Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
Programa de Maestría en Computación

Extracción semiautomática de atributos morfológicos de especies a partir de descripciones taxonómicas

Tesis para optar por el grado de
Magíster Scientiae en Computación

María Auxiliadora Mora

Cartago, Costa Rica
Junio, 2016

Aprobación de borrador de tesis

Extracción semiautomática de atributos morfológicos de especies a partir de descripciones taxonómicas

José Enrique Araya, Ph. D.
Profesor Asesor (Visto bueno)

Roberto Cortés Morales, Ph. D.
Director maestría (Refrendo)

Aprobación de tesis

Extracción semiautomática de atributos morfológicos de especies a partir de descripciones taxonómicas

Tribunal examinador

Manuel Vargas Del Valle, M.Sc. Profesional externo	Erick Mata Montero, Ph.D. Profesor
José Enrique Araya, Ph.D. Profesor Asesor	Roberto Cortés Morales, Ph.D. Director maestría (Refrendo)

Junio, 2016

Dedicatoria

A Alcides Víquez, Emilia Víquez Mora y Ricardo Víquez Mora, fuente de motivación constante para alcanzar mis objetivos.

Agradecimientos

Quiero expresar mi más sincero agradecimiento al director de esta investigación, el profesor José Enrique Araya por sus excelentes sugerencias durante la implementación del proyecto y todo el tiempo dedicado.

Gracias a Erick Mata y a Manuel Vargas por estar dispuestos a evaluar y realimentar los resultados de esta investigación.

Gracias a Hong Cui de la Universidad de Arizona, por su disposición a compartir conocimiento y aclarar mis dudas respecto al algoritmo implementado.

Gracias a Nelson Zamora por permitirme utilizar las descripciones de plantas de sus libros Árboles de Costa Rica volumen III, Árboles de Costa Rica volumen IV y el Manual de Plantas de Costa Rica y además, como experto botánico, realimentar el algoritmo de extracción de información.

Y finalmente, agradezco a mis compañeros del Instituto Nacional de Biodiversidad (INBio), los que durante los últimos 23 años, de una u otra forma, contribuyeron a desarrollar mi conocimiento y experiencia en el tema de informática para la biodiversidad.

Extracción semiautomática de atributos morfológicos de especies a partir de descripciones taxonómicas

Resumen

La literatura taxonómica permite documentar la biodiversidad del planeta y acceder al conocimiento necesario para su gestión sostenible. Sin embargo, mucha de la información taxonómica que existe se encuentra en publicaciones científicas en formato de texto. La cantidad de publicaciones generadas es grande lo que hace que procesarla manualmente sea una actividad compleja y muy costosa. La Biblioteca del Patrimonio de la Humanidad (BHL) estima que existe un acumulado de más de 120 millones de páginas distribuidas en 5,4 millones de libros publicados a partir de 1469, más alrededor de 800.000 monografías y 40.000 títulos de revistas (12.500 de estas son actuales) [6].

Es necesario desarrollar estándares y herramientas informáticas para extraer este conocimiento e integrarlo a los repositorios de datos libres existentes y apoyar así el avance de la ciencia, la educación y la conservación de la biodiversidad.

Este documento presenta un algoritmo basado en técnicas de lingüística computacional para extraer información estructurada a partir de las descripciones morfológicas de plantas escritas en español. El algoritmo desarrollado se basa en el trabajo de la Dra. Hong Cui de la Universidad de Arizona y fue aplicado al libro Árboles de Costa Rica volumen III (ACRv3) [8] y a un subconjunto de

descripciones del Manual de Plantas de Costa Rica (MPCR) con resultados muy competitivos (más del 94,1% de rendimiento promedio) anotando estructuras, caracteres, asociando caracteres a estructuras y procesando conjunciones. El software es libre[1], fue desarrollado en Java e integra tecnología existente como la biblioteca de herramientas de procesamiento de lenguaje natural FreeLing [9], la Ontología de Plantas (PO) [10], el Organizador de Términos Ontológicos (OTO) [11] y el Glosario español-inglés, inglés-español para la Flora Mesoamericana [12].

Palabras clave: Extracción de información, Procesamiento de lenguaje natural, Aprendizaje automático, Informática para la biodiversidad.

[1] El código fuente y ejemplos de descripciones estructuradas se encuentran en https://github.com/INBio.

Abstract

Taxonomic literature keeps records of the planet's biodiversity and gives access to the knowledge needed for its sustainable management. Unfortunately, most of the taxonomic information is available in scientific publications in text format. The amount of publications generated is very large; therefore to process it manually is a complex and very expensive activity. The Biodiversity Heritage Library (BHL) estimates that there are more than 120 million of pages published in over 5.4 million of books since 1469, plus about 800,000 monographs and 40,000 journal titles (12,500 of these are current) [6].

It is necessary to develop standards and software tools to extract and integrate this knowledge into existing free and open access repositories to support science, education, and biodiversity conservation.

This document presents an algorithm based on computational linguistics techniques to extract structured information from morphological descriptions of plants written in Spanish. The developed algorithm is based on the work of Dr. Hong Cui from the University of Arizona and was applied to the book Trees of Costa Rica Volume III (ACRv3) [8] and to a subset of descriptions of the Manual of Plants of Costa Rica (MPCR) with very competitive results (more than 94.1% of average performance) extracting structures, characters, associating characters to structures, and processing conjunctions. The implemented tool is free software, was developed using Java, and integrates existing technology as FreeLing [9], the

Plant Ontology (PO) [10], the Ontology Term Organizer (OTO) [11], and the Flora Mesoamericana English-Spanish Glossary [12].

Keywords: Information Extraction, Natural Language Processing, Machine Learning, Biodiversity Informatics.

Índice de contenido

Índice de figuras

Índice de tablas

Lista de acrónimos

ACRv3	Árboles de Costa Rica volumen III / Trees of Costa Rica Volume III
ACRv4	Árboles de Costa Rica volumen IV / Trees of Costa Rica Volume IV
BHL	Biblioteca del Patrimonio de la Humanidad / Biodiversity Heritage Library
Cmartt	Marcado de Tratamientos Taxonómicos en Chino / Chinese Markuper for Taxonomic Treatments
Eagles	Grupo Consultivo de Expertos en Estándares de Ingeniería del Lenguaje / Expert Advisory Group on Language Engineering Standards
EOL	Enciclopedia de la Vida / Encyclopedia of Life
FAT	Encontrar todos los Nombres Taxonómicos / Find All Taxonomic Names
Gate	Arquitectura General para la Ingeniería del Texto / General Architecture for Text Engineering
GBIF	Sistema Mundial de Información sobre Biodiversidad / Global Biodiversity Information Facility
IB	Informática para la Biodiversidad / Biodiversity Informatics
IE	Extracción de Información / Information Extraction
ITCR	Instituto Tecnológico de Costa Rica / Costa Rica Institute of Technology
LGPL	Licencia Pública General Reducida de GNU / GNU Lesser General Public License (LGPL)
Martt	Marcado de tratamientos taxonómicos / Markuper for Taxonomic Treatments
MPCR	Manual de Plantas de Costa Rica / Manual of Plants of Costa Rica
NLP	Procesamiento de Lenguaje Natural / Natural Language Processing
NLTK	Herramienta de Procesamiento de Lenguaje Natural / Natural Language Toolkit
OCR	Reconocimiento Óptico de Caracteres / Optical character recognition
OTO	Organizador de Términos de Ontología / Ontology Term Organizer

PO	Ontología de Plantas / Plant Ontology
POS	Etiquetado Gramatical / Part-of-speech
Post	Etiquetador gramatical / Part-of-speech Tagging
RDF	Marco de Descripción de Recursos / Resource Description Framework
Uima	Arquitectura de Gestión de Información no Estructurada / Unstructured Information Management Architecture
XML	Lenguaje de Marcas Extensible / Extensible Markup Language

Introducción

La trasformación de textos de la literatura taxonómica en datos estructurados continúa siendo un reto fundamental de la informática para la biodiversidad, reconocido así por iniciativas internacionales como el Sistema Mundial de Información sobre Biodiversidad (GBIF), la Enciclopedia de la Vida (EOL) y la Biblioteca del Patrimonio de la Humanidad (BHL) ([1], [2], [3]). Es necesario desarrollar estándares y herramientas informáticas para extraer el conocimiento sobre las especies e integrarlo a los repositorios de datos libres existentes para apoyar el avance de la ciencia, la educación y la conservación de la biodiversidad.

La literatura taxonómica permite documentar la biodiversidad del planeta y generar el conocimiento necesario para su gestión sostenible. Naturalistas europeos iniciaron el proceso de documentar la biodiversidad durante la edad media, proceso que se consolidó con la creación de la Taxonomía Linneana hace más de 250 años. Como resultado, la comunidad científica ha descrito más de 1,9 millones de especies lo que representa alrededor del 17% de la biodiversidad esperada del planeta [4].

El trabajo taxonómico, expresado de forma muy simplificada, consiste en organizar todas las formas de vida idealmente en una jerarquía, a cada taxón asignarle un nombre en latín, una categoría taxonómica que lo asocia a un nivel en la jerarquía, una descripción científica, una descripción diagnóstica que en ocasiones se acompaña de dibujos diagnósticos, la descripción del hábitat, información de la distribución, claves de identificación, entre otra información.

Figura 1: Secciones que documentan una familia y sus géneros en el Manual de Plantas de Costa Rica [5].

Plan del *Manual*

El objetivo de este *Manual* es incluir todas las especies de plantas con semillas definitivamente conocidas de todo el territorio dentro de las actuales fronteras políticas de Costa Rica, incluyendo la Isla del Coco, Isla del Caño y todas las otras islas cercanas a la costa. Para más detalles en cuanto al formato, véase el capítulo "Plan del *Manual*" en Vol. I: Introducción.

Dentro de las Gimnospermas, Monocotiledóneas y Dicotiledóneas (en este orden), las familias (con terminación "aceae"), géneros y especies siguen un orden **alfabético**.

Asteliaceae

M. H. Grayum

FC = Flora costaricensis, FM = Flora mesoamericana

Sumas mundiales

4 gén. y 50 spp., Chile, Bras., África, Australasia, Oceanía; 1 gén. y 1 sp. en CR. *FM* 6: 37 (Lott & García-Mendoza, 1994; sub Agavaceae).

Para cada género se cita la revisión taxonómica más reciente (en caso de que exista).

Cordyline

Fosberg, F. R. 1985. *Cordyline fruticosa* (L.) Chevalier [Agavaceae]. Baileya 22: 180–181.

Se toman en cuenta tanto las especies nativas como las exóticas naturalizadas, así como las exóticas cultivadas a gran escala.

Ca. 15 spp., Bras. (1 sp.), Asia–Australasia y Oceanía; 1 sp. introd. en CR.

Se cita la publicación original solamente para el nombre **aceptado** y su **basónimo**. La sinonimia está reducida.

Cordyline fruticosa (L.) A. Chev., *Jard. bot. Saigon* 66. 1919. *Convallaria fruticosa* L., *Herb. amb.* 16. 1754; *Cordyline terminalis* (L.) Kunth; *Taetsia fruticosa* (L.) Merr.; *T. f.* var. *ferrea* (L.) Standl.

CAÑA INDIA, CORNELINA, GRACENA.

Se indican los nombres comunes de amplio uso en Costa Rica.

Con familias monoespecíficas, se presenta solamente **una descripción** (las de familia, género y especie combinadas). Igualmente, para géneros monoespecíficos las descripciones de género y especie se combinan.

Arbolitos o arbustos ca. 1–3.5 m, escasamente ramificados. Hojas simples, dísticas, densamente agrupadas en los ápices de las ramas; pecíolo ca. 4–20 cm; lámina 14–57 × 3–14 cm, estrecha a ampliamente elíptica, lanceolada u oblanceolada, plana, verde o (frecuentemente) bordeada a fuertemente teñida o variegada con rojo o púrpura. Infl. terminal, una panícula de espigas bracteadas; pedúnculo ca. 5–35 cm. Fls. bisexuales, actinomorfas, virtualmente sésiles, con bractéolas subyacentes ovadas; miembros del perianto (tépalos) similares, ca. 10–18 mm, blancosas a rosados, basalmente connatos; estambres 6, separados, adnatos a la superficie del tubo floral; anteras basifijas; pistilo 1, compuesto; ovario súpero, trilocular; óvulos 2–20 por lóculo; placentación axilar; estilo 1. Fr. una baya, ca. 5–6 mm de diám., rojo brillante, con muchas semillas.

Las medidas de **largo por ancho** siempre se dan en ese formato. En los otros casos, cuando no se especifica la dimensión, se toma por "de largo (alto)". Se presenta la ilustración de por lo menos una especie por cada género nativo o naturalizado.

Las descripciones del hábitat siguen de manera flexible el Sistema de Zonas de Vida de Holdridge

Bosque húmedo y muy húmedo, cult., 0–1350+ m; ambas verts. Fl. ene., abr.–dic. India–Indonesia y Australasia, pero ampliamente introd. en los trópicos de todo el mundo. *Grayum et al.* 9498 (MO).

Para cada especie que se presenta formalmente, se cita una muestra de herbario recolectada en Costa Rica.

Los **herbarios** se indican según el *Index herbariorum*.

Distinta en su hábito arborescente y hojas pecioladas, por lo general brillantemente coloreadas; *C. fruticosa* se planta como ornamental y a veces para marcar linderos de propiedad. Por esto puede encontrarse en bosque ± primario, donde se le confunde como una sp. nativa.

Los datos de distribución se dan en el siguiente orden: Hábitat, rango de elevación dentro de Costa Rica; rango geográfico dentro de Costa Rica. Meses de floración en Costa Rica. Rango geográfico completo.

Después de cada descripción de familia, género y especie sigue una discusión breve. El objetivo principal de estas notas es presentar una "**pista**" para reconocer el taxón que se tiene en la mano

xiv

El trabajo se da a conocer por medio de publicaciones científicas en revistas arbitradas, manuales taxonómicos y guías de campo, entre otros productos. La figura 1 describe las secciones típicas que documentan una familia de plantas y sus géneros. El ejemplo fue tomado del Manual de Plantas de Costa Rica [5].

La cantidad de publicaciones generadas es muy grande lo que hace que procesarla manualmente sea una actividad compleja y muy costosa. BHL estima que existe un acumulado de más de 120 millones de páginas distribuidas en 5,4 millones de libros publicados a partir de 1469, más alrededor de 800.000 monografías y 40.000 títulos de revistas (12.500 de estas son actuales) [6]. Este acervo de información es utilizado por ecologistas, taxónomos, filo-genetistas, paleontólogos, médicos, gestores de los recursos naturales, ingenieros agrónomos, químicos, microbiólogos y la academia, entre otros, por ejemplo para: continuar generando conocimiento sobre la biodiversidad; establecer áreas silvestres protegidas efectivas para conservarla; identificar y combatir plagas de cultivos, especies invasoras y vectores de enfermedades; investigar sobre nuevos compuestos químicos útiles para la industria farmacéutica, cosmética, agrícola, entre otros muchos usos.

La informática para la biodiversidad (IB) brinda las técnicas y mecanismos para capturar, procesar, integrar y publicar datos e información sobre la biodiversidad del planeta. Iniciativas internacionales de IB, como GBIF, EOL, BHL y el Código de Barras de la Vida, trabajan en el descubrimiento, agregación, e intercambio libre y gratuito de datos genéticos, de presencia de especies, historia natural,

estado de conservación, manejo y conservación y datos geográficos, entre otros. Los datos integrados han permitido responder preguntas que tienen que ver con procesos que ocurren en el tiempo y el espacio, por ejemplo, los posibles efectos del cambio climático en especies particulares, efectos del cambio de uso del suelo en especies de una zona, predicción de posibles rutas de invasión de una especie, entre otros. A los tipos de datos antes mencionados, se han sumado más recientemente, las bases de datos de rasgos o características (*traits*) de las especies, almacenadas en forma de tripletas, extraídas de forma manual o semiautomática de textos de descripciones morfológicas, hábitat, historia natural, interacciones entre especies, distribución, entre otra información. Un ejemplo de estos repositorios es el TraitBank [7]**,** diseñado por EOL para integrar datos de múltiples bases de datos. Actualmente, TraitBank integra 50 recursos de datos con más de 11 millones de tripletas relacionadas con 330 caracteres que describen 1,7 millones de taxones.

El presente proyecto se ubica en el área de extracción de información (IE por sus siglas en inglés) y tiene como objetivo estructurar semiautomáticamente caracteres morfológicos de especies de plantas descritas en español, por medio de técnicas de análisis semántico, ontologías y un repositorio de conocimiento adquirido a partir del contenido de las mismas descripciones. Aunque muchos enfoques de IE han sido aplicados a documentos con información taxonómica, muy pocos han sido orientados a estructurar el contenido completo de la descripción morfológica de las especies y ningún esfuerzo documentado ha sido dirigido a este tipo de información pero escrita en español.

El algoritmo desarrollado se basa en el trabajo de la Dra. Hong Cui de la Universidad de Arizona y fue aplicado al libro Árboles de Costa Rica volumen III (ACRv3) [8] y a un subconjunto de descripciones del Manual de Plantas de Costa Rica (MPCR) con resultados muy competitivos (más del 94,1% de rendimiento promedio anotando estructuras, caracteres, asociando caracteres a estructuras y procesando conjunciones). El sistema fue desarrollado en Java e integra tecnología existente como la biblioteca de herramientas de procesamiento de lenguaje natural FreeLing [9], la Ontología de Plantas (PO) [10], el Organizador de Términos Ontológicos (OTO) [11] y el Glosario español-inglés, inglés-español para Flora Mesoamericana [12].

El resto de este documento está estructurado de la siguiente forma, el capítulo uno presenta los antecedentes e incluye el avance logrado a la fecha por la comunidad internacional de informática para la biodiversidad en la estructuración de información a partir de descripciones taxonómicas. El capítulo dos introduce el problema, presenta los objetivos y el alcance del proyecto. El capítulo tres detalla la metodología propuesta y describe el método de evaluación. El capítulo cuatro presenta los resultados y discusión. El capítulo cinco las conclusiones y recomendaciones. Al final del documento se adjuntan apéndices que incluyen una descripción del esquema de datos utilizado, el modelo de objetos, el modelo entidad relación, el proceso detallado de anotación semántica de las descripciones, entre otra información.

Capítulo 1: Antecedentes

1.1. Marco teórico

1.1.1. Extracción de información (IE)

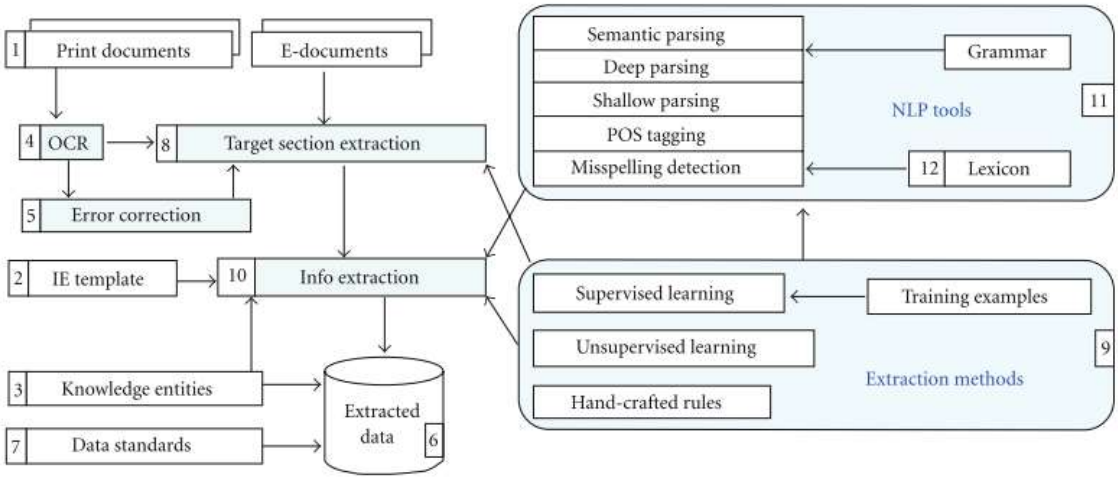
El área de IE desarrolla algoritmos para analizar el contenido de grandes volúmenes de texto no estructurado o semi-estructurado buscando documentar tipos predefinidos de eventos, entidades y relaciones. Su objetivo principal es identificar, recoger y normalizar información relevante para un usuario particular a partir de texto. La información se normaliza utilizando una representación estructurada, por ejemplo, una plantilla o un esquema. La IE utiliza herramientas de procesamiento de lenguaje natural (NLP), inteligencia artificial y aprendizaje automático, entre otras tecnologías.

1.1.2. Arquitectura de referencia de un sistema de IE

La figura 2 presenta una arquitectura de referencia de un sistema de extracción de información propuesta por Thessen, Cui y Mozzherin [13].

El color gris en los rectángulos indica que el tema es parte del proceso de IE. En la etapa de análisis y diseño del sistema se especifica claramente el objetivo de extracción y a partir de este, se define la plantilla de extracción (número 2 en el diagrama); se realiza un estudio del conocimiento disponible (3) fundamental para complementar el proceso de extracción (i.e. ontologías, vocabularios controlados, glosarios), los estándares existentes (7), los métodos y la tecnología a utilizar.

Figura 2. Arquitectura de referencia de un sistema de extracción de información.



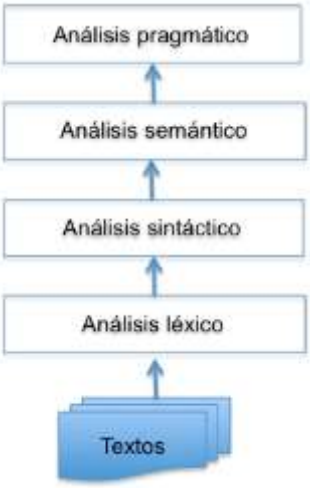
Fuente: Thessen, Cui y Mozzherin [13].

Las entradas del sistema pueden estar en formato electrónico o impreso (1). Si los documentos se obtienen en formato impreso se procesan antes con tecnología OCR (Optical Character Recognition) (4) y luego se corrigen los errores (5) generados durante el proceso de digitalización. Si es necesario, el sistema debe segmentar las secciones de interés dentro de los documentos (8) para esto se utilizan métodos de extracción para identificarlas. En etapas previas al proceso de extracción de información, frecuentemente se utilizan técnicas de NLP para, por ejemplo, etiquetar componentes del texto y generar una representación de este (i.e. un árbol) con más facilidades de procesamiento (11). Los métodos de extracción para anotar el texto pueden ser supervisados, semi-supervisados, no supervisados o basados en reglas (9). Entre los mecanismos de persistencia más utilizados están las base de datos, los documentos XML (Lenguaje de marcas extensible) o RDF (Marco de descripción de recursos) (6).

1.1.3. Técnicas de procesamiento de lenguaje natural (NLP)

El NLP es un área de investigación que abarca un conjunto de técnicas para la generación, manipulación y análisis del lenguaje natural. Aunque la mayoría de las técnicas son heredadas de la Lingüística y la Inteligencia Artificial, también han sido influenciadas por áreas relativamente nuevas como el Aprendizaje Automático, la Estadística Computacional y la Ciencia Cognitiva [14].

Figura 3. Etapas del análisis en NLP desde el punto de vista lingüístico.



Las técnicas de NLP pueden ser abordadas desde el campo de la lingüística (este enfoque es llamado simbólico) y se aplican de forma secuencial (en capas complementarias) como se muestra en la figura 3.

- **El análisis léxico**: Se centra en la forma en que las palabras son construidas a partir de pequeñas unidades de significado. De acuerdo a Hippisley [15], son tareas básicas del análisis léxico:
    - Tokenización: Consiste en la segmentación de una hilera en palabras o tokens individuales. Un token es una hilera de caracteres continuos formada por letras, números y símbolos, que se encuentran entre dos espacios o signos de puntuación.
    - Lematización: Consiste en asociar a un lema todas sus variantes morfológicas. Un lema es la palabra aceptada como representante de todas las formas flexionadas asociadas a esta (i.e. plurales, gerundios, femenino, entre otros). Ejemplo: lema = rojo; formas flexionadas= rojas, rojos, rojitos.
    - Asignación de etiquetas POS (part-of-speech tagging): Tarea que consiste en asociar una etiqueta a cada palabra de acuerdo al rol que juega en la oración, por ejemplo nombre, adjetivo, adverbio, determinante. Las etiquetas incluyen más información por ejemplo el género y número de la palabra.
- **Análisis sintáctico**: Por medio de este se procesa una hilera (típicamente una oración) para determinar su descripción estructural de acuerdo a una gramática [16]. El análisis sintáctico no tiene un fin como tal, es más bien un paso intermedio que genera una estructura (comúnmente jerárquica) que facilita el análisis semántico. El análisis puede realizarse de forma superficial (análisis parcial) o profundo (análisis completo utilizando

diferentes gramáticas y selección del árbol más adecuado para representar la oración).

- **Análisis semántico**: El fin último del NLP es entender el lenguaje, es decir, incorporar información provista por medio de este en una base de conocimiento o ejecutar una acción en respuesta a este [17]. Aplicaciones del análisis semántico pueden ser IE, recuperación de información, resúmenes de textos, minería, traducciones, entre otras.
- El **análisis pragmático** se refiere a cómo el contexto contribuye al significado. Este tema está fuera del alcance de esta investigación.

Las técnicas de NLP también pueden ser abordadas desde el análisis estadístico del lenguaje (este enfoque es llamado empírico). El NLP simbólico tiende a trabajar de arriba hacia abajo mediante la imposición de patrones gramaticales conocidos y asociaciones de significado a los textos. El NLP empírico trabaja de abajo hacia arriba a partir de los textos, buscando patrones y asociaciones para generar modelos, algunos de las cuales pueden no corresponder a relaciones sintácticas o semánticas. El enfoque empírico ha probado ser útil en el manejo de incertidumbre, portabilidad y robustez de los algoritmos.

1.1.4. Métodos de extracción de información

Los métodos de extracción de información puede estar basados en aprendizaje supervisado, semi-supervisado, no supervisado o reglas.

El aprendizaje supervisado requiere datos etiquetados (ejemplos de entrenamiento) para entrenar modelos que luego son aplicados a los procesos de extracción. Estos métodos son muy efectivos pero tienen la desventaja que se requiere mucho tiempo para generar los ejemplos de entrenamiento apropiados. Los algoritmos no supervisados requieren poco o ningún ejemplo de entrenamiento pero no son tan efectivos como los supervisados. El aprendizaje semi-supervisado busca formas innovadoras de reducir la cantidad de ejemplos requeridos para el entrenamiento procurando mantener la efectividad de los métodos. Estos algoritmos reciben datos no etiquetados y una cantidad mínima de datos etiquetados y si el proceso es exitoso, tienen un rendimiento comparable con los algoritmos de aprendizaje supervisados a un bajo costo.

Steven Abney en [18] presenta los cinco tipos más importantes de problemas en aprendizaje automático. Los cuatro primeros tienen que ver con la estimación de una función f(x) y están agrupados de acuerdo a si el algoritmo es supervisado o no supervisado y si la variable a predecir tiene un valor continuo o discreto.

En el extremo izquierdo de los algoritmos supervisados se encuentran los **algoritmos de clasificación** (reciben de entrada valores discretos). La meta para estos algoritmos consiste en asignar los valores de entrada a clases predeterminadas; en el extremo derecho se encuentran los **algoritmos de *Clustering*** que al igual que los anteriores realizan actividades de clasificación pero no supervisadas y con valores discretos, la meta es agrupar valores en clases no predeterminadas; el tercer tipo de algoritmo son los de **regresión** que

utilizan métodos supervisados pero difieren de los algoritmos de clasificación en que la función a ser aprendida recibe un valor continuo; en el cuarto tipo de algoritmo están los de **estimación de densidad** que utilizan métodos no supervisados con valor de entrada continuo; el quinto tipo es el **aprendizaje por refuerzo**, el objetivo de estos algoritmos es aprender a traducir situaciones en acciones para maximizar una señal de recompensa. Este problema difiere de los anteriores en que hay una entrada continua y en que la supervisión es indirecta.

Los **métodos basados en reglas** incorporan el conocimiento experto en un tema particular en forma de un conjunto de reglas. De acuerdo a Charniak y McDermott [19] una regla es una pieza de conocimiento que contribuye a solucionar los requerimientos de un grupo de usuarios. Las reglas tiene la forma: If <circunstancias> then <acciones>.

1.1.5. Pruebas y evaluación

Las métricas generalmente usadas en IE para evaluar los resultados son precisión y cobertura (*precision and recall*). Estas métricas miden el porcentaje de anotaciones correctas y lo completo del método de extracción, respectivamente. Además, se utiliza la medida F, que aplica la media armónica ponderada entre la precisión y la cobertura [20].

Las fórmulas se listan a continuación:

- Precisión = $\frac{\text{Número de instancias correctamente identificadas}}{\text{Total de instancias identificadas}}$

- Cobertura = $\frac{\text{Número de instancias correctamente identificadas}}{\text{Total de instancias correctas}}$

- F = $\frac{(b^2 + 1)\,(\text{Precisión} * \text{Cobertura})}{b^2 * (\text{Precisión} + \text{Cobertura})}$

  b representa la importancia relativa entre la precisión y la cobertura. Si b=1 ambas medidas tiene igual importancia.

1.1.6. Selección de la muestra:

Para la selección de la muestra se utilizó el algoritmo de la **rueda de la ruleta** [21] con el objetivo de darle más prioridad a las cláusulas con mayor cantidad de estructuras (indicador de complejidad).

