Pontificia Universidad Católica de Valparaíso
Facultad de Filosofía y Educación
Instituto de literatura y ciencias del lenguaje

Análisis y clasificación de nombres propios en artículos de geopolítica de la revista *Le Monde Diplomatique*: una aproximación desde la gramática del texto

Tesis para optar al grado académico de Licenciado en Lengua y Literatura Hispánica

Profesor guía:
Rogelio Nazar

Alumno:
Patricio Arriagada

Enmarcado en el proyecto FONDECYT 11140686

Viña del Mar, enero de 2016

Dedicatorias:

A Alicia Soto, mi madre, por su apoyo incondicional durante mis 25 años de vida. A Patricio Arriagada, mi padre, por su omnipresente protección desde el “más allá” y por heredarme su inquietud por intentar saberlo todo. Al profesor doctor Rogelio Nazar, por su noble amistad y por haberme enseñado tanto:

> Sé de cierto hombre, ya entrado en los sesenta, una eminencia en griego, latín, matemáticas, filósofo, médico, sin rival en todas estas cosas, que, al margen de todo, hace ya más de veinte años estruja su mollera y se atormenta con el estudio de la gramática. Pues bien, según me dice, sería feliz si pudiera vivir hasta establecer con certeza la distinción entre las ocho partes de la oración, algo que ni escritores griegos ni latinos lograron hacer de forma definitiva (Erasmo, 1511 [1984]: 97-98)

Contenido

1. Introducción

Según *Google Trends*[1], entre el año 2013 y el 2014, las tendencias de búsqueda correspondientes a nombres propios (NP) alcanzaron el 82% y 66%, respectivamente. Estas búsquedas se restringen solo a la clasificación canónica de los NP, que incluye las categorías de antropónimos, topónimos y nombres de organizaciones. Si se incluyeran otras clasificaciones de NP, el total de búsquedas superaría el 90%. Este interés de los usuarios demuestra que el estudio de los NP se constituye como un nicho atractivo para los investigadores del lenguaje natural, especialistas en minería de datos y recuperación de información, entre otras disciplinas.

El reconocimiento y clasificación de las entidades ayuda a mejorar las aplicaciones cuyo propósito es la recuperación/extracción de información y la búsqueda de respuestas (Fernández, Muñoz & Suárez, 2004; Manning et al 2009). Un antecedente importante en el marco de los sistemas de extracción de información es el trabajo realizado en las *Message Understanding Conferences* (MUC), celebradas desde 1987. A partir de la evaluación externa de los sistemas de extracción de información, el reconocimiento de entidades es un componente que requiere tanto del análisis léxico, sintáctico y, tal como proponemos en este trabajo, de la lingüística textual. Desde las primeras conferencias MUC, se ha avanzado considerablemente; sin embargo, el nicho de investigación continúa abierto hasta la actualidad, pues los sistemas de análisis aún no alcanzan su plenitud de desarrollo.

El propósito de esta investigación es determinar qué elementos pueden ser útiles en tanto predictores del tipo de entidad (NP) con la que coaparecen. El análisis se concentró en la conducta de los verbos y su capacidad de ofrecer pistas respecto al tipo de entidad que acompañan cuando están en una relación predicado/argumento con un NP. Por ende, este trabajo tendrá un carácter descriptivo y también predictivo, ya que se pretende determinar el potencial grado de predictibilidad de los verbos. Así, el sujeto de un verbo como *considerar* será, probablemente, un sujeto humano. Pero también otro tipo de elementos puede servir también como predictor. A modo ilustrativo, considérese el ejemplo de un fragmento de texto etiquetado por *Freeling* en

[1] https://www.google.es/trends/ [Con acceso: 20/11/2015]

la tabla 1. Buena parte de todas las ocurrencias de Ottawa y Toronto contendrán la palabra “ciudad” en su contexto, por lo tanto, su aparición en el contexto puede servir como predictor del tipo de entidad mencionada como un topónimo.

Su	Su	DP3CS0
Capital	Capital	NCFS000
Es	Ser	VSIP3S0
La	El	DA0FS0
Ciudad	ciudad	NCFS000
De	De	SPS00
Ottawa	ottawa	NP00G00
Y	Y	CC
La	El	DA0FS0
Ciudad	ciudad	NCFS000
Más	Más	RG
poblada	poblar	VMP00SF
Es	Ser	VSIP3S0
Toronto	toronto	NP00G00
.	.	Fp

Tabla 1 Ejemplo del análisis de Freeling en la clasificación de entidades

La nomenclatura utilizada en este sintético análisis corresponde al etiquetario *Eagles*[2], que se impone aún como estándar en el etiquetado de analizadores morfosintácticos en lexicones y corpus de lenguas europeas. Siguiendo este estándar, todo elemento etiquetado como *NP000G00* corresponderá a un topónimo, *NP000SP* corresponderá a una persona, *NP000O0* corresponderá a una organización y *NP000V0* aplica a la clasificación semántica “otros”.

La metodología empleada en esta tesina de pregrado está basada en una rutina de experimentación y observación del comportamiento de los verbos, en tanto elemento predictor de determinadas categorías de nombres propios que le acompañen (antropónimos, topónimos o nombre de organizaciones). La principal herramienta

[2] Mayor información en http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html [Con acceso: 20/11/15]

usada en la investigación es la suite de analizadores de idiomas *Freeling*[3]. A partir de un estudio de corpus, se observó el comportamiento de los elementos del contexto a nivel textual, lo que puede arrojar información acuciosa respecto al tipo de entidad que acompañan. Este trabajo se posiciona desde la gramática del texto, pues se consideró como objeto de estudio al texto y no solo a la oración (nivel sintáctico). La entidad a la que hace referencia el nombre propio siempre es la misma a lo largo de un texto, sin embargo, estas referencias no siempre señalan de manera explícita el objeto referido. Así, los mecanismos de correferencia textual son un elemento teórico y metodológico a considerar en la experimentación desarrollada con los nombres propios en el corpus.

Entre las potenciales aplicaciones de esta investigación se propone el tratamiento, extracción y desambiguación de nombres de entidades en corpus textuales de grandes dimensiones. Por ejemplo, un especialista que se encuentre recabando antecedentes sobre una organización comercial llamada "París" puede disponer de acceso al archivo digital de diversos periódicos y encontrar publicaciones referidas a esta entidad. Sin embargo, en primera instancia, una búsqueda utilizando los parámetros comunes de una base de datos de este tipo, arrojará miles de resultados que incluirán, además de algunas publicaciones sobre la organización en cuestión, artículos sobre la capital de Francia y sobre distintos sujetos que tienen ese apellido, con lo cual el trabajo de reconocimiento y clasificación manual de estos resultados resultaría extremadamente laborioso.

La presente tesina de pregrado tiene carácter exploratorio debido a la carencia de bibliografía específica respecto a la potencial capacidad predictiva que pueden poseer los verbos para determinar a qué tipo de entidad (antropónimo, topónimo, nombre de organización) pertenece un nombre propio. La pregunta de investigación *"¿puede el verbo que acompaña al nombre propio funcionar como predictor de una determinada categoría de nombre propio?"* será abordada a partir de la experimentación con datos y valores cuantitativos hallados en el corpus de geopolítica. En este sentido, este trabajo aborda la problemática desde un enfoque descriptivo, toda vez que se da cuenta de una eventual relación de interdependencia entre la posición gramatical de los nombres propios (considerando sus apariciones en el texto) y los verbos que les acompañan en

[3] http://nlp.lsi.upc.edu/freeling/ [Con acceso: 20/11/2015]

los enunciados en cuestión. La presente investigación se enmarca en el Proyecto FONDECYT 11140686: “Inducción automática de taxonomías de sustantivos generales y especializados a partir de corpus textuales desde el enfoque de la lingüística cuantitativa”, por lo que se agradece el apoyo del equipo y el liderazgo del doctor Nazar.

Por lo tanto, el objetivo general de esta tesina será determinar cuáles pueden ser los rasgos que faciliten la predicción sistemática del tipo de entidad al que pertenece un nombre, considerando la posición de los verbos. Un objetivo específico de la misma será determinar el grado de precisión con que actúa el verbo como elemento predictor. En este sentido, se ha empleado un corpus de geopolítica de la revista *Le Monde Diplomatique,* el cual es rico en nombres propios debido a la naturaleza del género. El sustento teórico pertinente permite la formulación de algunas hipótesis para abordar esta problemática y dilucidar si existe algún principio, patrón o recurrencia presente en los textos respecto a la relación del elemento verbo y el nombre propio que le antecede.

2. Marco teórico

El estudio de los nombres propios se constituye como un nicho de reciente exploración para los estudiosos del lenguaje que trabajan en el procesamiento del lenguaje natural. Sin embargo, esta temática ha suscitado el interés de diversos pensadores desde hace más de 2000 años. Un antecedente relevante para nuestra cultura occidental es el *Crátilo* de Platón (360 A. C.), diálogo en el cual se discute si el significado de las palabras se les atribuye de forma natural o si es arbitrario y depende de las costumbres (o creencias) de los sujetos que nominan objetos. Así, fueron planteadas las más primigenias interrogantes que la lingüística moderna se ha encargado de deliberar respecto al rol del significante, el significado, el referente y el nombre mismo.

Para contextualizar históricamente esta investigación, resulta importante destacar que entre los siglos V y VI las ciencias y las artes se escindieron. Las artes liberales que abrazaban el Trivium y el Quadrivium eran consideradas la mayor aproximación al esfuerzo del entendimiento humano. El Trivium abrazaba las artes que hoy conocemos como letras y humanidades, por su parte, el Quadrivium abarcaba lo que hoy conocemos como ciencias exactas. Ambas constituían aquello que los hombres más instruidos consideraban que conducía a la sabiduría.

Tal división de los saberes ha dado lugar a creencias que promulgan que el mundo de las humanidades y de los números son esferas de saber irreconciliables y contradictorias. Por otra parte, es usual advertir que especialistas de las ciencias exactas manifiesten ciertos cuestionamientos respecto del trabajo de los especialistas de las humanidades, por no emplear de manera acuciosa el método científico en sus aportes disciplinares; esta situación también suele presentarse cuando los profesionales de las artes y humanidades prescinden de la ciencia en su trabajo. Como enunció oportunamente Snow (1959), se abrió en el mundo occidental una profunda barrera de ignorancia y prejuicios entre una “cultura literaria” y una “cultura científica”. La cultura científica es la abanderada del futuro y la modernidad; por su parte, la cultura literaria obvia y desconoce los descubrimientos e innovaciones de la ciencia, pretendiendo continuar con la comprensión y administración de las sociedades desde

una óptima meramente cultural. Esta discusión ha transcurrido rauda durante décadas pero, afortunadamente, hoy es posible que un experto en lenguas colabore activamente con un experto en ciencia computacional.

Desde esta perspectiva, la lingüística computacional es un campo de estudio que integra tanto el conocimiento de los expertos en lingüística como las herramientas aportadas por la informática: es un área esencialmente interdisciplinar e integral, por lo que sus horizontes resultan promisorios al emplear los saberes de ambos "mundos".

2.1 Problemática

Los expertos en procesamiento del lenguaje natural se esfuerzan por desarrollar sistemas automáticos que permitan acceder de forma más eficaz y rápida al conocimiento. Esta tarea se torna compleja teniendo en cuenta la enorme cantidad de información que circula en la red, por lo tanto, no se puede dejar de considerar el importante papel que desempeñan los nombres propios en el procesamiento de textos pertenecientes a distintos géneros. Su relevancia y riqueza radica en que, a diferencia de los nombres comunes, los nombres propios refieren a una única entidad que hace alusión generalmente a personas, lugares o instituciones.

Para referirnos a este tipo de entidades hacemos uso de los nombres propios y cada vez que interactuamos con plataformas de acceso a información como *Google* hacemos uso de estos. Por ejemplo, cuando nos referimos a algún cuadro de Pablo Picasso, es común obviar el nombre de la obra y solo señalar que nos encontramos frente a "un Picasso" (siguiendo a la RAE, 2009); o cuando decimos "Valencia", podemos referirnos a la ciudad de España, a su equipo de fútbol, a una institución bancaria, al apellido de una persona natural, etc. Otro rasgo particular de los NP es su flexión fija, mediante la que se distingue del NC por carecer de flexión, salvo en casos muy específicos como los comentados por Coseriu (1982) donde el plural de los NP es indicio de que su uso se asimila al de un NC (ej: *En la galería se exhiben varios Picassos*).

Los seres humanos pueden hacer uso de su competencia comunicativa para adecuar las acepciones de sus enunciados según el contexto lingüístico en que se lleve

a cabo el acto comunicativo, sin embargo, los sistemas computarizados no gozan de esta facultad: en este punto el trabajo de los lingüistas y los ingenieros informáticos confluye para desarrollar herramientas más eficaces de acceso a la información.

