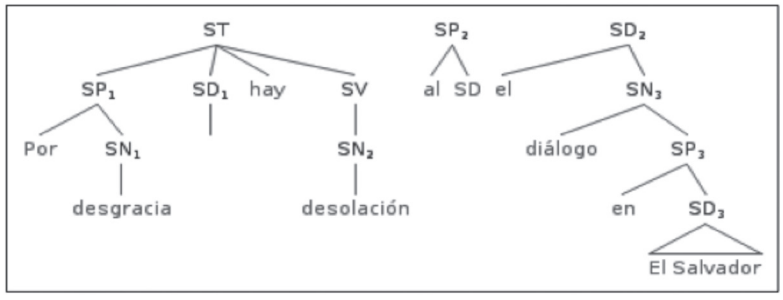


Para otras oraciones semianalizadas, en cambio, el árbol no se genera completo ni correctamente, como en:

XXS-599. *Por desgracia, hay desolación, al frustrarse el diálogo en El Salvador.*[25]

[ST[SP Por [SN desgracia]][SD] hay [SV [SN desolación]]]
[SP al [SD]]
[SD el [SN diálogo [SP en [SD El Salvador]]]]

Cuadro 10. Diagrama arbóreo de XXS-599

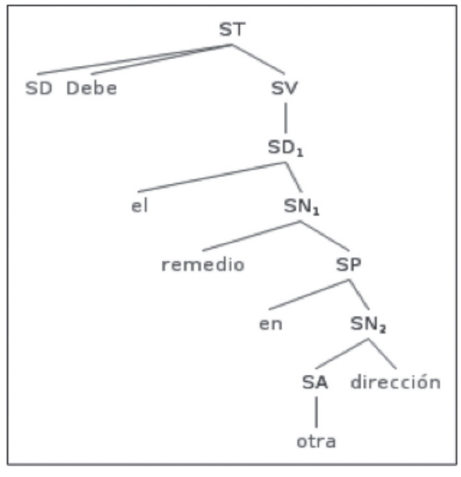


Sin embargo, también se dan casos en que la oración resulta en un análisis completo, como en el siguiente ejemplo, aunque como puede verse en el árbol, el análisis sintáctico no es correcto (el SP *en otra dirección* debió estar dominado por el nodo SV):

XX-149. *Debe buscarse el remedio en otra dirección.*[26]

[ST[SD] Debe [SV [SD el [SN remedio [SP en [SN[SA otra] dirección]]]]]]

Cuadro 11. Diagrama arbóreo de XX-149



Es de interpretarse que en estos casos la oración resulta analizada completamente (esto es, dominada toda la estructura por un único nodo superior ST) porque el constituyente nuclear de la oración es la forma modal (*debe*). Lo mismo sucede con todas las oraciones que tienen un modal como núcleo del ST, por ejemplo: *La importación **debe** restringirse; **Debe** nacionalizarse el aseguro; Las dificultades **pueden** vencerse*.