Este algoritmo es uno de los más utilizado en el área de algoritmos genéticos para seleccionar individuos con una probabilidad proporcional a su aptitud para sobrevivir, es decir, los que se desempeñan mejor en un ambiente particular tienen mayor probabilidad de ser seleccionados. En la rueda de la ruleta, cada individuo tiene una rebanada del círculo proporcional a su aptitud, es decir, cada individuo tiene una cantidad de números de la rifa proporcional a su desempeño. La rueda se hace girar N veces, con N = al tamaño de la muestra a seleccionar. El algoritmo se implementa de la siguiente forma:

- Se suma el valor total esperado de los individuos de la población (total = T).
- Se repite N veces:
  - Generar un entero aleatorio ‘r’ entre 0 y T.

- Sumar todos los valores esperados de los individuos, hasta que la suma sea mayor o igual a ‘r’. El individuo cuyo valor esperado coloca la suma sobre este límite es el seleccionado.

1.2. Trabajo relacionado

En los últimos años, varios métodos de extracción de información, incluyendo modelos probabilísticos, reglas generadas manualmente y autómatas, han sido aplicados en el campo de la informática para la biodiversidad para extraer información a partir de descripciones taxonómicas con diversos grados de automatización. En esta sección se presentan las tecnologías aplicadas de acuerdo al énfasis de la extracción. Un panorama muy completo fue publicado en el 2012 por Thessen y otros en [13]:

1. **Reconocimiento de nombres de entidades** (en este caso nombres taxonómicos): La asignación de nombres taxonómicos es una actividad regida por los Códigos Internacionales de Nomenclatura para Algas, Hongos y Plantas y Nomenclatura Zoológica. Estos códigos incluyen un conjunto de reglas para definir los nombres en todos los niveles de la jerarquía taxonómica. Por ejemplo, a nivel de división los nombre botánicos deben finalizar con el sufijo phyta (ej. Magnoliophyta), a nivel de clase finalizan con “opsida” (ej. Magnoliopsida), orden “ales” (ej. Fabales), familia “aceae” (ej. Fabaceae). Este conjunto de reglas facilita el proceso de anotación de nombres, por lo que esta

es la actividad de extracción de información en la que más se ha avanzado y el enfoque más utilizado es aplicar reglas nomenclaturales y diccionarios.

Entre las tecnologías que utilizan este enfoque están: TaxonFinder, se utiliza para identificar taxones a todos los niveles de la jerarquía taxonómica [22]; *Find All Taxonomic Names* (FAT) utiliza reglas y lógica difusa para mejorar la efectividad del sistema [23]; TaxonTagger identifica, anota, y extrae nombres científicos de páginas web y documentos PDF. Utiliza los servicios web del buscador nombre de GBIF como diccionario [24]; Linnaeus utiliza un autómata de estado finito determinista que selecciona los nombres de especies que coinciden con la palabra buscada y un conjunto de heurísticas para resolver menciones ambiguas [25].

Entre los sistemas que utilizan modelos probabilísticos para reconocer nombres de entidades están: NetiNeti que utiliza reglas nomenclaturales y aprendizaje probabilístico automático (Naive Bayes) para clasificar los nombres basados en las características estructurales y características derivadas del contexto [26].

Ejemplo de uso de los nombres de entidades: BHL tiene el objetivo de integrar y publicar por medio de Internet un corpus grande de publicaciones científicas asociadas a la biodiversidad formado principalmente por publicaciones históricas. BHL utiliza los servicios de TaxonFinder para luego de digitalizar (escanear) los volúmenes y procesarlos con OCR, marcar los nombres científicos que aparecen en cada página. El sistema permite que los usuarios hagan búsquedas por nombre científico y sus sinónimos, y devuelve todas las

páginas digitalizadas que contienen la información buscada [27].

2. **Estructuración de textos completos** (i.e. manuales y guías de campo) para anotar de forma semiautomática las secciones que los componen, por ejemplo en los manuales de plantas (se anota el nombre científico, los sinónimos, la descripción morfológica, la descripción diagnóstica, la distribución, claves, entre otras secciones). Ejemplos:
    - GoldenGATE [28] es un editor XML que apoya el marcado de texto automático con corrección manual. El proceso utiliza expresiones regulares y diccionarios para etiquetar descripciones taxonómicas. El marcado automático detecta correctamente los nombres y tratamientos taxonómicos. El sistema integra diferentes herramientas de procesamiento del lenguaje natural como Gate (General Architecture for Text Engineering) [29].
    - Curry y Connor [30] propusieron un sistema para estructurar documentos usando heurísticas para reconocer las secciones basadas en el estilo del texto, organización y puntuación. Usando información conocida sobre los documentos, por ejemplo que el nombre del taxón siempre se presenta en itálica (en caso de niveles sub-específicos), seguido por el nombre del autor en mayúsculas, luego la fecha y la página de la cita separados por comas.
    - Araya en [31] utilizó un enfoque similar al de Curry y Connor combinado con pre-procesamiento manual para estructurar los textos de 200 especies de árboles descritas en el Manual de Plantas de Costa Rica.

3. **Extracción de características morfológicas**: en esta área se ubica el presente proyecto. Ejemplos relevantes se listan a continuación en orden cronológico inverso:

   - Chinese Markuper for Taxonomic Treatments – Cmartt (2013): aplica y extiende las ideas de la Dra. Cui implementadas en Charaparser y fue aplicado a documentos escritos en chino. [32]
   - Charparser (2012): La aplicación se desarrolló para procesar descripciones en Inglés. Implementa un algoritmo de aprendizaje no supervisado utilizando Bootstrapping para anotar a nivel de cláusulas, análisis sintáctico, un integrador de ontologías (Ontology Terms Organizer – OTO) y heurísticas para anotar descripciones morfológicas completas [33].
   - Phenex (2010) desarrollado por Phenoscape, utiliza una ontología para anotar descripciones de fenotipos de organismos (cualquier característica observable de un organismo, como su morfología, desarrollo y comportamiento). Phenex es aplicado a documentación en inglés [34]. Recientemente el proyecto Phenoscape está utilizando también Charparser.
   - Markuper for Taxonomic Treatments - MARTT (2005) aplicado a descripciones en inglés, el sistema fue desarrollado basado en métodos de aprendizaje semi-automático inductivo y reforzado con reglas aprendidas durante el proceso. El sistema anota las descripciones a nivel de cláusulas [35].

- MultiFlora (2004) utiliza expresiones regulares, una ontología y la herramienta de procesamiento de lenguaje natural Gate. Para las pruebas se utilizaron descripciones en inglés [36].
- X-Tract y Terminator (1999) utilizan heurísticas y un diccionario de términos para estructurar el contenido de descripciones morfológicas a nivel de estructuras de organismos. Ambos se han aplicado a descripciones en inglés [37] [38].

Adicionalmente, la comunidad de informática para la biodiversidad ha mostrado interés en estructurar textos asociados a distribución, relaciones inter-específicas, hábitat y comportamiento de las especies. Con este fin, iniciativas internacionales apoyan proyectos de investigación en esta línea, un ejemplo de esto es el proyecto “EOL Rubenstein Fellows” [39].

Capítulo 2: Definición del problema y contribución

2.1. Definición del problema

Si bien, la información taxonómica publicada es cuantiosa, acceder a esta es complejo ya que mucha se encuentra en formato de texto no estructurado, dispersa en múltiples libros y revistas, lo que la hace difícil de procesar de forma automática para integrarla con otros tipos de información y generar nuevos productos. El problema que se plantea resolver con la presente investigación, consiste en generar semiautomáticamente información estructurada con alto valor semántico a partir de descripciones morfológicas de plantas.

Antes de describir en detalle el problema es necesario definir algunos conceptos como:

- **Descripción morfológica**: Las descripciones morfológicas documentan la apariencia física (forma y estructura) de las especies y pueden incluir estructuras, subestructuras, caracteres, estados y relaciones entre estructuras. Los caracteres permiten documentar los rasgos propios de las especies y están asociados a las estructuras o subestructuras que componen al individuo o espécimen. Ejemplo de estructuras son las hojas, el ápice, las flores o los frutos. Ejemplos de caracteres son largo, ancho, color, olor o arquitectura. Un ejemplo de parte de una descripción es la frase "hojas simples". En este caso la estructura que se describe es la hoja, el carácter es la arquitectura (no mencionado) y el estado de este carácter es "simples". Usualmente, las descripciones no incluyen el nombre del carácter que documentan por lo que se

requieren ontologías y vocabularios controlados para dado un estado seleccionar el o los caracteres a los que este podría corresponder. Una descripción completa de la especie *Quercus bumelioides* (libro ACRv4) se presenta a continuación:

> "Árbol pequeño a muy grande, 11-45 m de altura; tronco con la corteza escamosa, grisácea; ramitas con la corteza pálida y lenticelada; estípulas hasta 1,2 cm de largo, liguladas y cuando secas de color pardo. Hojas simples, alternas, 5-17 × 2,1-7,5 cm, elípticas, lanceolado-elípticas a obovadas o redondeadas, ápice redondeado, obtuso o agudo, base obtusa, redondeada o sucordada, glabras en el haz y esparcido-estrellado-pubescentes o con tricomas simples a lo largo de la vena central en el envés, margen entero; pecíolos 0,2-0,8 cm de largo. Inflorescencias masculinas y femeninas en amentos, de 4-10(-14) cm de largo, las femeninas más cortas. **Flores amarillentas, 4-6 mm de largo, sin sépalos ni pétalos y con 4-8(-12) estambres**. Frutos nueces, 2,2-2,8 × 1,8-2,2(-2,5) cm, ovoides, apiculadas en el ápice; rodeadas en la base por una cúpula o receptáculo leñoso y escamoso de 0,7-1,8 × 2-3 cm, campanuladas, grisáceas a pardo oscuro."

- **Cláusula**: Las descripciones morfológicas están formadas por oraciones o cláusulas que, en el caso de especies de plantas describen una estructura de forma completa. Las cláusulas contienen estructuras, subestructuras y los caracteres que documentan a estas. Una cláusula típica tomada de la descripción de *Quercus bumelioides* es "Flores amarillentas, 4-6 mm de largo, sin sépalos ni pétalos y con 4-8 (-12) estambres." En descripciones botánicas, las cláusulas inician casi siempre con el nombre de la estructura principal que documentan, en el ejemplo anterior "Flores".

El algoritmo propuesto generará información estructurada para describir estructuras (i.e. hoja, flor, tallo, fruto, entre otras) y subestructuras (i.e. ápice, pétalo, base, sépalo, eje, panícula, entre otras) por medio de caracteres (i.e. color,

forma, largo, arquitectura, pubescencia, entre otros) y estados (i.e. rojo, oblongo, 50 cm de largo, liguladas, tomentulosas, entre otros) extraídos a partir de las descripciones morfológicas. El nombre de los caracteres por lo general no se menciona en las descripciones y no se pretende que el algoritmo los infiera de estas. Para seleccionar el nombre del carácter apropiado se utilizará la Ontología de Plantas (PO) por medio de OTO [10]. Por ejemplo, el algoritmo aplicado a parte de la cláusula que describe las hojas de *Quercus bumelioides* "Hojas simples, alternas, 5-17 × 2,1-7,5 cm, elípticas, lanceolado-elípticas a obovadas o redondeadas…" generará la siguiente información estructurada para describir las hojas:

Carácter	**Estado**
architecture	simples
arrangement	alternas
length	5-17 cm (el algoritmo debe seccionar el rango, en este caso from=5 y to=17).
width	2,1-7,5 cm (el algoritmo debe seccionar el rango.)
shape	elípticas
shape	lanceolado-elípticas a obovadas
shape	redondeadas

Las descripciones morfológicas en general se caracterizan por:

- Emplear muchas abreviaturas y omitir palabras funcionales y verbos, haciendo que las oraciones se conviertan en frases telegráficas para ahorrar espacio en publicaciones científicas y guías de campo. Por ejemplo, la descripción de

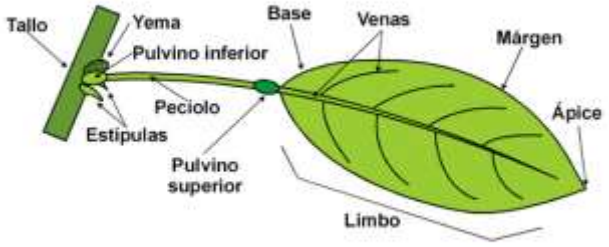
Persea donnell-smithii tomada del libro ACRv4 se presenta como sigue:

> Árbol pequeño, hasta 10 m de altura. Hojas simples, alternas, 8-18 × 4-8 cm, elípticas a elíptico-obovadas, agudas en el ápice, base obtusa, pardo-tomentosa en el envés, con 5-8 venas secundarias por lado, margen entero; pecíolos 2,5-3,5 cm de largo. Inflorescencias en panículas, hasta 18 cm de largo, con muchas flores. Flores doradas en apariencia por la pubescencia de los tépalos, 5-6 mm de largo. Frutos bayas, ca. 1,2 cm de largo.

En este texto “ca.” corresponde a “aproximadamente”.

- Los textos se presentan en un lenguaje muy técnico debido a que la terminología formal está basada en el latín.
- Los caracteres y estados varían mucho entre grupos biológicos (i.e. plantas, artrópodos, hongos y vertebrados). La figura 4 muestra algunas de las estructuras que permiten describir las hojas de las plantas.

Figura 4. Algunas de las estructuras que permiten describir las hojas de las plantas.[2]



- Los caracteres, en la mayoría de los casos, no son explícitos, por ejemplo la frase “Flores blancas” no hace mención del carácter “color”. Para asignar un

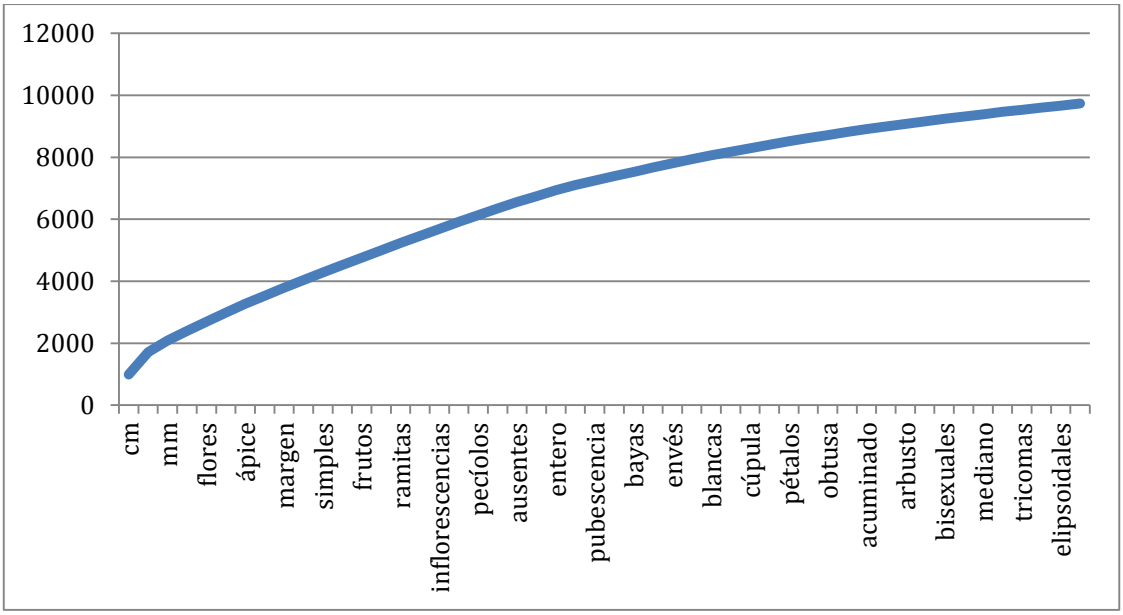
[2] Fuente: Permacultura México. En http://www.permacultura.org.mx/es/botanica/plantas/partes/hoja/.

nombre de carácter se utilizan ontologías y vocabularios controlados. El inconveniente con estas herramientas es que no siguen estándares generales, cada equipo de trabajo crea su propio conjunto de caracteres y vocabulario controlado para los posibles estados; son muy especializadas, es decir, se crean para grupos pequeños de taxones con ciertas excepciones, como el grupo de las plantas, para el que existen ontologías que abarcan todo el reino; muy pocas de estas están disponibles en español; su uso para buscar el carácter asociado a un estado puede ocasionar ambigüedad en el resultado, por ejemplo, en la cláusula "hojas simples , alternas , de (10) 1321 x 47 (8) cm , elípticas…" para el estado "elípticas" existen dos opciones de carácter en OTO "shape" y "arrangement".

- Las descripciones de plantas están formadas mayormente por nombres, adjetivos, números (medidas) y en menor cantidad adverbios. Los verbos se utilizan con muy poca frecuencia. Pueden incluir negaciones como "rara vez", "solo en ocasiones" o "sin". Ejemplo de ACRv4:

> Hojas simples, alternas, 7-15 × 3,5-9 cm, obovadas, **rara vez** elípticas, ápice obtuso, base cuneada, glabras y con el envés glauco, venas secundarias 7-10 pares, venas terciarias conspicuas, margen entero;

Figura 5. Frecuencia acumulativa de las 50 palabras más utilizadas en el libro de Árboles de Costa Rica volumen IV, estás representan alrededor de un 30% de las palabras. Solo se consideraron adjetivos, adverbios y nombres.



- El vocabulario utilizado es repetitivo. La figura 5 muestra un gráfico de frecuencia acumulativa de las 50 palabras más utilizadas en las descripciones morfológicas del libro ACRv4 teniendo en cuenta solamente nombres, adjetivos y adverbios. Estas representan alrededor de un 30% de todas las palabras utilizadas.

- Las descripciones morfológicas utilizan una sintaxis altamente estandarizada a pesar de que están escritas en lenguaje natural. La tabla 1 presenta ejemplos de tipos de frases utilizadas en las descripciones de libro ACRv4 y el número de repeticiones del patrón POS (*Part-of-speech*) en todo el conjunto de descripciones. En total, el documento contiene descripciones de 240 especies, las que contabilizan 2.457 cláusulas (la descripción de cómo se

segmentaron las cláusula se encuentra en la sección de metodología).

- En algunos casos, la estructura del texto de la descripción puede alejarse un poco del estándar y ser más parecidas a texto en lenguaje natural. Por ejemplo:

> Inflorescencias masculinas y femeninas en amentos, las masculinas 4,5-7 cm de largo, las femeninas mucho más cortas.

Ejemplo	POS a nivel de cláusula (una letra por token)	Cantidad cláusulas que presentan el POS	%
inflorescencias fasciculadas ,	Nombre+ Adjetivo + Signo de puntuación (NAF)	991	40.3
estípulas lineares , persistentes .	Nombre + Adjetivo + Signo de puntuación + Adjetivo + Signo de puntuación (NAFAF)	265	10.8
estambres 16-20 .	NZF	240	9.8
pecíolos 0,5-1,6 cm de largo .	NZNSRF	152	6.2
sépalos con glándulas ;	NSNF	110	4.5
árbol pequeño a grande , 8-35 m de altura ;	NASAFZNSNF	89	3.6
cúpula 0,3-1 cm de altura .	NZNSNF	50	2.0
flores blancas , 2-3 cm de diámetro .	NAFZNSNF	50	2.0
pecíolos hasta 1 cm de largo .	NSZNSRF	45	1.8
árbol pequeño o arbusto , 2-7 m de altura ;	NACNFZNSNF	42	1.7
flores blancas (bisexuales) , 8-10 mm de diámetro .	NAFNFFZNSNF	40	1.6
inflorescencias en panículas , 2,5-10 cm de largo .	NSNFZNSRF	37	1.5
frutos bayas , 1,5-2,5 cm de largo , elipsoidales ;	NAFZNSAFAF	25	1.0
inflorescencias paniculado-cimosas , 8-11 cm de largo .	NAFZNSRF	25	1.0
pecíolos 0,4-2,2 (-2,8) cm de largo .	NZFZFNSRF	24	1.0

Tabla 1. Ejemplos de tipos de cláusulas utilizadas en las descripciones morfológicas del libro ACRv4 (total de cláusulas = 2,457).

Nota: La segunda columna de la tabla incluye los marcadores POS (*part-of-speech*) de las cláusulas calculados concatenando el POS de cada *token* (N=Nombres, F=Signos de puntuación, A=Adjetivos, S=Preposiciones, Z=Numerales, C=Conjunciones, R=Adverbios, D=Determinantes, P=Pronombres)

2.2. Objetivo general

Desarrollar un algoritmo para extraer semiautomáticamente estructuras y caracteres morfológicos de especies de plantas descritas en español, presentes en guías de campo y manuales de flora, por medio de herramientas de análisis morfosintáctico, ontologías y un repositorio de conocimiento adquirido a partir del contenido de las mismas descripciones.

2.3. Objetivos específicos

1. Definir e implementar un algoritmo para extraer información de especies contenida en descripciones en español utilizando tecnología existente como analizadores morfosintácticos, ontologías y glosarios, entre otras herramientas.

2. Probar el algoritmo aplicándolo a descripciones científicas de las 240 especies de árboles contenidos en el libro de Árboles de Costa Rica volumen IV.

3. Evaluar el algoritmo utilizando el volumen III de la serie Árboles de Costa Rica.

4. Generar una base de conocimiento a partir de los conceptos aprendidos durante el proceso que constituya un lexicón base para continuar extrayendo información de descripciones morfológicas de plantas.

5. Proponer una generalización del algoritmo para solucionar el problema de extracción de conocimiento disponible en español, a partir de descripciones

morfológicas de plantas. Probar la generalización en un conjunto de descripciones del Manual de Plantas de Costa Rica.

2.4. Justificación del proyecto desde el punto de vista de innovación, impacto y profundidad

a. Innovación: El problema de extracción de conocimiento a partir de literatura taxonómica no ha sido aun resuelto. Como se mencionó anteriormente, los aportes realizados a la fecha por la comunidad internacional se han enfocado a tareas como reconocimiento de nombres taxonómicos (la mayor cantidad de investigaciones están asociadas con esta tarea), estructuración de textos completos para anotar de forma semiautomática las secciones que los componen y poca investigación se ha generado asociada a la extracción de características morfológicas de las especies a partir de descripciones taxonómicas. La tecnología descrita ha sido aplicada por lo general a documentos en inglés, ninguno de los proyectos publicados ha trabajado con documentación en español.

b. Impacto: El sistema implementado permitirá definir las bases para continuar el trabajo de procesar más de cien guías de campos de plantas y otros grupos biológicos como artrópodos, moluscos, vertebrados y hongos publicadas por la Editorial INBio y en un futuro apoyar a la comunidad latinoamericana en este proceso. El procesar esta información manualmente constituye un trabajo monumental y algoritmos como el propuesto en este proyecto alivianan en un alto porcentaje el esfuerzo requerido.

El proyecto permitió estructurar el conocimiento de 709 especies de plantas. En total se extrajeron 13.249 estructuras/subestructuras con 18.459 caracteres (46 caracteres diferentes). El algoritmo implementado es escalable, es decir puede ser aplicado a especies de plantas en general. Para documentar la escalabilidad, el algoritmo fue probado en un subconjunto de descripciones de especies del volumen VI del Manual de Plantas de Costa Ricas que incluyó plantas acuáticas, árboles, arbustos, epífitas, hierbas y lianas, entre otros tipos de plantas con un rendimiento muy bueno de 94,1% en promedio anotando estructuras, caracteres, asociando caracteres a estructuras y procesando conjunciones. La herramienta desarrollada es software libre por lo que la comunidad iberoamericana de informática para la biodiversidad podrá continuar el desarrollo de la tecnología para aplicarla a descripciones de otros grupos biológicos.

La información de descripciones morfológicas estructurada puede utilizarse para generar productos como:

a. Descripciones en lenguaje natural para usar en monografías, registros de especies o productos orientados a públicos meta con requerimientos diferentes.

b. Herramientas de apoyo en la identificación de especímenes (ej. claves electrónicas): las claves electrónicas se utilizan para apoyar a usuarios no expertos en la identificación de especímenes. El proceso de preparar una clave involucra definir una lista de caracteres que aplican al conjunto de

especies a documentar y asignar un estado a cada carácter por especie, es decir, la tarea consiste en llenar una matriz cuyas filas corresponden a las especies y la columnas a los caracteres que las describen. El preparar esta matriz es muy laborioso para un científico pero si la información existe de forma estructurada, un sistema de información permitiría simplificar el proceso. Existen herramientas que permiten cargar la matriz generada y por medio de una interfaz simple apoyan a usuarios en el proceso de identificación de especímenes. Un ejemplo de este tipo de herramientas es LucID [40].

c. Evaluación de calidad de manuscritos: información científica estructurada podría ser utilizada para realizar control de calidad de contenidos en información generada de forma no tan rigurosa, por ejemplo, podría ser utilizado para control de calidad en registros de especies.

d. Integración de la información a iniciativas como el TraitBank de EOL, GBIF y BHL para desarrollar mecanismos de búsquedas de información.

e. Análisis de datos de especies que tienen caracteres particulares para responder preguntas específicas, como por ejemplo, las planteadas por el proyecto “EOL Rubenstein Fellows” en [39]:

    - ¿Qué especies ocurren en zonas urbanas y suburbanas, y cuál es el subconjunto de las especies que prosperan en estas zonas (por ejemplo, adaptadores / explotadores urbanos o especies sinúrbicas o que se benefician de los humanos)? ¿Cuáles son los caracteres de las especies que prosperan en ambientes urbanos y suburbanos?

- ¿Cuáles son los rasgos físicos o de comportamiento más importantes asociados con la vulnerabilidad al cambio climático en los organismos ya evaluados por la Unión Internacional para la Conservación de la Naturaleza (UICN), y pueden estos resultados ser extendidos a organismos que aún no han sido evaluados? ¿Cuál es la velocidad relativa de la evolución en esos rasgos que son importantes para la adaptación de las especies al cambio global?
- ¿Es más probable que ocurra la coloración azul en especies de plantas y animales a gran altitud o a poca profundidad?
- ¿Cuáles son las características de las plantas y los animales que los hacen más susceptibles de ser registrados por el público general?

c. Profundidad: El algoritmo implementado permite extraer el conocimiento de las especies descritas en guías de campo y manuales de plantas en español tales como el libro de ACRv4, ACRv3 y el MPCR. Se procesó el texto completo de las descripciones lo que permitió generar información estructurada con alto valor semántico haciendo explícita la mayor parte de la información de las descripciones. El sistema genera estructuras, caracteres que describen las estructuras y relaciones entre estructuras. Para cada uno de estos objetos el sistema documenta el nombre, el valor, modificadores, restricciones, identifica el término y la ontología utilizada, entre otros atributos.

2.5. Alcance

El principal resultado de este proyecto es un algoritmo que permite estructurar

descripciones morfológicas de plantas escritas en español.

Figura 6. Parte del esquema estándar utilizado para dar formato a las descripciones morfológicas de plantas. El esquema fue propuesto por la Dra. Cui.

```
284         <xs:complexType name="biological_entity">
285                 <xs:sequence>
286                         <xs:element minOccurs="0" maxOccurs="unbounded" type="character" nam
287                 </xs:sequence>
288                 <xs:attribute name="alter_name"/>
289                 <xs:attribute name="constraint"/>
290                 <xs:attribute name="constraintid" type="xs:NCName"/>
291                 <xs:attribute name="geographical_constraint"/>
292                 <xs:attribute name="id" use="required" type="xs:ID"/>
293                 <xs:attribute name="in_brackets" type="xs:boolean"/>
294                 <xs:attribute name="name" use="required"/>
295                 <xs:attribute name="parallelism_constraint" type="xs:NCName"/>
296                 <xs:attribute name="taxon_constraint"/>
297                 <xs:attribute name="ontologyid" type="xs:string"/>
298                 <xs:attribute name="provenance" type="xs:string"/>
299                 <xs:attribute name="notes" type="xs:string"/>
300                 <xs:attribute name="name_original" type="xs:string"/>
301                 <xs:attribute name="type" type="biological_entity_type" use="required"/>
302         </xs:complexType>
303
304         <xs:complexType name="character">
305                 <xs:attribute name="char_type" type="xs:NCName"/>
306                 <xs:attribute name="constraint"/>
307                 <xs:attribute name="constraintid"/>
308                 <xs:attribute name="from"/>
309                 <xs:attribute name="from_inclusive" type="xs:boolean"/>
310                 <xs:attribute name="from_unit" type="xs:NCName"/>
311                 <xs:attribute name="geographical_constraint"/>
312                 <xs:attribute name="in_brackets" type="xs:boolean"/>
313                 <xs:attribute name="modifier"/>
314                 <xs:attribute name="name"/>
315                 <xs:attribute name="organ_constraint"/>
316                 <xs:attribute name="other_constraint"/>
317                 <xs:attribute name="parallelism_constraint" type="xs:NCName"/>
318                 <xs:attribute name="taxon_constraint"/>
319                 <xs:attribute name="to"/>
```

El sistema recibe las descripciones morfológicas en formato tabular (i.e. nombre científico y descripción), las procesa y genera como salida documentos en XML de acuerdo al esquema propuesto por la Dra. Cui [33][3]. El esquema permite documentar estructuras, caracteres y relaciones entre las estructuras. Una

[3] La última versión del esquema está disponible en https://github.com/biosemantics/schemas/blob/master/semanticMarkupOutput.xsd.

porción del esquema se presenta en la figura 6.

El procesamiento se realiza de forma semiautomática, se requiere muy poca intervención del usuario y esta está enfocada a las siguientes actividades:

- Traducir manualmente los adjetivos que no se encuentran en el Glosario español-inglés, inglés-español para la Flora Mesoamericana [12].
- Evaluar el contenido de la base de conocimiento. La evaluación consiste en verificar que las estructuras, estados de carácter y otros conceptos hayan sido tipificados bien en la base de conocimiento y verificar que las traducciones al inglés coincidan con los términos ontológicos disponibles en OTO.