Se pueden construir referencias mediante un sustantivo acompañado de otras palabras como preposiciones, adjetivos, adverbios u otras variaciones posibles: “el intendente de Valparaíso”, “la tercera marcha por la educación”, son algunos simples ejemplos de entidades nombradas que difieren de los nombres propios propiamente tales. En este punto, cabe señalar una característica prototípica de los NP: su potencial capacidad denotativa. Bertrand Russell plantea en su ensayo “On denoting” (1905) que solo es posible acceder a algunos conocimientos (entiéndase también objetos) a través de frases denotativas, a través del acercamiento directo hacia muchas cosas de las que no tenemos conocimiento directo (“*knowledge about”)*. “Cuando distinguimos entre significado (“*meaning”*) y denotación (*“denoting”*), debemos estar hablando acerca del significado: este tiene denotación y es un complejo, y no hay otra cosa aparte del significado que pueda ser llamado el complejo y de lo que pueda decirse que *tiene* significado y denotación. [...] Algunos significados tienen denotaciones” (Russell, 40). Entonces, al enunciar “el intendente de Valparaíso”, no es posible determinar a qué sujeto (antropónimo) se refiere específicamente el enunciado, a menos que se posea mayor información textual. Como señala Russell, “una frase denotativa es esencialmente una parte de una oración, y no posee, al igual que la mayoría de las palabras aisladas, una significación [*significance*] propia” (41). Son las características inherentes del objeto, sin valor léxico, las que lo revisten del carácter denotativo.

2.1.1 Tratamiento reciente de las investigaciones relacionadas con los nombres propios

Desde la década de 1980 han sido creadas varias iniciativas para clasificar los nombres en diferentes categorías que faciliten posteriores análisis de los textos. Las primeras instancias fueron difundidas en las conferencias MUC (Message Understanding Conferences) organizadas por el NIST (National Institute for Standards and Technology) que estaban dedicadas a promover nuevos métodos para la

extracción de información, teniendo como tarea esencial el poder identificar de manera automática ciertas unidades de información en los textos: entidades nombradas (personas, lugares, organizaciones). Las Conferencias MUC tenían por objetivo evaluar y fomentar la investigación en el análisis automatizado de mensajes de tipo militar que contenían información textual. La característica distintiva de las MUC no eran las propias conferencias, sino las evaluaciones a las que debían ser sometidos los participantes para poder acceder a estas reuniones (R. Grishman & Sundheim 1996).

El enorme crecimiento de la información digital en la actualidad hace necesario el desarrollo de sistemas que puedan procesar, analizar y explotar el potencial que esta posee. Tales tareas se realizan por medio de aplicaciones dirigidas a la extracción de información (*information extraction)*, recuperación de información (*information retrieval*) y la búsqueda de respuestas (*question answering)*. A esta tarea que se encarga de reconocer y clasificar las entidades nombradas se le conoce como NER (*named-entity recognition*), la cual consiste en identificar de manera automática todos los nombres de personas, organizaciones y lugares geográficos en un texto; estos serán los tipos de entidades que se pueden encontrar. En otras palabras, los sistemas de reconocimiento de entidades NER, se encargan de detectar y clasificar las entidades que aparecen en los textos, las cuales son categorizadas según tipo de entidad (Ferrández et al 2005).

Existen numerosos programas computacionales desarrollados para la clasificación de estas entidades, es decir, de asignarles una categoría predefinida: al procesar la frase “Michelle Bachelet visitó Temuco en una visita presidencial”, se identificarían dos entidades nombradas. “Michelle Bachelet” correspondería a una entidad de persona y “Temuco” a una entidad de lugar; sin embargo, la tarea no resulta tan clara. Hay muchas clases de entidades que son ambiguas porque un mismo nombre puede estarse refiriendo a entidades distintas, lo cual puede conducir a error (anteriormente se señala las distintas referencias que puede tener el nombre propio “Valencia”, por ejemplo).

Este problema de identificación se extiende hasta la actualidad porque los sistemas automáticos carecen de la competencia para identificar cuándo se hace alusión a un nombre propio si se le llama de distintas maneras: “Maradona”, “Diego Armando”, “el pibe de oro” se refieren a la misma persona pero se torna difícil poder

programar un sistema para que pueda captar estas relaciones. Aunque no forma parte de este proyecto de investigación, el problema se agudiza aún más cuando se intenta trabajar con nombres propios en plataformas multilingües, pues las traducciones no logran adaptar los términos a otra lengua con exactitud: por ejemplo, el nombre propio del teórico “Mijaíl Bajtín” resulta problemático ya que en inglés su apellido suele aparecer como “Bakhtin” y en francés como “Bakhtine” (en el original cirílico se escribiría Михаи́л Бахти́н).

Para resolver estos problemas al momento de clasificar correctamente las distintas clases de nombres propios al que puede enfrentarse un sistema, se han implementado dos métodos: los sistemas basados en reglas y los estadísticos. En caso de los estadísticos se encuentran los supervisados y los no supervisados. Los sistemas basados en reglas se crean de forma manual de acuerdo a la situación investigativa particular (según lengua, tipos de textos, registros de habla, etc.), y los estadísticos emplean corpus textuales que ya han sido analizados con anterioridad, a fin de inducir patrones o reglas que faciliten la automatización de la tarea (mediante ensayo y error). Actualmente se trabaja con sistemas semi supervisados (como se aprecia en la bibliografía empleada) los cuales a partir de unos ejemplos previos (con resultados exitosos) buscan extender las reglas para ir mejorándolas paulatinamente.

2.2 Discusión teórica

2.2.1 Sobre el nombre propio, algunas definiciones y aproximaciones

Una aproximación teórica al nombre propio ha sido realizada por Fernández Leborans, (1999). Resulta necesario diferenciar el nombre propio del nombre común, pues existen ciertos rasgos característicos del primero que se repiten con regularidad. A saber: introducción mediante mayúscula, flexión fija, falta de significado léxico (existe primacía de la designación o referencia por sobre la significación), ausencia de determinante (en la función referencial prototípica) e imposibilidad de traducción. No obstante, respecto a su imposibilidad de traducción, es posible identificar ciertos matices cuando un “Miguel” puede pasar a “Michael” o a “Mijail” inclusive, los cuales tienen un carácter arbitrario, en tanto convenciones traductológicas. Existen autores

que difieren de esta caracterización de tipo taxonómico. A saber, Jonasson (1994) considera que, cualquiera que sea el criterio adoptado, no se logra distinguir claramente los nombres propios de los nombres comunes, debido a la falta de precisión de la noción de nombre propio.

Respecto a la categorización de los nombres propios, es posible identificar tres subclases: antropónimos (o nombres de persona), topónimos (o nombres de lugares) y nombres de organizaciones. Los nombres propios “puros” son los antropónimos y los topónimos, pues están constituidos por formas léxicas especializadas en la función de nombre propio. Los nombres de organizaciones pueden estar acompañados de nombres comunes con determinación, eventualmente acompañados por modificadores adjetivos o prepositivos (ej. Real Academia Española) por lo que difieren de los nombres propios “puros”.

Los nombres propios también poseen género y si bien este no está respaldado por una pauta de formación regular en cada lengua, es posible señalar que se determina normalmente por el género del nombre apelativo que especifican. Así, los NNPP de ciudades son normalmente femeninos (cuando coinciden el significado y la terminación, ejemplo: Sevilla, La Habana). Por otra parte, los NNPP de accidentes geográficos son masculinos siempre y cuando estos accidentes denominen lugares tales como ríos, montes, países, lagos, océanos, etc. En tercer lugar, los NNPP de instituciones u organizaciones ven determinado su género según el nombre apelativo que especifican: la PUCV (universidad), el Santiago Bernabeu (estadio), el María Guerrero (teatro).

También pueden tener lugar variaciones de número los nombres de pila cuando se “pluralizan”, adoptando los alomorfos “-s” o “-es” (las Teresas, las Pilares, etc.). En cambio, cuando se trata de apellidos, existe clara tendencia a no modificar el nombre y realizar el plural solo en el artículo (los Pérez, los González, etc.). Sin embargo, Fernández Leborans deja constancia de que el nombre propio no puede tener “semánticamente” plural aunque pueda aplicarse a una pluralidad de objetos: “la pluralidad es tal desde el punto de vista de los objetos, y no desde el punto de vista de la designación: en cuanto nombrada por un nombre propio, la pluralidad se vuelve un individuo…” (Coseriu, 1955:281 en op cit).

2.2.2 Contenido de los nombres propios

Uno de los atributos que distinguen al NP del nombre común es su modo de designación, ya que el primero designa únicamente personas, lugares o instituciones de un modo único y propio de manera que no puedan designar a otro objeto diverso del designado; en tanto el nombre común, incluye en su conjunto a todos los seres de una misma especie (Charadeau, 1992). Por otra parte, los NNPP no pueden ser objeto de definición léxica, lo que sí puede realizarse con el nombre común.

A partir de 1960 se adoptó la teoría del racimo, la cual plantea que el sentido de un nombre propio no se asocia con una sola descripción, sino con un conjunto -o racimo- inespecificado e indefinido de descripciones que convienen al referente. Además, Fernández Leborans recoge la propuesta de Searle (1967) en la que se da cuenta del sentido impreciso de los NNPP. Su valor informativo no incluye ni cuántas ni cuáles han de ser las características descriptivas que constituyen la identidad de su referente: “los nombres propios funcionan no como descripciones, sino como ganchos para colgar descripciones”. Lo que importa al gramático es si el contenido de los NNPP es expresable en términos de rasgos léxico-semánticos, que permitan distinguirlo del nombre común.

Es posible señalar que los nombres propios poseen naturaleza cognitiva, pues siguiendo el trabajo de Jonasson (1994), el nombre propio no puede ser adecuadamente caracterizado por la referencia, ni por el significado; sino que se revela en un plano más profundo, el cognitivo. La función cognitiva del NP consiste en nombrar, afirmar y mantener una individualidad; los NNPP son depositados en la memoria a largo plazo, asociados a un conocimiento específico, directamente a la imagen de un particular. Por su parte, los nombres comunes al poseer significado léxico codificado, se almacenan asociados a un conocimiento general, a un concepto, aplicándose a un número indefinido de particulares.

2.2.3 Sobre su capacidad de designación y denotación

Mill (1843) señaló que, si bien la designación de un objeto a través de su NP no dice nada respecto al objeto, da lugar a la existencia de informaciones asociadas al mismo NP. Dicho de otro modo, aunque el NP carece de significado léxico, evoca lo que el interlocutor ya sabe de su referente. Por su parte, Russell (1912) explicó la descripción en tanto relación con un tipo de conocimiento, con una relación cognitiva no directa con el objeto; de este manera, el referente del NP es conocido por descripción como el objeto que es "el tal-y-tal", como "el *x* que tiene tal propiedad y tal otra". En este sentido, el interés del gramático respecto al contenido de los nombres propios y propiedades de designación, obedecen a evidenciar que la diferencia entre NP y NC guarda relación con su intensión: el NP es un concepto que se asocia a un individuo, el NC se asocia a un conjunto o clase (pleno de información léxica pero no de propiedades de designación).

Respecto a la designación del NP, Kripke (1972) define a un designador rígido como un designador que designa el mismo objeto en todos los mundos posibles. Un NP se asocia a su referente en el ámbito de un enunciado determinado. El designador rígido no es el que siempre designa al mismo individuo, sino el que designa el mismo referente en cualquier mundo posible asociado a un a enunciado (Gary-Prieur, 1994). Un ejemplo de este fenómeno se aprecia a continuación:

a) Si *Napoleón* no hubiera sido francés, el destino de Europa hubiera sido diferente.
b) Si *Napoleón* continúa persiguiendo al gato, tendré que encerrarlo en su jaula.

Es posible advertir que el NP *Napoleón* designa en los casos a) y b) un mismo referente en el mundo real –correspondiente al enunciado- y en el mundo posible –representado por la construcción hipotética- y, por otro parte, el NP no refiere al mismo individuo en a) y en b); o sea, no hay lugar para ambigüedad referencial. La función de designación rígida implica irreductibilidad a una descripción definida cualquiera, que conviene a un individuo designado. A diferencia de un NP, las descripciones definidas no pueden designar a un mismo individuo invariablemente, pues son designadores no-rígidos; cambian de referencia de un mundo posible a otro. Las descripciones definidas no importan tanto por la adecuación de la propiedad descrita a un determinado

referente, como por la capacidad de designarlo de modo único (Fernández Leborans, *op cit*).

En otro ámbito de la designación, al "multiplicar" un NP este pasa a usarse como NC. Al enunciar "todos los Teun van Dijk que han estudiado el discuso", "van Dijk" adquiere el sentido de NC, toda vez que se presenten distintas facetas suyas con recursos sintácticos que representan a distintas entidades, distintos atributos que constituyen su evolución intelectual a lo largo del tiempo.

2.2.4 Los verbos y sus clasificaciones semánticas

De Miguel (1999), señala que el aspecto informa sobre la manera en que un evento se desarrolla u ocurre: implicando un cambio (por ejemplo, en el caso de madurar) o la ausencia de cambio (por ejemplo en el caso de *estar azul*); alcanzando un límite (por ejemplo, *llegar*) o careciendo de él (*viajar*); de forma única (por ejemplo, *disparar*) o repetida (*ametrallar*); de forma permanente (*ser chileno*), habitual (*cortejar*) o intermitente (*parpadear*). Además, señala que el aspecto puede aportar información respecto a la extensión temporal del evento (periodo no acotado de tiempo, acotado o un instante), sobre cuál es la fase principal del evento descrito (fase inicial, fase media o fase final) o sobre la intensidad con que un evento tiene lugar (*peinar* tiene intensidad neutra respecto a *repeinar* que es intensivo incrementativo o *atusar* que corresponde a un atenuativo), entre diversas clasificaciones.