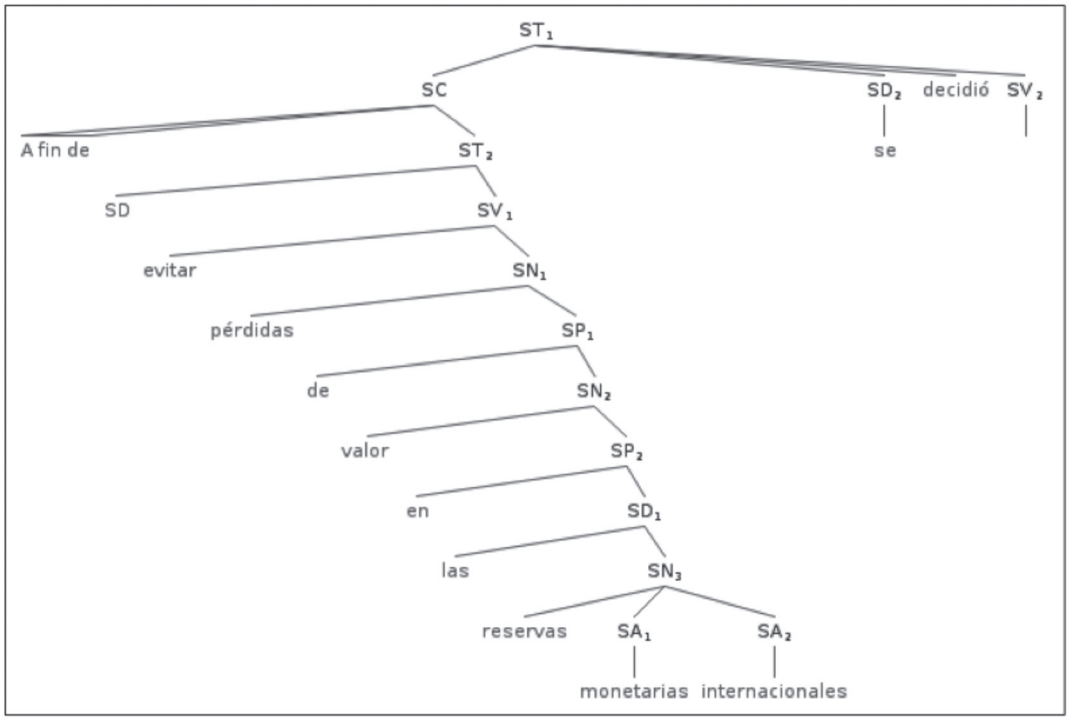
Ahora bien, cuando el pronombre enclítico es distinto de **se**, la oración resulta analizada completamente; sin embargo, al igual que en los casos anteriores el *parser* elimina el infinitivo, como en el siguiente ejemplo, en donde en el análisis desaparece la palabra *diversificarlas*.

Nótese, sin embargo, que en el árbol queda vacío el espacio en donde debería aparecer, como elemento terminal del SV_2.

XX-2379. *A fin de evitar pérdidas de valor en las reservas monetarias internacionales se decidió diversificarlas*.[27]

[ST[SC A fin de [ST[SD] [SV evitar [SN pérdidas [SP de [SN valor [SP en [SD las [SN reservas [SA monetarias][SA internacionales]]]]]]]]][SD se] decidió [SV]]

Cuadro 12. Diagrama arbóreo de XX-2379

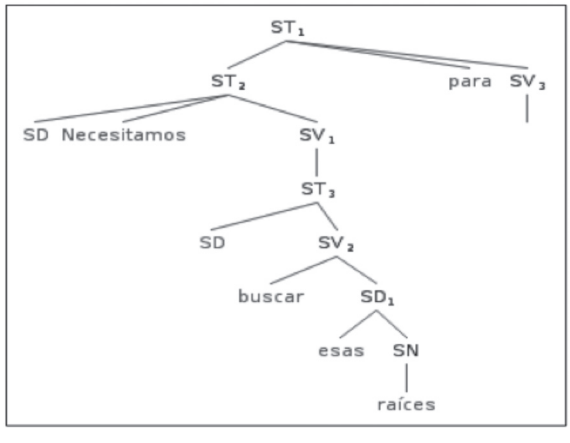


Un ejemplo semejante es:

XX-2796. *Necesitamos buscar esas raíces para fortalecerlas*.[28]

2796- [ST[ST[SD] Necesitamos [SV [ST[SD] [SV buscar [SD esas [SN raíces]]]]]] para [SV]]

Cuadro 13. Diagrama arbóreo de XX-2796



Nótese que en este caso, el SP *para fortalecerlas* resulta erróneamente identificado como el ST principal, al parecer tomando el vocablo *para* como núcleo.

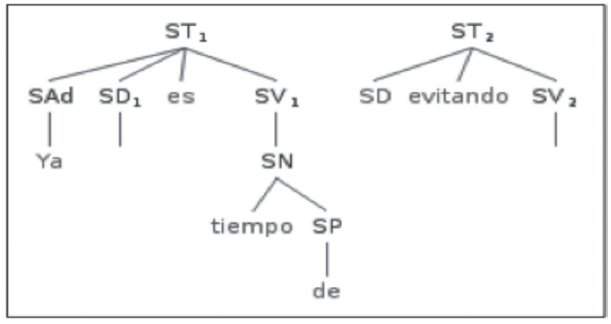
Como último ejemplo de la elisión del infinitivo con pronombre enclítico, véase el siguiente ejemplo semianalizado:

XXS-139- *Ya es tiempo de irlo evitando*.[29]

[ST[SAd Ya][SD] es [SV [SN tiempo [SP de]]]]

[ST[SD] evitando [SV]]