El algoritmo de estructuración fue aplicado a las descripciones contenidas en los libros ACRv3, ACRv4 y MPCR.

Capítulo 3: Metodología

3.1. Introducción

El proyecto utiliza técnicas de procesamiento de lenguaje natural, reglas morfo-sintácticas y ontologías para extraer caracteres morfológicos de especies de plantas a partir de descripciones científicas en español. El algoritmo construye sobre tecnología existente como la biblioteca de NLP Freeling, el Organizador de Términos de Ontología (OTO), la Ontología de Plantas (PO) y el Glosario inglés-español, español-inglés para la Flora Mesoamericana (descritos en la sección 3.2).

El objetivo principal es convertir las descripciones, que inicialmente se encuentran en formato de texto, en registros de una base de datos. La meta de extracción es identificar estructuras, subestructuras, estado de los caracteres, restricciones, relacionar los caracteres con la estructura/subestructura que corresponde y determinar relaciones entre estructuras.

El algoritmo no procesa todos los sintagmas[4] preposicionales ni verbales, sin embargo, como una prueba de concepto y de cara a un refinamiento posterior, se estructuraron los sintagmas preposicionales que inician con los tokens “sin“ o “con“. El resto de sintagmas preposicionales o verbales sólo son delimitados como *constraint_preposition* y *constraint_verb* respectivamente. De los sintagmas no estructurados el algoritmo extrae únicamente las estructuras o subestructuras presentes, para asegurarse que estas se toman en cuenta en el proceso de

[4] Un sintagma es un grupo de palabras que tienen una función determinada en la oración.

asociación de los siguientes caracteres de la cláusula.

Figura 7. a) Ejemplo de estructuración de una cláusula parte de la descripción de la especie *Quercus salicifolia* del libro ACRv4. b) Gráfico que ilustra el proceso de asociación de caracteres a estructuras utilizando la concordancia en género y número entre tokens.

Figura 7a

Cláusula original: hojas simples , alternas , elípticas , ápice acuminado , caudado o agudo , base caudada u obtusa , glabras o a veces con tricomas dispersos a lo largo de la vena central por el envés ;

Chunks: [hojas simples] , [alternas] , [elípticas] , [ápice acuminado] , [caudado o agudo] , [base caudada u obtusa] , [glabras o a veces con tricomas dispersos a lo largo de la vena central por el envés] ;

Anotación en XML

```
<statement id="T10L5">
    <biological_entity id="T10L5S1-95456" name="hojas" type "structure">
        <character name="architecture" value="simples"/>
        <character name="arrangement" value="alternas"/>
        <character name="shape" value="elípticas" notes="Carácter repetido"/>
        <character name="arrangement" value="elípticas" notes="Carácter repetido"/>
        <character name="pubescence" value="glabras" constraint_conjunction="o"
         constraint_preposition="a veces con tricomas dispersos a lo largo de la vena central por el envés" />
    </biological_entity>
    <biological_entity id="T10L5S4-95457" name="ápice" type="structure">
        <character name="shape" value="acuminado"/>
        <character name="shape" value="caudado" constraint_conjunction="o"/>
        <character name="shape" value="agudo"/>
    </biological_entity>
    <biological_entity id="T10L5S6-95458" name="base" type= "structure">
        <character name="shape" value="caudada" constraint_conjunction="u"/>
        <character name="shape" value="obtusa"/>
    </biological_entity>
    <biological_entity id="T10L5S7-95459" name="tricomas" type= "structure">
    ....
</statement>
```

Figura 7b.

La figura 7a muestra un ejemplo de estructuración de una cláusula del libro ACRv4. En este ejemplo, la frase “hojas simples“ genera un objeto de tipo *biological_entity* para la estructura “hojas“ y un objeto de tipo *character* para el carácter “architecture” con estado “simples“. El nombre del carácter “architecture“ se toma de la PO debido a que comúnmente estos componentes de la descripción no aparecen. La PO se utiliza a través de OTO para dado un estado, buscar los caracteres que tengan definido en su vocabulario controlado este estado. Los caracteres que cumplen esta condición pueden ser varios, en el ejemplo anterior, el estado “elípticas” está asociado a “arrangement“ y a “shape“. El sistema agrega una nota de “Carácter repetido“ para que en una etapa posterior sea el experto el que determine el carácter que corresponde.

El algoritmo debe asociar cada carácter con la estructura o subestructura a la que corresponde utilizando el orden de aparición de los tokens y su concordancia en género y número. La figura 7b muestra para la misma descripción, las entidades biológicas reconocidas (en rojo), los estados de los caracteres (en verde) y la relación entre entidades biológicas y caracteres (flecha). En este ejemplo, los caracteres “simples“, “alternas“ y “elípticas” deben ser asociados a la estructura “hojas”; “acuminado” y “caudado o agudo“ a la subestructura “ápice”; y “caudada u obtusa” a la estructura “base”; en cambio “glabras“ debe ser asociado a la estructura “hojas“ porque no coincide en género y número con las subestructuras más cercanas. La estructura “tricomas“ es parte de un sintagma preposicional, sin embargo, el sistema la extrae para que esté disponible en el proceso de asociación de los siguientes caracteres (si existen).

El algoritmo analiza el rol de cada token y las dependencias entre tokens en una cláusula y crea o modifica, a partir de esos roles, objetos que pueden ser estructuras (*bilogical_entity* de acuerdo al esquema), caracteres (*character*) o relaciones (*relation*). Una descripción detallada del análisis se encuentra en el apéndice IV.

El esquema de datos utilizado para la estructuración fue propuesto por Cui en [33][5]. El esquema incluye secciones asociadas a metadatos generales como información sobre el documento del cual se extraen las descripciones, la persona que realizó la extracción, los recursos utilizados (i.e. software, ontologías, servicios), entre otros. El esquema define además conceptos que permiten estructurar la taxonomía superior y conceptos asociados a las descripciones morfológicas como *biological_entity*, *character* y *relation*. El apéndice I contiene una descripción detallada de los conceptos del modelo de datos, el apéndice II presenta el modelo de objetos y apéndice III el diagrama entidad-relación.

El algoritmo propuesto fue implementado utilizando Java porque facilitaba la integración de la tecnología seleccionada descrita en la sección 3.2. El sistema almacena los resultados de la estructuración en una base de datos PostgreSQL lo que permite presentarlos en muchos formatos, en esta investigación se presentan en XML utilizando el esquema propuesto por Cui.

[5] La versión actualizada del documento está disponible en https://github.com/biosemantics/schemas/blob/master/semanticMarkupOutput.xsd.

3.2. Descripción del sistema

La figura 8 presenta el diagrama de flujo del algoritmo y la figura 9 un diagrama entidad-relación simplificado. La entrada inicial del sistema está compuesta por documentos en formato tabular que contienen las descripciones morfológicas y los nombres científicos de las especies a procesar. Estas descripciones se importan en la base de datos en la tabla TAXON_DESCRIPTION (un registro por descripción).

Figura 8. Diagrama de flujo del algoritmo implementado.

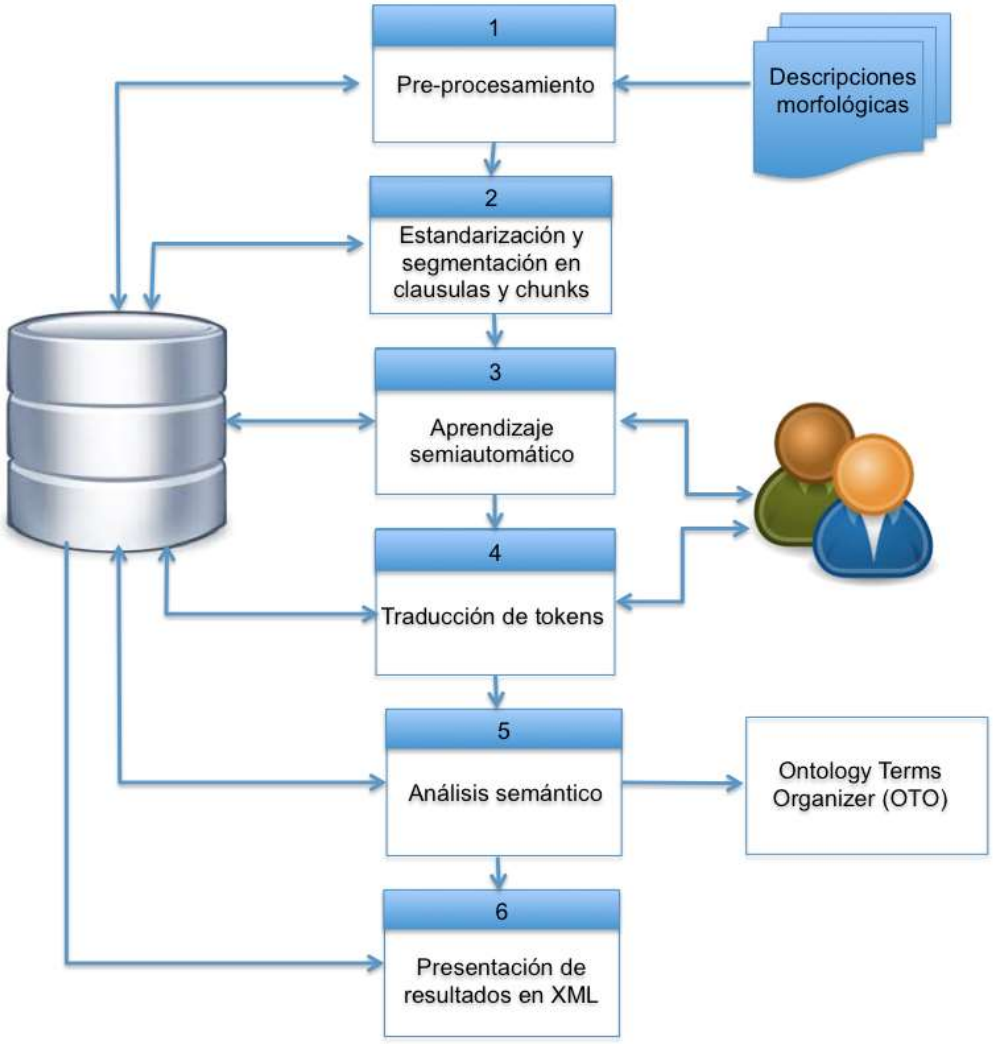
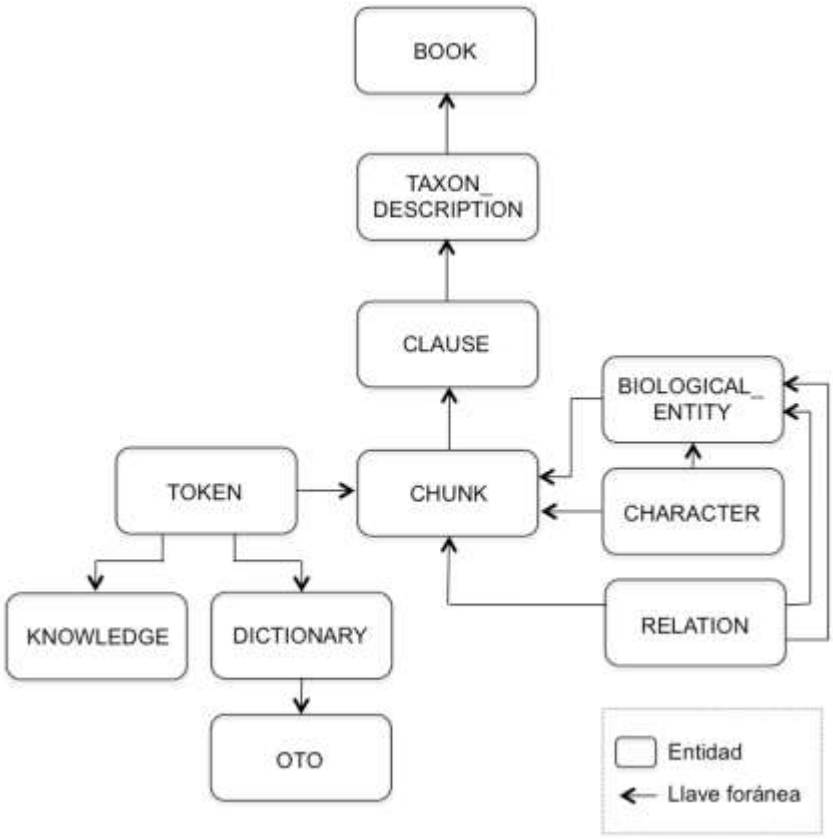


Figura 9. Diagrama entidad-relación simplificado.



El algoritmo consta de las siguientes etapas:

Etapa- 1. Pre-procesamiento de los textos de las descripciones morfológicas.

Etapa- 2. Estandarización y segmentación de las descripciones en cláusulas y chunks.

Etapa- 3. Aprendizaje semiautomático de nombres de estructuras y estado de los caracteres.

Etapa- 4. Traducción de tokens al inglés para que coincidan con entradas en la PO.

Etapa- 5. Anotación semántica de las descripciones.

Etapa- 6. Generación de resultados en XML.

Etapa-1. Pre-procesamiento de los textos de las descripciones morfológicas.

El pre-procesamiento de los textos de las descripciones morfológicas consiste en:

- Eliminar de las descripciones comillas dobles y simples.
- Agregar un punto al final de todas las descripciones (en caso de que no lo tengan), requerido por Freeling para segmentar los textos en cláusulas.

Etapa-2. Estandarización y segmentación de las descripciones en cláusulas y chunks.

Las descripciones morfológicas se segmentan en cláusulas utilizando como separador el punto, los dos puntos y el punto y coma. Las cláusulas se construyen de forma estandarizada a partir de los tokens que son parte de estas, todos escritos en letra minúscula y separados por un espacio. La tabla CLAUSE contiene las cláusulas asociadas a cada descripción. Cada cláusula debe iniciar con un nombre de estructura para que sea procesada por el algoritmo.

Con el objetivo de simplificar las hileras a analizar y corregir errores en árboles de dependencia generados por Freeling, cada cláusula a su vez es segmentada en chunks utilizando como separador la coma. Por ejemplo, la siguiente cláusula (código T8L5) genera 12 chunks como se presentan en la tabla 2.

T8L5: hojas simples , alternas , (8,5) 14,5-33 × (4) 6-14 cm , oblongas a obovadas , ápice redondeado , obtuso a abrupto-acuminado , base redondeada , obtusa , truncada o levemente subcordada , glabras en el haz y con una pubescencia tomentosa sedosa con tricomas fasciculados en el envés , margen entero , crenado o distalmente denticulado ;

contents
hojas simples
alternas
(8,5) 14,5-33 × (4) 6-14 cm
oblongas a obovadas
ápice redondeado
obtuso a abrupto-acuminado
base redondeada
obtusa
truncada o levemente subcordada
glabras en el haz y con una pubescencia tomentosa sedosa con tricomas fasciculados en el envés
margen entero
crenado o distalmente denticulado ;

Tabla 2. Lista de chunks generados a partir de la cláusula T8L5 de la descripción de *Quercus insignis* del libro ACRv4.

El proceso de estandarización de chunks realiza lo siguiente:

- Elimina los espacios entre guiones y números. Los números y guiones deben estar siempre contiguos, por ejemplo debe ser (-8) y no (- 8). Para eliminar los espacios se utilizan expresiones regulares.
- Elimina guiones entre un número y el paréntesis. Es decir los casos como (6-) deben ser reemplazados como (6). Los números entre paréntesis son utilizados en las descripciones de plantas para documentar rangos atípicos (explicados en el apéndice IV).

Etapa-3. Aprendizaje semiautomático de nombres de estructuras y estado de los caracteres.

La base de conocimiento es clave para el buen funcionamiento del sistema porque en ella se tipifican todas las estructuras, estados de caracteres y otro tipo de conocimiento del área de aplicación. El contenido de la base de conocimiento se utiliza para corregir el rol asignado por Freeling a los diferentes tokens.

La base de conocimiento se actualiza durante el proceso de aprendizaje del sistema, el usuario debe al final del proceso, verificar que el conocimiento adquirido por el sistema sea correcto. Al finalizar el proceso el usuario realiza las siguientes tareas:

- Seleccionar el o los tokens que se utilizan en la definición de áreas en un libro particular, por ejemplo, en "(8,5) 14,5-33 × (4) 6-14 cm" se debe tipificar el token "×" como indicador de área (type = G).
- Seleccionar los nombres o adjetivos que pueden actuar como modificadores de una estructura. Un modificador, de acuerdo a Cui en [33], delimita el conjunto de objetos a los que aplican los caracteres y estados. Ejemplo: en la frase "ramitas jóvenes," el adjetivo jóvenes actúa como modificador ya que restringe el ámbito de las "ramitas" a solo las "jóvenes".
- Verificar que el conocimiento adquirido haya sido tipificado correctamente por el sistema.

La tabla KNOWLEDGE mantiene el conocimiento adquirido por el sistema. Los tipos de conocimiento que pueden ser incluidos se listan a continuación:

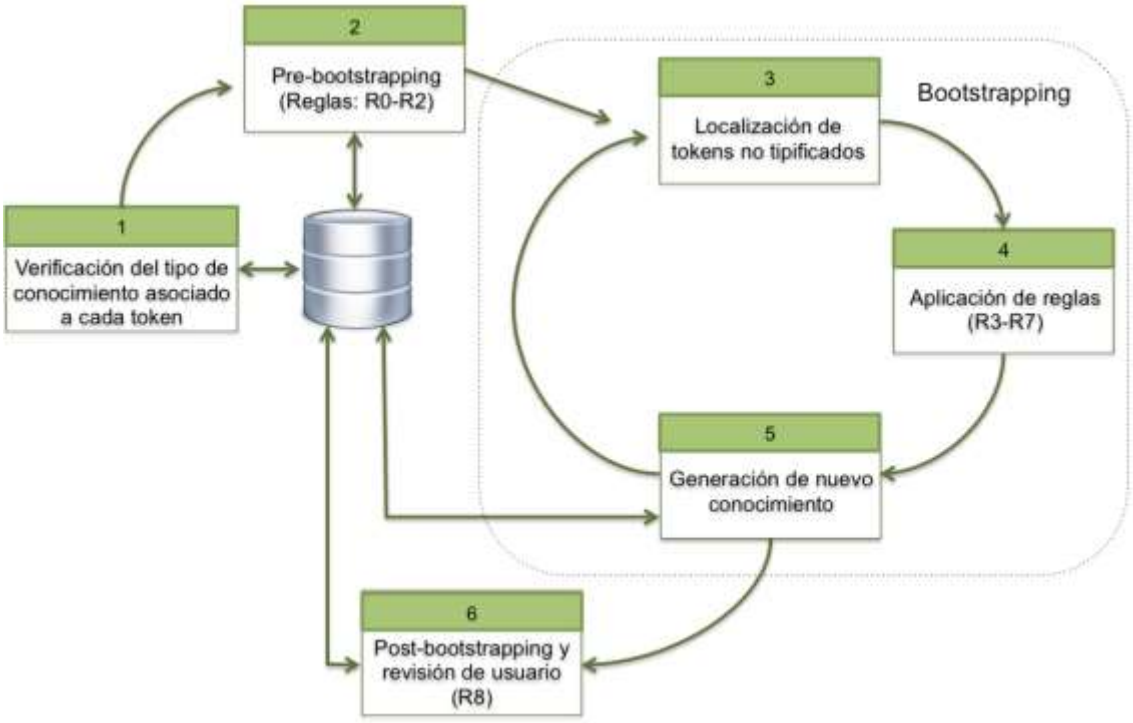
- E = Estructura o subestructura.
- A = Estado de un carácter, en muchos casos es un adjetivo. Por ejemplo “redondeado”, “obtuso” o “abrupto-acuminado”.
- M = Modificadores de estructura. Ejemplos: “adultas”, “jóvenes”, “femeninas” o “masculinas”.
- U = Unidad de medida. Por ejemplo “cm.”, “mm.” o “m.”
- G = Área. Ejemplo “x” o “por”.
- V=Verbo. Ejemplo: “cubiertas”, “descritos” o “son”.
- T = Nombre de carácter. Ejemplo “altura”, “color” o “grosor”.
- R = Restricciones de estructura o caracteres. Por lo general son adverbios como “frecuentemente” o “longitudinalmente”.

Para realizar exitosamente la estructuración de los textos se requiere que el resultado del proceso de tokenización y asignación de etiquetas POS realizado por Freeling sea muy bueno. Sin embargo, debido al lenguaje semi-estructurado de las descripciones, Freeling asigna correctamente etiquetas POS solo a adverbios, determinantes, pronombres, conjunciones, preposiciones, numerales y fechas. Las etiquetas POS asignadas a verbos, nombres y adjetivos deben ser revisadas y en algunos casos corregidas. El realizar este proceso de revisión manualmente consume mucho tiempo por lo que se implementó un algoritmo de aprendizaje muy simple basado en reglas. El algoritmo utiliza *bootstrapping* para implementar un proceso de aprendizaje incremental sin ejemplos.

Bootstrapping

El *bootstrapping* es un proceso de aprendizaje iterativo e incremental en el que en cada ciclo se integra nuevo conocimiento a partir del conocimiento existente [42]. El flujo de datos durante el proceso de aprendizaje se muestra en la figura 10.

Figura 10. Flujo de datos para la generación de conocimiento por medio de *bootstrapping*.



El algoritmo toma la salida del proceso de tokenización y para cada token verifica en la base de conocimiento el tipo que le corresponde utilizando el conocimiento adquirido previamente (etapa 1 en el dibujo). Luego aplica las reglas de *pre-bootstrapping (2)* que se listan a continuación a los tokens no tipificados:

- R0 → Todos los token etiquetados como adjetivos por Freeling se incluyen como estado de carácter en la base de conocimiento (type = A). Además, todos los tokens que poseen un guión en medio de dos o más palabras son incluidos como estado de carácter, verificando antes que alguna de las palabras sea un adjetivo. Ejemplo: “diminuto-ferrugíneo-puberulentas”.
- R1 → Para todos los tokens que inician una cláusula, si la etiqueta POS asignada por Freeling es de tipo nombre, se incluyen en la base de conocimiento como estructuras. Ejemplo: “flores verdoso amarillento con algo de rojizo”. En este caso el token “flores“ se incluye en la base de conocimiento como estructura (type = E).
- R2 → Todos los verbos en forma de participio (terminados en ado, edo, ido, to, so, cho, su forma femenina y plural) se incluyen como estado de carácter (type = A). Ejemplo “ramitas con una pubescencia de pelos estrellados“. En este caso “estrellado“ es tipificado como estado de carácter.

Luego de aplicar las reglas de pre-bootstrapping el algoritmo aplica las reglas de aprendizaje (bootstrapping) a los tokens no tipificados (etapas 3, 4 y 5 de la figura 10):

- R3 → Dos tokens separados por la preposición “a” si al menos uno de ellos es un estado de carácter (type= A) el otro se etiqueta también como estado de carácter. Solo se toman en cuenta tokens etiquetados por Freeling como verbo, nombre o adjetivo. Ejemplo: “redondeada a cordada a

subcordada“. En este caso el token “redondeada“ fue etiquetado en R2 con type = A por lo que en R3 “cordada“ y “subcordada”, que inicialmente Freeling etiquetó como nombres, son etiquetados con type = A.

- R4 → Dos tokens separados por una conjunción “o” o “u” si al menos uno de ellos es un estado de carácter (type = A) el otro se etiqueta también como estado de carácter. Solo se toman en cuenta tokens etiquetados por Freeling como verbo o adjetivo. Ejemplo: “retuso o corto-acuminado“. En este caso en R0 se asignó type = A a “corto-acuminado” por lo que con R4 se asigna a “retuso“ type = A.
- R5 → Dos palabras con el mismo lema (palabra aceptada como representante de todas las formas flexionadas asociadas a esta) se etiquetan con el mismo tipo.
- R6 → Si un token fue etiquetado por Freeling como nombre y está seguido de un estado de carácter (type = A) implica que el primer nombre es una estructura (type = E). Ejemplo: “pubescencia fina blanquecina o no visible a simple vista“. En este caso “fina“ fue etiquetado por Freeling como adjetivo y “pubescencia“ como nombre por lo que con R6 se asigna a “pubescencia“ el type = E.
- R7 → Dos tokens etiquetados por Freeling como nombres separados por una conjunción “o” o “u” si uno de ellos es una estructura (type = E) el otro también. Ejemplo: “con una punta o ápice de 1-1,5 mm“. En este caso tanto “ápice“ como “punta” fueron etiquetados como nombres por Freeling.

“ápice” fue etiquetado con type = E con R1 en “ápice acuminado“ por lo que en R7 se etiqueta “punta“ con type = E.

Cuando el algoritmo después de un ciclo, no genera nuevo conocimiento el proceso finaliza y se aplican las reglas de post-bootstrapping (6) que se listan a continuación:

- R8 → Los tokens restantes se incluyen en la base de conocimiento de la siguiente forma:
    - Si el token fue etiquetado por Freeling como nombre se incluye como estructura.
    - Si el token fue etiquetado por Freeling como verbo se incluye como verbo.

Luego de ejecutar la regla número ocho (R8) un usuario calificado debe revisar los tipos asignados al conocimiento incluido en la base de datos.

Etapa-4. Traducción de tokens al inglés.

Las ontologías son un recurso muy valioso y limitado en el campo de la IB y entre las que están disponibles, no muchas manejan traducciones de términos al español. La PO incluye traducciones al español en forma de sinónimos, sin embargo, en esta investigación se decidió utilizar OTO debido a la gran ventaja que tiene de ser un agregador de diferentes ontologías. Sin embargo, OTO no integra sinónimos por lo que, para hacer coincidir los estados de caracteres con los incluidos en OTO fue necesario traducir los términos utilizando en primera

instancia *Google Translator* (con un 82,5% de éxito en la traducción de los términos de las descripciones del libro ACRv4) y luego se integró al sistema el Glosario inglés-español, español-inglés para la Flora Mesoamericana. Al finalizar el proceso, un usuario debe traducir manualmente los términos no traducidos.

Etapa-5. Anotación semántica de las descripciones.

Las reglas y condiciones que guían el proceso de extracción se definieron a partir del análisis morfosintáctico realizado a las estructuras gramaticales más utilizadas en las descripciones del libro ACRv4. La tabla 3 presenta la estructura gramatical de los tipos de chunks más comúnmente utilizados. Los 34 tipos de chunks en la tabla representan el 73,72% de todos los chunks del libro ACRv4. La columna "POS corregido" incluye una representación de los árboles de relaciones construidos por Freeling, corregidos luego con la base de conocimiento. Las etiquetas se construyen concatenando el tipo asignado al token de cada nodo. El árbol se recorre en profundidad primero.

Los chunks que forman una cláusula son procesados en orden de izquierda a derecha. El orden es importante porque no todos los chunks incluyen la estructura a la que se debe asociar los caracteres por lo que el algoritmo debe buscar esta en los chunks procesados previamente.

Chunk (ejemplo)	POS corregido (obtenido al recorrer el árbol en profundidad primero*)	Cantidad de chunks	%
margen entero	Estructura + Estado de carácter (EA)	1152	17,05
deciduas	Estado de carácter (A)	828	12,25
6-30 m de altura	Unidad de medida + número + preposición + carácter (UZST)	613	9,07
6-30 × 2-10,5 cm	ZGUZ	289	4,28
frutos bayas	EE	214	3,17
pecíolos 0,2-1 cm de largo	EUZST	212	3,14
estambres 62-80	EZ	192	2,84
ovadas a lanceolado-oblongas	ASA	146	2,16
ápice obtuso a corto-acuminado	EASA	132	1,95
axilares o subterminales	CAA	114	1,69
inflorescencias en panículas	ESE	108	1,60
árbol pequeño a mediano	HASA	92	1,36
de 3-4 mm de largo	SUZST	91	1,35
flores blancas (bisexuales)	EAAFF	75	1,11
pecíolos hasta 10 mm de largo	ESUZST	58	0,86
margen glandular-crenado o glandular-aserrado	ECAA	58	0,86
glabras en el haz	ASED	56	0,83
ampliamente elípticas	AR	53	0,78
árbol pequeño o arbusto	CHAH	51	0,75
con 10-13 venas secundarias por lado	SEZMSE	43	0,64
con 6-10 glándulas	SEZ	41	0,61
ramitas usualmente glabras	EAR	41	0,61
5-8 (-10) mm de largo	ZZFFUST	39	0,58
lanceoladas o lanceolado-oblongas	ACA	39	0,58
inflorescencias en panículas cimosas	ESEA	32	0,47
corteza fibrosa y fuerte	EACA	29	0,43
árbol mediano	HA	29	0,43
pecíolos 0,8-1,1 (-1,3) cm de largo	EZZFFUST	27	0,40
con la superficie reticulado-perforada	SEDA	25	0,37
abrupto-acuminadas a largo-	ASASED	21	0,31

Chunk (ejemplo)	POS corregido (obtenido al recorrer el árbol en profundidad primero*)	Cantidad de chunks	%
acuminadas en el ápice			
aproximadamente 7 mm de diámetro	RUZST	21	0,31
flores verde amarillento	EAA	21	0,31
árbol grande a muy grande	HASRA	20	0,30
amarillentas cuando maduras	ACM	20	0,30
Total		**4982**	**73,72%**

Tabla 3: Ejemplos de tipos de chunks presentes en las descripciones morfológicas del libro ACRv4 (total de chunks = 6758)

*Nota: E=estructuras, A=Adjetivos, S=Preposiciones, Z=Numerales, C=Conjunciones, R=Adverbios, D=Determinantes, P=Pronombres, M=modificador, T=atributo, F=signo de puntuación, U= unidad de medida, G=área.