En español, la información relativa al evento puede contenerse en la raíz verbal, tal como en *llegar* frente a *viajar*; en este caso, el comportamiento sintáctico del verbo es el que permite discriminar su información aspectual. Ciertos elementos del contexto como morfemas derivativos (como ocurre en *repeinar* respecto a *peinar*), morfemas flexivos (*condujeron* respecto a *conducir*), por perífrasis (*ando buscando la cartera* respecto a *andarán buscando la cartera*), entre otros, pueden entregar información aspectual respecto al verbo. Por otra parte, y en el caso concreto del español, la información aspectual puede ser proporcionada por las unidades léxicas cuando funcionan como predicados, no solo los verbos.

En este sentido, los verbos portan, por el contenido semántico de su raíz, la información relacionada con el modo en que tiene lugar el evento que describen (con o sin límite, con o sin duración, de manera única o repetida, entre otras informaciones que puede entregar). Esta denominación léxico-semántica se conoce tradicionalmente como *aktionsart* o "modo de acción" o inclusive como aspecto léxico (De Miguel, op cit). El término *aktionsart* fue empleado por primera vez por Sigurd Agrell en 1908, a fin de describir el sistema temporal de la lengua polaca. Con este término indicó las funciones de significado de la composición verbal, las cuales expresan de una manera más exacta cómo la acción se realizará y que marcan la clase y el modo de su realización. Además, Agrell apunta dos perspectivas diferentes del análisis del verbo: la morfológica y la semántica (Pawlak, 2008).

A su vez, De Miguel (op cit) diferencia el aspecto léxico del aspecto flexivo, señalando que este último corresponde a la información relativa al modo en que tiene lugar un evento que viene proporcionada por los morfemas flexivos del verbo. Por ejemplo, en *Juan viajaba por Europa*, el evento denotado por el predicado está sin delimitar; pero en *Juan viajó mucho hasta que conoció a Teresa*, el evento de *viajar* se presenta como un todo concluido y delimitado, debido a la información entregada por el pretérito perfecto *viajó*.

En síntesis, el aspecto corresponde a la función discursiva que explica cómo se desarrolla o sucede un evento, informando al mismo tiempo de la extensión temporal del mismo. Por otra parte, el aspecto realiza un tratamiento del tiempo como una propiedad innata del mismo evento, mostrando su desarrollo o su distribución en el tiempo, a la vez que no se hace referencia al tiempo de habla (se entiende que este está contenido en el mismo verbo). El aspecto puede manifestarse de dos modos: mediante la morfología flexiva (aspecto flexivo) o a través del contenido semántico de la raíz del verbo (aspecto léxico o *aktionsart*). La información referida a la estructura interna del verbo, en la que tiene lugar un evento o estado y que viene proporcionada por los morfemas flexivos del verbo, corresponden al aspecto flexivo. A su vez y a modo de ejemplo, la información del *sufijo /-aba/* en *jug/-aba/* indica que la acción aún no ha concluido en relación con un punto en el pasado (De Miguel, 1999).

La discusión teórica referida al aspecto cobró mayor relevancia a partir del trabajo de Zeno Vendler (1967), filósofo del lenguaje que profundizó los estudios del aspecto léxico, cuantificadores y la nominalización. Este autor elaboró una categorización semántica de los verbos, compuesta de cuatro clases: *states* (estados)*, activities* (actividades)*, accomplishments* (realizaciones) y *achievements* (logros). Como se ha señalado anteriormente, esta forma de la aspectualidad también ha sido denominada *aktionsart* (Bibiloni, 1999). Los verbos que corresponden a *states* y *activities* (atélicos) pueden continuar indefinidamente y se les puede encontrar frecuentemente en su forma imperfectiva (*María juega con Pedro*). Por su parte, los verbos de *accomplishments* y *achievements* (télicos) tienen un punto final inherente a ellos, por lo que se les suele hallar en su forma perfectiva (*Pedro encendió la luz*). Los verbos de *accomplishments* se presentan como durativos, mientras que los verbos de *achievements* gozan de puntualidad; los verbos de *states* y *activities* contrastan en cuanto a su dinamicidad (Delgado Díaz & Ortiz López, 2012).

En la siguiente tabla se sintetiza la clasificación aspectual de los verbos (Vendler, 1957):

Predicado	Estado	Actividad	Realización	Logro
	Saber	*Correr*	*Escribir*	*Encontrar*
Dinamicidad	-	+	+	+
Telicidad	-	-	+	+
Puntualidad	-	-	-	+

Tabla 2 Clasificación aspectual de los verbos según Vendler (1957)

A partir de la observación de la tabla 2, se puede colegir que existe una relación intrínseca entre el aspecto y el tiempo, pues ambos conceptos se vinculan con la temporalidad de los eventos verbales. Como advierte Vendler (op cit), el tiempo sitúa el evento verbal en un tiempo externo y el aspecto aborda la estructura interna del evento o estado, lo cual implica que se puede aludir a un evento no terminado en el pasado. Esta distinción se explicita con el siguiente ejemplo: *Ayer hablé con María mientras comía el almuerzo*.

En esta oración queda clara la distinción entre el tiempo verbal *hablé*, el cual denota el tiempo en que se efectuó la acción, frente a la estructura del evento *comía*. Por otra parte, en el caso del aspecto flexivo frente al tiempo, se distingue la oposición entre el pretérito y el imperfecto en los siguientes ejemplos: *Comí manzanas (a); Comía manzanas (b)*. En el ejemplo *(a)*, queda de manifiesto que el evento de comer refiere a un caso único; en *(b)* el evento refiere a un evento habitual o durativo. Ambas distinciones pueden establecerse a partir de la información ofrecida por el aspecto léxico, contenido en la flexión verbal.

Como señalan Gibrán Delgado & Ortiz López (op cit), ciertos contextos están favorecidos por el pretérito y el imperfecto como se muestra en la siguiente tabla:

Forma temporal	**Contexto**
Pretérito	[+ pasado, + terminado]
Imperfecto	[+ pasado, - terminado]
Ambiguo	[+/- pasado, +/- terminado]

Tabla 3 Formas temporales y su relación con el contexto

Otros autores como Montrul & Slabakova (2000), interpretan el imperfecto como habitual (*Almorzaba al mediodía todos los días*) o durativo (*Comía el almuerzo mientras escuchaba la radio*). Ambas referencias omiten el punto inicial y final del evento; así, un evento habitual se define dentro de una serie de eventos terminada, y un evento durativo se desarrolla dentro de un determinado periodo de tiempo. El valor semántico de los verbos incide en la forma del pasado en que este aparecerá, los verbos atélicos (*María fue maestra; Juan corrió mucho; Pedro era maestro; Pablo corría mucho*) suelen presentarse en el imperfecto, mientras que los verbos télicos (*Pedro escribió un poema; Carlos prendió la luz; Mario escribía un poema; Alberto prendía la luz*) ocurren en el pretérito.

2.2.5 Análisis sintáctico de dependencias

En esta investigación se ha empleado el programa de código abierto *Freeling* (http://nlp.lsi.upc.edu/freeling/), el cual realiza diversas tareas asociadas al tratamiento del lenguaje natural. En este apartado, se enunciará el análisis sintáctico de dependencias realizado por *Freeling* y se explicará brevemente en qué consiste la tarea del análisis de dependencias.

El estudio de las relaciones formales entre un conjunto de palabras que forman enunciados o estructuras más grandes constituye el objeto de estudio de la sintaxis, la cual ha tenido desde los aportes del generativismo chomskiano, un amplio desarrollo en las ciencias de la computación y actualmente, en el tratamiento del lenguaje natural. Como señala Villayandre Llamazares (2010), el punto de partida son las categorías gramaticales o clases de palabras tales como nombres, verbos, preposiciones, entre otros, las cuales se agrupan en clases abiertas, clases cerradas, palabras funcionales, etc. El proceso en el que se asignan etiquetas a las palabras que componen un texto se llama *tagging* y experimenta complicaciones al realizarse de manera automática, pues las palabras presentan ambigüedad al momento de ser etiquetadas.

Al pasar del etiquetado morfológico al sintáctico se presentan una infinidad de dificultades, las cuales son subsanadas mediante el ingreso de nuevas reglas al sistema automatizado de *tagging*. Estas reglas pueden confeccionarse de manera manual (por ejemplo, una palabra ambigua será un nombre si está precedida por un determinante), mediante estadísticas obtenidas de manera automática a partir de un corpus de entrenamiento del cual el etiquetador infiere las reglas (por ejemplo, asignando a las palabras ambiguas la categoría más frecuente en dicho corpus, entre otras pistas), o empleando un método semi automatizado en el que se combinan ambos procedimientos. Así, será posible mejorar los resultados del análisis mientras se elaboran más reglas aplicadas a las características de los corpus empleados en los experimentos (Villayandre Llamazares, op cit).

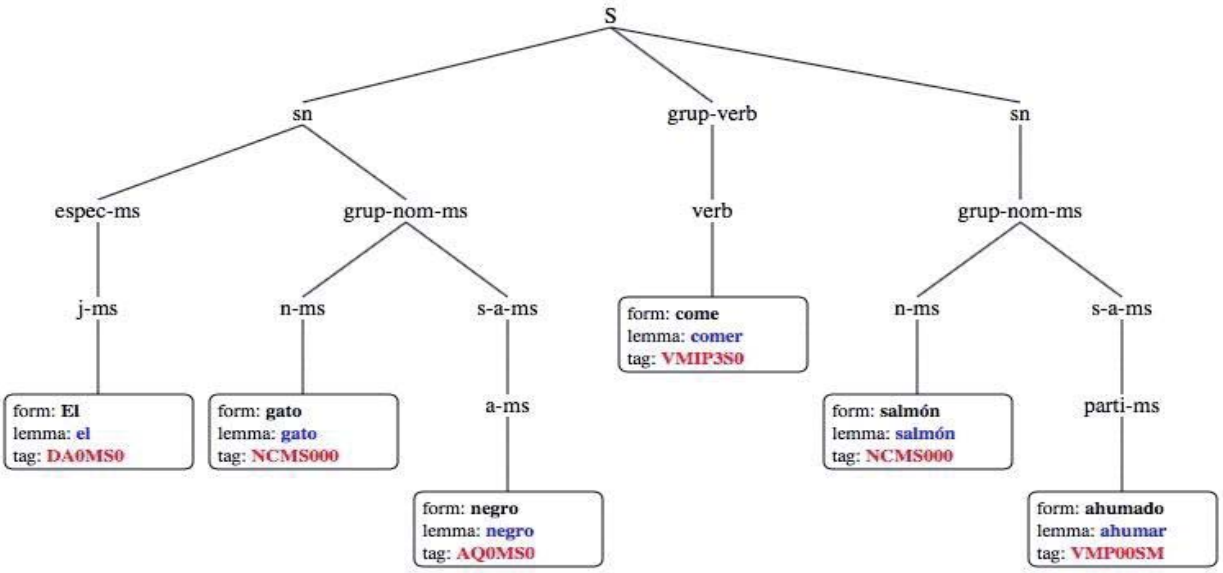

Ilustración 1 Demostración del análisis sintáctico de Freeling

En su tesis de Magister, Miguel Ballesteros (2010) experimentó con *Maltparser*, un generador de analizadores de dependencias basado en el aprendizaje automático mediante el cual se obtiene una precisión cercana al 80% en el análisis sintáctico. Desde su investigación, aborda la dependencia como las relaciones binarias asimétricas que existen entre las palabras de una frase, a nivel de estructura sintáctica. La tarea a realizar consiste en establecer criterios para definir cuáles relaciones de dependencia existen, a fin de distinguir de qué manera están relacionadas dos palabras en una frase y si estas relaciones están o no etiquetadas en el sistema.

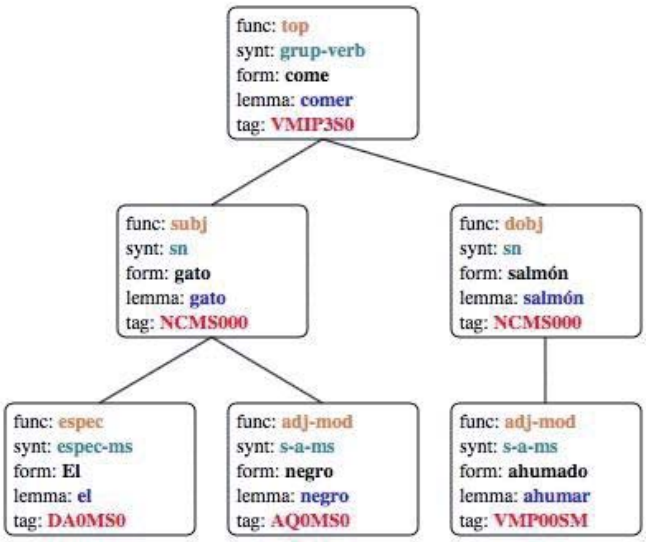

Ilustración 2 Ejemplo de una rutina con analizador de dependencias

2.2.6 Gramática del texto

Con la inclusión de la gramática del texto, el análisis textual ha superado la propuesta normativa de reducir el estudio del texto a la mera relación entre elementos de la cadena a nivel oracional (relaciones morfológicas y morfosintácticas). Por lo tanto, tienen lugar también los elementos contextuales que operan en un discurso, enriqueciendo y complejizando el análisis. Estos estudios se enmarcan dentro de la Lingüística en general, ya sea la que estudia la lengua en tanto sistema o bien la que estudia la competencia de un hablante-oyente ideal.