Cuadro 14. Diagrama arbóreo de XXS-139

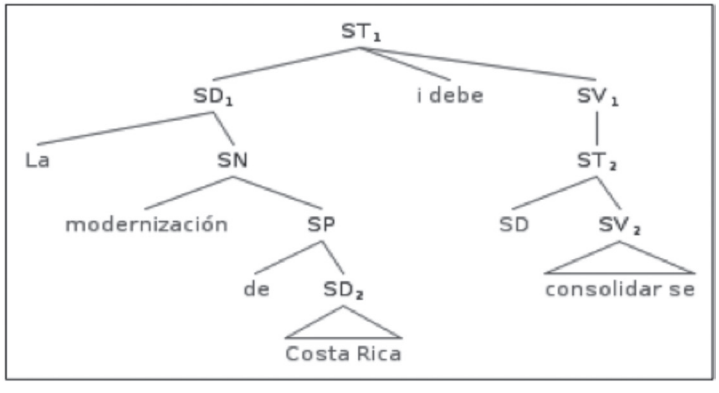


Este caso demuestra la imposibilidad para el *parser* de analizar sintagmas que contengan infinitivo con pronombre enclítico. Sin embargo, debe agregarse que si en la gran mayoría de los casos ocurre la elisión del tal infinitivo, unos pocos casos muestran que el *parser* lo analiza correctamente. Un ejemplo es el siguiente:

XX-3669. *La modernización de Costa Rica debe consolidarse*.[30]

[ST[SD La [SN modernización [SP de [SD Costa Rica]]]]i debe [SV [ST[SD] [SV consolidar se]]]]

Cuadro 15. Diagrama arbóreo de XX-3669

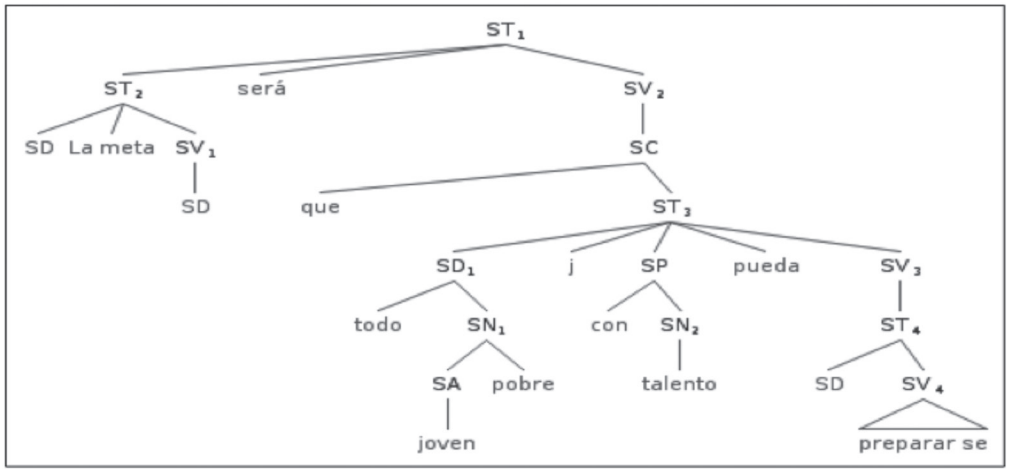


Por último, también debe investigarse el caso en que un infinitivo con **se** enclítico se da en una perífrasis modal dentro de un SC (sintagma complementizador), situación en la cual el infinitivo no se elimina, como demuestra el siguiente ejemplo:

XX-2214. *La meta será que todo joven pobre con talento pueda prepararse*.[31]

[ST[ST[SD] La meta [SV [SD]]] será [SV [SC que [ST[SD todo [SN[SA joven] pobre]]j[SP con [SN talento]] pueda [SV [ST[SD] [SV preparar se]]]]]]]

Cuadro 16. Diagrama arbóreo de XX-2214



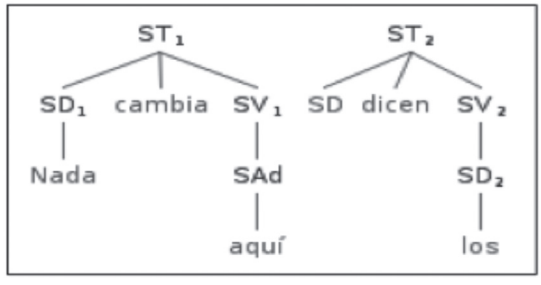
7.2.2. Otras elisiones

Es necesario hacer una revisión exhaustiva del *treebank* para determinar otros tipos de elisiones, por ejemplo de un sustantivo, como ocurre en el siguiente ejemplo semianalizado:

XXS-225. ¡Nada cambia aquí -dicen los jóvenes-[32]

[ST[SD Nada] cambia [SV [SAd aquí]]] [ST[SD] dicen [SV [SD los]]]

Cuadro 17. Diagrama arbóreo de XXS-225



Es probable que la elisión del sustantivo se relacione con la trasposición del sujeto *los jóvenes*, como se observa en el árbol al aparecer dominado el SD_2 por un SV. Los fenómenos de movimiento de constituyentes representan un notable problema para el *parser*, como se muestra en la siguiente sección.

7.2.3. Dislocaciones

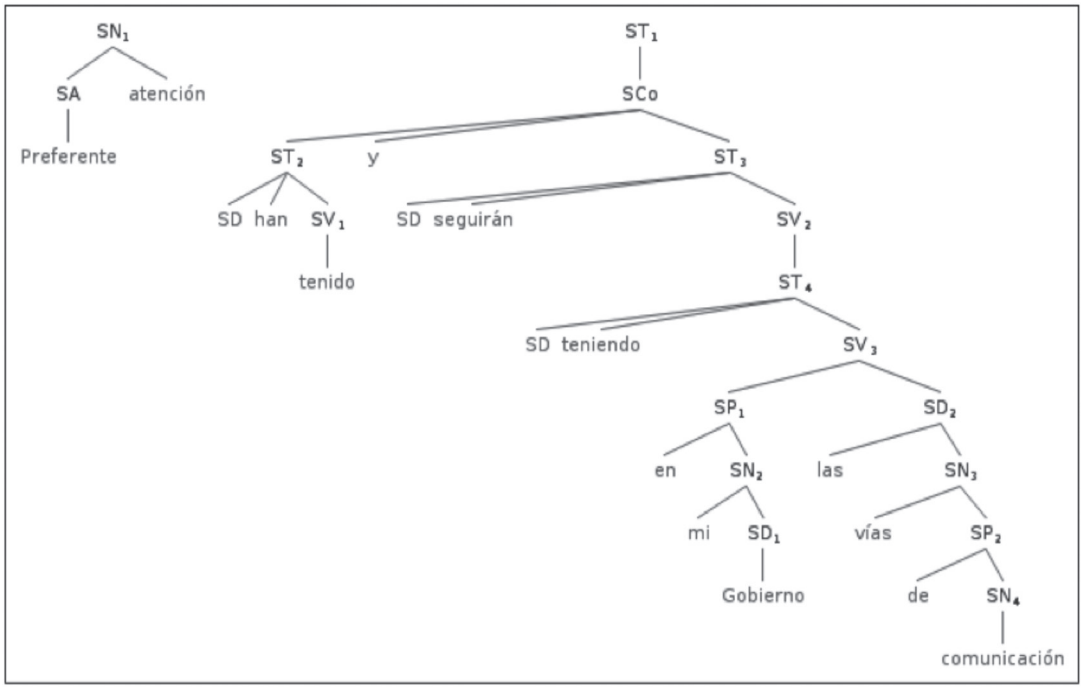
Es una tendencia generalizada del *parser* no analizar completamente las oraciones que presentan dislocaciones sean a la derecha, a la izquierda o ambas. En la oración: *Preferente atención han tenido y seguirán teniendo en mi Gobierno las vías de comunicación*, se ha movido a posición inicial el SN de objeto directo *preferente atención* y se ha traspuesto el SN de sujeto *las vías de comunicación*. El análisis resulta incompleto, como puede verse en el diagrama arbóreo:

XXS-8. *Preferente atención han tenido y seguirán teniendo en mi Gobierno las vías de comunicación*.[33]

[SN[SA Preferente] atención]

[ST [SCo[ST[SD] han [SV tenido]] y [ST[SD] seguirán [SV [ST[SD] teniendo [SV [SP en [SN mi [SD Gobierno]]][SD las [SN vías [SP de [SN comunicación]]]]]]]]]