El algoritmo debe generar tres tipos de objetos: estructuras, caracteres y relaciones. Los tokens que no generan estructuras, caracteres o relaciones modifican a alguno de estos tipos de objetos. Por ejemplo, los adverbios si bien no instancian un objeto, modifican la estructura o carácter con la cual tienen una relación de dependencia. Por ejemplo, en el chunk “semillas aladas dorsalmente“ el adverbio dorsalmente debe modificar el carácter “aladas“. El texto estructurado para este ejemplo se presenta en la figura 11.

Figura 11. Texto estructurado para el chunk “semillas aladas dorsalmente“

```
−<biological_entity id="T12L8S1-69357" name="semillas" type="structure">
    <character name="architecture" value="aladas" constraint="dorsalmente"/>
  </biological_entity>
```

El algoritmo analiza las dependencias entre tokens a partir del árbol de dependencia generado por Freeling, la base de conocimiento y un conjunto de

condiciones relacionadas con la posición del token en el árbol y nodos aledaños (ancestro, hermanos e hijos).

El algoritmo recorre el árbol de dependencia de cada chunk en profundidad primero y para cada nodo instancia o modifica los objetos correspondientes. La tabla 4 presenta los chunks generados para la cláusula "inflorescencias paniculado-cimosas, 5-23 cm de largo , ejes densamente estrigulosos ." (código T173L6).

Código	Chunk
T173L6S1	inflorescencias paniculado-cimosas
T173L6S2	5-23 cm de largo
T173L6S3	ejes densamente estrigulosos

Tabla 4. Chunks generados para la cláusula "inflorescencias paniculado-cimosas, 5-23 cm de largo , ejes densamente estrigulosos ." (código T173L6).

Para cada uno de los chunks se genera el árbol de dependencia como se muestran de forma simplificada en la figura 12. Las letras dentro del círculo corresponden al tipo de token en la base de conocimiento (E=estructuras, A=Adjetivos, S=Preposiciones, Z=Numerales, C=Conjunciones, R=Adverbios, D=Determinantes, P=Pronombres, M=modificador, T=Atributo, F=Signo de puntuación, U= Unidad de medida y G=Área).

Figura 12. Árboles de dependencia simplificados generados para los chunks de la cláusula "inflorescencias paniculado-cimosas, 5-23 cm de largo , ejes densamente estrigulosos ." (código T173L6). Los árboles utilizan la simbología de la base de conocimiento para nombrar los nodos.

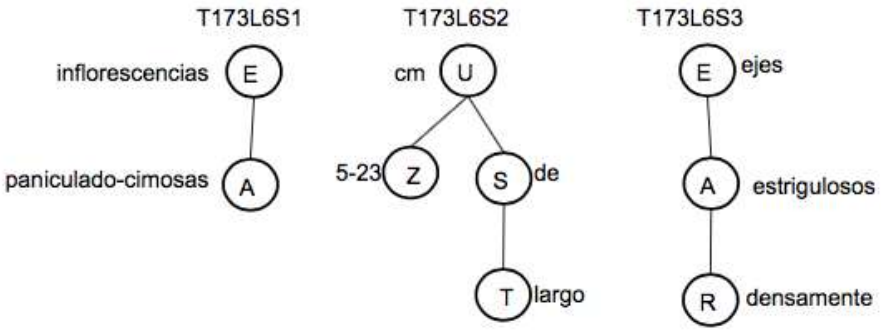


Figura 13. Texto estructurado para la cláusula "inflorescencias paniculado-cimosas, 5-23 cm de largo , ejes densamente estrigulosos ." (código T173L6).

```
−<statement id="T173L6" text=" inflorescencias paniculado-cimosas , 5-23 cm de largo , ejes densamente estrigulosos .">
 −<biological_entity id="T173L6S1-171020" name="inflorescencias" type="structure">
    <character name="arrangement" value="paniculado-cimosas"/>
    <character name="size_or_quantity" value="5-23" char_type="range_value" from="5" from_unit="cm" to="23"
    to_unit="cm" constraint_preposition="de largo"/>
  </biological_entity>
 −<biological_entity id="T173L6S3-171021" name="ejes" type="structure">
    <character name="pubescence" value="estrigulosos" constraint="densamente"/>
  </biological_entity>
 </statement>
```

Por ejemplo, para estructurar la cláusula T173L6 el algoritmo recorre en orden los tres árboles de la figura 12. Primero recorre el árbol T173L6S1 y genera inicialmente una instancia de la clase Biological_entity de tipo estructura (E) para el token “inflorescencias” y asociada a esta un carácter (A) con estado “paniculado-cimosas“. El recorrido del árbol T173L6S2 genera una instancia de carácter para la unidad de medida (U) “cm“ que luego es actualizada con el valor “5-23”; la preposición (S) “de“ y el resto del chunk se delimitan como constraint_preposition y se asigna al carácter creado anteriormente. El carácter creado se asocia a la estructura “inflorescencias”. El recorrido del árbol T173L6S3 genera una estructura con el nombre “ejes“ y un carácter con estado

“estrigulosos“ y restricción “densamente“. La cláusula estructurada se muestra en la figura 13. El detalle del análisis de árboles de dependencia se encuentra en el apéndice IV.

Los caracteres se asocian a la estructura que describen de acuerdo a las siguientes reglas:

1. Se asocian a la estructura con la que tienen una relación de dependencia en el mismo chunk.
2. Se asocian a la estructura procesada previamente si coinciden en género y número.
3. Si no coinciden en género y número con ninguna estructura procesada previamente en la misma cláusula entonces se asocian a la estructura principal (inicio de la cláusula).

Etapa-6. Presentación de resultados en XML.

El contenido de la base de datos al final del proceso es presentado como documentos en XML de acuerdo al esquema propuesto por Cui. Para cada descripción de taxón se crea un documento independiente. El apéndice VII presenta una descripción completa en formato XML.

3.3. Pruebas y evaluación

El algoritmo implementa dos técnicas: aprendizaje no supervisado (*bootstrapping*) y anotación semántica de las descripciones. Para ambos algoritmos se evaluó la precisión, cobertura y la medida F.

La estructuración de las cláusulas fue evaluada usando el concepto “razonable” de acuerdo a lo propuesto por Cui. La medida razonable y estricta se calculó de la siguiente forma:

- El algoritmo identifica las estructuras de forma:
    - Razonable => si identifica el nombre de la estructura bien.
    - Estricta => si además identifica bien el modificador y las restricciones.
- Identifica los caracteres y estados de forma:
    - Razonable => si existe el nombre en la ontología (si la traducción está bien) e identifica bien el valor, la unidad de medida y los rangos.
    - Estricto => si procesa además restricciones.

Adicionalmente se evaluó:

- Si el algoritmo asocia bien los caracteres a la estructura correcta (de acuerdo a lineamientos descritos en el Apéndice IV)
- Si estructura bien las conjunciones.

La selección de la muestra de cláusulas por libro a evaluar (5% de cláusulas por

libro) fue realizada utilizando el método de selección por la rueda de la ruleta. Se escogieron aleatoriamente los chunks de modo que se obtuvieran casos representativos tanto de cláusulas simples como de cláusulas complejas.

3.4. Tecnología utilizada

El software utilizado durante la presente investigación fue seleccionado por medio de un proceso de evaluación de la tecnología disponible en Internet (las variables tomadas en cuenta para la evaluación están disponibles en el apéndice V). Para la implementación del algoritmo se utilizó la siguiente tecnología:

Freeling 3.1 [9]: Constituye una biblioteca de herramientas para el procesamiento de lenguaje natural aplicable al castellano, catalán, vasco, inglés e italiano. La herramienta es desarrollada por el Centro de Investigación TALP de la Universidad Politécnica de Cataluña con el apoyo de una comunidad de desarrolladores bajo la licencia GNU Lesser General Public License (LGPL).

Algunas de las herramientas que incluye Freeling se listan a continuación:

- Identificación del idioma de un texto.
- Tokenización: El analizador morfológico para el castellano de Freeling segmenta los tokens y utiliza un conjunto de etiquetas para representar la información morfológica de estos basado en las etiquetas propuestas por el

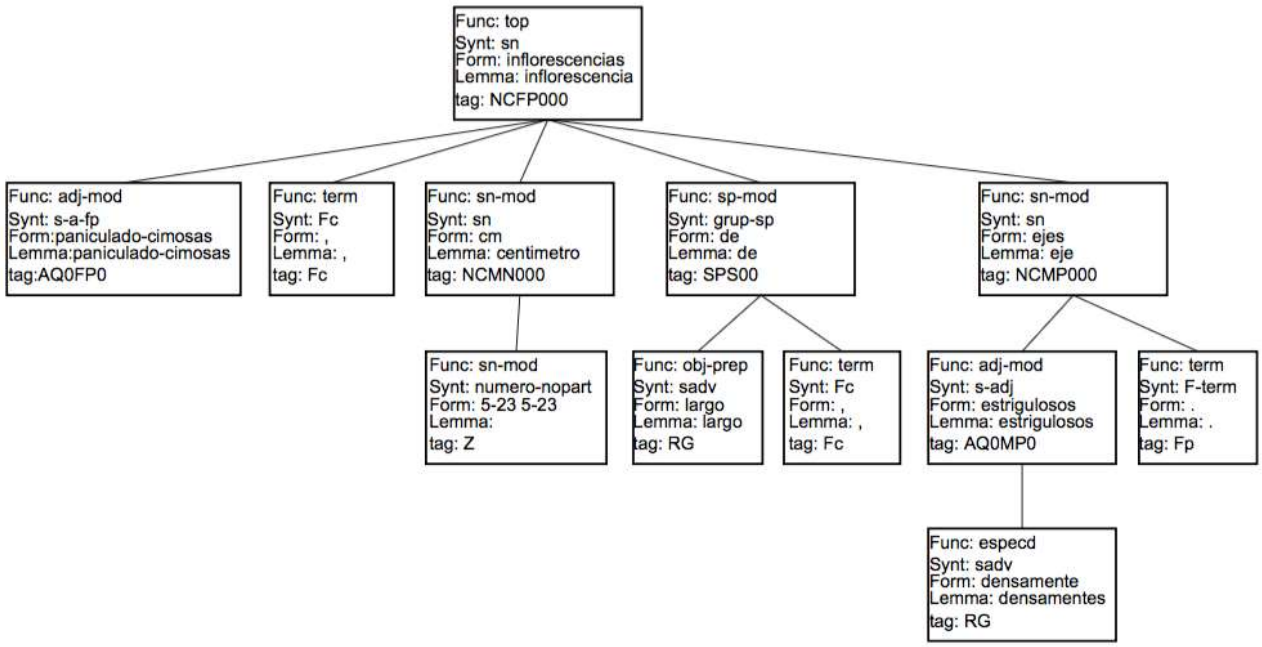
grupo Eagles[6]. Por ejemplo, al token “hojas“ que es parte del chunk “hojas simples”, se le asocia la etiqueta NCFP000. En este caso la “N” corresponde a la categoría Nombre; la “C“ al tipo “Común” (podría ser “Propio“); la “F“ corresponde al género, en esta caso Femenino (otras opciones “Masculino” y “Común“); la “P“ corresponde al número, en este caso “Plural“ (otras opciones “Singular“ e “Invariable“). Los siguientes dos ceros corresponden a la clasificación semántica (i.e. persona, lugar, organización u otros) y el último cero documenta el grado que puede ser “Apreciativo“ (i.e. aumentativos, despectivos, diminutivos, entre otros).

- Segmentación de oraciones: El usuario puede configurar los servicios de forma flexible, por ejemplo para definir los tokens indicadores de inicio de una nueva oración.
- Desambiguación del sentido de las palabras y etiquetado POS.
- Reconocimiento de multi-palabras, fechas/horas, expresiones numéricas (i.e. números, porcentajes, entre otros), expresiones de moneda, expresiones de medidas físicas (i.e. longitud, precisión, velocidad, frecuencia, temperatura, entre otras), nombres propios, entre otros.
- Análisis sintáctico superficial y completo o profundo.
- Análisis de dependencias: Este módulo recibe un párrafo y devuelve el árbol de dependencia asociado a este. Los árboles se construyen por medio de un algoritmo basado en reglas. Este enfoque establece que entre dos palabras siempre existe una relación de dependencia donde una

[6] La documentación de las etiquetas utilizadas está disponible en http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

es la regente y la otra la subordinada. Por lo tanto, si cada dos palabras en una oración tienen una relación de dependencia, la oración completa puede ser representada en forma de un árbol. Freeling genera los árbol de dependencia y asigna etiquetas sintácticas a cada uno de los nodos para describir la relación entre las palabras[7]. La figura 14 muestra el árbol sintáctico de la cláusula "inflorescencias paniculado-cimosas, 5-23 cm de largo , ejes densamente estrigulosos ." (código en la base de datos T173L6) tomada del libro ACRv4.

Figura 14. Árbol de dependencia generado por Freeling para la cláusula "inflorescencias paniculado-cimosas, 5-23 cm de largo , ejes densamente estrigulosos ." (código T173L6).



Cada nodo de árbol contiene datos asociados a un token particular. Los datos incluyen:

[7] La lista completa de etiquetas está disponible en http://devel.cpl.upc.edu/freeling/svn/trunk/doc/grammars/esCHUNKtags

- Func: Función sintáctica del token con respecto al ancestro. El significado de cada una de la etiquetas está descrito en la documentación de Freeling. Por ejemplo, “obj-prep” = objeto de la preposición.
- Synt: Corresponde al nombre del componente encabezado por esta palabra. Por ejemplo, “sn” o sintagma nominal.
- Form: Contiene el token en su forma original.
- Lemma: El lema o palabra aceptada como representante de todas las formas flexionadas asociadas a esta (i.e. plurales, gerundios, femenino, entre otros).
- Tag: Corresponde a la etiqueta asignada por Freeling durante el proceso de tokenización.

Ontology Term Organizer (OTO) [11]: Es una herramienta basada en web para integrar y administrar ontologías en el dominio de la biología. Uno de los servicios más importante que brinda OTO es la posibilidad de establecer consenso y asistir a los especialistas de distintos grupos de trabajo en el proceso de desarrollo de ontologías. Se prevé que este sea el mecanismo para definir al menos vocabularios controlados para procesar textos de descripciones morfológicas de otros grupos biológicos que no cuentan con ontologías.

OTO se utiliza para sugerir nombres de carácter asociados a los estados, debido a que los nombres de los caracteres que describen una especie son omitidos de las

descripciones. Por ejemplo la frase “hojas simples” no incluye en nombre del carácter, en este caso de “architecture“.

Plant Ontology (PO) [41]: El desarrollo de la Ontología de Plantas es liderado por un consorcio de instituciones con el objetivo de producir vocabularios controlados que puedan ser aplicados a bases de datos y literatura botánica. La ontología incluye términos asociados a desarrollo, anatomía, morfología, genómica y proteómica.

La PO actualmente está integrada a OTO por lo que puede ser accedida a partir de los servicios web que OTO provee.

Glosario inglés-español, español-inglés para la Flora Mesoamericana [12]**:** Es una herramienta basada en web desarrollada por el *Missouri Botanical Garden* como parte de la publicación Flora Mesoamericana. El glosario se utilizó para traducir al inglés los tokens antes de realizar una búsqueda en OTO ya que la información en OTO solo está disponible en inglés.

Capítulo 4: Resultados y discusión

4.1. Resultados

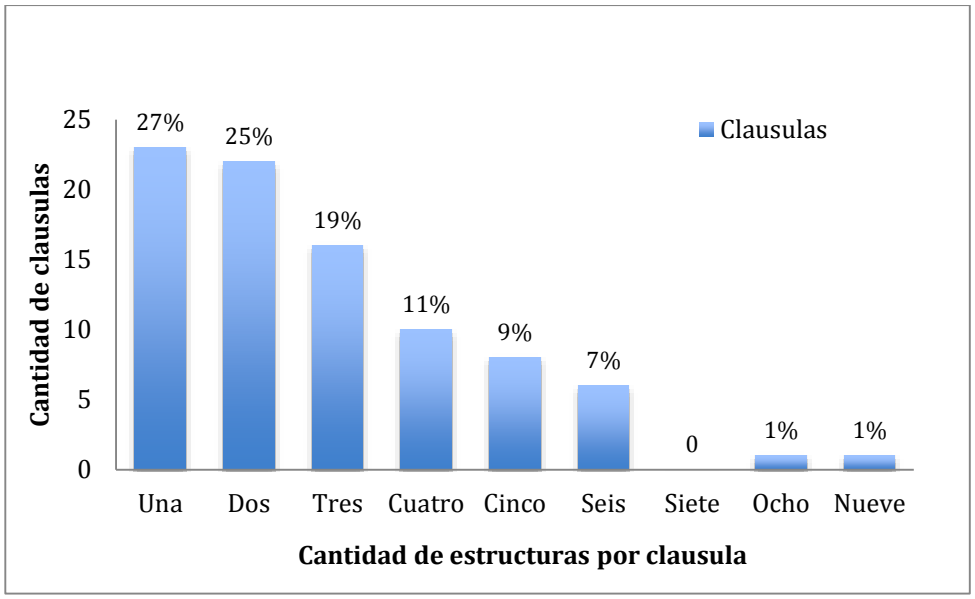
Los textos utilizados en esta investigación provienen de los libros ACRv3 y ACRv4, proporcionados por el autor principal de ambas publicaciones Nelson Zamora y una selección de descripciones tomadas del MPCR proporcionadas por el Dr. José Enrique Araya con la autorización de Nelson Zamora (uno de los autores del manual.) Las descripciones del MPCR se extrajeron del volumen VI del manual de forma semiautomática en un proyecto de investigación de extracción de conocimiento de literatura biológica liderado por el Dr. Araya, financiado por el Instituto Tecnológico de Costa Rica (ITCR). ACRv4 fue utilizado para el desarrollo del algoritmo. Los libros ACRv3 y MPCR fueron utilizados para la evaluación.

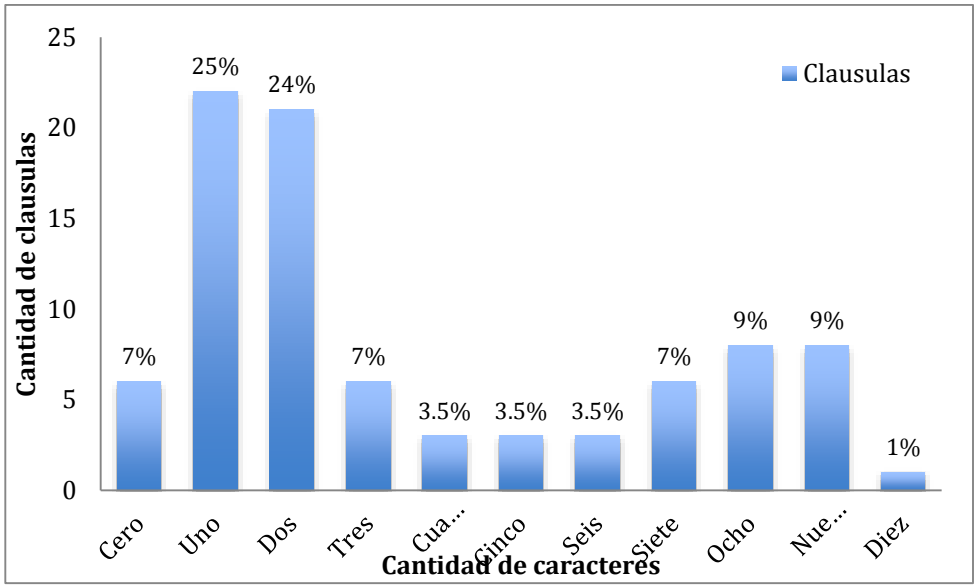
En esta sección se describen los resultados de ejecutar el algoritmo en una muestra aleatoria de cláusulas extraídas de los libros ACRv3 y MPCR (5% del total de cláusulas disponibles). Los datos de la muestra fueron escogidos utilizando el algoritmo de selección por la rueda de la ruleta, algoritmo que permite asignar más prioridad a las cláusulas con mayor cantidad de estructuras (como un indicador de complejidad). La tabla 5 presenta la cantidad promedio de estructuras y caracteres por cláusula para cada libro.

Libro	Cantidad de descripciones	Cláusulas	Cláusulas (muestra)	Promedio (en la muestra)	
				Estructuras	Caracteres
ACRv3	233	1.738	87 (5%)	2,85	3,62
MPCR	237	2.230	106 (5%)	3,42	3,69

Tabla 5: Cantidad promedio de estructuras y caracteres en las cláusulas evaluadas de los libros ACRv3 y MPCR.

Figura 15. Complejidad (cantidad de estructuras) y cantidad de caracteres en las cláusulas de la muestra del libro ACRv3.

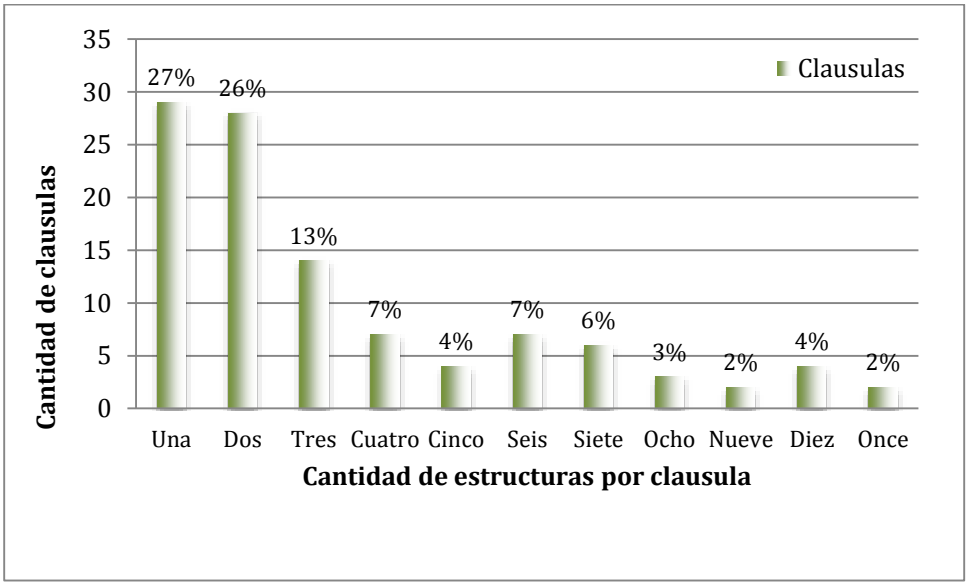


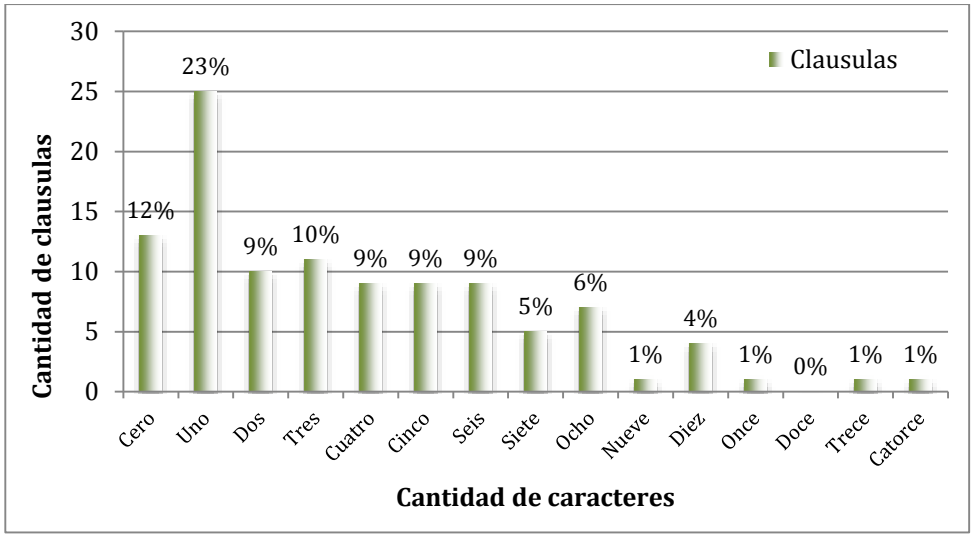


El libro ACRv3 incluye información de 233 especies con 1.738 cláusulas, de estas se seleccionaron 87 (5%). Del volumen VI del MPCR se seleccionaron 237 descripciones de especies de las cuales se extrajo una muestra aleatoria de 106 (5%).

La complejidad de las cláusulas en la muestra tomada del libro ACRv3 estuvo bien distribuida. Un 52% de las cláusulas eran simples y un 48% complejas. Se estima que una cláusula es simple si tiene dos o menos estructuras y complejas con más de dos estructuras. La figura 15 muestra la cantidad de estructuras y caracteres por cláusula en la muestra del libro ACRv3. Una cláusula puede contener cero caracteres debido a que todos se encuentran dentro de un sintagma preposicional o verbal no procesado.

Figura 16: Complejidad (cantidad de estructuras) y cantidad de caracteres en las cláusulas de la muestra del MPCR.





La complejidad de la muestra del MPCR estuvo también muy bien distribuida. Un 53% de las cláusulas eran simples y un 47% complejas. La figura 16 muestra la cantidad de estructuras y caracteres por cláusula en la muestra tomada del MPCR.

Se calculó la precisión, cobertura y medida F para la muestra de cada libro de forma individual (las tablas 6, 7 y 8 muestran los resultados). El Apéndice V contiene los datos completos de la evaluación de ambos libros.

Libro	Identificación de estructuras (precisión)		Estructuración de caracteres (precisión)		Asociación caracteres a estructuras (precisión)	Asociación conjunciones (precisión)
	Razonable	Estricto	Razonable	Estricto		
ACRv3	98,7% (245/248)	97,9% (243/248)	98,7% (314/318)	98,1% (312/318)	98,7% (312/316)	96,4% (27/28)
MPCR	99,7% (363/364)	98,1% (357/364)	97,9% (385/393)	92,8% (365/393)	86,4% (338/391)	92,4% (61/66)

Tabla 6: Precisión del algoritmo al ser aplicado a la muestra de los libros ACRv3 y MPCR.

Libro	Identificación de estructuras (cobertura)		Estructuración de caracteres (cobertura)		Asociación caracteres a estructuras (cobertura)	Asociación conjunciones (cobertura)
	Razonable	**Estricto**	**Razonable**	**Estricto**		
ACRv3	98,7% (245/248)	97,9% (243/248)	99,6% (314/315)	99% (312/315)	98,7% (312/316)	96,4% (27/28)
MPCR	99,7% (363/364)	98,1% (357/364)	98,4% (385/391)	93.0% (365/391)	86,4% (338/391)	92,4% (61/66)

Tabla 7: Cobertura del algoritmo al ser aplicado a la muestra de los libros ACRv3 y MPCR.

Libro	Identificación de estructuras (F)		Identificación de caracteres (F)		Asociación caracteres a estructuras (F)	Asociación conjunc. (F)	Promedio (valores razonable)
	Razonable	**Estricto**	**Razonable**	**Estricto**			
ACRv3	98,7	97,9	99,1	98,5	98,7	96,4	98,2
MPCR	99,7	98,1	98,1	93,3	86,4	92,4	94,1

Tabla 8: Rendimiento (F) del algoritmo al ser aplicado a la muestra de los libros ACRv3 y MPCR.

Los resultados (precisión y cobertura) de evaluar el algoritmo de aprendizaje no supervisado (bootstrapping) con datos de los libros ACRv4, MPCR y ACRv3 se muestran en la tabla 9. El algoritmo clasifica nombres, verbos y adjetivos en estados de carácter (A), estructuras (E) y verbos (V) que son almacenados en la base de conocimiento.

Libro	Tipo de token	Precisión	Cobertura	Rendimiento (F)
ACRv4	Valor de carácter (A)	99,6% (486/488)	94,2% (486/516)	96,8
ACRv4	Estructura (E)	83,1% (138/166)	98,6% (138/140)	90,1
ACRv4	Verbo (V)	89,5% (17/19)	100% (17/17)	94,4
MPCR	Valor de carácter (A)	99,8% (628/629)	96,3% (628/652)	98
MPCR	Estructura (E)	86% (130/151)	98,5% (130/132)	91,8
MPCR	Verbo (V)	83,3% (35/42)	100% (35/35)	90,9
ACRv3	Valor de carácter (A)	100% (183/183)	87,1% (183/210)	93,1
ACRv3	Estructura (E)	57,8% (37/64)	100% (37/37)	73,2
ACRv3	Verbo (V)	100%	100%	100%

Tabla 9: Resultados (precisión, cobertura y F) de evaluar el algoritmo de aprendizaje no supervisado (bootstrapping) en los libros ACRv4, MPCR y ACRv3.

El costo computacional de ejecutar el algoritmo completo en un computador MacBook Pro Core i7 con 8GB de RAM es de 17,41 segundos en promedio por descripción (para los libros ACRv3 y ACRv4) sin tomar en cuenta el tiempo requerido para realizar los procesos manuales (i.e. evaluar los resultados del algoritmo de aprendizaje y traducir al inglés los términos que no se encuentran en el diccionario).

4.2. Análisis y discusión

De acuerdo a los resultados de la evaluación presentados en la sección anterior se puede concluir que por la naturaleza semi-estructurada de las descripciones morfológicas de plantas es factible implementar con excelentes resultados un algoritmo de análisis semántico simple basado en reglas utilizando la tecnología disponible (Freeling, OTO, PO y el Glosario inglés-español, español-inglés para la Flora Mesoamericana). El algoritmo obtuvo más de un 94,1% de rendimiento promedio anotando estructuras, caracteres, asociando caracteres a estructuras y procesando conjunciones.