Cuenca (2010) señala que “la gramática del texto se ocupa de los mecanismos formales, fundamentalmente gramaticales y léxicos, que hacen que un conjunto de oraciones formen una unidad superior desde un punto de vista semántico y comunicativo […] entendiendo a “la gramática del texto como equivalente a la cohesión o sintaxis textual (que no se contrapone a la sintaxis oracional, sino que la incluye y la supera” (11). Además plantea que “los mecanismos de cohesión, como la pronominalización o la conexión (oracional y textual) manifiestan relaciones que pertenecen al texto como unidad semántica (coherencia o semántica textual) y como unidad comunicativa (adecuación o pragmática textual” (op cit.).

La presente investigación guarda relación con la gramática del texto en tanto considera al texto como unidad de análisis. Debido a que el trabajo de análisis y clasificación de NP opera en un corpus compuesto de diversos artículos de geopolítica, es posible observar distintas menciones a un mismo referente (NP). Al clasificar una instancia regular de un NP a partir de las pistas contextuales presentes en el texto –como un verbo, por ejemplo-, se puede colegir que mediante mecanismos de correferencia, aunque no se mencione explícitamente el NP, el texto aportará la información suficiente para poder identificar estos mecanismos.

2.2.7 *Named entities* (NEE) y *named entities recognition* (NER)

Como se señaló anteriormente, el término “entidad nombrada” surgió en las conferencias de entendimiento de mensajes (conferencias *MUC*), en las que se buscaba promover y evaluar la investigación en el área de extracción de información. Se las definió como: “palabra o secuencia de palabras que se identifican como nombre de persona, organización, lugar, fecha, tiempo, porcentaje a cantidad” (Solorio, 2005). El reconocimiento de estas entidades es una tarea lingüístico-computacional en la cual se busca clasificar cada palabra de un documento en una determinada categoría de una lista de categorías definida (persona, lugar, organización, varios). En la presente investigación, se empleó en analizador/clasificador de entidades nombres (NEE) de *Freeling,* cuyos resultados posteriormente fueron mejorados mediante una rutina de aprendizaje automático que mejoró la tasa de aciertos en la clasificación de entidades realizada por *Freeling.*

2.2.8 Síntesis de los avances de la lingüística computacional

Siguiendo a Domínguez Burgos (2002), la lingüística computacional (LC) es una ciencia interdisciplinaria que se ubica entre la lingüística y la informática, con énfasis en la primera. Su fin es la elaboración de modelos computacionales que reproduzcan uno o más aspectos del lenguaje humano. La lingüística computacional no es exclusiva de

centros académicos pues actualmente recibe importantes impulsos de la industria privada. Las empresas dedicadas a la informática han reconocido desde hace tiempo que el procesamiento automático del lenguaje humano será uno de los principales campos de desarrollo en las próximas décadas.

Algunos de los programas de investigación en que trabajan los lingüistas computacionales son utilizados para:

- Elaborar modelos de teorías lingüísticas (Freeling).
- Enseñar idiomas extranjeros (Duolingo, *E-learning*).
- Corregir la sintaxis y la ortografía de textos en un idioma dado (correctores de herramientas ofimáticas).
- Reconocer la voz humana y procesar la información contenida en frases pronunciadas naturalmente por cualquier persona (Siri para móviles de Apple).
- Crear sistemas expertos que respalden la labor de especialistas en un área dada.
- Elaborar juegos digitales que usen de una u otra forma el lenguaje humano (juegos de karaoke).
- Producir traducciones automáticas de textos o ayudar a traductores humanos en su trabajo (*Google translator*).
- Generar voz artificial con alto grado de naturalidad para la transmisión de información por teléfono, etc. (plataformas call center para asistencia a usuarios las 24 horas del día).

La trayectoria seguida por la LC da cuenta del trabajo mancomunado entre la informática y la lingüística. La formación lingüística tradicional en Chile no ha integrado en la formación académica de los estudiantes los aportes de la LC. Sin embargo, actualmente es normal en todo el mundo el empleo de herramientas como el buscador *Google,* traductores *on line,* las correcciones ortográficas de *Microsoft Word,* las herramientas computacionales para aprender idiomas, la interacción oral con un sistema de GPS en un automóvil que orienta al conductor, etc. El conocimiento que la

lingüística ha acuñado durante el siglo XX y XXI ha contribuido a la informática, desarrollando herramientas de uso cotidiano por las personas.

Eventualmente, el desarrollo de las tecnologías cibernéticas ha encontrado un importante apoyo desde el sector militar y la LC no ha sido la excepción. Ya desde fines de la segunda guerra mundial se comenzaron a elaborar herramientas de traducción inglés-ruso con propósitos geopolíticos; los servicios de inteligencia (EEUU/URSS) sabían que podían realizar importantes avances de espionaje con las computadoras, por lo que realizaron considerables aportes de capital a estas investigaciones. Como señala Hutchins (1998), grandes aportes a las tecnologías de procesamiento del lenguaje humano tuvieron lugar en las décadas del 40' y del 50' pues la teoría de los autómatas y los modelos probabilísticos o teoría de la información supusieron toda una revolución que provocó cambios trascendentales.

La teoría de los autómatas fue desarrollada por Alan Turing, matemático británico precursor de la informática moderna que obtuvo prestigio internacional por crear un prototipo que podría determinar si una máquina podía presentar indicios de inteligencia artificial al ser sometida a un test (*test de turing*) realizado en paralelo a seres humanos. Esta teoría propone que es posible determinar la cantidad de operaciones posibles a realizar por una máquina abstracta (una máquina como una computadora, aunque Turing no conocía las computadoras como las conocemos hoy). Para establecer los límites entre lo que una máquina podía y no podía calcular, propuso describir los lenguajes formales (lenguajes cuyos símbolos y reglas están formalmente especificados para ser unidos y combinados) a fin de especificar procesos de cálculo algorítmicos. Este tipo de "máquina abstracta", hoy denominada "autómatas finitos" fue empleada en primera instancia para modelar el funcionamiento del cerebro (Chomsky, 2000).

En términos simples, se debe entender un algoritmo como un conjunto dado de instrucciones o reglas plenamente definidas, ordenadas y finitas que posibilitan la realización de una actividad a través de pasos sucesivos que no dejen lugar a dudas

por parte de quien deba realizar dicha actividad. Si bien las personas presumen desconocer la algoritmia, la resolución de problemas mediante algoritmia es habitual, pues seguir las instrucciones de un manual para armar un mueble, las instrucciones específicas que recibe un trabajador dentro de su labor o incluso el procedimiento para realizar multiplicaciones y divisiones en la escuela constituyen procedimientos algorítmicos.

Por otra parte, los modelos probabilísticos tienen como exponente destacado al estadounidense Claude Shannon (1948), quien aplicó *la teoría de la probabilidad de procesos* de Markov para desarrollar una teoría de la información o de la comunicación. Esencialmente, esta teoría de probabilidad de procesos propone que frente a una sucesión de variables aleatorias (estocásticas) que evolucionan en función de otra variable, es posible estimar la probabilidad de que ocurra un evento a partir del evento inmediatamente anterior. Shannon contribuyó a la evolución de la lingüística computacional con sus ideas sobre el canal de ruido y decodificación del lenguaje. A partir de estos avances de Shannon en el campo probabilístico, se cimentó el camino para que Noam Chomsky iniciara sus intentos por describir posibles gramáticas. Sus trabajos en el área de la teoría de lenguajes formales propulsaron en medida significativa el desarrollo de una teoría de los lenguajes de programación.

Posteriormente, a finales de los años 50, las investigaciones en LC fueron concentrándose en dos áreas: un campo simbólico y uno estocástico. El enfoque simbólico, a su vez produjo dos corrientes importantes: una liderada por Chomsky, científicos de la computación y de la lingüística formal interesados en el análisis sintáctico, y otra interesada en la inteligencia artificial. El campo estocástico estuvo representado principalmente por ingenieros eléctricos que trabajaban con estadísticas y probabilidades. Para estas últimas investigaciones emplearon el teorema de Bayes (1763) en el reconocimiento óptico de caracteres y para probar la autoría de textos (Mosteller & Wallace, 1984). Ya para los años 40, los ingenieros desarrollaron el espectrógrafo de sonido, el cual permite el análisis de las ondas sonoras y, en los años 50, obtuvieron los primeros reconocedores artificiales de voz.

Respecto a los trabajos de lingüística de corpus, el corpus Brown de inglés americano fue el primero de gran envergadura y motivó en los años 60 diversas

investigaciones en el área. Este contenía 500 muestras de texto en inglés, las que alcanzaban la extensión de aproximadamente un millón de palabras compiladas a partir de obras publicadas en EEUU en 1961.

Los primeros antecedentes de la máquina que hoy se conoce como *Cleverbot* y que posibilita a cualquier persona poder conversar con una aplicación que simula inteligencia artificial, se remontan a los años 60. El programa informático *Eliza*, desarrollado en el Massachusetts Institute of Technology (MIT) entre 1964 y 1966 por Joseph Weizenbaum (López Cabello, 2010), dio cuenta de una precaria aptitud para mantener una conversación coherente con los usuarios y fue uno de los primeros programas en procesar lenguaje natural (o al menos intentarlo). El sistema era sencillo pero podía crear la impresión de que poseía algún tipo de inteligencia, debido a que estaba basado en el reconocimiento de expresiones regulares.

Las expresiones regulares (también conocidas como *regex*) son un tipo de lenguaje formal que nos permiten hacer búsquedas y sustituciones en los textos detectando secuencias de caracteres que cumplan una condición dada. Por ejemplo, rastrear palabras que tengan cierto prefijo o cierto sufijo, verbos de primera, segunda o tercera conjugación, etc.). Una gran ventaja de las expresiones regulares es que pueden ser creadas y empleadas por cualquier usuario que esté familiarizado con la sintaxis de las mismas; es decir, que pueda dominar el significado de los distintos caracteres que la conformen y así crear patrones de búsqueda de información lingüística. Para esta tarea se requiere aprender rudimentos de programación para manejar y elaborar ficheros *regex*. Uno de los lenguajes de programación más usados en español para las labores de un lingüista que trabaja con expresiones regulares es *Perl,* puesto que fue diseñado para tales fines.

En la década de 1970 los trabajos relacionados con estadística y probabilidades condujeron al desarrollo de algoritmos de reconocimiento de voz que superaron los trabajos anteriores; estos fueron financiados principalmente por IBM y AT&T Bell Laboratories. El interés por las posibilidades de la lógica condujo al desarrollo de *PROLOG*, un lenguaje de programación declarativo muy usado en distintas áreas de la Inteligencia Artificial y la LC. Para el desarrollo de estas revolucionarias herramientas fue significativo el aporte disciplinar de la *gramática funcional* y la *gramática léxica*

funcional. Se consideró que la *gramática funcional* era un aporte pues su objetivo es descubrir el "sistema de la lengua" mediante los usos que se hacen de las expresiones lingüísticas. Para esto considera el estudio del empleo de las palabras, los sintagmas (grupos de palabras que forman otros sub constituyentes, siendo al menos uno de ellos un núcleo sintáctico), las oraciones y la adecuación de estos elementos.

Para finales de la década de 1980, los modelos probabilísticos dejaron de ser dominio exclusivo de los ingenieros en el área de reconocimiento de voz y comenzaron a ser empleados por lingüistas para el análisis morfológico y sintáctico, para la traducción automática, etc. Con la masificación de Internet en la década de 1990, el estudio y desarrollo de nuevas tecnologías que mejoren el procesamiento automático del lenguaje ha posibilitado que todos los grandes conglomerados tecnológicos, de una u otra forma, trabajen en el área de la LC. Como señala López Morrás (2004):

> "Todavía se sabe relativamente poco sobre el lenguaje humano. Los lingüistas llevan décadas intentando descifrar cómo funciona esta capacidad única de la especie humana. Muchos animales tienen formas complejas de comunicación pero, que se sepa, ninguno de estos "lenguajes" cumple la característica más significativa del lenguaje humano natural: la infinitud discreta.
>
> *El lenguaje humano natural es discreto en cuanto a sus unidades, pero infinito en cuanto a las combinaciones que pueden hacerse con estas unidades.* Por ejemplo, las palabras son unidades discretas y finitas de la lengua. Sin embargo, combinando un número limitado de palabras podemos construir infinitas frases. Y esa es la razón por la que un niño o un adulto construyen continuamente frases que no han escuchado jamás a partir de palabras que sí ha tenido que escuchar y memorizar con anterioridad. Así, hablar es inventar continuamente nuevas combinaciones" (Morrás, op cit).

Por lo tanto, es posible concluir que parte importante de la formación de un docente de lengua, de un lingüista, o de un investigador del lenguaje es ampliamente vinculable con el quehacer de un profesional del área de la ingeniería y la tecnología.

Principales áreas de la LC:

a. Etiquetado morfológico o *Tagging*

Este corresponde al análisis y etiquetado morfológico automático de los elementos de la cadena de palabras en una frase. Sin tener en cuenta el resto de la oración, la palabra *como,* por ejemplo, puede ser un verbo, una conjunción o un adverbio. A fin de disminuir la cantidad de ambigüedades detectadas, se ingresan datos de carácter sintáctico y/o semántico para mejorar el proceso de etiquetamiento. El *tagging* se va enriqueciendo a medida que se colectan datos estadísticos que se almacenan en algún *data bank*. En esta investigación se recurre al etiquetamiento automático realizado por el software *Freeling*. Esta tarea reviste de nuevas funcionalidades el trabajo que normalmente es realizado por un gramático, pues posibilita el análisis y etiquetado de corpus de miles o incluso millones de palabras.

b. Análisis sintáctico o *Parsing*

El *Parsing* corresponde al análisis automático de una oración dada, en el campo de la inteligencia artificial (o en el de la LC) y ofrece la posibilidad de realizar análisis semántico además del sintáctico que se realizaba anteriormente. Los tipos de algoritmos que se crean para análisis sintáctico pueden tener dos enfoques: *bottom-up* (ascendente) y *top-down* (descendente).