Cuadro 18. Diagrama arbóreo de XXS-8

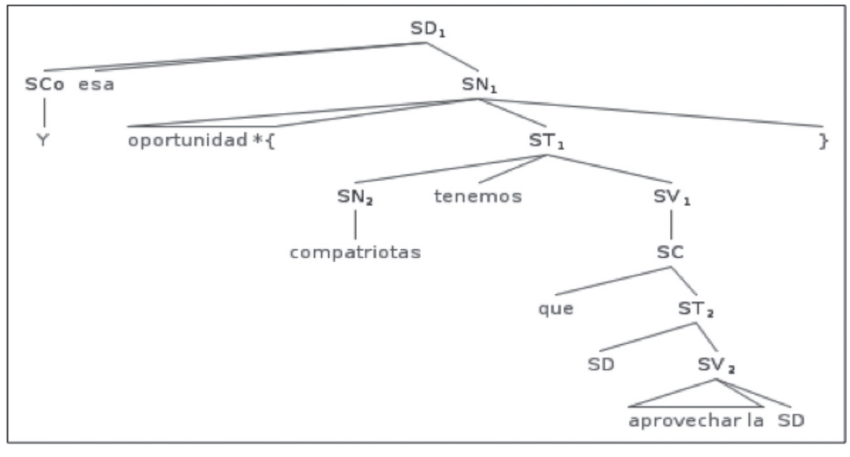


Ahora bien, en unos pocos casos se logró un análisis completo, por ejemplo:

XX-3988. *Y esa oportunidad, compatriotas, tenemos que aprovecharla*.[34]

[SD[SCo Y] esa [SN oportunidad *{[ST[SN compatriotas] tenemos [SV [SC que [ST[SD] [SV aprovechar la [SD]]]]]}]]

Cuadro 19. Diagrama arbóreo de XX-3988



Obsérvese en este ejemplo el análisis correcto del infinitivo con pronombre enclítico (*aprovecharla*), cuando, como se comentó al final de la sección 7.2.2, aparece dentro de un sintagma complementizador (SC).

7.2.4. Otras construcciones problemáticas

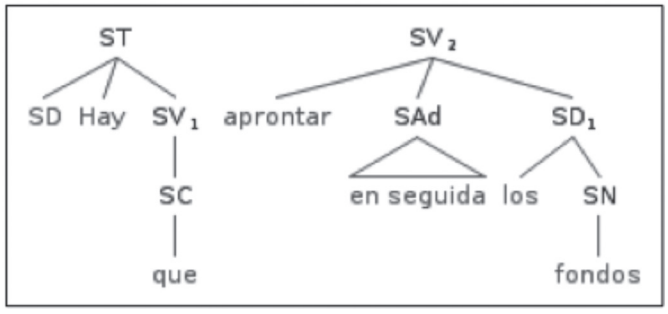
Muchas oraciones impersonales deónticas introducidas por *hay que* + infinitivo resultan semianalizadas:

XXS-93. *Hay que aprontar en seguida los fondos*.[35]

[ST[SD] Hay [SV [SC que]]]

[SV aprontar [SAd en seguida][SD los [SN fondos]]]

Cuadro 20. Diagrama arbóreo de XXS-93



También algunas oraciones exclamativas resultaron semianalizadas; por ejemplo:

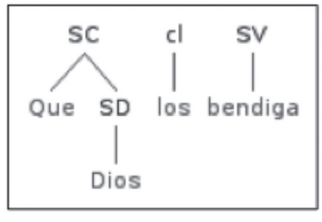
XXS-667. ¡Que Dios los bendiga![36]

[SC Que [SD Dios]]

[cl los]

[SV bendiga]