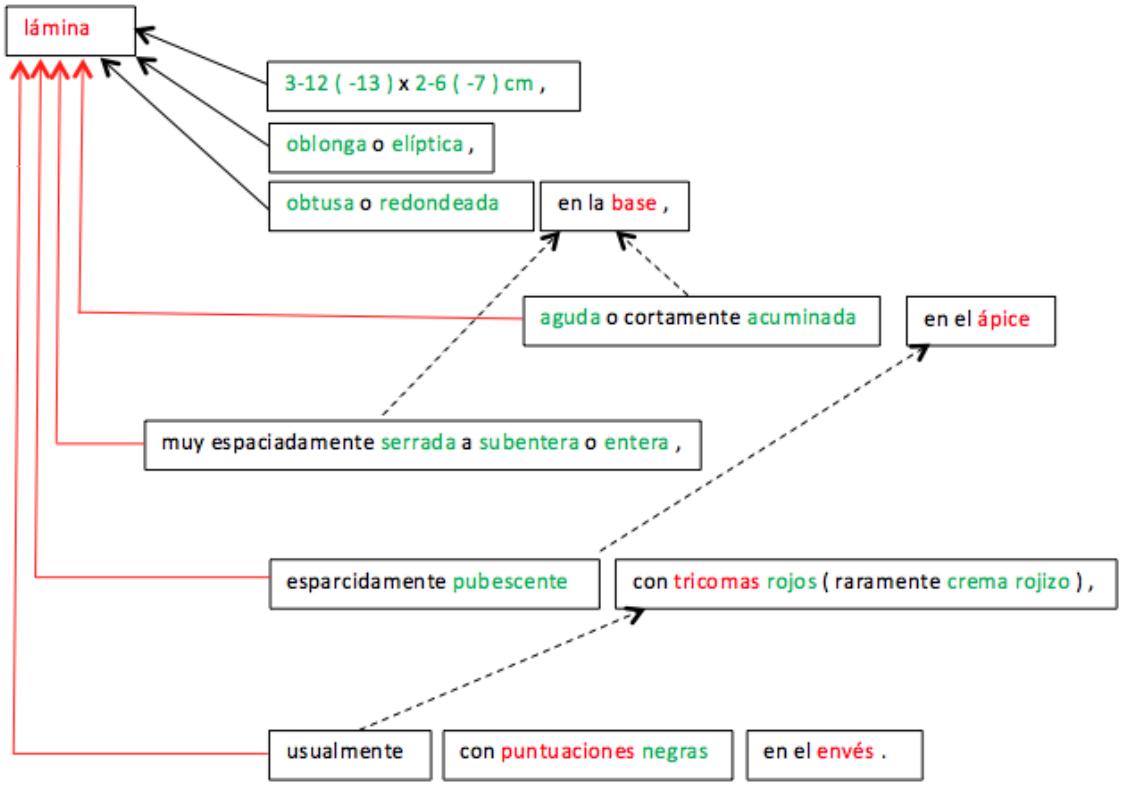
El algoritmo es escalable (dentro del grupo biológico de las plantas) como se demostró al evaluarlo no solo en registros de árboles (ACRv3) sino también en registros de plantas acuáticas, arbustos, epífitas, hierbas y lianas descritas en el MPCR. Las cláusulas del MPCR son un poco más complejas como lo muestra la figura 16, sin embargo la degradación del algoritmo fue aceptable. El rendimiento pasó de 98,2 en promedio al evaluar el libro ACRv3 a 94,1 al evaluar el MPCR. El mayor inconveniente al procesar el MPCR se dio al asociar caracteres a estructuras (F=86,4) lo que evidencia que la simple heurística de asociación carácter/estructura por concordancia en género y número debe ser complementada con el uso de ontologías u otros recursos (ej. una base de conocimiento ampliada).

La figura 17 ejemplifica el problema de asociar caracteres a estructuras para la descripción de la lámina del arbusto *Hydrangea asterolasia* (incluida en el MPCR).

> "lámina 3-12 (-13) x 2-6 (-7) cm , oblonga o elíptica , obtusa o redondeada en la base , aguda o cortamente acuminada en el ápice , muy espaciadamente serrada a subentera o entera , esparcidamente pubescente con tricomas rojos (raramente crema rojizo) , usualmente con puntuaciones negras en el envés ."

Figura 17[8]. a) Diagrama y b)extracto del documento XML que muestra el error al asignar caracteres a estructuras utilizando la heurística simple de concordancia en género y número en la cláusula T520L3 que describe la especie *Hydrangea asterolasia* (MPCR).

Figura 17a



[8] Figura preparada por el professor José Enrique Araya.

Figura 17b

```
−<statement id="T520L3" text=" lámina 3-12 ( -13 ) x 2-6 ( -7 ) cm , oblonga o elíptica , obtusa o redondeada en la
base , aguda o cortamente acuminada en el ápice , muy espaciadamente serrada a subentera o entera ,
esparcidamente pubescente con tricomas rojos ( raramente crema rojizo ) , usualmente con puntuaciones negras en
el envés .">
 −<biological_entity id="T520L3S1-202057" name="lámina" type="structure">
    <character name="length" value="3-12" char_type="range_value" from="3" from_unit="cm" to="12"
    to_unit="cm"/>
    <character name="atypical_range" value="-13" char_type="range_value" from="12" from_unit="cm"
    to="-13" to_unit="cm"/>
    <character name="width" value="2-6" char_type="range_value" from="2" from_unit="cm" to="6"
    to_unit="cm"/>
    <character name="atypical_range" value="-7" char_type="range_value" from="6" from_unit="cm"
    to="-7" to_unit="cm"/>
    <character name="shape" value="oblonga" constraint_conjunction="o"/>
    <character name="shape" value="elíptica" notes="Caracter repetido"/>
    <character name="arrangement" value="elíptica" notes="Caracter repetido"/>
    <character name="shape" value="obtusa" constraint_conjunction="o"/>
    <character name="shape" value="redondeada" constraint_preposition="en la base"/>
  </biological_entity>
 −<biological_entity id="T520L3S3-202058" name="base" type="structure">
    <character name="shape" value="aguda" constraint_conjunction="o"/>
    <character name="shape" value="acuminada" constraint="cortamente -" constraint_preposition="en el
    ápice"/>
    <character name="architecture" value="entera" notes="Caracter repetido"/>
    <character name="shape" value="entera" notes="Caracter repetido"/>
    <character name="architecture" value="serrada" constraint="muy espaciadamente -"
    other_constraint="a" constraint_preposition="a subentera o entera"/>
  </biological_entity>
 −<biological_entity id="T520L3S4-202059" name="ápice" type="structure">
    <character name="pubescence" value="pubescente" constraint="esparcidamente -"
    constraint_preposition="con tricomas rojos ( raramente crema rojizo )"/>
  </biological_entity>
  <biological_entity id="T520L3S6-202060" name="tricomas" constraint="usualmente"
  constraint_preposition="con puntuaciones negras en el envés ." type="structure"/>
  <biological_entity id="T520L3S7-202061" name="puntuaciones" type="structure"/>
  <biological_entity id="T520L3S7-202062" name="envés" type="structure"/>
</statement>
```

En este ejemplo todos los chunks deberían ser asociados a la estructura principal (figura 17a, línea sólida roja si existe error de asociación). Sin embargo, fueron asociados a subestructuras siguiendo la heurística de coincidencia en género y número con la estructura previamente procesada (línea punteada). Por ejemplo “aguda o cortamente acuminada en el ápice“ se asoció a base en lugar de a lámina. La figura 17b muestra el texto en formato XML.

La figura 18 muestra un ejemplo en el que la heurística funcionó bien, es decir asignó caracteres a la última estructura procesada de forma correcta utilizando el género y número.

Figura 18: a) Diagrama y b)documento XML que muestra un ejemplo correcto de asignar caracteres a estructuras utilizando el género y número en la cláusula T524L1 que describe la especie Hydrangea steyermarkii (MPCR).

Figura 18a

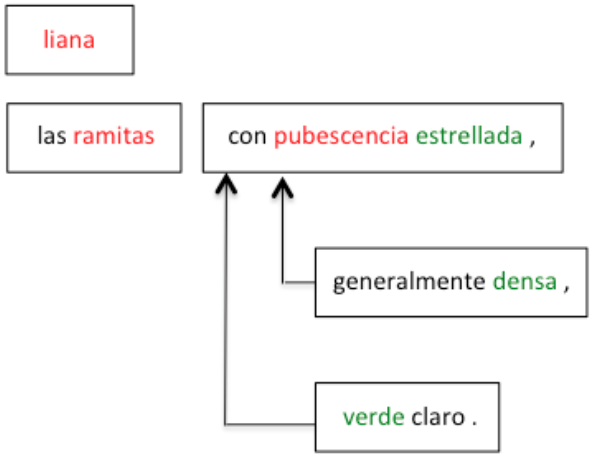


Figura 18b

```
−<statement id="T524L1" text=" liana , las ramitas con pubescencia estrellada , generalmente densa , verde claro .">
   <biological_entity id="T524L1S1-202133" name="liana" type="structure"/>
   <biological_entity id="T524L1S2-202134" name="ramitas" constraint_preposition="con pubescencia
   estrellada" type="structure"/>
  −<biological_entity id="T524L1S2-202135" name="pubescencia" type="structure">
     <character name="density" value="densa" constraint="generalmente"/>
     <character name="coloration" value="verde" constraint="claro"/>
   </biological_entity>
 </statement>
```

El rendimiento del algoritmo puede ser mejorado con el uso de ontologías (que incluyen jerarquías de estructuras/subestructuras y vocabularios controlados de caracteres válidos para describir una estructura) o con información adicional en la base de conocimiento, de la siguiente forma:

1. Evaluar el contexto en el chunk y determinar a partir de este la estructura a la que se deben asociar los caracteres. Por ejemplo, en la cláusula T520L3 que se muestra en la figura 17b, el carácter (“shape”, “aguda“) se asoció a la estructura “base” y si se evalúa el chunk completo “aguda o cortamente acuminada en el ápice“ es evidente que el “ápice” no es parte de la “base” sino de la “lámina“ de la hoja. A esta conclusión se puede llegar utilizando jerarquías de estructuras/subestructuras (manejadas por ontologías o incluidas en la base de conocimiento).
2. Antes de asociar un carácter a una estructura se debe evaluar si la asociación es correcta. Por ejemplo, la arquitectura “serrada” no es opción para describir la base de las hojas, es más bien una opción para describir la “lámina”. A esta conclusión se puede llegar utilizando vocabularios controlados de caracteres válidos para describir una estructura.

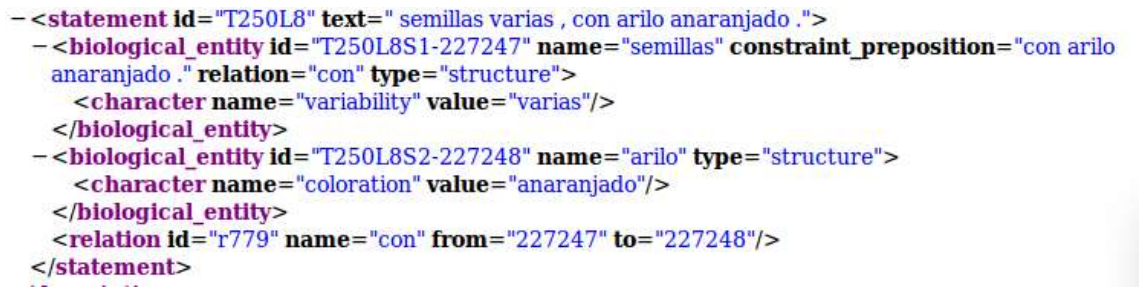
Este resultado debe ser tomado en cuenta si se desea extender el algoritmo para aplicarlo a otros grupos biológicos (i.e. vertebrados, artrópodos) ya que no existen aun ontologías en todos los casos. El caso de las plantas es un caso particular al igual que el de los hongos. Otros grupos biológicos no necesariamente cuentan con una ontología general.

Manejo de preposiciones y verbos:

El algoritmo no procesa todos los sintagmas preposicionales ni verbales, sin embargo, como una prueba de concepto se estructuraron los sintagmas preposicionales que inician con los tokens “sin“ o “con“. El resto de sintagmas

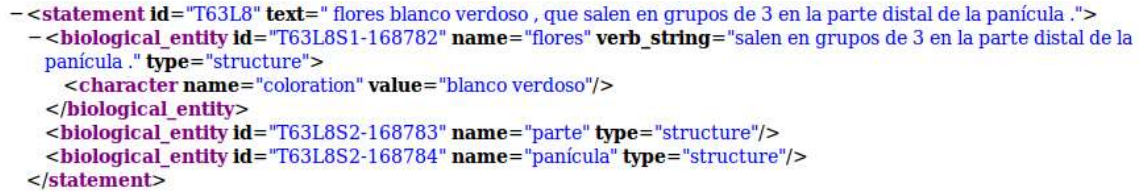
preposicionales o verbales solo son delimitados como *constraint_preposition* y *constraint_verb* respectivamente. Con los resultado de esta investigación y en un refinamiento posterior del algoritmo, se debe definir más detalladamente el alcance de la meta de extracción en estos casos. El refinamiento deberá tomar en cuenta el significado de cada preposición y verbo y de acuerdo a este anotar el chunk.

Figura 19. Resultado de la estructuración de la cláusula T250L8 (ejemplo de sintagma preposicional que inicia con el token "con").



Por ejemplo, la figura 19 muestra el resultado de estructurar la cláusula T250L8 "semillas varias , **con** arilo anaranjado". En este caso, el algoritmo anotó a partir de la preposición "con" el resto de la frase como constraint_preposition y además creó una relación de nombre "con" entre las estructuras "semillas" y "arilo".

Figura 20. Resultado de estructurar la cláusula T63L8 (ejemplo de uso de verbos).



La figura 20 presenta un ejemplo del manejo de los verbos realizado por el sistema. En el ejemplo se utilizó la cláusula “flores blanco verdoso , que **salen** en grupos de 3 en la parte distal de la panícula”. En este caso, el algoritmo procesa el verbo “salen“ delimitando el resto de la frase con la etiqueta verb_string. Otros ejemplos de chunks que incluyen preposiciones o verbos se listan en la tabla 10.

Código	**Chunk**
T166L3S9	**con** domacios (**en** forma **de** huecos o perforaciones) **en** las axilas **de** las venas secundarias distales
T174L4S12	**sin** domacios o **a** veces **con** grupitos **de** tricomas **en** la axila **de** las venas
T101L4S8	**con** 3 venas conspicuas **unidas** o casi **unidas arriba_de** la base y domacios o grupitos **de** tricomas **en** la axila **de** las venas principales
T19L11S4	**abriéndose en** 3 valvas rojas o moradas cuando maduras
T83L4S9	**con** 2 venas secundarias basales prominentes y ascendentes **conectadas** cerca (del) **de** el margen **formando** una vena submarginal
T27L2S1	la corteza se **exfolia en** placas circulares
T47L5S5	corimbiformes y **semejando** una escoba
T47L5S3	**hasta** 35 cm cuando **están en** fruto
T156L2S4	ocasionalmente huecas **hacia** el ápice y **a** veces **albergando** hormigas

Tabla 10: Algunos ejemplos de chunks que incluyen preposiciones o verbos (resaltados en negrita).

Clasificación de tokens:

El algoritmo de *bootstrapping* clasifica nombres, verbos y adjetivos en estados de carácter (A), estructuras (E) y verbos (V) utilizando heurísticas simples relacionadas con el contexto en el que se encuentran los tokens. Los adjetivos por lo general son bien etiquetados por Freeling. El mayor problema se presenta en los nombres ya que muchos de los tokens con rol no reconocido por Freeling se agrupan en esta categoría. El algoritmo se desempeñó muy bien para los libros ACRv4 y MPCR. El caso del libro ACRv3 es particular porque el texto que se procesó presentaba errores de falta de guiones entre adjetivos compuestos por

más de una palabra, por ejemplo en “diminutopubescentes” (debería ser diminuto-pubescentes), “pardopubescentes” (pardo-pubescentes) y “abruptoacuminado” (abrupto-acuminado) lo que ocasionó un exceso de errores de precisión al clasificar estructuras.

Capítulo 5: Conclusiones y trabajo futuro

5.1. Conclusiones

En esta investigación se presenta un algoritmo basado en el trabajo de la Dra. Cui adecuado a los requerimientos del idioma español, que alcanza resultados muy competitivos en la anotación de descripciones morfológicas de plantas logrando un rendimiento promedio del 94,1% en la anotación de estructuras, anotación de caracteres, asociación de caracteres a estructuras y procesamiento de conjunciones.

El algoritmo está basado en reglas que fueron definidas luego de analizar los patrones sintácticos de los árboles de dependencia de las estructuras gramaticales más utilizadas en el libro ACRv4. Para definir las reglas se analizó el 73,72% de los chunks del libro.

El buen resultado del sistema depende fuertemente de que el rol asignado a las palabras dentro de un chunk sea el correcto, por lo que fue necesario implementar un algoritmo de aprendizaje no supervisado (*bootstrapping*) para corregir el etiquetado POS asignado por el analizador morfosintáctico de Freeling. El lenguaje telegráfico de las descripciones botánicas, que está lleno de nombres, adjetivos y adverbios con pocos verbos, hace que el analizador morfosintáctico de Freeling no asigne bien etiquetas POS a nombres, adjetivos y verbos.

El algoritmo implementado se basa en el lenguaje telegráfico utilizado por la

comunidad de expertos botánicos. Sin embargo, este puede generalizarse a otros grupos biológico pre-procesando los textos de las descripciones para omitir algunas palabras funcionales (ejemplo los verbos ser y estar) que acerquen a estas al lenguaje telegráfico utilizado por los botánicos y extendiendo la funcionalidad del algoritmo.

Las ontologías son un recurso importante para la estructuración de las descripciones ya que permiten definir los caracteres que describen una estructura/subestructura y mejorar la asociación de caracteres a estructuras. Sin embargo, no todos los grupos biológicos cuentan con ontologías generales como la PO, por lo que el trabajo que realiza el equipo de la Dra. Cui con el desarrollo de OTO es muy importante para avanzar en la estructuración de las descripciones morfológicas de otros grupos biológicos. Es importante integrar a OTO el servicio de sinónimos que incorporen la traducción de los términos a otros idiomas para evitar la duplicidad de esfuerzo al tener que traducirlos.

Los objetivos específicos de la presente investigación definidos originalmente se cumplieron ya que:

1. Se implementó con muy buenos resultados, un algoritmo para extraer información de especies contenida en descripciones en español utilizando tecnología existente.

2. El algoritmo fue desarrollado utilizando las descripciones científicas de las 240 especies de árboles contenidos en el libro ACRv4.

3. El algoritmo fue evaluado utilizando el libro ACRv3. Para la evaluación se seleccionó el 5% de las cláusulas del libro para las cuales se calculó precisión, cobertura y medida F. Los datos de la muestra fueron escogidos utilizando el algoritmo de selección por la rueda de la ruleta, algoritmo que permite asignar más prioridad a las cláusulas con mayor cantidad de estructuras (como un indicador de complejidad). La evaluación se realizó en conjunto con el Dr. José Enrique Araya.

4. Se generó una base de conocimiento a partir de los conceptos aprendidos que puede ser utilizada para apoyar futuros proyectos de extracción de información en descripciones morfológicas de plantas.

5. El algoritmo fue generalizado y evaluado para resolver el problema de extracción de conocimiento disponible en español, a partir de descripciones morfológicas de plantas (no sólo de árboles). La prueba se realizó con un subconjunto de descripciones del Manual de Plantas de Costa Rica que incluyó plantas acuáticas, árboles, arbustos, epífitas, hierbas y lianas, entre otros tipos de plantas con un rendimiento muy bueno de 94,1% en promedio anotando estructuras, caracteres, asociando caracteres a estructuras y procesando conjunciones. Los datos completos de la evaluación se encuentran en el Apéndice V.

5.2. Trabajo futuro

Si bien el algoritmo tiene un rendimiento muy bueno es importante refinarlo y realizar mejoras en algunas de las etapas del proceso que se listan a continuación:

- Pre-procesamiento de los textos de las descripciones morfológicas. Es recomendable que el pre-procesamiento utilice los glosarios de los manuales y libros a procesar para reemplazar todas las abreviaturas antes de ejecutar el algoritmo. Esto mejoraría el etiquetado POS y los árboles de dependencias generados por Freeling.
- Traducción de tokens al inglés para que coincidan con entradas en la PO. Este es el proceso manual que requiere más atención del usuario. Posibles mejoras incluyen el uso de los sinónimos de la PO; agregar otros glosarios; incorporar nuevamente el servicio web del traductor de Google (*Google Translator*) que demostró ser una herramienta útil (82,5% de éxito) y utilizar los sinónimos que maneja Google; incorporar el Wiktionary[9] que en la mayoría de las definiciones incluye la acepción botánica del término.
- Anotación semántica de las descripciones.
    - Antes de iniciar el proceso de anotación semántica se debe etiquetar cada chunk con el nombre de la estructura o subestructura que este describe. Este refinamiento permitirá simplificar y estructurar mejor el algoritmo de análisis semántico. Por ejemplo, luego de segmentar en chunks la cláusula T8L5, el algoritmo debería asociar una estructura o subestructura a cada chunk de la siguiente forma:

[9] Disponible en http://en.wiktionary.org/wiki

Cláusula T8L5: hojas simples , alternas , (8,5) 14,5-33 × (4) 6-14 cm , oblongas a obovadas , ápice redondeado , obtuso a abrupto-acuminado , base redondeada , obtusa , truncada o levemente subcordada , glabras en el haz y con una pubescencia tomentosa sedosa con tricomas fasciculados en el envés , margen entero , crenado o distalmente denticulado ;

Chunks etiquetados con una estructura o subestructura:

Etiqueta	**Chunk**
hojas	hojas simples ,
hojas	alternas ,
hojas	(8,5) 14,5-33 × (4) 6-14 cm ,
hojas	oblongas a obovadas ,
ápice	ápice redondeado ,
ápice	obtuso a abrupto-acuminado ,
base	base redondeada ,
base	obtusa ,
base	truncada o levemente subcordada ,
hojas	glabras en el haz y con una pubescencia tomentosa sedosa con tricomas fasciculados en el envés ,
margen	margen entero ,
margen	crenado o distalmente denticulado ;

- Mucha información no fue extraía por ser parte de un sintagma preposicional o verbal. Se debe definir el alcance del proceso de extracción de información para estructurar al máximo la información contenida en estos sintagmas.
- La selección del carácter apropiado entre los caracteres repetidos podría realizarse por medio de heurísticas parecidas a las utilizadas en el *bootstrapping*.
- La asociación de caracteres a una estructura debe ser mejorada como se propone en la sección de Análisis y discusión.

Apéndices

Apéndice I - Esquema de datos

En esta sección se describen brevemente los conceptos *description, statement, biological_entity* y *character* incluidos en el esquema de datos[10]. La descripción morfológica completa de un taxón se delimita por el concepto *description*. Cada *description* contiene una o más cláusulas (*statements)* con todas sus estructuras (*biological_entity*) y caracteres (*character*). La estructura de datos de *description* y *statement* se presentan en las figuras 21 y 22:

Figura 21. Concepto *description* incluido en el esquema de datos.

```
<xs:complexType name="description" mixed="true">
        <xs:sequence>
                <xs:element minOccurs="0" maxOccurs="unbounded" type="statement" name="statement"/>
        </xs:sequence>
        <xs:attribute name="type" type="description_type" use="required"/>
        <xs:attribute name="scope"/>

</xs:complexType>
```

Figura 22. Concepto *statement* incluido en el esquema de datos.

```
<xs:complexType name="statement">
        <xs:sequence>
                <xs:element name="text" type="xs:string"/>
                <xs:choice minOccurs="0" maxOccurs="unbounded">
                        <xs:element type="relation" name="relation"/>
                        <xs:element type="biological_entity" name="biological_entity"/>
                        <xs:element type="nonEmptyString" name="value"/>
                </xs:choice>
        </xs:sequence>
        <xs:attribute name="id" use="required" type="xs:ID"/>
        <xs:attribute name="provenance" type="xs:string"/>
        <xs:attribute name="notes" type="xs:string"/>
</xs:complexType>
```

[10] La última versión del esquema está disponible en https://github.com/biosemantics/schemas/blob/master/semanticMarkupOutput.xsd.

Las **entidades biológicas** (*Biological_Entity*) en el esquema de Cui pueden ser de tipo *morphology*, *phenology*, *habitat*, *elevation*, *ecology*, *distribution*, *morphology-from-key* y *other*. La sección del esquema que permite estructurar las entidades biológicas se presenta a continuación (figura 23):

Figura 23. Concepto *biological_entity* definido en el esquema de datos.

```
<xs:complexType name="biological_entity">
        <xs:sequence>
                <xs:element minOccurs="0" maxOccurs="unbounded" type="character" name="character"/>
        </xs:sequence>
        <xs:attribute name="alter_name"/>
        <xs:attribute name="constraint"/>
        <xs:attribute name="constraintid" type="xs:NCName"/>
        <xs:attribute name="geographical_constraint"/>
        <xs:attribute name="id" use="required" type="xs:ID"/>
        <xs:attribute name="in_brackets" type="xs:boolean"/>
        <xs:attribute name="name" use="required"/>
        <xs:attribute name="parallelism_constraint" type="xs:NCName"/>
        <xs:attribute name="taxon_constraint"/>
        <xs:attribute name="ontologyid" type="xs:string"/>
        <xs:attribute name="provenance" type="xs:string"/>
        <xs:attribute name="notes" type="xs:string"/>
        <xs:attribute name="name_original" type="xs:string"/>
        <xs:attribute name="type" type="biological_entity_type" use="required"/>
</xs:complexType>
```

En esta investigación todas las estructuras generadas son del tipo *morphology*. Cada estructura tiene asociado un conjunto de caracteres que la describen e incluye los siguientes atributos (solo se listan los atributos utilizados en la investigación):

- **name**: Presenta el lema asociado al token. Las estructuras son extraídas de los nombres que se encuentran en las descripciones.
- **constraint**: Incluye modificadores de estructura en su mayoría nombres (i.e. femeninas, masculinas) y adverbios que califican la estructura.
- **other_constraint**: Contiene preposiciones (en un refinamiento posterior del algoritmo las preposiciones deben convertirse en relaciones entre estructuras).
- **id**: Está formado por los códigos secuenciales asignados por la base de datos a las descripciones morfológicas (T), las cláusulas (L), los chunks (S) y el secuencial asignado a la estructura. Un ejemplo de código puede ser T8L5S1 – 68262. Donde el T8 = código de la descripción morfológica, L5 = número de cláusula (línea) dentro de la descripción morfológica, en este caso la oración número 5 y S1 = código del chunk, en este caso la primer frase antes de la primera coma.

- **in_brackets**: Indicador de si la palabra se presenta entre paréntesis en el texto original.
- **notes**: Presenta algunos comentarios relacionados con la cláusula.
- **name_original**: Delimita el contenido del token original.

Para la investigación y teniendo en consideración que el algoritmo de análisis morfosintáctico requiere un refinamiento se crearon los siguientes nuevos conceptos:

- **constraint_preposition**: Contiene partes completas de chunk que inician con la preposición que se está procesando. El objetivo del concepto es delimitar la información no procesada asociada a las preposiciones.
- **verb_string**: Contiene la sección completa del chunk a partir de un verbo.

Cada carácter debe estar asociado a la estructura que corresponde. La sección del esquema que documenta caracteres se presenta en la figura 24.

Figura 24. Concepto *character* definido en el esquema de datos.

```
<xs:complexType name="character">
        <xs:attribute name="negation" use="required" type="xs:boolean"/>
        <xs:attribute name="char_type" type="xs:NCName"/>
        <xs:attribute name="constraint"/>
        <xs:attribute name="constraintid"/>
        <xs:attribute name="from"/>
        <xs:attribute name="from_inclusive" type="xs:boolean"/>
        <xs:attribute name="from_unit" type="xs:NCName"/>
        <xs:attribute name="geographical_constraint"/>
        <xs:attribute name="in_brackets" type="xs:boolean"/>
        <xs:attribute name="modifier"/>
        <xs:attribute name="name"/>
        <xs:attribute name="organ_constraint"/>
        <xs:attribute name="other_constraint"/>
        <xs:attribute name="parallelism_constraint" type="xs:NCName"/>
        <xs:attribute name="taxon_constraint"/>
        <xs:attribute name="to"/>
        <xs:attribute name="to_inclusive" type="xs:boolean"/>
        <xs:attribute name="to_unit" type="xs:NCName"/>
        <xs:attribute name="type"/>
        <xs:attribute name="unit"/>
        <xs:attribute name="upper_restricted" type="xs:boolean"/>
        <xs:attribute name="value"/>
        <xs:attribute name="ontologyid" type="xs:string"/>
        <xs:attribute name="provenance" type="xs:string"/>
        <xs:attribute name="notes" type="xs:string"/>
        <xs:attribute name="is_modifier" type="xs:boolean"/>
</xs:complexType>
```

Los atributos utilizados en la investigación se listan a continuación:

- char_type: Tipo de carácter, por ejemplo: *range_value, size_or_quantity.*
- constraint: Incluye los adverbios y otras restricciones aplicados al carácter.
- from: Utilizado en caso de rangos para definir el inicio de este.
- from_inclusive: Indicador de si el rango es abierto (0) o cerrado (1).
- from_unit: Unidad de medida que se aplica al inicio del rango.
- in_brackets: Indicador de si el estado del carácter va entre paréntesis.
- name: Nombre del carácter tomado de la PO.
- other_constraint: Contiene preposiciones (en un refinamiento posterior del algoritmo las preposiciones deben convertirse en relaciones entre estructuras).
- to: Utilizado en caso de rangos para definir el final de este.
- to_inclusive: Indicador de si el rango es abierto (0) o cerrado (1).
- to_unit: Unidad de medida que se aplica al valor final del rango.
- value: Estado del carácter.
- ontologyid: ID del carácter en la ontología utilizada.
- notes: Campo utilizado por ejemplo para indicar si un carácter es repetido, entre otros comentarios.