En el enfoque *bottom-up* se procede a comenzar el análisis desde unidades mínimas para ir reconociendo estructuras cada vez más complejas hasta llegar a frases, y de allí construir una o varias oraciones. Por su parte, el *top-down* parte de que la información que se obtiene es una oración dada, y realiza hipótesis sobre cuáles frases pueden constituir la oración dada y cómo están organizadas estas frases, para así llegar hasta las unidades mínimas. Los problemas de ambigüedad resultan relativamente sencillos a nivel morfológico en comparación con los problemas de identificación que se producen a nivel sintáctico. Se usa cada vez más la estadística para resolver ambigüedades sintácticas y semánticas. Abney (1994) planteó que los

problemas de desambiguación disminuyen considerablemente si se emplea la técnica de *chunking* (cortar las frases en trozos), pues analizar el contexto completo de una oración resulta complejo y los resultados del *parsing* no siempre son óptimos.

c. El reconocimiento de voz y la conversión de texto a voz

El reconocimiento de voz y la conversión de texto a voz consisten en la transcripción automática de la voz humana en datos que puedan ser procesados por la computadora. El desafío consiste en traducir una señal acústica continua en una serie de símbolos discretos, un texto equivalente a la fiel representación de la señal en cuestión. La comprensión automática del habla (*automatic speech understanding)* busca llegar a producir algún tipo de procesamiento semántico de la oración, por lo que el conocimiento gramatical es fundamental en estas investigaciones. Entre los exponentes teóricos en el área, se encuentran Jurafsky & Martin los cuales proponen un enfoque empírico para el procesamiento del lenguaje, basado en la aplicación de algoritmos de aprendizaje automático estadísticos y de otro tipo para grandes corpus (2000).

Los reconocedores de voz se basan en modelos que representan la probabilidad de las oraciones (entendida una oración como una serie *n* de palabras o *n*-gramas dados), en modelos que representan la probabilidad de que una palabra determinada sea realizada como un grupo de fonemas dados (tal como los modelos de Markov) y en modelos que expresan la relación probabilística entre fonemas y características acústicas o espectrales dadas. Estos sistemas suelen usarse actualmente en las plataformas de *call center,* estaciones de trenes, sistemas de atención a personas con discapacidades que no pueden usar un computador pero sí le pueden hablar para que este ejecute órdenes, etc.).

d. Recuperación inteligente de información o *Information retrieval*

Este es un campo muy amplio que incluye todas las formas de almacenamiento y envío digital de datos de cualquier índole. En el caso de la LC, se trata principalmente

de técnicas para la extracción de datos contenidos en textos y su transmisión a los usuarios. Para ello se usan actualmente métodos de procesamiento estadísticos y simbólicos diversos. Los buscadores de Internet se basan en uno o más de estos métodos de recuperación de información. Uno de los propósitos de esta investigación es colaborar con el proceso de aprendizaje automático de una máquina que pueda reconocer cuándo se está en presencia de un nombre propio y a qué tipo de entidad corresponde (nombre, organización o lugar).

e. Sistemas de diálogo y sistemas expertos

Estos sistemas permiten la comunicación entre uno o más usuarios y la computadora. Los sistemas de diálogo son empleados en el desarrollo de interfaces más amigables al momento de realizar compras *on line* o para guiar a un usuario al momento de instalar software en su computadora, etc. Por su parte, los sistemas expertos son entendidos como representaciones del conocimiento de expertos en un campo dado que han sido almacenados de manera digital. Un sistema experto se compone de un *software* de manipulación, un banco de datos que contiene hechos y reglas válidas para el área de conocimiento que se representa, un componente de generación de inferencias a partir del *software* y del banco datos, así como una interfaz con el usuario.

Dos usos comunes para este tipo de sistemas es el que se emplea en las clínicas y laboratorios ya que a partir de resultados obtenidos, el software entrega una plantilla al facultativo con las potenciales causas médicas que están relacionadas con los datos obtenidos. Además, se usa en las industrias para la inspección mecanizada de equipos altamente sofisticados. Inclusive, los avances de la LC han podido construir sistemas en los que se entregan resultados por escrito con una evaluación automática del análisis realizado, redactado y representado tal como si lo hubiese escrito un experto.

f. Traducción automática

El desarrollo de herramientas de traducción sin duda es una tarea ardua pues, el traductor no solo requiere de la comprensión profunda de dos sistemas lingüísticos dados, sino también de dos culturas y técnicas de comunicación (además de todos los componentes pragmáticos involucrados). Por lo tanto, aunque se ha evolucionado de manera significativa, aún no se puede decir que la labor del traductor ha podido ser totalmente sustituida por un sistema de traducción automática. Como se señala en Hutchins & Sommers (1992), la traducción de textos mediante computadoras fue una de las primeras metas que los pioneros de la informática y la inteligencia artificial se fijaron. La traducción automática crece cada vez más como campo de investigación y desarrollo, al tiempo que la necesidad de traducciones de textos técnicos y comerciales crece mucho más allá de la capacidad de los traductores.

g. Text mining

Debido a la colosal cantidad de datos que se producen a diario, surge la necesidad de procesamiento automático de la información. Para tal efecto, se realiza un proceso semi automatizado de estos, en busca de patrones que pueden ser extraídos desde grandes bases de datos. Según García Reyes (2012) el *text mining* corresponde a metodologías activas cuyo propósito es la extracción de la información y el conocimiento de los datos. La minería de datos incluye modelos de reconocimiento de patrones, aprendizaje en cuanto a las máquinas y constituye un proceso enfocado al análisis de grandes bases de datos con el propósito de extraer información y conocimiento que pueda ser útil para la toma de decisiones y resolución de problemas en diversas investigaciones.

h. Presente de la LC

Al encontrarse aún en sus inicios, la LC no es una carrera profesional tradicional. Actualmente en Chile ninguna universidad imparte la carrera y solo se pueden estudiar algunas materias afines en el ámbito de la lingüística aplicada o según las necesidades de los departamentos de investigación de universidades o empresas. Queda de manifiesto que el campo de estudio de un lingüista computacional está compuesto esencialmente de informática, matemáticas (lingüística algorítmica) y lingüística, por lo que el actual perfil del lingüista formado en Chile dista mucho de la formación impartida en los países desarrollados. Un estudiante de LC debe tener conocimiento de teorías estructuralistas, teorías generativistas clásicas, la gramática de casos de Fillmore, la gramática funcional y la gramática del texto, entre muchas otras materias. Respecto a la gramática del texto, el análisis lingüístico de los sistemas automatizados suele realizarse a nivel oracional; sin embargo, con la inclusión de la gramática del texto se podrá analizar el texto como unidad completa, lo cual implicará una revolución en el procesamiento del lenguaje natural, toda vez que las plataformas de aprendizaje automática mejorarán considerablemente su desempeño.

i. Perl

Cabe señalar que un lingüista computacional debe saber programar y tener óptimos conocimientos de los algoritmos fundamentales usados en informática. En el caso de esta investigación, el trabajo se ha realizado con el lenguaje de programación *Perl*, el cual está diseñado para fines investigativos y para procesamiento y manipulación de textos. Este puede ser descargado de manera gratuita desde http://www.perl.com y funciona en prácticamente todos los sistemas operativos (*Windows, Apple OSX, Linux, etc.).*

Perl es un lenguaje de programación diseñado por Larry Wall en 1987. El propósito para el que originalmente fue desarrollado es la manipulación de texto, aunque hoy se lo utiliza para administración de sistemas, desarrollo web, programación en red, etc. Es un lenguaje de alto nivel, es decir, no opera al nivel de la máquina

(código binario), lo cual ofrece flexibilidad al programador al momento de elaborar abstracciones (o de ser literal) asimilándose a la expresión oral en la escritura del programa y su posterior compilación por un intérprete.

Como advierten Wall, Christiansen & Orwant (2000) este lenguaje de programación está optimizado para exploración y extracción de información desde ficheros de texto o desde reportes de información basada en este proceso. Una de sus ventajas comparativas es que posee un potente conjunto de expresiones regulares, con las que los programadores pueden crear sus propios patrones de análisis de textos para aplicarlos en los mismos. Puede ser descargado desde su sitio oficial www.perl.org.

j. Freeling

Freeling es una librería de código abierto para el procesamiento multilingüe, que proporciona una amplia gama de funcionalidades de análisis para varios propósitos. Al ser de código abierto, debiese posibilitar avances más rápidos en proyectos de investigación y costes más reducidos en el desarrollo de aplicaciones industriales de procesamiento del lenguaje natural. Tiene soporte para asturiano, catalán, castellano, galés, gallego, inglés, italiano, portugués y ruso (Atserias et al, 2006).

Freeling es un *software* que está compuesto por una librería que ofrece distintas alternativas para realizar análisis lingüístico (análisis morfológico, reconocimiento de fechas, etiquetado gramatical -o *PoS tagging, part of speech tagging*-, etc.) La última versión (3.1) ofrece servicios tales como reconocimiento de idiomas,

	as	ca	cy	en	es	gl	it	pt	ru
Tokenization	X	X	X	X	X	X	X	X	X
Sentence splitting	X	X	X	X	X	X	X	X	X
Number detection		X		X	X	X	X	X	X
Date detection		X		X	X	X		X	X
Morphological dictionary	X	X	X	X	X	X	X	X	X
Affix rules	X	X	X	X	X	X	X	X	
Multiword detection	X	X	X	X	X	X	X	X	
Basic named entity detection	X	X	X	X	X	X	X	X	X
B-I-O named entity detection				X	X	X			
Named Entity Classification				X	X				
Quantity detection		X		X	X	X		X	X
PoS tagging	X	X	X	X	X	X	X	X	X
WN sense annotation		X		X	X				
UKB sense disambiguation		X		X	X				
Shallow parsing	X	X		X	X	X		X	
Full/dependency parsing	X	X		X	X	X			
Coreference resolution					X				

Servicios de análisis disponibles para cada lengua.

tokenizing (sustitución de datos sensibles a fin de retener la información relevante pero sin poner en peligro la seguridad de estos), división y análisis de oraciones (*splitting)*, análisis morfológico, detección y clasificación de entidades nombradas (fundamental para esta investigación), reconocimiento de fechas, números, magnitudes físicas, recurrencia de datos, codificación fonética, análisis sintáctico, análisis de dependencia, análisis de ambigüedad y sentido de los términos en un texto, etc. Una de sus ventajas es que no es necesario ser un programador experto para poder usar las funciones señaladas, no obstante, tampoco está diseñado para legos en informática.

El programa aún no está diseñado para ser empleado de manera sencilla por cualquier usuario, se espera que en el futuro mejore su funcionamiento (tiene *bugs* o agujeros importantes, por ejemplo al intentar analizar cantidades de texto extensas), además de incorporarle nuevas funciones. Actualmente no cuenta con una interfaz gráfica común, por lo que para ejecutarlo es necesario tipear comandos desde una terminal. *Freeling* no es un programa liviano pues requiere diversas librerías para realizar con éxito los diversos servicios lingüísticos que ofrece y está principalmente orientado a desarrolladores, expertos tales como los lingüistas computacionales que posean los conocimientos y herramientas para desarrollar nuevos procesamientos lingüísticos. En otras palabras, no es posible determinar cuántos tipos de uso puede ofrecer *Freeling*, pues depende del nivel de competencias informáticas de los usuarios.

3. Marco metodológico

3.1 Tipo de investigación

Tal como ya se anticipó en la introducción, en primera instancia esta investigación tiene carácter exploratorio pues no existe bibliografía relevante respecto a la potencial capacidad predictiva de los verbos para determinar a qué tipo de entidad pertenece un nombre propio. Posteriormente, el trabajo adquiere carácter descriptivo pues se busca someter a análisis las relaciones entre la posición sintáctica del verbo y los nombres propios que les acompañan en los textos. Además de la dimensión exploratoria y descriptiva de esta tesina, también se pretende realizar una propuesta metodológica para determinar cuantitativamente el potencial predictivo del verbo, lo cual exige no solo limitarse a la exploración y descripción del fenómeno, sino al aspecto predictivo de la investigación. Existen diversos estudios previos en el tratamiento de los nombres propios y la Gramática de Bosque y Demonte (1999) ofreció un sustento teórico sólido para tal análisis.

3.2 Pregunta de investigación

¿Puede el verbo que acompaña al nombre propio funcionar como predictor de una determinada categoría de nombre propio?

De acuerdo a los rasgos prototípicos de un nombre propio, y tal como un software realiza la detección y clasificación de estos, la posición sintáctica del verbo, en tanto elemento predictor de los nombres propios no ha sido estudiada e integrada a los sistemas automáticos de extracción de información empleados en esta investigación.

3.3 Objetivos

Objetivos

Objetivo general: Determinar cuáles pueden ser los rasgos que faciliten la predicción de manera sistemática del tipo de entidad al que pertenece un nombre propio.