Cuadro 21. Diagrama arbóreo de XXS-667



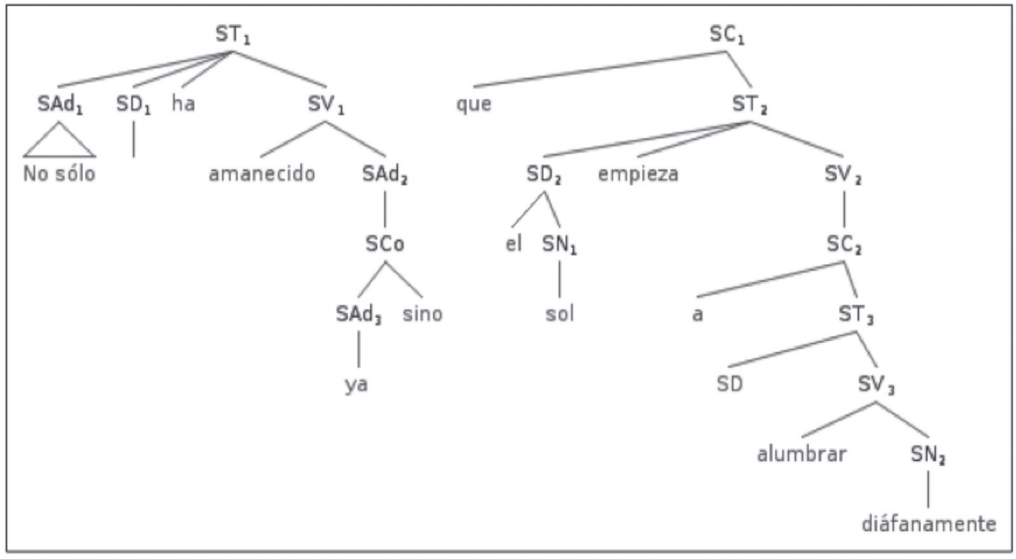
Finalmente, las coordinaciones copulativas con *no (solo)... sino (que)*, resultaron siempre en análisis incompletos, como muestran las oraciones XXS-558 y XXS-632,

XXS-558. *No sólo ha amanecido ya, sino que el sol empieza a alumbrar diáfanamente*.[37]

[ST[SAd No sólo][SD] ha [SV amanecido [SAd [SCo[SAd ya] sino]]]]

[SC que [ST[SD el [SN sol]] empieza [SV [SC a [ST[SD] [SV alumbrar [SN diáfanamente]]]]]]]

Cuadro 22. Diagrama arbóreo de XXS-558

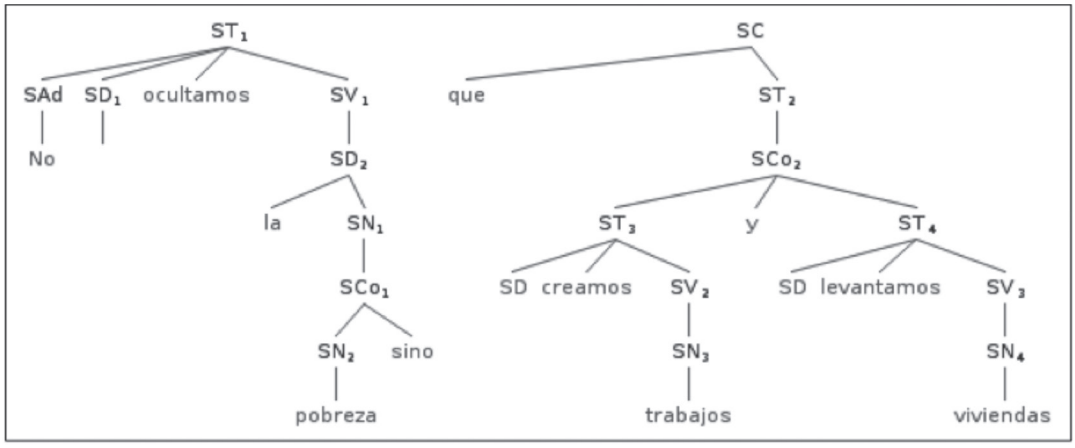


XXS- 632. *No ocultamos la pobreza, sino que creamos trabajos y levantamos viviendas*.[38]

[ST[SAd No][SD] ocultamos [SV [SD la [SN [SCo[SN pobreza] sino]]]]]

[SC que [ST [SCo[ST[SD] creamos [SV [SN trabajos]]] y [ST[SD] levantamos [SV [SN viviendas]]]]]]

Cuadro 23. Diagrama arbóreo de XXS-632



Como puede observarse en estos últimos ejemplos de construcciones coordinadas, si bien el *parser* produce análisis incompletos (árboles separados), el análisis sintáctico de ambas cláusulas es perfecto. Esto nos conduce a una observación fundamental y que justifica nuestra división del *treebank* en los componentes "analizado" y "seminanalizado". La observación es que el análisis incompleto (componente semianalizado) no implica un análisis sintáctico incorrecto. El *parser* produce análisis completos que no son necesariamente correctos, así como análisis incompletos que pueden ser correctos. En los últimos ejemplos, lo que sucede es que queda sin explicitarse un nivel superior que conjunta las dos estructruras (ST_1 y SC).