Conceptos adicionales:

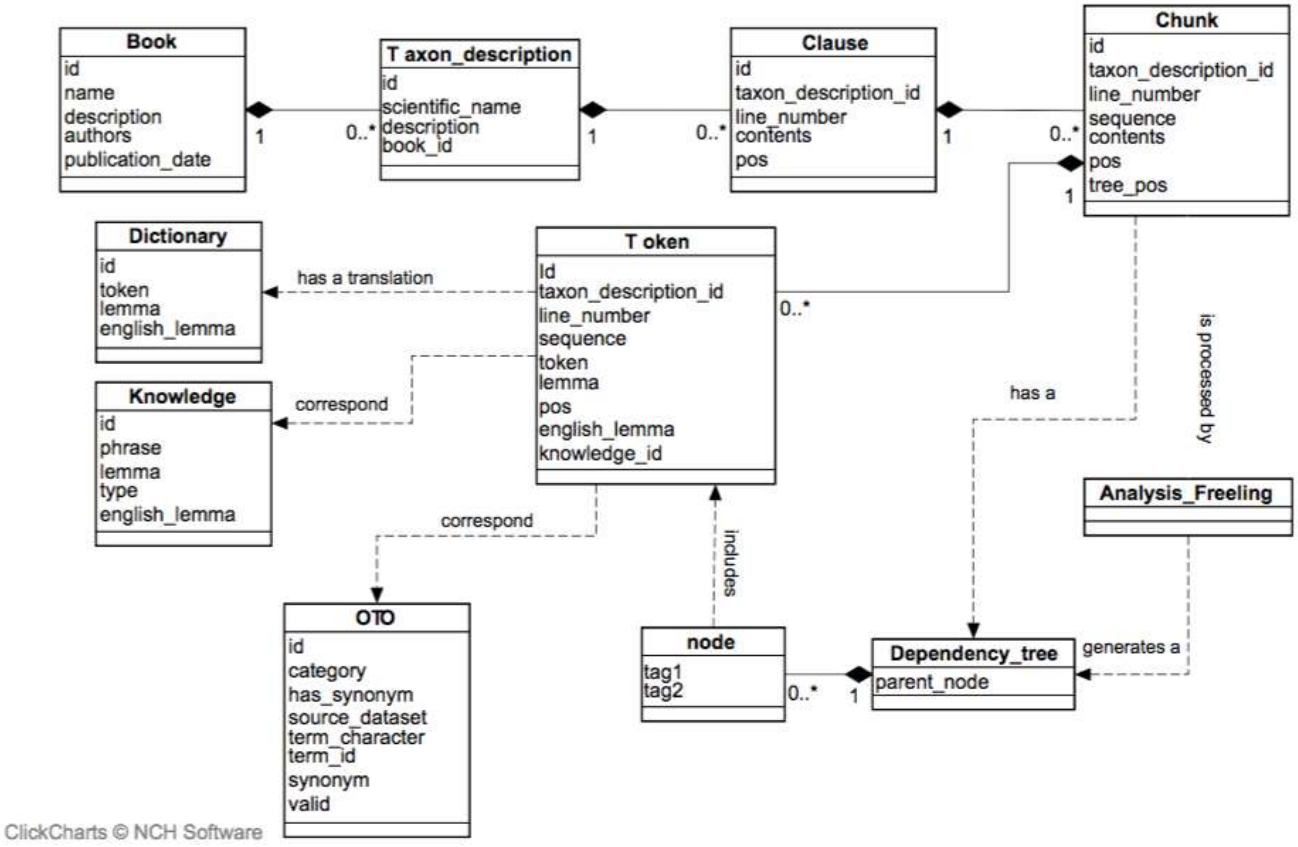
- **constraint_preposition**: Contiene partes completas de chunk que inician con la preposición que se está procesando. El objetivo del campo es delimitar la información no procesada.
- **verb_string**: Contiene la sección completa del chunk a partir de un verbo.

Figura 25. Extracto del resultado de estructurar la cláusula T8L5.

Un ejemplo de los resultados de estructurar la cláusula T8L5 se muestra en la figura 25. Los estados de carácter que poseen más de una entrada en la PO originan que el valor se repita para cada uno de los posibles caracteres. Por ejemplo: el estado “truncada“ (subestructura “base“) está asociado en la ontología a “architecture“ y a “shape“. Un especialista debe realizar la revisión final de los resultados para seleccionar la opción que aplica en cada caso. Los caracteres repetidos poseen una nota que lo indica.

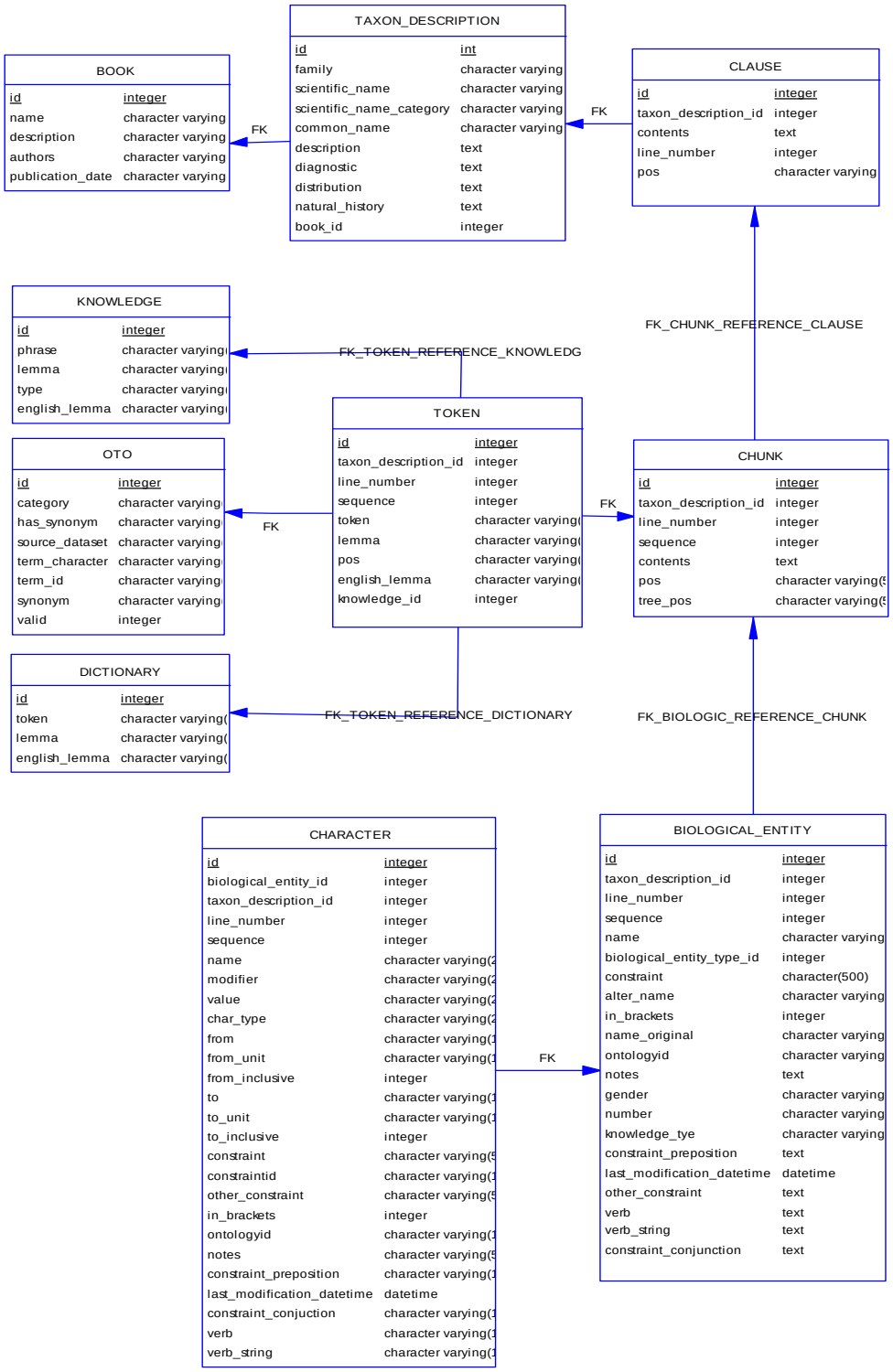
Apéndice II - Modelo de objetos

Figura 26. Modelo de objetos.



Apéndice III – Diagrama entidad – relación

Figura 27. Diagrama entidad – relación



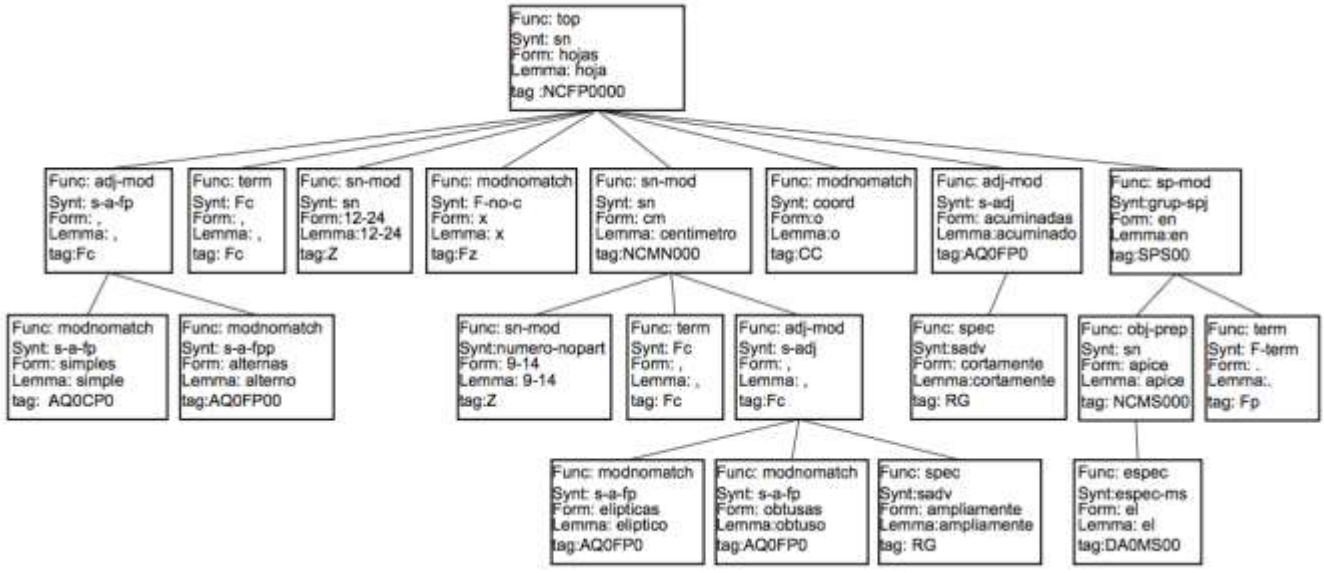
Apéndice IV - Anotación semántica de las descripciones

Esta sección describe el algoritmo de anotación de las descripciones desarrollado usando 240 descripciones de especies de árboles de Costa Rica incluidas en el libro Árboles de Costa Rica volumen 4 (ACRv4). El algoritmo analiza para cada token su tipo y posición en el árbol de dependencia para así instanciar o modificar un objeto (estructura o carácter) que luego es almacenado en la base de datos.

Como se mencionó en la sección de metodología, las descripciones morfológicas se segmentan en cláusulas utilizando como separador el punto, los dos puntos y el punto y coma. Las cláusulas se construyen de forma estandarizada a partir de los tokens que son parte de estas, todos escritos en letra minúscula y separados por un espacio (inclusive entre las palabras y signos de puntuación). Cada cláusula debe iniciar con un nombre de estructura para que sea procesada por el algoritmo.

Con el objetivo de simplificar las hileras a analizar y corregir errores en árboles de dependencia generados al aplicar Freeling a oraciones con un lenguaje telegráfico, cada cláusula a su vez es segmentada en chunks utilizando como separador la coma. La figura 28 presenta un árbol de dependencia generado para parte de la cláusula T160L4 “hojas simples , alternas , 12-24 × 9-14 cm , ampliamente elípticas , obtusas o cortamente acuminadas en el ápice“. El árbol presenta un error al aplicar un adverbio de forma incorrecta.

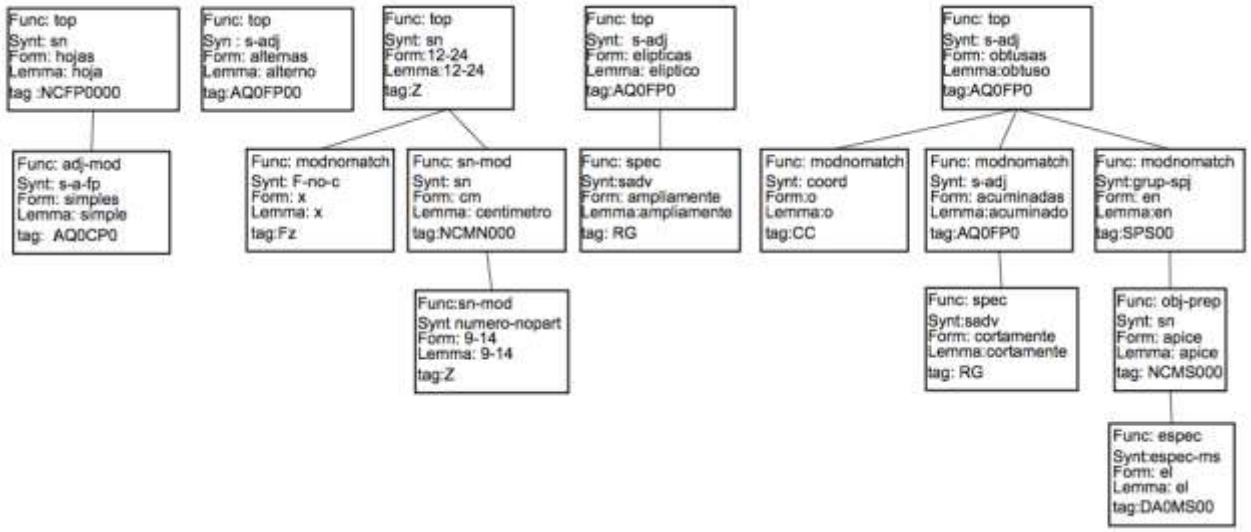
Figura 28. Árbol de dependencia construido por Freeling para parte de la cláusula T160L4 “hojas simples , alternas , 12-24 × 9-14 cm , ampliamente elípticas , obtusas o cortamente acuminadas en el ápice.“



El error se presenta al procesar “… ampliamente elípticas , obtusas …“, el token “ampliamente“ debe ser aplicado a “elípticas“ y no a “obtusas“. La cláusula se

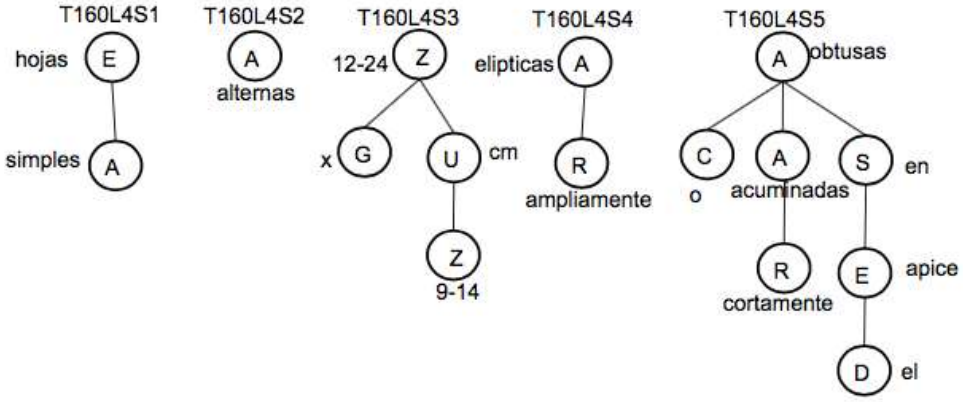
simplifica y se evita el error si esta se divide en chunks y se procesan cinco segmentos individuales.

Figura 29. Árboles de dependencia construidos por Freeling para cada uno de los chunks de la cláusula T160L4 “hojas simples , alternas , 12-24 × 9-14 cm , ampliamente elípticas , obtusas o cortamente acuminadas en el ápice.“. No se tomaron en cuenta los signos de puntuación porque no aportan al ejemplo.



La figura 29 muestra los árboles generados individualmente para cada chunk de la cláusula anterior. Como se puede observar el error fue corregido porque “elípticas” y “obtusas“ se procesaron en árboles separados. La complejidad se redujo y las dependencias se visualizan de forma más directa. Este enfoque tiene el inconveniente que se deben incorporar reglas de dependencia entre árboles, por ejemplo, el algoritmo debe manejar el que “elípticas” tiene una dependencia con “hojas“. Los árboles simplificados y corregidos a partir de la base de conocimiento se presentan la figura 30.

Figura 30. Árboles de dependencia simplificados y actualizados con tipos de token tomados de la base de conocimiento para cada uno de los chunks de la cláusula T160L4 “hojas simples , alternas , 12-24 × 9-14 cm , ampliamente elípticas , obtusas o cortamente acuminadas en el ápice.“.



La figura 31 presenta la cláusula estructurada en XML.

Figura 31. Texto estructurado para la cláusula T160L4 “hojas simples , alternas , 12-24 × 9-14 cm , ampliamente elípticas , obtusas o cortamente acuminadas en el ápice.“.

```
−<statement id="T160L4" text=" hojas simples , alternas , 12-24 × 9-14 cm , ampliamente elípticas , obtusas o
 cortamente acuminadas en el ápice , base obtusa a redondeada , esparcido-pubescente y glabrescente en el haz ,
 esparcida o más frecuente denso-pubescente en el envés , con 7-8 venas secundarias por lado , sin domacios ,
 margen entero ;">
 −<biological_entity id="T160L4S1-170750" name="hojas" constraint_preposition="con 7-8 venas
  secundarias por lado - sin domacios" type="structure">
    <character name="architecture" value="simples"/>
    <character name="arrangement" value="alternas"/>
    <character name="length" value="12-24" char_type="range_value" from="12" from_unit="cm"
    to="24" to_unit="cm"/>
    <character name="width" value="9-14" char_type="range_value" from="9" from_unit="cm" to="14"
    to_unit="cm"/>
    <character name="arrangement" value="elípticas" constraint="ampliamente" notes="Caracter
    repetido"/>
    <character name="shape" value="elípticas" notes="Caracter repetido"/>
    <character name="shape" value="obtusas" constraint_conjunction="o"/>
    <character name="shape" value="acuminadas" constraint="cortamente -"
    constraint_preposition="en el ápice"/>
  </biological_entity>
  <biological_entity id="T160L4S5-170751" name="ápice" type="structure"/>
```

El estado “elípticas” aparece como carácter repetido porque la ontología asocia al estado elíptica esos dos caracteres. El experto debe decidir si el carácter corresponde a “arrangement” o a ”shape”.

Las reglas o condiciones que sigue el proceso de análisis semántico se listan a continuación:

1. Estructuras (E)

- Instancian un objeto de la clase Biological_entity de tipo estructura.
- Dos estructuras juntas deben ser parte del mismo objeto. La segunda estructura es un modificador de la primera. Ejemplo “frutos nueces” (en T3L9S1). El objeto adquiere el género y número de la segunda estructura. Ejemplo: “frutos bayas , 1,5 cm de largo , esféricas”.

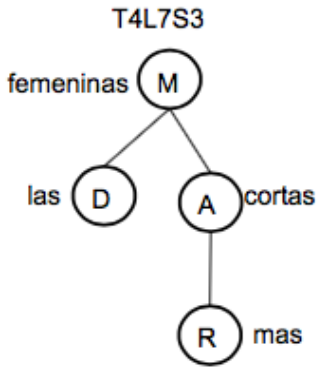
2. Hábito (H)

- Instancia un objeto de la clase Biological_entity de tipo de estructura (tipo H).

3. Modificadores (M)

- Un modificador, de acuerdo a Cui en [33], delimita el conjunto de objetos a los que aplican los caracteres y estados. En el ejemplo “ramitas jóvenes con una pubescencia diminuta”, el token “jóvenes” restringe el conjunto de ramitas.
- Los modificadores se procesan como *constraint* de la última estructura procesada.
- Para modificadores que tengan como hijo izquierdo un determinante se debe crear una nueva estructura. Por ejemplo en: “inflorescencias masculinas y femeninas en amentos“(T4L7S1), “las femeninas más cortas” (T4L7S3) el algoritmo crea una instancia de estructura adicional que corresponde a “inflorescencias femeninas“. El árbol de dependencia y el texto estructurado se muestran en la figura 32.

Figura 32. Árbol de dependencia y texto estructurado para el chunk “las femeninas más cortas” código T4L7S3.



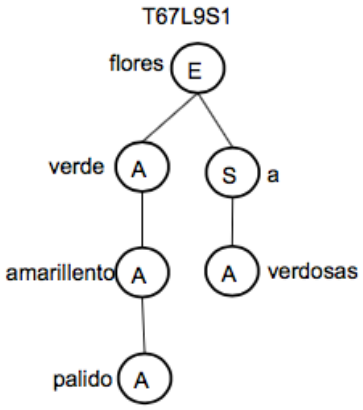
```
−<statement id="T4L7" text=" inflorescencias masculinas y femeninas en amentos , de 4-10 ( -14 ) cm de largo , las femeninas
más cortas ."> 
 −<biological_entity id="T4L7S1-167385" name="inflorescencias" constraint_preposition="de 4-10 ( -14 ) cm de largo"
  type="structure">
    <character name="reproduction" value="masculinas" constraint_conjunction="y"/>
    <character name="reproduction" value="femeninas" constraint_preposition="en amentos"/>
  </biological_entity>
  <biological_entity id="T4L7S1-167386" name="amentos" type="structure"/>
 −<biological_entity id="T4L7S3-167387" name="inflorescencias" constraint="femeninas" type="structure">
    <character name="size" value="cortas" constraint="más" notes="Caracter repetido"/>
    <character name="height" value="cortas" notes="Caracter repetido"/>
    <character name="length" value="cortas" notes="Caracter repetido"/>
  </biological_entity>
</statement>
```

Esta condición debería aplicarse a todos los adjetivos para que se procese igual por ejemplo el caso de “las rojas más cortas“.

4.Adjetivos (A)

- Los adjetivos crean una instancia de la clase Character.
- Dos o más adjetivos que están juntos debe ser incluidos en el mismo objeto en caso de que complementen su significado. Complementan su significado si la categoría (*category*) de la ontología es la misma. Ejemplo (en T67L9S1) “flores verde amarillento pálido a verdosas”. En este caso el *category* en todos los casos es “color”. El árbol de dependencia y el texto estructurado se muestran en la figura 33.

Figura 33. Árbol de dependencia y texto estructurado para el chunk código T67L9S1 “flores verde amarillento pálido a verdosas”.




```
−<statement id="T67L9" text=" flores verde amarillento pálido a verdosas ;">
 −<biological_entity id="T67L9S1-168877" name="flores" type="structure">
    <character name="coloration" value="verde amarillento pálido" other_constraint="a"
    constraint_preposition="a verdosas ;"/>
  </biological_entity>
</statement>
```

5. Numerales (Z)

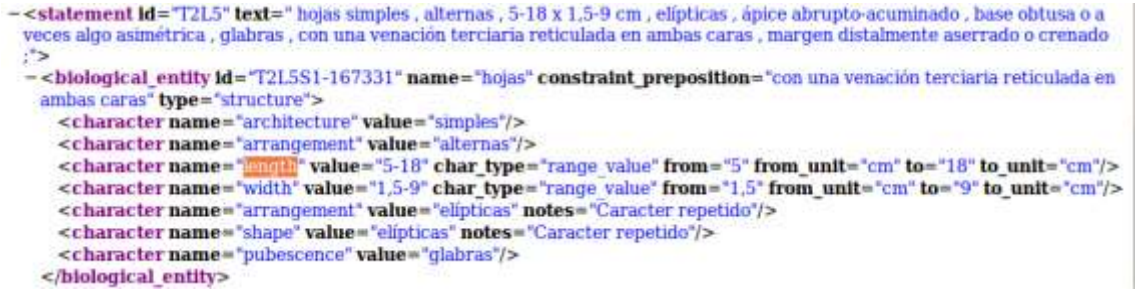
- Los numerales instancian un objeto de la clase Character.
- Los numerales definen cantidades, tamaños o áreas. La tabla 11 muestra algunos ejemplos de chunks que incluyen numerales:

Chunk	Contenido	Comentario
T2L5S3	5-18 x 1,5-9 cm	Por convención los botánicos definen la primera medida como largo y la segunda como ancho. Entonces en este ejemplo, largo = 5-18 por ancho = 1,5-9 cm.
T57L9S1	sépalos (4) 5	Sépalos raras veces 4 (cantidad atípica) comúnmente 5.
T79L6S1	sépalos (1) 4 (-6)	Sépalos raras veces [1-4[, comúnmente 4, raras veces]4-6].
T3L5S3	9,5-19 (-22) × 4-7 (-8) cm	Largo 9,5 – 19 (raras veces hasta 22) por ancho 4-7 (raras veces hasta 8) cm.
T18L11S1	semillas 1-3,2 × 1,5-2,2 mm	

Tabla 11. Ejemplos de uso de numerales en el libro ACRv4.

- Los rangos numéricos llenan las propiedades from y to del objeto carácter. Ejemplo, “5-18 x 1,5-9 cm”. En este caso se crea un carácter con atributos name = “length“, value = “5-18”, char_type = “range_value”, from = “5”, from_unit = “cm”, to = “18”, to_unit = “cm”. El resultado de la estructuración se muestra en la figura 34.

Figura 34. Resultado de estructurar el chunk “5-18 x 1,5-9 cm”.



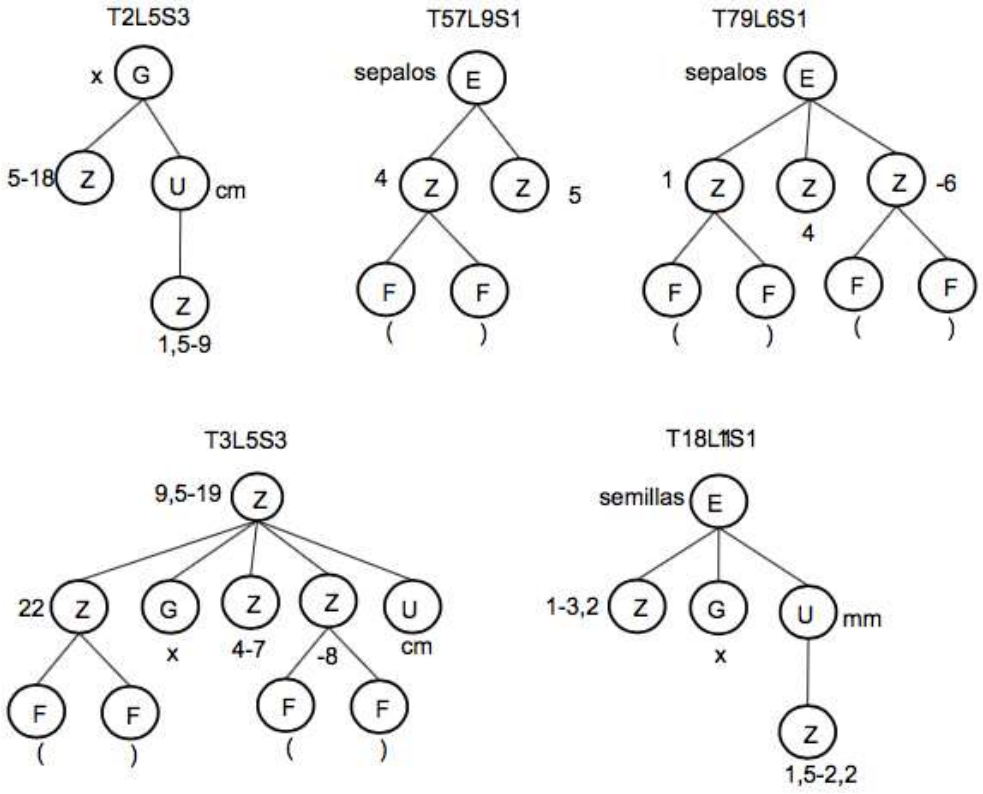
- Los rangos atípicos se presentan entre paréntesis antes o después de un numeral. En el ejemplo T3L5S3 “9,5-19 (-22) × 4-7 (-8) cm“, (-22) define un rango atípico que va de 19 a 22 cm (el 19 no incluido). Los rangos atípicos pueden estar antes o después de un número y deben procesarse dependiendo si están a la izquierda o derecha de este. En el ejemplo anterior hay dos rangos atípicos uno para el largo y otro asociado al ancho. El resultado de la estructuración del ejemplo se muestra en la figura 35.