Objetivos específicos: Determinar el grado de precisión del verbo como predictor de la categoría de nombre propio.

3.4 Instrumento de recolección de datos y procedimiento de análisis

Corpus

Se utilizó un extenso corpus de geopolítica del diario *Le Monde Diplomatique*: un 50% de este se usó para análisis de datos y elaboración de reglas, el otro 50% para comprobar las mismas. Este corpus abarca un total de más de 8.000.000 de palabras. Corresponde a los archivos completos de *Le Monde Diplomatique edición Cono Sur*, los cuales comprenden desde su fundación en 1999 hasta diciembre de 2010. Como se señaló anteriormente, este fue dividido en dos partes, usándose ambas para levantamiento de hipótesis y verificación de estas, respectivamente.

3.5 Experimentación y evaluaciones

En la presente investigación se realizaron tres instancias de experimentación, las cuales abordan desde distintas perspectivas la problemática planteada respecto al potencial rol predictor del verbo que acompaña al nombre propio.

3.5.1 Evaluación de la detección y clasificación de nombres propios con Freeling

En primer lugar, se procedió al análisis de corpus con el propósito de establecer la tasa de efectividad de *Freeling* frente a la tarea de detección y clasificación de nombres propios. Los resultados obtenidos por el software fueron analizados a fin de detectar aciertos y errores en la ejecución de la tarea, para posteriormente detallar en cuáles tareas *Freeling* siempre podría fallar. Cabe señalar que se realizó un análisis manual de los resultados obtenidos de manera automática, el cual se constituyó como un instrumento de validación o refutación de los resultados obtenidos por el programa.

3.5.1.1 Observaciones al proceso de detección y clasificación de nombres propios realizado por *Freeling*, *Citius Tagger*, *Semantria* y *Alchemy Language*

Hay nombres propios que son reconocidos como distintos tipos de entidad dentro de un mismo texto, lo cual guarda relación con las tasas de error halladas en la evaluación realizada. Esto se explica porque Freeling y los demás programas empleados en la evaluación no integran herramientas de análisis propios de la gramática del texto: analizan cada oración por separado pero nunca el texto como una sola unidad. Por lo tanto, al no considerar elementos como el referente o mecanismos de correferencia que establezcan enlaces dentro de texto, no procede obtener resultados precisos respecto al reconocimiento de los distintos nombres propios presentes. Por otra parte, no todos los textos pueden ser evaluados, debido a una falla de diseño del software ya que al superar una frase de más de 10-12 palabras (sin puntos ni comas), entra en un bucle o *loop* (realiza la evaluación de manera infinita sin poder finalizar el proceso). Se ha procedido a evaluar en primera instancia 5 textos de *Le Monde Diplomatique* para detectar alguna regularidad entre los rasgos de los verbos y el tipo de entidad de nombre propio que el software clasifica.

En el caso de los textos de Geopolítica, se produce un fenómeno muy interesante que guarda relación con las distintas concepciones que se tienen de algunas instituciones. A saber, el Estado puede ser concebido como el territorio, como una organización, como un agente político que realiza acciones. Por ende, se puede

proponer que en geopolítica el análisis de los nombres propios no resulta tan certero (computacionalmente) como en otros géneros porque las acciones son realizadas por Estados, instituciones complejas.

Posiblemente los distintos resultados en la clasificación de NEE obedecen a que a veces algunos nombres propios operan como agentes de acciones y no como pacientes. Por ejemplo, "EEUU movilizó sus tropas". Aquí claramente no es el territorio el que se está movilizando, sino es una acción realizada por EEUU como institución / organización y así lo entiende el clasificador. Por lo tanto, a nuestro juicio, no todos los resultados que sean diversos de la acepción común de cada nombre propio, necesariamente están erradas. Un error que siempre comete el programa es reconocer como nombre propio a cualquier elemento que aparece con mayúscula, por lo tanto, suele reconocer como un nombre propio al primer elemento de una oración. Como es sabido, uno de los rasgos que todo nombre propio posee es que se introduce con mayúscula (RAE, 2009).

El reconocimiento más certero es el de los nombres de personas y de topónimos. Debido a la naturaleza de este género textual, las clasificaciones de las organizaciones suelen ser más imprecisas o ambiguas; no por eso erradas. Se evidencia un óptimo reconocimiento de las siglas como organizaciones, sin embargo, todas las siglas son clasificadas como organizaciones (incluso el síndrome “HIV”). En lo que más falla es en la correcta detección de nombres propios (de antropónimos específicamente). Por último, cuando hay dos sustantivos con mayúscula lo clasifica inmediatamente como antropónimo.

La evaluación formal de los resultados de reconocimiento y clasificación de nombres propios es una tarea difícil de realizar ya que el software interpreta los textos a nivel oracional; carece de herramientas para identificar un referente dentro de los textos (es necesario incluir herramientas propias de los hallazgos de la gramática textual), así suele confundir un antropónimo con el nombre de una organización, el de un topónimo con un nombre de organización o el nombre de una obra de arte con un antropónimo u organización (debiese clasificarlo como "otros" tipos de nombre propio). Por ende, la sola evaluación del rendimiento de Freeling no constituye una evidencia realmente precisa u objetiva.

Como señalan Manning et al (2009), el valor-F (o *F-score*) en estadística corresponde al índice de precisión que tiene un test. Da cuenta del valor único pondera entre la precisión y la exhaustividad. Tal como en este experimento, suele usarse en la fase de pruebas de clasificación de documentos, recuperación de información, entre diversos usos.

La fórmula que sintetiza esta operación es $F_1 = 2 * \frac{Precisión * Exhaustividad}{Precisión + Exhaustividad}$

En esta investigación, la estimación del Valor-F dará cuenta de la tasa de precisión con que cada uno de los *softwares* detectó y clasificó los nombres propios presentes en el corpus de *Le Monde Diplomatique*.

Con el software "Semantria" la cantidad errores de detección de NEE es menor, pues no incurre en el error de "Freeling" de considerar como NEE cualquier término que comience con mayúscula. Un error detectado en los resultados de "Citius Tagger" es que al reconocer un nombre propio compuesto o un nombre propio compuesto de varias palabras, las reconoce como nombres por separado, lo cual constituye un error no menor en la detección y clasificación de NEE. De los 4 programas de extracción utilizados, el único que ofrece un análisis acabado es Freeling pues etiqueta cada elemento dentro de la cadena. Por otra parte, el programa con la mayor precisión resultó ser el software de pago "Semantria".

Los textos de carácter literario son los que arrojaron la mayor tasa de errores, posiblemente debido a la disposición retórica de los mismos (plenamente estéticos / estetizantes). Un error generalizado de los extractores es que no consideran el carácter de denotación que poseen los NP (Russell, op cit), solo un elemento único en el universo puede ser nombrado mediante un nombre propio. Por ejemplo, reiteradamente en el corpus se identifica como un nombre propio el término "Código Penal" solo porque se presenta con mayúsculas iniciales y no considerando que este término refiere a una entidad abstracta.

Cabe señalar que debido a la naturaleza de este breve trabajo (tesina de pregrado), el análisis solo evaluó la efectividad de los programas en la detección de nombres propios, y se excluyó la tarea de clasificación según nombre de persona, lugar u organización. Si bien la detección de nombres propios de manera automática es una tarea compleja, lo es mucho más si se pretende clasificarlos según tipo de entidad y

esperar resultados con una alta tasa de exactitud. Inclusive para un ser humano que domine herramientas de análisis morfosintáctico, semántico y de gramática del texto, algunos ejercicios de reconocimiento podrían dar lugar a resultados e interpretaciones diversas. Una investigación centrada en la clasificación de entidades nombradas resulta una interesante propuesta para trabajos futuros.

Luego de una evaluación preliminar, se ha observado que los motores que mejor realizan la labor de detección y clasificación de NEE pertenecen a *Freeling* (75,19% de rendimiento considerando porcentajes de aciertos y cobertura) y *Semantria* (83,12% de rendimiento considerando porcentajes de aciertos y cobertura). Sin el cálculo de la cobertura no hubiera sido posible determinar con mayor exactitud los valores de acierto, pues se presume que la cantidad de NEE detectados por el software será menor que la cantidad detectada manualmente. En este sentido, queda de manifiesto que los dos softwares señalados tienen mejores herramientas gramaticales que *Citius Tagger* y *Alchemy Language*; posiblemente esto se deba a que originalmente fueron diseñados para analizar la lengua inglesa (*Freeling* y *Semantria* tienen más versiones previas y en el caso de Semantria, este tiene una versión de pago y un amplio equipo de profesionales mejorando el software)". Debido a que *Semantria* es un programa de pago (y con código fuente cerrado) al que no se puede acceder ni mejorar sin autorización de los propietarios, se ha preferido utilizar Freeling como herramienta principal de esta experimentación.

3.5.2 Mejora de los resultados obtenidos por *Freeling* mediante un *script Perl*

Posteriormente, se realizó un análisis empleando un script *Perl*, el cual opera con una base poco precisa, pues clasifica las entidades a partir de las palabras (incluidos verbos) que ocurren en el contexto. Debido a los problemas señalados anteriormente (resultados poco precisos y dificultad de análisis de oraciones largas), no se utilizó el analizador de dependencias

La aplicación del trabajo investigativo se ha realizado en dos momentos. Primeramente, se realizó un análisis y evaluación preliminar de la tasa de clasificación correcta de NP con los *softwares* listados en la tabla 4.

Freeling	http://nlp.lsi.upc.edu/freeling/
Semantria	https://semantria.com/
Alchemy Language	http://www.alchemyapi.com/products
Citius Tagger	http://gramatica.usc.es/pln/tools/CitiusTools.html

Tabla 4 Softwares usados en la evaluación y sus links de descarga

Como ya se ha señalado anteriormente, el corpus escogido fue una colección de artículos publicados por el periódico francés *Le Monde Diplomatique*. Debido a que la investigación requiere evaluar la clasificación de NP en antropónimos, topónimos y organizaciones, se escogió un corpus de geopolítica, disciplina en la que la referencia a figuras políticas, territorios y organizaciones son lugar común en este género textual. El ejercicio de evaluación mostró algunas rutinas de difícil resolución, como por ejemplo la siguiente:

"[…] Estados Unidos alienta también a Saddam, haciendo que el enemigo de ayer acceda al rango de aliado […]"

De acuerdo a una definición convencional, Estados Unidos debería ser automáticamente etiquetado como un NP de lugar geográfico como sujeto paciente. Un ejemplo típico de topónimo como sujeto paciente sería "EEUU se ubica en América del Norte". Sin embargo, en el contexto de estos artículos de geopolítica, actúa como una organización (como sujeto agente de la acción "alentar"). Tras el procesamiento de este mismo corpus mediante los programas señalados en la tabla 4, se procedió a registrar los resultados a fin de obtener la tasa de precisión, cobertura y F1.

Lo interesante de este procedimiento es que los elementos predictores son extraídos automáticamente del corpus de entrenamiento; es decir que no son condiciones introducidas por un programador, como hubiera sido el caso por ejemplo de un sistema de reglas del tipo “si se encuentra la marca “Dr.” antes de la entidad, entonces se clasifica al NP como antropónimo. Al contrario, un sistema como el implementado en esta investigación tendrá mayor cobertura ya que extrae los datos directamente del corpus y no de la introspección del investigador. Los datos que son considerados por el sistema para producir este modelo y mejorar la clasificación son los siguientes:

- **Palabras del contexto.** Las palabras del contexto que rodea a una entidad son tenidas en cuenta como factor predictor, ya sean de categoría gramatical nombre, adjetivo o sustantivo. Por ejemplo, si un tipo de entidad aparece con frecuencia en el corpus de entrenamiento antes de un verbo como “considerar”, el sistema aprenderá que tal entidad refiere a un humano. Lo mismo ocurrirá en otros casos, como cuando la entidad está precedida por palabras como “Señor”, “Profesor” o “Doctor”.
- **Almacenamiento y posterior reconocimiento de los NP evaluados previamente.** El algoritmo aprende a reconocer el tipo de entidad que corresponda porque ya ha clasificado anteriormente otras entidades que tienen los mismos componentes. En el caso de un antropónimo, por ejemplo, sería el caso de los nombres de pila o los apellidos. Así, si en el corpus de entrenamiento se ha observado que en reiteradas ocasiones que el componente “Vincent” forma parte de entidades que han sido clasificadas como antropónimos, entonces el script tendrá un importante elemento de decisión cuando se encuentre frente a la tarea de clasificar una entidad como “Vincent Heredia”.
- **Etiquetado de *Freeling*.** El elemento descrito en el punto anterior es útil solamente en el caso de los elementos poliléxicos, pero cuando estamos ante el nombre de una entidad compuesta por un solo elemento, tal como “Argentina”, entonces podrá recurrir al recuento de cuál ha sido la etiqueta que *Freeling* asignó con mayor frecuencia a esta entidad en el corpus de entrenamiento.

- **Filtrado de nombres comunes.** De manera similar al caso anterior, pero a la inversa, se ha aplicado un filtro que consiste en recopilar todas las palabras que han sido clasificadas como nombres comunes por Freeling en el corpus de entrenamiento. Si un elemento ha sido etiquetado como NC con más frecuencia que como NP, el sistema rechazará entonces la propuesta de Freeling de tratar tal elemento como NP en el corpus de prueba.