8. Conclusiones

En este artículo se han abordado varias tareas: en primer lugar se hizo una exposición del concepto de *treebank* (corpus oracionales anotados con información lingüística a varios niveles) y se reseñó su desarrollo en el ámbito de la lingüística de corpus. Asimismo se hizo una revisión de los proyectos actuales en el área de construcción de corpus y *treebanks* del español para ubicar en ese contexto el desarrollo del TREEBANK IPROCOLDI. También se describió la interfaz que alberga tanto este *treebank* como el CODIMEP-CR/IPROCOLDI, corpus en crudo del cual se extrajeron los datos para la anotación.

El último apartado se dedicó a exponer algunos de los problemas que se encontraron en la anotación automática del corpus. Debe enfatizarse que tal exposición no pretende ser exhaustiva, sino que se señalan algunos de los problemas que se fueron detectando mediante una inspección general, conforme se avanzaba en la construcción del *treebank*. La tarea de sistematizar los errores y proponer las reglas que puedan mejorar el análisis automático queda pendiente. Es de esperar que mediante proyectos futuros se pueda perfeccionar el corpus anotado para que sirva su objetivo último: contribuir al desarrollo de mejores y más exactas herramientas automatizadas para el análisis lingüístico.

Notas

1. Los investigadores agradecemos al Dr. Éric Wehrli, del Laboratoire d'Analyse et de Technologie du Langage de la Universidad de Ginebra, su gentil anuencia a que utilizáramos liberalmente la herramienta FIPS.
2. Más detalles sobre FIPS se pueden encontrar en: http://www.latl.unige.ch/
3. El conjunto de etiquetas EAGLES (*EAGLES tagset*) para el español puede consultarse en la dirección: http://www.lsi.upc.edu/~nlp/tools/parole-sp.html
4. Este paquete de herramientas para el análisis lingüístico se distribuye bajo la GNU General Public License de la Free Software Foundation. Ver: http://nlp.lsi.upc.edu/freeling
5. Del documento CODIMEP-CR: 2001_Rodríguez_Echeverría.
6. En: http://www.scs.leeds.ac.uk/amalgam/amalgam/multi-parsed.html
7. En: http://www.dlsi.ua.es/projectes/3lb/index.html
8. Más información sobre *treebanks* se encuentra en Lüdelin & Kytö (2008), y en particular Wallis (2008) y Nivre (2008).
9. Tomado de: http://www.rae.es/rae/gestores/gespub000019.nsf/voTodosporId/DBC9D1B343D484B0C1257164003C8BFE? OpenDocument. Consultado: 10/12/12.
10. Véase también el sitio antiguo: http://www.lllf.uam.es/~sandoval/UAMTreebank.html
11. Ver detalles en: http://www.cl.uzh.ch/research/paralleltreebanks/smultron_en.html
12. Ubicado en: http://nlp.lsi.upc.edu/freeling/index.php?option=com_content&task=view&id=13&Itemid=42
13. Del documento CODIMEP-CR: 1827_Mora_Fernández.
14. Ubicado en: http://code.google.com/p/phpsyntaxtree/
15. Del documento CODIMEP-CR: 2003_Pacheco_de_la_Espriella.
16. Del documento CODIMEP-CR: 1926_Jiménez_Oreamuno.
17. Del documento CODIMEP-CR: 1909_González_Víquez.
18. Del documento CODIMEP-CR: 1916_González_Flores.
19. Del documento CODIMEP-CR: 1910_González_Víquez.
20. Del documento CODIMEP-CR: 1928_Jiménez_Oreamuno.
21. Del documento CODIMEP-CR: 1873_Guardia_Gutiérrez.