Figura 35. Resultado de estructurar el chunk T3L5S3 “9,5-19 (-22) × 4-7 (-8) cm“.


```
- <statement id="T3L5" text=" hojas simples , alternas , 9,5-19 ( -22 ) × 4-7 ( -8 ) cm , elípticas a lanceolado-elípticas o amplio-obovadas , ápice acuminado a caudado , base cuneada , glabras en ambas caras o a veces pubescentes solo a lo largo de la vena central por el envés , margen entero ;">
  - <biological_entity id="T3L5S1-167349" name="hojas" type="structure">
      <character name="architecture" value="simples"/>
      <character name="arrangement" value="alternas"/>
      <character name="length" value="9,5-19" char_type="range_value" from="9,5" from_unit="cm" to="19" to_unit="cm"/>
      <character name="atypical_range" value="-22" char_type="range_value" from="19" from_unit="cm" to="-22" to_unit="cm"/>
      <character name="width" value="4-7" char_type="range_value" from="4" from_unit="cm" to="7" to_unit="cm"/>
      <character name="atypical_range" value="-8" char_type="range_value" from="7" from_unit="cm" to="-8" to_unit="cm"/>
      <character name="shape" value="elípticas" notes="Caracter repetido"/>
      <character name="arrangement" value="elípticas" other_constraint="a" notes="Caracter repetido"
      constraint_preposition="a lanceolado-elípticas o amplio-obovadas"/>
      <character name="pubescence" value="glabras" constraint_preposition="en ambas caras o a veces pubescentes solo a lo largo de la vena central por el envés"/>
    </biological_entity>
```


- En caso de áreas, por convención en manuales de planta, la primera parte corresponde al largo y la segunda al ancho.
- Los árboles de dependencia simplificados para todos los ejemplos de la tabla 11 se muestran en la figura 36.

Figura 36. árboles de dependencia simplificados asociados a los chunks de la tabla 11.



6. Unidad de medida (U)

- La unidad de medida en muchos casos es el nodo padre o regente de un número por lo que se procesa primero y crea un objeto Character que debe ser actualizado por el nodo subordinado o hijo si este es un número.

7. Áreas (G)

- El token tipo área no genera ningún cambio en el conjunto de objetos, solo es un indicador que de que el número siguiente corresponde a un valor del atributo ancho.

8. Adverbios (R)

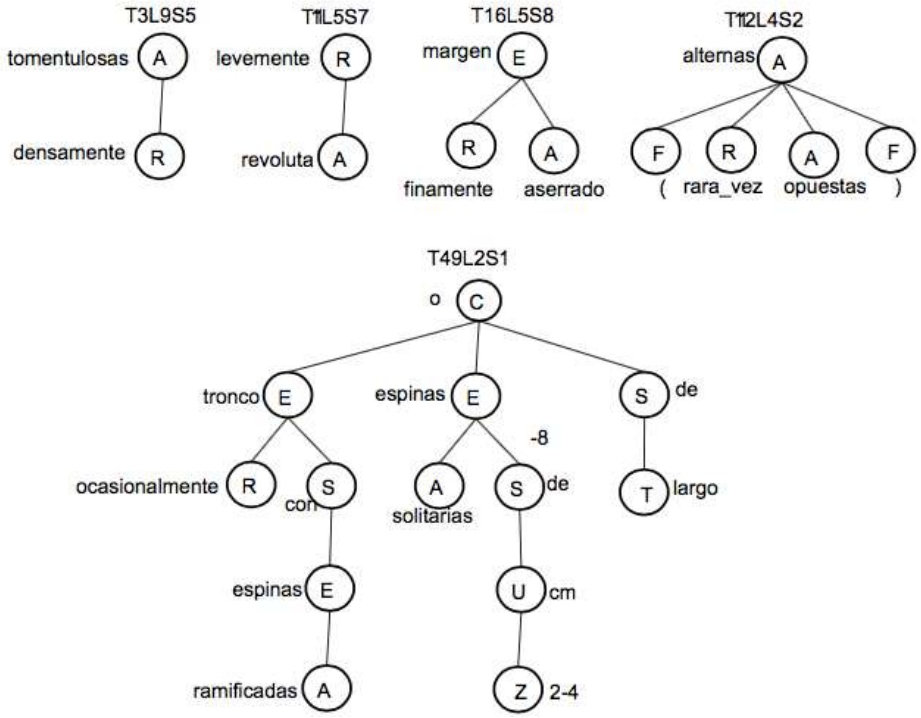
- Los adverbios modifican la estructura o carácter con la cual tienen una relación de dependencia. Modifican mayormente a caracteres, sin embargo en algunas ocasiones, en chunks complejos, pueden modificar estructuras. La tabla 12 muestra algunos ejemplos de uso de adverbios.

Chunk	Contenido
T3L9S5	densamente tomentulosas
T11L5S7	levemente revoluta
T16L5S8	margen finamente aserrado
T49L2S1	tronco ocasionalmente con espinas ramificadas o espinas solitarias de 2-4 cm de largo
T112L4S2	alternas (rara vez opuestas)

Tabla 12. Ejemplos de uso de adverbios en el libro ACRv4.

Los árboles de dependencia para los ejemplos anteriores se muestran en la figura 37.

Figura 37. Árboles de dependencia para los ejemplos de uso de adverbios de la tabla 12.



- Los adverbios se almacenan en el atributo *constraint* de la estructura o caractere al que modifican.
- La aplicación a uno u otro tipo de objeto (Biological_entity o Character) obedece a:
    - Si el adverbio tiene un hijo que es un adjetivo, verbo, número, unidad de medida, estructura o adverbio entonces se aplica a este objeto. Ejemplo T11L5S7. El texto estructurado se muestra en la figura 38.

Figura 38. Texto estructurado que ejemplifica el uso de adverbios en el chunk T11L5S7.

<character name="shape" value="revoluta" constraint="levemente" notes="Caracter repetido"/>

    - Si el adverbio tiene un hermano que es un adjetivo, verbo, número, unida de medida , estructura o adverbio entonces se aplica a este objeto. Ejemplo T112L4S2. El texto estructurado se muestra en la figura 39.

Figura 39. Texto estructurado que ejemplifica el uso de adverbios en el chunk T112L4S2.

<character name="arrangement" value="opuestas" constraint="rara_vez"/>

    - Si el adverbio tiene un padre que es un adjetivo, verbo, número, unida de medida, estructura o adverbio entonces se aplica a este objeto. Ejemplo T3L9S5. El texto estructurado se muestra en la figura 40.

Figura 40. Texto estructurado que ejemplifica el uso de adverbios en el chunk T3L9S5.

<character name="pubescence" value="tomentulosas" constraint="densamente"/>

- Un ejemplo de aplicarlo a carácter es (T16L5S8) “margen finamente aserrado“ se muestra en la figura 41. En este caso se crea el carácter “aserrado” y el atributo *constraint* = “finamente”.

Figura 41. Texto estructurado que ejemplifica el uso de adverbios en el chunk T16L5S8.

<biological_entity id="T16L5S8-167678" name="margen" type="structure">
<character name="architecture" value="aserrado" constraint="finamente" notes="Caracter repetido"/>

- Si el adverbio aparece en una frase entre paréntesis (T112L4S2) el algoritmo lo aplica al token que corresponde dentro de la frase como lo muestra la figura 42. Ejemplo “alternas (rara vez opuestas)”.

Figura 42. Texto estructurado que ejemplifica el uso de adverbios en el chunk T112L4S2.

```
−<biological_entity id="T112L4S1-169770" name="hojas" type="structure">
    <character name="architecture" value="simples"/>
    <character name="arrangement" value="alternas"/>
    <character name="arrangement" value="opuestas" constraint="rara_vez"/>
```

- Las conjunciones (o,y,u,e) generan el mismo efecto que los paréntesis.

9. Caracteres (T)

- El token tipo carácter no genera ningún cambio en el conjunto de objetos porque ya es tomado en cuenta con el uso de la ontología.

10. Determinantes (D)

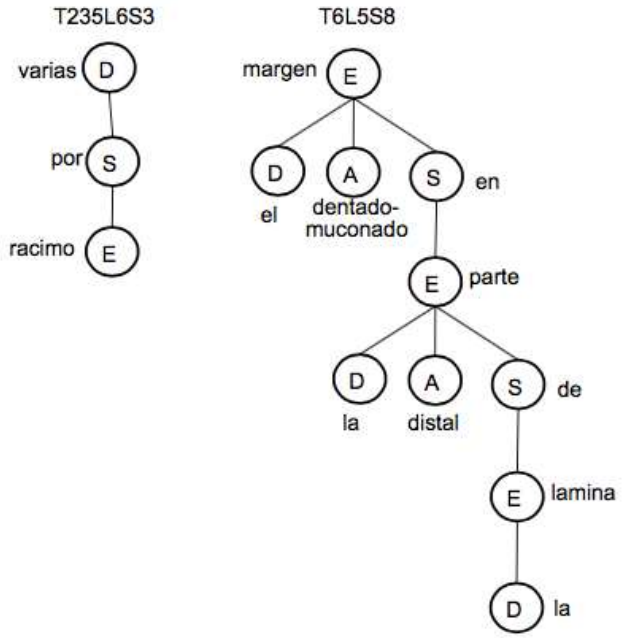
- El algoritmo procesa determinantes indefinidos (i.e. alguno, bastante, ninguna, todas, todos, varias) y artículos (i.e. el, la, los, las) como una prueba del concepto. Sin embargo, se debe analizar el uso de otros tipos de determinantes que juegan un rol importante en descripciones de otros grupos biológicos como vertebrados, artrópodos, entre otros. La tabla 13 presenta algunos ejemplos de chunks que utilizan determinantes.

Chunk	Contenido
T235L6S3	varias por racimo
T6L5S8	el margen dentado-mucronado en la parte distal de la lámina

Tabla 13. Ejemplos de chunks que incluyen determinantes.

Los árboles de dependencia para los ejemplos anteriores se muestran en la figura 43.

Figura 43. Árboles de dependencia para los ejemplos de uso de determinantes de la tabla 13.



- Los determinantes indefinidos son aplicados como modificadores de estructuras. El texto estructurado del ejemplo T235L6S3 se muestra en la figura 44.

Figura 44. Texto estructurado que ejemplifica el uso de determinantes en el chunk T235L6S1.

```
- <statement id="T235L6" text=" inflorescencias racemosas , 7-14 cm de largo , varias por racimo ."> 
  - <biological_entity id="T235L6S1-172539" name="inflorescencias" constraint="varias -" constraint_preposition="por racimo ." 
    type="structure">
      <character name="arrangement" value="racemosas"/>
      <character name="size_or_quantity" value="7-14" char_type="range_value" from="7" from_unit="cm" to="14" to_unit="cm" 
      constraint_preposition="de largo"/>
    </biological_entity>
```

- Los artículos modifican el género y número de la estructura regente.
- Los artículos que acompañan a un modificador de estructura (i.e. el hijo de un modificador) generan una instancia de Biological_entity. El nombre de la estructura se toma de la última estructura procesada que coincide en género y número con el artículo. En el ejemplo T4L7S3 “las femeninas más cortas” se crea una estructura adicional para inflorescencias con un constraint = femeninas.

11. Pronombres (P)

- El algoritmo procesa únicamente pronombres indefinidos (i.e. algo, bastante, mucho, nadie). Sin embargo, es importante analizar el uso de otros tipos de pronombres que juegan un rol importante en descripciones de

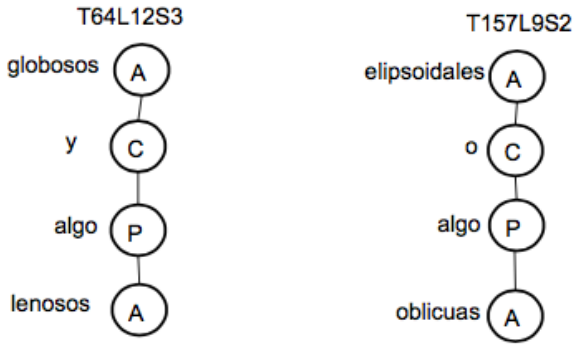
otros grupos biológicos como vertebrados, artrópodos, entre otros. La tabla 14 presenta algunos ejemplos del uso de pronombres en el libro ACRv4.

Chunk	Contenido
T64L12S3	globosos y algo leñosos
T157L9S2	elipsoidales o algo oblicuas

Tabla 14. Ejemplos de chunks que incluyen pronombres en el libro ACRv4.

Los árboles de dependencia para los ejemplos anteriores se muestran en la figura 45.

Figura 45. Árboles de dependencia para los ejemplos de uso de pronombres de la tabla 14.



- Los pronombres actúan como modificadores del carácter con el que tienen relación de dependencia.

12. Conjunciones (C)

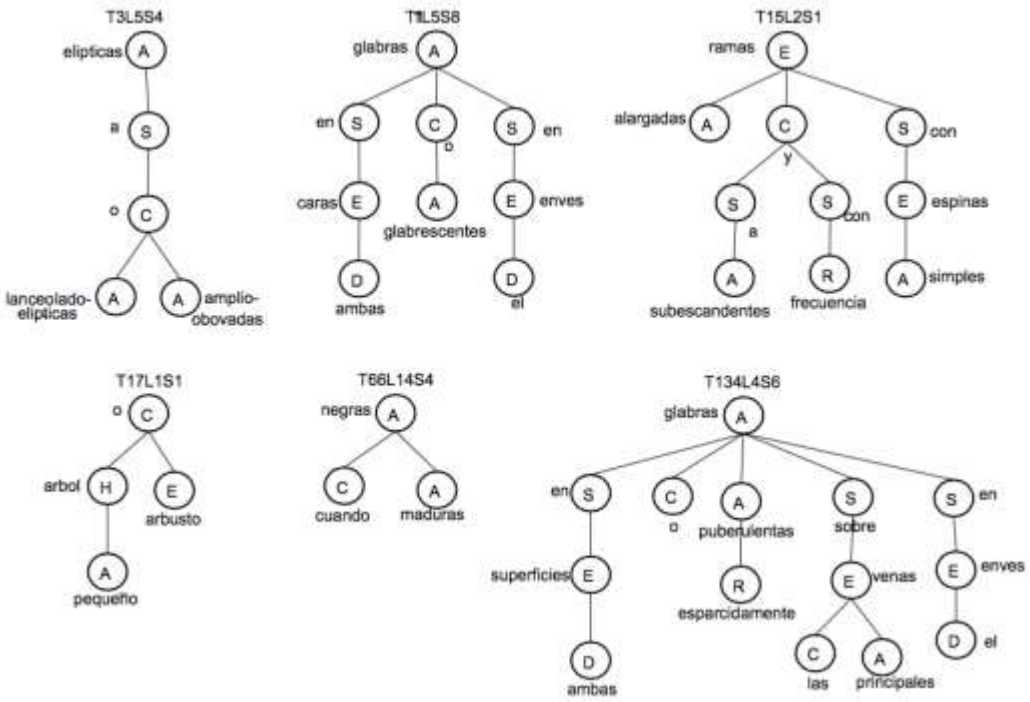
- La conjunciones se asocian a un carácter o estructura dependiendo del objeto de la conjunción. Si el objeto de la conjunción es una estructura la conjunción se asocia a esta. En el siguiente ejemplo: “árbol pequeño o arbusto” (arbusto debería asociarse a árbol y no a pequeño. La tabla 15 presenta algunos ejemplos del uso de conjunciones en el libro ACRv4.

Chunk	Contenido
T3L5S4	elípticas a lanceolado-elípticas o amplio-obovadas
T11L5S8	glabras en ambas caras o glabrescentes en el envés
T15L2S1	ramas alargadas a subescandentes y con frecuencia con espinas simples
T17L1S1	árbol pequeño o arbusto
T66L14S4	negras cuando maduras
T134L4S6	glabras en ambas superficies o esparcidamente puberulentas sobre las venas principales en el envés

Tabla 15. Ejemplos de chunks que incluyen conjunciones en el libro ACRv4.

Los árboles de dependencia para los ejemplos anteriores se muestran en la figura 46.

Figura 46. Árboles de dependencia para los ejemplos de uso de conjunciones de la tabla 15.



- Si la conjunción no es “y”, “o”, “u”, ”e” entonces se debe definir el atributo Constraint_conjunction igual a toda la tira a partir de la conjunción y hasta el final del chunk. En el ejemplo “negras cuando maduras” (T66L14S4) la tira “cuando maduras” se asigna al atributo Constraint_conjuction del objeto Character o Biological_entity como se muestra en la figura 47.

Figura 47. Texto estructurado que ejemplifica el uso de conjunciones en el chunk T66L14S4.

```
<character name="coloration" value="negras"
constraint_conjunction="cuando maduras"/>
```

- El atributo definido se asocia a un objeto Character o Biological_entity de acuerdo a las siguientes reglas:
    - Si el nodo de la conjunción tiene 2 hijos que no sean signos de puntuación, la conjunción debe ser aplicada al hijo de más a la izquierda. El texto estructurado del ejemplo T17L1S1 se muestra en la figura 48**.**

Figura 48. Texto estructurado que ejemplifica el uso de conjunciones en el chunk T17L1S1.

```
- <statement id="T17L1" text=" árbol pequeño o arbusto , 1,5-12 m de altura ;">
  - <biological_entity id="T17L1S1-167697" name="árbol" constraint_conjunction="o" type="structure">
      <character name="size" value="pequeño"/>
    </biological_entity>
  - <biological_entity id="T17L1S1-167698" name="arbusto" type="structure">
      <character name="size_or_quantity" value="1,5-12" char_type="range_value" from="1,5" from_unit="m" to="12"
      to_unit="m" constraint_preposition="de altura ;"/>
    </biological_entity>
  </statement>
```

    - Si el nodo tiene un hijo la conjunción debe ser aplicada al ultimo objeto procesado del tipo del hijo (carácter o estructura).
    - Si el nodo no tiene hijos se busca al hermano derecho y se aplica la conjunción al ultimo objeto procesado del tipo del hermano (estructura o carácter).
- Los adjetivos que aparecen luego de la conjunción deben asociarse a la ultima estructura utilizada es decir a la que se asignó la primera parte del chunk. Por ejemplo en T134L4S6 “glabras en ambas superficies o esparcidamente puberulentas sobre las venas principales en el envés”. En este caso el adjetivo “puberulentas” debe ser asociado a la estructura que se asoció “glabras” y no a “superficies”.

13. Preposiciones (S) y verbos (V)

- El algoritmo no extrae información en sintagmas verbales o preposicionales pero sí identifica estas secciones para ser procesadas en un refinamiento posterior.
- El sintagma preposicional o verbal se asocia a la estructura o carácter que corresponda (actualizando el atributo Constraint_preposition o Constraint_verb). La estructura a la que corresponde un sintagma

preposicional o verbal se calcula de acuerdo a la proximidad, el género y el número de la siguiente forma:

- Si el nodo de la preposición o verbo es la raíz se aplica a la estructura principal.
- Si no, se aplica al ultimo carácter o estructura incluido en la base de datos (se verifica por medio del atributo last_modification_datetime).
- Además, para definir rangos de caracteres no numéricos (definidos por “a” y “hasta”) se almacena el token en el atributo Other_constraint.

Apéndice V - Datos completos de la evaluación de los libros ACRv3 y MPCR.

Datos de evaluación del libro Árboles de Costa Rica v3.

Fecha: 10 de diciembre de 2015

Evaluadores: José Enrique Araya y María Mora

Descrip. Código	Cláusula	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Preposiciones y verbos no procesados	Explicación del error
		Estructuras	Caracteres	Razonable	Estricto	Mal	Razonable	Estricto	Bien	Mal	Bien	Mal		
242	4	2	3	2	2		3	3	3		0	0	1	
243	3	9	9	9	9		9	9	9		1	0	2	
248	1	2	2	2	2		2	2	2		1	0	0	
250	4	3	6	3	3		6	6	6		0	0	1	
256	4	2	5	2	2		5	5	5		0	0	1	
256	7	3	3	3	3		3	3	3		1	0	2	
261	3	1	1	1	1		1	1	1		0	0	0	
262	3	1	1	1	1		1	1	1		0	0	0	
263	4	3	8	3	3		8	8	8		1	0	1	
271	6	2	1	2	2		1	1	1		0	0	1	
278	6	2	3	2	2		3	3	3		0	0	2	
280	7	2	2	2	2		2	2	2		0	0	1	

Descrip. Código	Cláusula	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Preposiciones y verbos no procesados	Explicación del error
		Estructuras	Caracteres	Razonable	Estricto	Mal	Razonable	Estricto	Bien	Mal	Bien	Mal		
285	3	1	1	1	1		1	1	1		0	0	1	
286	4	5	6	5	5		6	6	6		0	0	2	
288	4	6	8	6	6		8	8	8		1	0	2	
303	5	1	1	1	1		1	1	1		0	0	1	
304	1	2	2	2	2		1	1	2		1	0	0	Cuando la traducción del término no esta bien no coincide con la ontología.
304	2	1	1	1	1		1	1	1		0	0	0	
304	4	5	8	5	5		8	8	8		0	0	3	
306	3	6	8	6	6		8	8	8		1	0	2	
310	4	2	1	2	2		1	1	1		0	0	1	
311	3	4	9	4	4		9	9	9		1	0	0	
312	9	2	0	2	2		0	0	0		0	0	2	
317	2	2	2	2	2		2	2	2		0	0	0	
318	2	2	2	2	2		2	2	2		0	0	0	
319	7	2	4	2	2		4	4	4		0	0	2	
320	3	1	1	1	1		1	1	1		0	0	0	

Descrip. Código	Cláusula	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Preposiciones y verbos no procesados	Explicación del error
		Estructuras	Caracteres	Razonable	Estricto	Mal	Razonable	Estricto	Bien	Mal	Bien	Mal		
320	8	2	0	2	2		0	0	0		0	0	2	Se perdió el "como" porque estaba asociado a un sintagma preposicional. No es lo mismo "de 1 cm de largo" que "como de 1cm...".
324	4	5	8	5	5		8	8	8		0	0	2	
326	4	5	8	5	5		8	8	8		0	0	3	
329	6	3	5	3	2		5	5	5	1	0	0	2	Fallo por comisión, asignó una estructura como carácter (debió ser un constraint) y un carácter que no debió haber procesado.
333	1	1	1	1	1		1	1	1		0	0	0	
339	3	1	1	1	1		1	1	1		0	0	0	
343	3	5	7	5	5		7	7	7		0	0	2	

Descrip. Código	Cláusula	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Preposiciones y verbos no procesados	Explicación del error
		Estructuras	Caracteres	Razonable	Estricto	Mal	Razonable	Estricto	Bien	Mal	Bien	Mal		
345	5	1	2	1	1		2	2	2		0	0	0	
345	6	4	3	4	4		3	3	3		0	0	1	
350	4	3	2	3	3		2	2	2		1	1	2	Error de asignación de conjunción a la estructura principal.
353	2	1	2	1	1		2	2	2		0	0	0	
356	7	2	1	2	2		1	1	1		0	0	1	
359	4	3	5	3	3		5	5	5		0	0	1	
362	1	2	1	2	2		1	1	1		1	0	1	
362	2	1	2	1	1		2	1	2		0	0	0	Se contabiliza error por repetir un adverbio como constraint.
362	6	1	1	1	1		1	1	1		0	0	0	
363	7	3	1	3	3		1	1	1		0	0	2	
364	5	3	3	3	3		3	3	3		0	0	1	
366	3	3	6	2	2		6	6	6		1	0	3	Asignó mal la segunda estructura y la incluyó como

Descrip. Código	Cláusula	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Preposiciones y verbos no procesados	Explicación del error
		Estructuras	Caracteres	Razonable	Estricto	Mal	Razonable	Estricto	Bien	Mal	Bien	Mal		
														un carácter.
366	5	2	3	2	1		3	3	3		1	0	1	Asignó mal la segunda estructura y la incluyó como un carácter (frutos bayas).
372	1	2	2	2	2		2	2	2		1	0	0	
373	8	4	2	4	4		2	2	2		0	0	3	
374	7	3	2	3	3		2	2	2		0	0	3	
384	4	3	9	2	2		9	9	8	1	1	0	0	Asignó un carácter a la estructura anterior. Asignó una estructura como carácter de la anterior.
386	7	3	1	3	3		1	1	1		0	0	2	
389	4	3	7	3	3		7	7	7		1	0	2	
397	6	4	2	4	4		2	2	2		0	0	2	
398	8	4	4	4	4		4	4	4		0	0	3	
400	8	4	7	4	4		7	7	7		1	0	3	
404	6	1	2	1	1		2	2	2		0	0	0	

Descrip. Código	Cláusula	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Preposiciones y verbos no procesados	Explicación del error
		Estructuras	Caracteres	Razonable	Estricto	Mal	Razonable	Estricto	Bien	Mal	Bien	Mal		
405	1	2	2	2	2		2	2	2		1	0	0	
407	4	4	9	4	4		9	8	8	1	0	0	3	Un carácter se asoció a una estructura incorrecta. Duplicó una preposición.
410	2	2	0	2	2		0	0	0		0	0	2	
415	4	6	9	6	6		9	9	9		1	0	2	
416	4	6	9	5	5		9	9	8	1	1	0	3	Un carácter se asoció a la estructura principal indebidamente. Borde se estructuró como carácter y se asoció a la estructura principal con todos sus caracteres.
417	4	5	8	5	5		8	8	8		1	0	2	
418	1	1	1	1	1		1	1	1		0	0	0	
418	5	1	1	1	1		1	1	1		0	0	1	
420	4	8	9	8	8		9	9	9		0	0	5	

Descrip. Código	Cláusula	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Preposiciones y verbos no procesados	Explicación del error
		Estructuras	Caracteres	Razonable	Estricto	Mal	Razonable	Estricto	Bien	Mal	Bien	Mal		
420	7	3	0	3	3		0	0	0		0	0	2	
422	4	5	10	5	5		10	10	10		2	0	2	
426	4	5	7	5	5		7	7	7		0	0	3	
435	3	1	2	1	1		2	2	2		0	0	0	
437	3	1	1	1	1		1	1	1		0	0	0	
438	4	3	9	3	3		9	9	9		2	0	2	
441	3	1	1	1	1		1	1	1		0	0	0	
446	4	4	8	4	4		8	8	8		0	0	3	
447	2	4	2	4	4		2	2	2		0	0	2	
448	4	6	7	6	6		7	7	7		0	0	3	
452	4	6	7	6	6		7	7	7		0	0	2	
452	7	4	0	4	4		0	0	0		0	0	2	
457	7	2	1	2	2		1	1	1		0	0	1	
458	5	1	4	1	1		4	4	4		2	0	1	
460	7	1	2	1	1		2	2	2		0	0	1	
462	7	1	2	1	1		2	2	2		0	0	0	
472	1	2	1	2	2		1	1	1		0	0	0	
473	3	1	2	1	1		2	2	2		1	0	0	

Descrip. Código	Cláusula	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Preposiciones y verbos no procesados	Explicación del error
		Estructuras	Caracteres	Razonable	Estricto	Mal	Razonable	Estricto	Bien	Mal	Bien	Mal		
473	5	3	0	3	3		0	0	0		0	0	1	
424	1	2	2	2	2		2	2	2		0	0	0	
298	1	1	1	1	1		1	1	1		0	0	0	
Total		**248**	**315**	**245**	**243**		**314**	**312**	**312**	**4**	**27**	**1**	**113**	

Datos de evaluación del Manual de plantas de Costa Rica volumen VI

Fecha: 21 de enero de 2016

Evaluadores: José Enrique Araya y María Mora

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
479	3	planta	1	1	1	1		1	1	1		0			
482	2	arbusto	2	0	2	2		0	0	0		0		1	
484	6	árbol	1	1	1	1		1	1	1		0			
486	8	árbol	1	3	1	1		2	2	3		0			No incluyó el valor del número.
488	5	árbol	2	0	2	2		0	0	0		0		1	
490	6	arbustos	9	7	9	9		7	6	6	1	0		4	Hay una repetición de preposición en el campo constraint_preposition.
493	6	árbol	2	0	2	2		0	0	0		0		1	
497	4	arbusto	5	5	5	5		5	5	4	1	0		3	Hay una asociación mal debido a ambigüedad en la cláusula.

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
499	8	hierba	1	2	1	1		2	2	2		0			
500	3	arbustos	1	2	1	1		2	2	2		0		2	
500	6	arbustos	3	0	3	3		0	0	0		0		1	
503	15	lianas	1	3	1	1		3	3	3		1			
508	5	árboles	3	6	3	3		6	5	6		1		2	Token “±” tipificado correctamente como un número. El sistema debe tratarlo como un adverbio.
517	4	hierba	3	2	3	3		2	2	2		0		2	
520	3	arbusto	6	13	6	6		13	13	8	5	3		5	Error en la asignación de caracteres a estructura debido a ambigüedad en la cláusula.
521	1	arbusto	2	4	2	2		4	4	4		1			
521	4	arbusto	2	1	2	2		1	1	1		0		1	
523	3	arbusto	4	14	4	4		14	14	8	6	4	1	3	

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
523	2	arbusto	2	0	2	2		0	0	0		0		1	
527	9	hierbas	2	2	2	2		2	2	2		0		2	
532	6	hierbas	3	0	3	3		0	0	0		0		1	
532	11	hierbas	1	2	1	1		2	2	2		0			
537	8	subarbusto	2	3	2	2		3	2	3		0		1	No expresó bien un rango por el carácter “x” adicional en “4-6x”. Caso nuevo.
540	4	arbusto	6	10	6	6		10	10	8	2	0		4	Mala asociación debido a ambigüedad en la cláusula.
540	6	arbusto	1	1	1	1		1	1	1		0			
543	1	arbusto	2	4	2	2		4	4	4		0	1		
547	4	liana	6	6	6	6		6	6	6		0	1		
547	10	liana	2	4	2	2		4	4	4		0		1	
548	4	árbol	5	7	5	5		7	7	4	3	2		3	Mala asociación debido a ambigüedad en la cláusula.