La decisión final del sistema de clasificar o de rechazar una determinada entidad se basa en la combinación de estos diferentes tipos de pistas.

3.5.2.1 Algunos resultados de la evaluación

Con el fin de calcular precisión y cobertura, realizamos previamente una detección manual de todas las entidades mencionadas en nuestros textos, lo que arrojó una cantidad de 398 entidades. A partir de allí comparamos el resultado de los distintos sistemas para realizar nuestra evaluación preliminar.

Sistema	**Detecta**	**Errores**	**Precisión**	**cobertura**
Freeling	331	146	56%	53%
Semantria	262	5	97%	60%
Alchemy Language	161	16	90%	39%
Citius Tagger	468	134	71%	83%

Tabla 5: Resultado de la evaluación preliminar de los diversos sistemas

A partir de los resultados de Freeling, se aplicó el sistema creado para esta etapa de la investigación realizando varios experimentos con el mismo corpus de prueba, pero variando la proporción del corpus de entrenamiento utilizado, desde un 5% hasta un 70%, obteniendo los resultados que se muestran en la tabla 5. En este caso no evaluamos cobertura porque nuestro script solo vuelve a clasificar las entidades seleccionadas por *Freeling*, pero no ejecuta el proceso de selección.

Sistema		**Precisión**
Freeling		56%
Nuestro sistema, con distintas proporciones del corpus de entrenamiento	5%	73%
	10%	72%
	15%	74%
	30%	73%
	50%	73%
	70%	71%

Tabla 6: Comparación del desempeño de Freeling con nuestro sistema utilizando distintas proporciones del corpus de entrenamiento

Respecto a la información de la tabla 6, se ha observado que los resultados del sistema implementado superan ampliamente el desempeño de *Freeling,* llegando esta diferencia hasta casi a los 20 puntos. Luego, resulta sorprendente notar que este desempeño sin embargo no cambia significativamente en función del tamaño del corpus de entrenamiento utilizado, puesto que la variación puede aducirse simplemente a error de muestreo. Estos resultados no coinciden con lo que cabe esperar, ya que al aumentar la información ofrecida por el corpus de entrenamiento, la tasa de clasificaciones correctas debería haber arrojado resultados superiores. Esta circunstancia es sin embargo positiva, ya que basta un pequeño corpus de entrenamiento para tener un rendimiento superior al observado al inicio de esta investigación.

En cuanto al análisis de los errores que se observaron en los resultados de la experimentación, resaltó la dificultad de desambiguar palabras que se introducen con una mayúscula pero no ejercen ningún oficio sustantivo en la oración, además de dificultar la correcta clasificación de las entidades cuando no se manifiesta claramente si un NP actúa como un topónimo o una organización, tal como en ejemplo referido en que EEUU actúa como una organización. A su vez, el algoritmo tuvo dificultades al momento de clasificar entidades correspondientes a organizaciones que también aplican a nombres de personas, como por ejemplo en el caso de *Louis Vuitton.* También se presentó el problema de la desambiguación al clasificar entidades abstractas como “Biblia”, “Corán”, “Yihad”, “Flores”. Este último caso es muy

característico de aquellos en los que el NP se puede confundir con un NC. En otras palabras, se ha realizado una prueba a grandes rasgos con *Freeling* probando todas las palabras del contexto, a fin de determinar que efectivamente el contexto puede ayudar a predecir el tipo de entidad al cual corresponde un determinado NP.

Al encontrar en reiteradas instancias a un antropónimo acompañando al verbo *escribir*, por ejemplo, se propone que dentro de un texto o dentro de un corpus, el NP –explicitado o mediante correferencias- que acompañe a este verbo corresponderá a un tipo de entidad de persona, lo cual podría constituir una regla aplicable a un análisis predictivo del tipo de entidad. Ciertos verbos como *escribir*, *pensar*, *oler*, *amar*, etc., normalmente van acompañados por un antropónimo, pues constituyen acciones exclusivas de los seres humanos. Las reglas de predicción se plantean a partir de la recurrencia de casos, los que se extraen del análisis morfosintáctico de las oraciones, y del texto completo mediante la información contextual entregada por la correferencia a los NP (mediante pronombres, hiperónimos, entre otros). En este trabajo, el *input* es el texto; en este plano se van a analizar los NP, con énfasis en la relación verbo-nombre propio. Por lo tanto, la lingüística textual se constituye como el enfoque teórico de la investigación, debido a que el carácter del análisis del corpus no se satisface solo desde el nivel sintáctico.

Si bien los porcentajes conseguidos no son óptimos, la tasa de mejora es sustancial respecto a lo obtenido con *Freeling*. Con respecto a otros programas, como *Semantria* o *Citius Tagger*, que muestran un desempeño superior, resulta difícil comparar resultados con los obtenidos por este sistema porque se trata de aplicaciones con algoritmos propietarios, los cuales no pueden ser modificados ni tampoco puede estudiarse su funcionamiento. Se ofrecerá este script *Perl* como open source desde el servidor web del Instituto de Literatura y Ciencias del Lenguaje.

3.5.3 **Evaluación manual del elemento verbo en tanto elemento predictor del tipo de entidad al que acompaña**

A continuación, a modo de profundización de las evaluaciones en torno al objetivo de esta investigación y debido a los problemas con el analizador de

dependencias de *Freeling,* se procedió a realizar un análisis manual del comportamiento de los verbos, con la finalidad de determinar posibles reglas que puedan predecir el tipo de entidad que aparece junto a determinado verbo. El propósito de esta tercera instancia es controlar con mayor precisión solo el papel que desempeñan los verbos y cuando estos mantienen una relación de dependencia sintáctica con el nombre propio.

Luego de un análisis completo del corpus con *Freeling* se calcularon la cantidad total de verbos que operan como sujeto y como objeto de nombres propios de personas, de lugares y de organizaciones. A partir de esta lista completa, se escogieron algunos verbos que tenían la más alta cantidad de apariciones según el tipo de entidad al que acompañan. A continuación, se escogieron algunos verbos que respondían a la frecuencia media de apariciones en el corpus. Habiendo realizado el levantamiento de estos datos, se procedió a buscar esta información en el corpus con la herramienta de análisis y concordancias textuales *AntConc* (http://www.laurenceanthony.net/software.html).

Esta última instancia de la investigación tuvo como propósito reafirmar los hallazgos encontrados en los ejercicios de experimentación anterior y aproximarse al objetivo específico de la tesina: cómo determinar el grado de precisión del verbo como predictor de la categoría de nombre propio. En primer lugar, como se señaló en el apartado referido a la gramática del texto, *Freeling* solo pudo identificar los nombres propios según tipo de entidad (persona, lugar, organización) cuando estos aparecían acompañando al verbo de manera explícita. Los mecanismos de correferencia y anáforas (entre otros) aún no han sido integrados a los sistemas de análisis de corpus, estos solo operan a nivel oracional. Por lo tanto, la cantidad de casos detectados por *Freeling* fue inferior a la cantidad real de referencias a nombres propios, lo cual deja en evidencia la necesidad de investigar en profundidad cómo hacer que un sistema automático procese información considerando al texto completo como unidad de análisis y no solo a nivel sintáctico.

Un antecedente teórico de esta tarea puede encontrarse en Renau & Nazar (2011), en el cual se propone una metodología para la creación automática de patrones léxicos basada en el *Corpus Pattern Analysis* (CPA) de Hanks. Esta propuesta se

fundamenta en la carencia de significado intrínseco de las palabras, pues estas lo adquieren en contexto. En este sentido, “el CPA se presenta como un método basado en corpus para la detección de los patrones normales de uso de una palabra, en relación con sus rasgos sintácticos y semánticos”.

A continuación, se contabilizaron la cantidad de apariciones de los verbos seleccionados en el corpus 1 y en el corpus 2 para observar si existía alguna regularidad, al menos en la cantidad de apariciones y en la conducta sintáctica de la relación entre nombre propio / verbo. En tanto elementos predictivos del tipo de entidad, los verbos que entregaron mayor información relevante fueron los de frecuencia media, pues su regularidad de comportamiento sintáctico permitió plantear algunas reglas que fueron puestas a prueba evaluando el corpus 2. La tabla que se presenta a continuación da cuenta de una propuesta metodológica que pretende determinar el potencial predictor del tipo de entidad que acompaña a un verbo y la tasa de recurrencia entre la cantidad de apariciones exitosas de la regla planteada en ambos corpus.

Verbo / Clasificación Aktionsart Zeno Vendler	**Cantidad total de apariciones del verbo junto a NP de entidad según *Freeling***	**Regla formulada a partir de corpus 1**	**Apariciones en corpus 1 de acuerdo a regla formulada**	**Apariciones en corpus 2 de acuerdo a regla formulada**	**% de relación de recurrencia de la regla planteada a partir de la cantidad de eventos afines entre ambos corpus**	**% de relación de recurrencia de la regla planteada respecto a la cantidad de apariciones del verbo junto a NP de entidad según *Freeling***
Ver / Proceso	525	Ver opera como predictor para NP de lugar en posición argumento 2	30 apariciones del caso planteado en la regla	30 apariciones del caso planteado en la regla	**100%** recurrencia de la regla a partir de la cantidad de casos encontrados en ambos corpus	60 casos exitosos de un total de 525 detectados automáticamente = **11,4%** de potencial predictivo del tipo de entidad que acompaña al verbo
Destinar / Realización	61	Destinar opera como predictor para NP de lugar en posición argumento 2	36 apariciones del caso planteado en la regla	20 apariciones del caso planteado en la regla	**55,6%** de recurrencia de la regla a partir de la cantidad de casos encontrados en ambos	56 casos exitosos de un total de 61 detectados automáticamente = **91,8%** de potencial predictivo del

					corpus	tipo de entidad que acompaña al verbo
Estar / Estado	1314	Estar opera como predictor para NP de organización en posición argumento 1	267 apariciones del caso planteado en la regla	320 apariciones del caso planteado en la regla	**83,4%** de recurrencia de la regla a partir de la cantidad de caso encontrados en ambos corpus	587 casos exitosos de un total de 1314 detectados automáticamente = **44,7%** de potencial predictivo del tipo de entidad que acompaña al verbo
Instalar / Realización	104	Instalar opera como predictor para NP de organización en posición argumento 1	36 apariciones del caso planteado en la regla	31 apariciones del caso planteado en la regla	**86,1%** de recurrencia de la regla a partir de la cantidad de caso encontrados en ambos corpus	67 casos exitosos de un total de 104 detectados automáticamente = **64,4%** de potencial predictivo del tipo de entidad que acompaña al verbo
Decir / Proceso	588	Decir opera como predictor para NP de persona en posición argumento 2	131 apariciones del caso planteado en la regla en posición argumento 2	158 apariciones del caso planteado en la regla en posición argumento 2	**82,9%** de recurrencia de la regla a partir de la cantidad de casos encontrados en ambos corpus	289 casos exitosos de un total de 588 detectados automáticamente = **49,1%** de potencial predictivo del tipo de entidad que acompaña al verbo
Agregar / Proceso	74	Agregar opera como predictor para NP de persona en posición argumento 1	36 apariciones del caso planteado en la regla	27 apariciones del caso planteado en la regla	**75%** de recurrencia de la regla a partir de la cantidad de casos encontrados en ambos corpus	63 casos exitosos de un total de 74 detectados automáticamente = **85,1%** de potencial predictivo del tipo de entidad que acompaña al verbo
Suceder / Proceso	60	Suceder opera como predictor para NP de lugar en posición argumento 2	51 apariciones del caso planteado en la regla	45 apariciones del caso planteado en la regla	**88,2%** de recurrencia de la regla a partir de la cantidad de casos encontrados en ambos corpus	96 casos exitosos de un total de 60 detectados automáticamente = **160%** de potencial predictivo del tipo de entidad que acompaña al verbo
Entregar / Logro	101	Entregar opera como predictor para NP de	34 apariciones del caso planteado en	41 apariciones del caso planteado en	**82,3%** de recurrencia de la regla a partir de la	75 casos exitosos de un total de 101 detectados

		organización en posición argumento 2	la regla	la regla	cantidad de casos encontrados en ambos corpus	automáticamente = **74,3%** de potencial predictivo del tipo de entidad que acompaña al verbo
Entender / Logro	75	Entender opera como predictor para NP de persona en posición argumento 1	24 apariciones del caso planteado en la regla	13 apariciones del caso planteado en la regla	**54,1%** de recurrencia de la regla a partir de la cantidad de casos encontrados en ambos corpus	37 casos exitosos de un total de 75 detectados automáticamente = **49,3%** de potencial predictivo del tipo de entidad que acompaña al verbo

Tabla 7: Variables y resultados obtenidos de la evaluación de la información arrojada por Freeling y el potencial predictivo de los verbos

3.5.3.1 Explicación de los resultados obtenidos con cada verbo:

- **Ver**: es un verbo con alta recurrencia (525 respecto a la media de 61,3 apariciones detectadas por automáticamente por Freeling en la lista de verbos que acompañan al nombre propio del lugar) por lo que no se considera como un evento problemático el irregular resultado obtenido. Puede colegirse que a mayor distancia de la frecuencia media, menor será el potencial predictivo del tipo de entidad que acompaña al verbo (solo 11,4% en este caso). No obstante, es interesante la relación de recurrencia de apariciones de la regla planteada en ambos corpus (en ambos la regla se presentó en 30 ocasiones) pues da cuenta de un "comportamiento" armónico del verbo "ver" en el corpus.