22. Del documento CODIMEP-CR: 1973_Figueres_Ferrer.
23. Del documento CODIMEP-CR: 1994_Calderón_Fournier.
24. Del documento CODIMEP-CR: 1973_Figueres_Ferrer.
25. Del documento CODIMEP-CR: 1988_Arias_Sánchez.
26. Del documento CODIMEP-CR: 1910_González_Víquez.
27. Del documento CODIMEP-CR: 1979_Carazo_Odio.
28. Del documento CODIMEP-CR: 1985_Monge_Álvarez.
29. Del documento CODIMEP-CR: 1928_Jiménez_Oreamuno.
30. Del documento CODIMEP-CR: 1994_Calderón_Fournier.
31. Del documento CODIMEP-CR: 1975_Oduber_Quirós.
32. Del documento CODIMEP-CR: 1935_Jiménez_Oreamuno.
33. Del documento CODIMEP-CR: 1907_González_Víquez.
34. Del documento CODIMEP-CR: 1999_Rodríguez_Echeverría.
35. Del documento CODIMEP-CR: 1925_Jiménez_Oreamuno.
36. Del documento CODIMEP-CR: 1992_Calderón_Fournier.
37. Del documento CODIMEP-CR: 1985_Monge_Álvarez.
38. Del documento CODIMEP-CR: 1990_Arias_Sánchez.

Bibliografía

Abeillé, Anne (Ed.). 2003. *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic Publishers.

Baker, P., A. Hardy y T. McEnery. 2006. *A Glossary of Corpus Linguistics*. Edingburg: Edinburgh University Press.

Göhring, Anne. 2009. "Spanish Expansion of a Parallel Treebank". Lizentiatsarbeit der Philosophischen Fakultät der Universität Zürich. http://www.cl.uzh.ch/studies/theses/lic-master-theses/lizGoehringAnne.pdf.

Jara Murillo, Carla Victoria. 2011. *CODIMEP-CR: Corpus Digital de Mensajes Presidenciales de Costa Rica*. https://sites.google.com/site/mensajepresidencialcr/.

Leech, G. 2004. "Adding Linguistic Information". En: M. Wynne (Ed.). *Developing linguistic corpora. A guide to Good Practice. Arts and Humanities Data Service*. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm.

Leoni de León, Jorge Antonio, Sandra Schwab y Éric Wehrli. 2008. "Análisis sintáctico profundo del español: un ejemplo del procesamiento de secuencias idiomáticas". *Procesamiento de Lenguaje Natural*. 41: 37-44. http://hdl.handle.net/10045/8062.

Lemnitzer, L. y Zinsmeister, H. 2006. *Korpuslinguistik. Eine Einführung*. Tuebingen: Gunter Narr Verlag.

Llisterri, Joaquim. 2012. "Corpus Linguistics and Written Language Resources – Bibliography". http://liceu.uab.es/~joaquim/language_resources/lang_res/biblio_corpus.html.

Lüdeling, A. and M. Kytö. (Eds.). 2008. *Corpus Linguistics: An International Handbook*. Handbücher zur Sprache und Kommunikationswissenschaft series. Berlin: Mouton de Gruyter.

Moreno, Antonio, Susana López y Fernando Sánchez. 2003. "Developing a Syntactic Annotation Scheme and Tools for a Spanish Treebank". En: Anne Abeillé. (Ed.). *Treebanks: Building and Using parsed Corpora*.

Navarro Colorado, F. B. 2007. Metodología, construcción y explotación de corpus anotados semántica y anafóricamente. Tesis doctoral. Universidad de Alicante. http://rua.ua.es/dspace/bitstream/10045/7736/1/tesis_doctoral_francisco_de_borja.pdf

Nivre, Joakim. 2008. "Treebanks". En: A. Lüdeling and M. Kytö (Eds.). *Corpus Linguistics: An International Handbook*. http://stp.lingfil.uu.se/~nivre/docs/hsk.pdf.

Padró, Lluís y Evgeny Stanilovsky. 2012. "FreeLing 3.0: Towards Wider Multilinguality". Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. http://nlp.lsi.upc.edu/freeling/index.php?option=com_content&task=view&id=20&Itemid=49. (Ver además el sitio FreeLing 3.0, http://nlp.lsi.upc.edu/freeling)

Sampson, G. 2003. "Thoughts on Two Decades of Drawing Trees". En: Anne Abeillé. (Ed.). *Treebanks: Building and Using Parsed Corpora*.

Schütze, Hinrich. 1999. *Foundations of statistical natural language processing*. Cambridge: MIT.

Subirats, Carlos y Marc Ortega. 2012. *Corpus del Español Actual*. http://sfncorpora.uab.es/CQPweb/cea/

Wallis, Sean. 2008. "Searching treebanks and other structured corpora". En: A. Lüdeling and M. Kytö (Eds.). *Corpus Linguistics: An International Handbook*.