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
557	1	árbol	4	5	4	4		5	5	5		1		2	
562	1	hierbas	4	2	4	4		2	2	2		1		2	
568	1	arbusto	2	2	2	2		2	2	2		1			
569	5	arbusto	1	1	1	1		1	0	1		0			No procesó un número.
570	1	arbustos	6	10	6	6		10	9	10		5		2	
570	5	arbustos	7	8	7	7		8	7	8		2	1	5	Falló el procesamiento del símbolo ± (hay que reemplazarlo en el pre-procesamiento).
571	2	arbustos	2	3	2	2		3	3	3		1			
572	1	árbol	4	4	4	4		4	3	4		1		1	Un adverbio repetido.
573	1	árbol	4	5	4	4		5	4	5		1			Falto al procesar ±.
575	1	árbol	4	5	4	4		5	5	4	1	2			
575	2	árbol	2	3	1	1		3	2	2	1	0			Fallo al procesar un articulo + un adjetivo que

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
															actúa como modificador.
577	5	árbol	1	1	1	1		1	1	1		0			
577	7	árbol	5	5	5	5		5	5	5		2			
589	3	árbol	1	5	1	1		5	5	5		1			
590	8	árbol	1	1	1	1		1	1	1		0			
591	5	arbustos	4	2	4	4		2	2	2		0		2	
595	4	árbol	1	1	1	1		1	1	1		0			
595	5	árbol	2	2	2	2		2	2	1		1		2	
595	6	árbol	1	1	1	1		1	1	1		0			
596	1	arbusto	3	5	3	3		5	5	5		1			
601	3	arbusto	7	6	7	7		6	6	6		1		3	
602	8	arbusto	1	1	1	1		1	1	1		0			
611	2	árbol	2	0	2	2		0	0	0		0		1	
611	3	árbol	6	6	6	6		6	6	5	1	0		4	Carácter mal asociado debido a ambigüedad en la cláusula..
611	4	árbol	1	1	1	1		1	1	1		0			

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
612	2	árbol	2	0	2	2		0	0	0		0		1	
613	2	arbusto	2	0	2	2		0	0	0		0		1	
613	3	arbusto	10	6	10	10		6	6	4	2	1			Caracteres mal asociados debido a ambigüedad en la cláusula.
614	2	árbol	2	0	2	2		0	0	0		0		1	
619	3	árbol	6	7	6	6		7	7	5	2	2		4	Caracteres mal asociados debido a ambigüedad en la cláusula.
619	7	árbol	1	1	1	1		0	0	1		0			
623	3	arbusto	7	7	7	7		7	6	7		1		6	El símbolo ± debió ser reemplazado en el pre-procesamiento.
623	5	arbusto	3	6	3	3		6	3	6		0			Repetición de adverbios y el símbolo ± debió ser reemplazado en el pre-procesamiento

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
															.
626	2	arbusto	2	0	2	2		0	0	0		0		1	
626	7	arbusto	1	1	1	1		0	0	1		0		0	El valor del carácter no fue procesado (numero con punto decimal).
630	3	arbusto	8	8	8	8		8	8	5	3	2		5	Caracteres mal asociados debido a ambigüedad en la cláusula.
632	3	hierba	2	8	2	2		8	8	8		1		4	
633	3	árbol	10	10	10	10		10	10	5	5	2		7	Caracteres mal asociados debido a ambigüedad en la cláusula.
633	8	árbol	1	1	1	1		1	1	1		0		0	
636	1	árbol	2	3	2	1		3	2	3		0		0	El símbolo ± debió ser reemplazado en el pre-procesamiento. El modificador

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
															jóvenes no fue aplicado a la estructura.
638	1	árbol	2	3	2	1		3	2	3		0		0	
638	3	árbol	8	9	8	8		9	9	6	3	1		6	Caracteres mal asignados debido a ambigüedad en la cláusula.
640	3	árbol	10	5	10	10		5	5	5		0		4	
643	3	arbusto	6	6	6	6		6	6	5	1	1		5	Caracteres mal asignados debido a ambigüedad en la cláusula.
644	3	arbusto	7	8	7	7		8	8	7	1	2		5	Caracteres mal asignados debido a ambigüedad en la cláusula.
646	8	arbusto	1	1	1	1		0	0	1		0		0	Valor no procesado correctamente.
648	3	arbusto	10	8	10	10		8	8	8		2		5	
650	8	árbol	1	1	1	1		1	1	1		0		0	

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
654	4	arbusto	1	3	1	1		3	3	3		0		0	
658	7	arbusto	1	1	1	1		0	0	1		0		0	Valor no procesado correctamente.
659	1	arbusto	3	4	3	2		4	4	4		0		0	Modificador no procesado 'jóvenes'.
666	3	árbol	7	7	7	7		7	7	7		1		6	
666	8	árbol	1	1	1	1		0	0	1		0		0	
669	1	árbol	2	3	2	1		3	3	3		0		0	Modificador no procesado 'jóvenes'.
670	3	arbusto	10	10	10	10		10	9	10		3		6	Un adverbio repetido.
671	7	árbol	1	1	1	1		1	1	1		0		0	
673	3	arbusto	11	5	11	11		5	5	5		0		4	
675	1	arbusto	3	4	3	2		4	4	4		0		0	
678	3	arbusto	11	6	11	11		6	6	5	1	1		4	Caracteres mal asignados debido a ambigüedad en la cláusula.
680	1	sufrútices	3	1	3	3		1	1	1		1		0	

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
683	8	árbol	1	0	1	1		0	0	0		0		1	
684	5	arbusto	2	0	2	2		0	0	0		0		1	
685	1	hierba	5	8	5	5		8	8	6	2	0		5	Me parece que "hasta aproximadamente 2 cm de diám" se refiere más bien a tallo y no a hierba.
685	4	hierba	1	4	1	1		4	4	4		0		1	
687	7	hierba	7	11	7	7		11	10	10	1	0		3	
691	14	hierba	2	1	2	2		1	1	1		0		1	
693	3	hierbas	3	6	3	3		6	6	1	6	3		2	Creó un carácter adicional erróneamente (spp. acuaticas).
693	16	hierbas	2	3	2	2		3	2	3		0	1	1	
696	9	semiterrestre	2	1	2	2		1	0	1		0		1	El símbolo ± debió ser reemplazado en el pre-procesamiento.

Descrip. Código	Cláusula	Hábito	Cantidad de		Estructuras bien procesadas			Caracteres bien procesados		Caracteres bien asociados		Conjunciones bien asignadas		Prep. y verbos no procesados	Explicación del error
			Estruct.	Caract.	Razonab.	Estricto	Mal	Razonab.	Estricto	Bien	Mal	Bien	Mal		
702	10	acuática	3	4	3	3		4	3	2	2	0		2	Carácter con un adverbio repetido. Creo un carácter de mas.
703	9	epifítica	8	8	8	8		8	8	6	2	1		4	Caracteres mal asignados debido a ambigüedad en la cláusula.
706	11	epífita	3	3	3	3		3	3	2	1	0		3	Caracteres mal asignados debido a ambigüedad en la cláusula.
709	3	Hierba	3	1	3	3		1	1	1		1		0	
709	5	Hierba	1	4	1	1		4	4	4		1		0	
710	1	hierbas	3	1	3	2		1	1	1		1		1	Una restricción mal aplicada en estructura.
710	12	hierbas	1	2	1	1		2	2	2		0		0	
710	14	hierbas	2	1	2	2		1	1	1		0		1	
Total			**364**	**391**	**363**	**357**	**0**	**385**	**365**	**338**	**53**	**61**	**5**	**165**	

Apéndice VI - Evaluación de la tecnología disponible

	Phyton – Nltk	nlp.stanford	Gate	opennlp.apache	LingPipe	Apache Uima	Freeling
URL	http://www.nltk.org/	http://nlp.stanford.edu/	gate.ac.uk	http://opennlp.apache.org/	http://alias-i.com/lingpipe/	http://uima.apache.org	http://nlp.lsi.upc.edu/freeling/
Tipo de herramienta	General Framework	NLP herramientas y modelos	General Framework	NLP herramientas y modelos	NLP herramientas y modelos	General Framework	NLP herramientas y modelos
Origen (fecha)	2001 Department of Computer and Information Science at the University of Pennsylvania.	2010	1995. University of Sheffield (originalmente)	2010		2006.	2003. Universidad Politécnica de Cataluña
Documentación y soporte	- Natural Language Processing. OReilly 2009 (http://nltk.org/book/) - Pyhton Text Processing with NLTK Cookbook - Active community	- Sí. Comunidad activa.	- Sí. Comunidad activa. Vídeos, tutoriales, etc.	- Sí. Comunidad activa.	- Sí. Comunidad activa.	- Sí. Comunidad activa.	- Sí. Comunidad activa.
Idiomas soportados	Corpus etiquetado para otros idiomas, incluyendo chino, hindi, portugués, español, holandés y catalán. Hay un foro de discusión para los usuarios NLTK trabajan en Español.	Inglés, alemán, chino, italiano, portugués, búlgaro.	Inglés, español, chino, árabe, búlgaro, francés, alemán, hindi, italiano, cebuano, rumano, ruso.	Inglés, español, y otros.	Inglés, español y holandés.	Inglés, francés, alemán, italiano, portugués, ruso, español, sueco	Español, catalán, francés, gallego, italiano, Inglés, Ruso, Portugués, galés y asturiana. Checa y Eslovenia tienen apoyo parcial.
Lenguaje de programación	Phyton	Java	Java	Java	Java	Java, C++ y otros	C++ y Java

GUI	Si	No	Sí	No	No	No	No
Licencia	Apache	GPL + Licencia Comercial	Open source	Apache license	Open source	Apache license	Open source

Tabla 16. Evaluación de herramientas para procesamiento de lenguaje natural.

Variable	Plant Ontology Consortium (POC)	Flora of North America (FNA)	Phenotypic quality ontology (PATO)	Ontology Term Organizer (OTO)
Fecha inicio	2007	2001	2006	2012
Tipo	Ontología	Glosario	Ontología	Agregador de Ontologías (consenso)
Documenta	Anatomía, genes y fenotipos	Morfología y anatomía	Fenotipos	Morfología y anatomía
Grupos biológicos	Plantas	Plantas de Norteamérica. No incluye algas		Amplio
Interfaz gráfica	Sí	Sí	No	Sí
API	Sí	No parece	ftp para bajar	Sí
Idioma	Inglés con sinónimos español/	Inglés	Inglés	Inglés
Licencia	Libre	Libre	Libre	Libre

Tabla 17. Evaluación de ontologías disponibles para el área de aplicación

Apéndice VII - Ejemplo de una descripción completa estructurada en XML.

```
<treatment xmlns="http://www.github.com/inbio" xmlns:xsi="http://www.w3.o
rg/2001/XMLSchema-
instance"xsi:schemaLocation="http://raw.githubusercontent.com/biosemantic
s/schemas/0.0.1/semanticMarkupOutput.xsd">

<meta source="Nelson Zamora (Autor), Quírico Jiménez (Autor), Luis J.
Poveda (Autor), Claudia Aragón (Ilustradora), Diana Ávila (editor),
Rodrigo Gámez (prefacio) . Árboles de Costa Rica vol. IV. Primera edición
(26 de Augosto, 2011)">

 <processor>
  <processed_by operator="Maria Mora" date="Tue Dec 22 17:26:31 CST
  2015">
  <resource type="Ontology Terms Organizer (OTO) Glossary"/>
  <resource type="Plant Ontology"/>
  <resource type="Language analysis tool: FreeLing" version="3.1"/>
  </processed_by>

 </processor>

</meta>

<taxon_identification taxon_name="Quercus salicifolia
Nee" rank="species"/>
<description taxon_description=" Árbol pequeño a mediano, 10-25 m de
altura; tronco con la corteza escamosa, grisáceo pálido; ramitas
ennegrecidas, glabras; estípulas hasta 8 mm de largo, liguladas y
deciduas. Hojas simples, alternas, 4-10(-14) × 1-3,3 cm, elípticas,
angosto-elípticas a linear-elípticas, ápice acuminado, caudado o agudo,
base caudada u obtusa, glabras o a veces con tricomas dispersos a lo
largo de la vena central por el envés, margen entero; pecíolos hasta 1 cm
de largo. Inflorescencias masculinas y femeninas en amentos, las
masculinas 3-8,5 cm de largo, las femeninas mucho más cortas. Flores
amarillentas o cremosas, 4-6 mm de largo, sin sépalos ni pétalos y con 4-
8(-12) estambres. Frutos nueces, 0,6-1,2(-15) × 0,7-1,2(-1,3) cm,
ovoides, apiculadas en el ápice y tomentulosas; rodeadas en la base por
una cúpula o receptáculo leñoso y escamoso de 0,5-1,5 x 1,2-1,6 cm,
cónico o campanulado, pardo y con una pubescencia muy
fina."description_type="morphology">

  <statement id="T10L1" text=" árbol pequeño a mediano , 10-25 m de
  altura ;">
    <biological_entity id="T10L1S1-
    167530" name="árbol" type="structure">
      <character name="size" value="pequeño" other_constraint="a" constr
      aint_preposition="a mediano"/>
      <character name="size_or_quantity" value="10-
      25" char_type="range_value" from="10" from_unit="m" to="25" to_uni
      t="m"constraint_preposition="de altura ;"/>
    </biological_entity>
  </statement>

  <statement id="T10L2" text=" tronco con la corteza escamosa , grisáceo
  pálido ;">
```

```
  <biological_entity id="T10L2S1-
  167531" name="tronco" constraint_preposition="con la corteza
  escamosa" type="structure">
    <character name="coloration" value="grisáceo pálido"/>
  </biological_entity>

  <biological_entity id="T10L2S1-
  167532" name="corteza" type="structure"/>
</statement>

<statement id="T10L3" text=" ramitas ennegrecidas , glabras ;">
  <biological_entity id="T10L3S1-
  167533" name="ramitas" type="structure">
    <character name="coloration" value="ennegrecidas"/>
    <character name="pubescence" value="glabras"/>
  </biological_entity>
</statement>

<statement id="T10L4" text=" estípulas hasta 8 mm de largo , liguladas
y deciduas .">
  <biological_entity id="T10L4S1-
  167534" name="estípulas" constraint_preposition="hasta 8 mm de
  largo" other_constraint="hasta"type="structure">
    <character name="architecture" value="liguladas" constraint_conjun
    ction="y"/>
    <character name="duration" value="deciduas"/>
  </biological_entity>
</statement>

<statement id="T10L5" text=" hojas simples , alternas , 4-10 ( -14 ) ×
1-3,3 cm , elípticas , angosto-elípticas a linear-elípticas , ápice
acuminado , caudado o agudo , base caudada u obtusa , glabras o a veces
con tricomas dispersos a lo largo de la vena central por el envés ,
margen entero ;">
  <biological_entity id="T10L5S1-
  167535" name="hojas" type="structure">
    <character name="architecture" value="simples"/>
    <character name="arrangement" value="alternas"/>
    <character name="length" value="4-
    10" char_type="range_value" from="4" from_unit="cm" to="10" to_uni
    t="cm"/>
    <character name="atypical_range" value="-
    14" char_type="range_value" from="10" from_unit="cm" to="-
    14" to_unit="cm"/>
    <character name="width" value="1-
    3,3" char_type="range_value" from="1" from_unit="cm" to="3,3" to_u
    nit="cm"/>
    <character name="arrangement" value="elípticas" notes="Caracter
    repetido"/>
    <character name="shape" value="elípticas" notes="Caracter
    repetido"/>
    <character name="arrangement" value="angosto-
    elípticas" notes="Caracter repetido"/>
    <character name="shape" value="angosto-
    elípticas" other_constraint="a" notes="Caracter
    repetido" constraint_preposition="a linear-elípticas"/>
```

```
      <character name="pubescence" value="glabras" other_constraint="a"
      constraint_conjunction="o" constraint_preposition="a veces con
      tricomas dispersos a lo largo de la vena central por el envés"/>
    </biological_entity>

    <biological_entity id="T10L5S6-
    167536" name="ápice" type="structure">
      <character name="shape" value="acuminado"/>
      <character name="shape" value="caudado" constraint_conjunction="o"
      />
      <character name="shape" value="agudo"/>
    </biological_entity>

    <biological_entity id="T10L5S8-167537" name="base" type="structure">
      <character name="shape" value="caudada" constraint_conjunction="u"
      />
      <character name="shape" value="obtusa"/>
    </biological_entity>

    <biological_entity id="T10L5S9-
    167538" name="tricomas" type="structure"/>

    <biological_entity id="T10L5S9-
    167539" name="vena" type="structure"/>

    <biological_entity id="T10L5S9-
    167540" name="envés" type="structure"/>

    <biological_entity id="T10L5S10-
    167541" name="margen" type="structure">
      <character name="architecture" value="entero" notes="Caracter
      repetido"/>
      <character name="shape" value="entero" notes="Caracter repetido"/>
    </biological_entity>
  </statement>

  <statement id="T10L6" text=" pecíolos hasta 1 cm de largo .">
    <biological_entity id="T10L6S1-
    167542" name="pecíolos" constraint_preposition="hasta 1 cm de largo
    ." other_constraint="hasta"type="structure"/>
  </statement>

  <statement id="T10L7" text=" inflorescencias masculinas y femeninas en
  amentos , las masculinas 3-8,5 cm de largo , las femeninas mucho más
  cortas .">
    <biological_entity id="T10L7S1-
    167543" name="inflorescencias" type="structure">
      <character name="reproduction" value="masculinas" constraint_conju
      nction="y"/>
      <character name="reproduction" value="femeninas" constraint_prepos
      ition="en amentos"/>
      <character name="reproduction" value="masculinas"/>
    </biological_entity>

    <biological_entity id="T10L7S1-
    167544" name="amentos" type="structure"/>
```

```
  <biological_entity id="T10L7S2-
  167545" name="inflorescencias" constraint="masculinas" type="structu
  re">
    <character name="size_or_quantity" value="3-
    8,5" char_type="range_value" from="3" from_unit="cm" to="8,5" to_u
    nit="cm"constraint_preposition="de largo"/>
    <character name="reproduction" value="femeninas"/>
  </biological_entity>

  <biological_entity id="T10L7S3-
  167546" name="inflorescencias" constraint="femeninas" type="structur
  e">
    <character name="height" value="cortas" constraint="mucho-
    más" notes="Caracter repetido"/>
    <character name="length" value="cortas" notes="Caracter
    repetido"/>
    <character name="size" value="cortas" notes="Caracter repetido"/>
  </biological_entity>
</statement>

<statement id="T10L8" text=" flores amarillentas o cremosas , 4-6 mm de
largo , sin sépalos ni pétalos y con 4-8 ( -12 ) estambres .">
  <biological_entity id="T10L8S1-
  167547" name="flores" constraint_preposition="sin sépalos ni pétalos
  y con 4-8 ( -12 ) estambres ."type="structure">
    <character name="coloration" value="amarillentas" constraint_conju
    nction="o"/>
    <character name="coloration" value="cremosas"/>
    <character name="size_or_quantity" value="4-
    6" char_type="range_value" from="4" from_unit="mm" to="6" to_unit=
    "mm"constraint_preposition="de largo"/>
  </biological_entity>

  <biological_entity id="T10L8S3-
  167548" name="sépalos" type="structure"/>

  <biological_entity id="T10L8S3-
  167549" name="pétalos" type="structure"/>

  <biological_entity id="T10L8S3-
  167550" name="estambres" type="structure"/>
</statement>

<statement id="T10L9" text=" frutos nueces , 0,6-1,2 ( -15 ) × 0,7-1,2
( -1,3 ) cm , ovoides , apiculadas en el ápice y tomentulosas ;">
  <biological_entity id="T10L9S1-
  167551" name="frutos" constraint="nueces" type="structure">
    <character name="length" value="0,6-
    1,2" char_type="range_value" from="0,6" from_unit="cm" to="1,2" to
    _unit="cm"/>
    <character name="atypical_range" value="-
    15" char_type="range_value" from="1,2" from_unit="cm" to="-
    15" to_unit="cm"/>
    <character name="width" value="0,7-
    1,2" char_type="range_value" from="0,7" from_unit="cm" to="1,2" to
    _unit="cm"/>
```

```
      <character name="atypical_range" value="-
      1,3" char_type="range_value" from="1,2" from_unit="cm" to="-
      1,3" to_unit="cm"/>
      <character name="shape" value="ovoides"/>
      <character name="architecture" value="apiculadas" notes="Caracter
      repetido"/>
      <character name="shape" value="apiculadas" notes="Caracter
      repetido" constraint_preposition="en el ápice y tomentulosas ;"/>
    </biological_entity>

    <biological_entity id="T10L9S4-
    167552" name="ápice" type="structure"/>
  </statement>

  <statement id="T10L10" text=" rodeadas en la base por una cúpula o
  receptáculo leñoso y escamoso de 0,5-1,5 x 1,2-1,6 cm , cónico o
  campanulado , pardo y con una pubescencia muy fina .">
    <biological_entity id="T10L10S1-
    167553" name="base" type="structure"/>

    <biological_entity id="T10L10S1-
    167554" name="cúpula" type="structure"/>

    <biological_entity id="T10L10S1-
    167555" name="receptáculo" type="structure">
      <character name="shape" value="cónico" constraint_conjunction="o"/
      >
      <character name="shape" value="campanulado"/>
      <character name="coloration" value="pardo" constraint_conjunction=
      "y" constraint_preposition="con una pubescencia muy fina ."/>
    </biological_entity>
    <biological_entity id="T10L10S3-
    167556" name="pubescencia" type="structure"/>
  </statement>
</description>

</treatment>
```

Referencias bibliográficas

[1] D. Hobern, A. Apostolico, E. Arnaud, J. C. Bello, D. Canhos, G. Dubois, D. Field, E. García, A. Hardisty, J. Harrison, B. Heidorn, L. Krishtalke, E. Mata, R. Page, C. Parr, J. Price, and S. Willoughby, "Global Biodiversity Informatics Outlook: Delivering Biodiversity Knowledge in the Information Age," 2012.

[2] A. E. Thessen and C. S. Parr, "Knowledge Extraction and Semantic Annotation of Text from the Encyclopedia of Life," *PLoS One*, vol. 9, no. 3, p. e89550, 2014.

[3] J. La Salle, Q. Wheeler, P. Jackway, S. Winterton, D. Hobern, and D. Lovell, "Accelerating taxonomic discovery through automated character extraction," *Zootaxa*, vol. 55, pp. 43–55, 2009.

[4] A. D. Chapman, "Numbers of Living Species in Australia and the World," *Heritage*, vol. 2nd, no. September, p. 84, 2009.

[5] B. E. Hammel, *Manual de plantas de Costa Rica, Volume 1*. Missouri Botanical Garden Press, 2004, 2004.

[6] C. N. C. Rinaldo, "The Biodiversity Heritage Library: an Expanding International Collaboration," *Nat. Preced.*, 2009.

[7] C. S. Parr, N. Wilson, K. Schulz, P. Leary, J. Hammock, J. Rice, and R. J. Corrigan Jr., "TraitBank: Practical semantics for organism attribute data," 2014.

[8] N. (Author) Zamora, Q. (Author) Jiménez, L. J. (Author) Poveda A., and C. (Illustrator) Aragón, *Árboles de Costa Rica vol. III (Spanish Edition)*. Editorial INBio, 2004.

[9] L. Padró Cirera and E. Stanilovsky, "FreeLing 3.0: Towards Wider Multilinguality," *Int. Conf. Lang. Resour. Eval.*, pp. 2473–2479, 2012.

[10] S. Avraham, C. W. Tung, K. Ilic, P. Jaiswal, E. a. Kellogg, S. Mccouch, A. Pujar, L. Reiser, S. Y. Rhee, M. M. Sachs, M. Schaeffer, L. Stein, P. Stevens, L. Vincent, F. Zapata, and D. Ware, "The Plant Ontology Database: A community resource for plant structure and developmental stages controlled vocabulary and annotations," *Nucleic Acids Res.*, vol. 36, pp. 449–454, 2008.

[11] J. Macklin, R. A. Morris, and P. J. Morris, "OTO : Ontology Term Organizer," 2012.

[12] F. Chiang and M. Sousa, "Glosario Inglés-Español, Español-Inglés para Flora Mesoamericana." [Online]. Available: http://www.mobot.org/MOBOT/TROPICOS/Meso/Glossary/termfr.html.

[13] A. E. Thessen, H. Cui, and D. Mozzherin, "Applications of natural language processing in biodiversity science," *Adv. Bioinformatics*, vol. 2012, 2012.

[14] N. Madnani, "Getting started on natural language processing with Python," *Crossroads*, pp. 1–16, 2007.

[15] A. Hippisley, "Lexical Analysis," in *Handbook of Natural Language Procesing-Second Edition*, 2010, pp. 31–58.

[16] P. (University of G. Ljunglof and M. (Stockholm U. Wirén, “Syntactic Parsing,” in *Handbook of Natural Language Procesing- Second Edition*, 2010.

[17] C. Goddard and Andrea C. Schalley, “Semantic Analysis,” in *Handbook of Natural Language Procesing- Second Edition*, I. Nitin and F. Damerau, Eds. CRC Press`, 2010, pp. 93–120.

[18] A. Steven, *Semisupervised Learning for Computational Linguisticts*. Chapman & Hall/CRC, 2007.

[19] E. Charniak and D. McDermott, *Artificial Intelligence*. Addison-Wesley Publishing Company, 1985.

[20] J. Hobbs and E. Riloff, “Information Extraction,” in *Handbook of Natural Language Procesing- Second Edition*, Second edi., I. Nitin and F. Damerau, Eds. CRC Press`, 2010, pp. 511–526.

[21] G. Algorithms, *Introduction to Genetic Algorithms*. 2008.

[22] P. Leary, “TaxonFinder org.” [Online]. Available: http://taxonfinder.org/about.

[23] G. Sautter, K. Böhm, and D. Agosti, “A combining approach to Find All taxon names (FAT) in legacy biosystematics literature,” *Biodivers. Informatics*, vol. 3, pp. 46–58, 2006.

[24] GBIF, “DarwinCore Archive Spreadsheet Processor- Online tool,” *GBIF Secretariat*. [Online]. Available: http://www.gbif-uat.org/resource/81273.

[25] M. Gerner, G. Nenadic, and C. M. Bergman, “LINNAEUS: a species name identification system for biomedical literature.,” *BMC Bioinformatics*, vol. 11, p. 85, 2010.

[26] L. M. Akella, C. N. Norton, and H. Miller, “NetiNeti: discovery of scientific names from text using machine learning methods.,” *BMC Bioinformatics*, vol. 13, p. 211, 2012.

[27] Q. Wei, P. B. Heidorn, and C. Freeland, “Name Matters: Taxonomic Name Recognition (TNR) in Biodiversity Heritage Library (BHL),” *Methods*, pp. 3–7, 2010.

[28] C. Klingenberg, G. Sautter, D. Agosti, and T. Catapano, “GoldenGATE XML Markup Editor Introduction and Manual for the Generation of TaxonX-based Legacy Literature Documents using the GoldenGATE Editor,” p. 34.

[29] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, “GATE: an architecture for development of robust HLT applications,” *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, pp. 168–175, 2002.

[30] G. B. Curry and R. J. Connor, “Automated Extraction of Biodiversity Data from Taxonomic Descriptions,” in *Biodiversity Databases Techniques, Politics, and Applications*, 2007.

[31] J. E. Araya, “Informe final - Reestructuración descripción árboles maderables de Costa Rica,” 2009.

[32] H. Duan, Y. Hei, Z. Cui, “Heuristics based semantics annotation of biodiversity documents in Chinese,” *Chinese J. Libr. Inf. Sci.*, 2013.

[33] H. Cui, “CharaParser for Fine-Grained Semantic Annotation of Organism Morphological Descriptions,” *J. Am. Soc. Inf. Sci. Technol.*, 2012.

[34] J. P. Balhoff, W. M. Dahdul, C. R. Kothari, H. Lapp, J. G. Lundberg, P. Mabee, P. E. Midford, M. Westerfield, and T. J. Vision, “Phenex: Ontological annotation of phenotypic diversity,” *PLoS One*, vol. 5, no. 5, pp. 1–10, 2010.

[35] H. Cui and M. Studies, “MARTT : A General Approach to Automatic Markup of Taxonomic Descriptions with XML,” *CAIS*, pp. 1–11, 2005.

[36] M. M. Wood, L. S.J., V. Tablan, D. Maynard, and H. Cunningham, “Populating a Database from Parallel Texts using Ontology-based Information Extraction,” 2004.

[37] A. Rocio and S. Afredo, “X-tract: Structure extraction from botanical textual descriptions,” *Proc. string Process. Inf. Retr. Symp. Int. Work. Groupw.*, pp. 2–7, 1999.

[38] J. Diederich, R. Fortuner, and J. Milton, “Computer-assisted data extraction from the taxonomical literature,” *Department of Mathematics, University of California, Davis*, 1999. [Online]. Available: https://www.math.ucdavis.edu/~milton/genisys/terminator.html. [Accessed: 19-Mar-2015].

[39] EOL, “2013 Rubenstein Competition,” 2013. [Online]. Available: http://eol.org/info/rubenstein_2013_competition.

[40] G. A. Norton, D. J. Patterson, and M. Schneider, “LucID : A Multimedia Educational Tool for Identification and Diagnostics,” *Int. J. Innov. Sci. Math. Educ. (formerly CAL-laborate Int.*, 2012.

[41] PO-Consortium, “The Plant Ontology Consortium and plant ontologies.,” *Comp. Funct. Genomics*, vol. 3, no. 2, pp. 137–42, 2002.

[42] R. Jones, A. Mccallum, K. Nigam, and E. Riloff, “Bootstrapping for Text Learning Tasks,” *IJCAI-99 Work. Text Min. Found. Tech. Appl.*, pp. 52–63, 1999.