- **Destinar**: La cantidad de apariciones detectadas automáticamente por Freeling coincide con la frecuencia media de apariciones de verbos que acompañan al tipo de entidad "lugar" (61,3), por lo que no resulta sorpresivo que el potencial predictivo del verbo sea alto (91,8%).

- **Estar**: el potencial predictivo del verbo en este corpus no es óptimo pues su recurrencia es de 1314 apariciones junto a NP de organización, respecto a una frecuencia media de 102,7 recurrencias. Sin embargo, la cantidad de apariciones de la

regla planteada (“estar” opera como predictor para NP de organización en posición argumento 1) alcanza un 83,4%, lo cual evidencia una regularidad en el comportamiento de los verbos dentro de un determinado género textual.

- **Instalar**: en el caso de este verbo, el potencial predictivo es solo de un 64,4% pues en el contexto de este corpus de geopolítica, tanto una persona como una organización pueden ser agentes de la acción "instalar", lo que genera ambigüedad referencial. Por su parte, la recurrencia del comportamiento del verbo respecto de la regla planteada es de un 86,6% debido a que este verbo aparece muy cerca de la frecuencia media de apariciones (102,71 frente a 104 apariciones detectadas automáticamente por Freeling).

- **Decir**: en el contexto del corpus, este verbo posee poco potencial predictivo (49,1%) debido a que está considerablemente distante de la frecuencia media de apariciones de verbos que acompañan al NP de persona (73,7 apariciones como media, respecto a 588 apariciones del verbo en el corpus, acompañando a un NP de persona). Cabe señalar que la recurrencia de la regla es de un 82,9% en ambos corpus, posiblemente por la correcta adecuación de la regla planteada dentro del contexto del corpus.

- **Agregar**: dentro del contexto del corpus, "agregar" se constituye como un verbo con un óptimo potencial predictivo del tipo de entidad al que acompaña (85,1%). Además, la regla planteada se presenta de manera uniforme en ambos corpus, alcanzando un 75% de similitud en la cantidad de apariciones de la regla en ambos corpus.

- **Suceder**: Al igual que "agregar", este verbo posee una coocurrencia similar (60 apariciones) a la frecuencia media de apariciones del tipo de verbos que acompañan al NP de lugar (61,3 apariciones de media), por lo que la recurrencia de la regla en ambos corpus fue óptima (88,2%). Respecto al potencial predictivo del verbo, este alcanzó un 160% debido a que en diversas ocasiones el NP adyacente se encontraba elidido o pronominalizado (“*Si bien a nivel local la autocrítica es posible dentro de las instituciones culturales públicas, no sucede lo mismo en la capital*”: “La capital” hace

referencia a Tokio dentro del texto) y Freeling no se encuentra habilitado para inferir/detectar mecanismos de correferencia textual. Este análisis se realizó de manera manual, debido a esto la detección de casos exitosos fue superior a la detectada automáticamente por el software.

- **Entregar**: en el caso de "entregar", su cantidad de apariciones junto a un NP de organización (101) es similar a la frecuencia media de apariciones detectadas automáticamente por Freeling (102,7) por lo que la recurrencia de la regla planteada a partir del corpus 1 es óptima (82,3%) y el potencial predictivo del verbo alcanza el 74,3%.

- **Entender**: En el caso de este verbo, los resultados obtenidos no son óptimos ya que la regla planteada aplica tanto para el NP de persona en posición argumento 1 ("*José Saramago entendía la solidaridad como un hecho consustancial a vivir*") y en posición argumento 2 ("*Ahora bien, para entender a Frondizi habrá que analizar también a los otros protagonistas*"). Para alcanzar un potencial predictivo óptimo (sobre el 80%) la regla debiese plantear que *"entender" opera como predictor para NP de persona en posición argumento 1 y 2*.

4. Conclusiones

En esta tesina de pregrado se abordó el análisis y clasificación de nombres propios en un corpus de geopolítica de la revista *Le Monde Diplomatique,* desde la gramática del texto. El tratamiento computacional de la investigación posibilitó la elaboración de una propuesta metodológica para determinar el potencial predictivo del tipo de entidad que acompaña a algunos verbos, la cual a pesar de su precariedad y perfectibilidad, se constituye como una respuesta a la carencia de bibliografía referida a esta temática. Por otra parte, la experimentación realizada dio cuenta de la necesaria (y poco desarrollada aún) vinculación entre la gramática del texto y la lingüística computacional para el análisis y tratamiento del lenguaje natural. El análisis del discurso y la gramática textual deben profundizar sus acercamientos con la ciencia computacional, así como los especialistas de la informática debiesen trabajar arduamente en optimizar los procesos de programación y modelado de la comunicación humana en sistemas automáticos. Los resultados finales de esta tesina demostraron que efectivamente el verbo sí puede funcionar como elemento predictor del tipo de entidad al que acompaña, además de evidenciar que el potencial predictivo de algunos verbos es más alto que otros.

La presente investigación dejó en evidencia la necesidad de incentivar la enseñanza del modelado y confección de patrones lingüísticos al momento de estudiar el texto. Como política educativa a nivel mundial, la programación computacional será una de las áreas de mayor relevancia para una óptima inserción en la sociedad de la información. Por lo tanto, el mero análisis teórico y descriptivo de la lengua ya no es suficiente si se pretende formar especialistas del lenguaje con competencias profesionales adecuadas para las actuales y futuras necesidades del mercado laboral: es necesario que el lingüista del siglo XXI pueda modelar sus hallazgos lingüísticos a través de código, manipulando grandes cantidades de texto en un entorno *Big Data*. El lingüista del siglo XXI debe saber programar, el profesor de lengua debe saber programar para hacer del lenguaje una instancia de reflexión y aprendizaje significativo en la educación escolar de la nueva era, altamente enfocada en el desarrollo tecnológico y la innovación.

Otro experimento interesante es analizar y procesar automáticamente corpus de diversos géneros textuales, siempre desde la perspectiva del estudio del NP. Si bien los textos de geopolítica se caracterizan por componerse de una alta recurrencia de entidades nombradas, la organización retórica y estilística del corpus de *Le Monde Diplomatique* no abarca toda la riqueza y diversidad lingüística que sí podrían aportar otros géneros como el periodístico, narrativo, académico, etc. En este sentido, se propone la inclusión de nuevos cursos en los planes de estudio de las carreras de lengua, con el propósito de entregar a los alumnos más herramientas para el análisis cuantitativo del texto y no solo cualitativo como se ha venido realizando desde hace décadas.

Como trabajo futuro, existe particular interés en estudiar el poder predictor de las unidades del contexto en función de su categoría gramatical. Además de esto, se propone extender esta investigación a fin de no estudiar solamente los elementos que coocurren con las entidades sino también aquellos que contraen una determinada relación sintáctica con el NP, como sería el caso prototípico del verbo. Esta investigación requiere realizar un análisis sintáctico de dependencias, y si bien *Freeling* en principio puede hacer tal análisis, esto resulta computacionalmente mucho más complejo, requiere mayor tiempo de procesamiento y un sistema de control más exhaustivo para no entrar en bucles infinitos cuando procesa oraciones muy largas. El estudio de la colocación verbo-nominal es un ámbito sumamente interesante desde el punto de vista teórico y actualmente constituye otro atractivo nicho de investigación. Finalmente, otra tarea a realizar será compilar una base de datos de aquellos verbos que funcionen mejor como predictores del tipo de entidad. Probablemente sería de interés para algún equipo de ciencia computacional interesado en la recuperación de información y en mejorar las tecnologías de búsqueda en la web, entre otras potenciales aplicaciones.

5. Referencias

Abney, S. P. (1994). Parsing By Chunks. *Bell Communications Research*, 1-18. Retrieved Septiembre 13, 2015, from http://www.vinartus.net/spa/90e.pdf

Atserias, J., Casas, B., Comelles, E., Gonzalez, M., Padró, L., & Padró, M. (2006). FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006).* Genoa: ELRA.

Ballesteros, M. (2010). *Mejora de la precisión para el análisis de dependencias usando Maltparser para el castellano.* Tesis doctoral. Madrid: Universidad Complutense de Madrid.

Bayes, T. (1763). An essay towards Solving a Problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*(53), 370-418. Recuperado el 27 de Diciembre de 2015, de http://www.stat.ucla.edu/history/essay.pdf

Bibiloni, E. G. (1999). Narratividad y aspectualidad. *Anclajes: Revista del Instituto de Análisis Semiótico del Discurso, 3*(3), 17-56.

Colenguando. (31 de Julio de 2015). *Colenguando*. Obtenido de La lengua, las humanidades y las expresiones regulares: http://encomienda.colenguando.com/la-lengua-las-humanidades-y-las-expresiones-regulares/

Charaudeau, Patrick. (1992): *Grammaire du sens et de l'expression*, París, Hachette.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind.* Cambridge : University Press.

Coseriu, E. (1982). El plural en los nombres propios. En *Teoría del Lenguaje y Lingüística General* (págs. 261-281). Madrid: Gredos.

Cuenca, M. J. (2010). *Gramática del texto.* Madrid : Arco Libros.

De Miguel, E. (1999). El aspecto léxico. En I. Bosque, & V. Demonte, *Gramática Descriptiva de la Lengua Española* (págs. 2977-3060). Madrid: Espasa Calpe.

Delgado-Díaz, Gibran. (2014). Teoría versus uso: análisis sobre el pretérito y el imperfecto. *Boletín de filología*, *49*(1), 11-36. Recuperado en 04 de diciembre de 2015, de http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-93032014000100001&lng=es&tlng=es. 10.4067/S0718-93032014000100001.

Delgado Díaz, Gibran & Ortiz López, Luis A. *Selected Proceedings of the 14th Hispanic Linguistics Symposium*, ed. Kimberly Geeslin and Manuel Díaz-Campos, 165-178. Somerville, MA: Cascadilla Proceedings Project, 2012. Visto en http://www.lingref.com/cpp/hls/14/paper2663.pdf , recuperado el 30 de noviembre de 2015.

Domínguez Burgos, A. (2002). Lingüística computacional: un esbozo. *Boletín de Lingüística*, 104-119.

Ferrández, Oscar, y otros. «NERUA: sistema de detección y clasificación de entidades utlizando aprendizaje automático.» *Procesamiento del Lenguaje Natural* 35 (2005): 37-44. Digital.

García Reyes, R. (2012). *Minería de datos para la toma de decisiones e inteligencia de negocios: aplicaciones en la mercadotecnia.* Universidad Nacional Autónoma de México.

Gary-Prieur, Marie-Noëlle (1994). *Grammaire du nom propre*, París, PUF, 1994.

Hutchins, W. John & Sommers, Harold L. 1992. *An Introduction to Machine Translation*. London. Academic Press.

Hutchins, J. (1998). The origins of the translator's workstation. *Machine Translation, 13*(4), 287-307. Recuperado el 15 de Noviembre de 2015, de http://www.hutchinsweb.me.uk/MTJ-1998.pdf

Jonasson, Kerstin. *Le nom prope. Constructions et interprétations*. Lovaina: Duculot, 1994.

Jurafsky, Daniel & Martin, James. 2000. *Speech and Language Processing,An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. New Jersey. Prentice Hall. Upper Saddle River.

Kripke, Saul (1972). *El nombrar y la necesidad*. México, UNAM, 1985. Traduccido al español.

Le Monde Diplomatique edición Cono Sur (2010). El Dipló - archivos completos (cd). Santiago. doi:ISBN 978-987-614-278-6

López Morrás, J. (2004). *¿Qué es la Lingüística Computacional o PLN?* Recuperado el 5 de septiembre de Aùcel Digital http://www.aucel.com/pln/k-es.html

López Cabello, V. (2010). *Inteligencia Artificial Eliza.* Barcelona. Recuperado el 14 de Octubre de 2015, de http://lab.uvic.cat/sites/default/files/memories/Inteligencia%20Artificial_Victoria%20Lopez_.pdf

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An introduction to information retrieval.* Cambridge: Cambridge University Press.

Mill, John Stuart (1843): *A system of logic, Ratiocinative and Inductive:* Londres, Routledge and Kegan Paul.

Montrul, Sylvina & Slabakova, Roumyana. 2000. Acquiring Semantic Properties of Preterite and Imperfect Tenses in L2 Spanish. *Proceedings from BUCLD*, 24.

Morala Rodríguez, J. (1986) El nombre propio, ¿objeto de estudio interdisciplinar?. *Revista Contextos*, 49-62.

Mosteller, F., & Wallace, D. (1984). *Applied Bayesian and classical inference the case of the Federalist papers.* New York: Springer.

Pawlak, A. (2008). Sobre los orígenes y las confusiones terminológico-conceptuales de los términos de aspecto y de aktionsart. *Studia Romanica Posnaniensia, 35*, 257-266.

RAE. (2009). *Nueva gramática de la lengua española.* Madrid: Espasa Libros.

Rebollo, M. (1995). El nombre propio y su significado. Anuario De Estudios Filológicos, 399-406.

Renau, I., & Nazar, R. (2011). Propuesta metodológica para la creación automática de patrones léxicos usando el Corpus Pattern Analysis. *Actas del 27º Congreso de la SEPLN*.

Russell, B. (1905). ”Sobre el denotar” en T.M. Simpson (ed.) (1973) *Semántica filosófica: problemas y discusiones* Madrid: s.xxi, pp. 29-48.
- (1912): *The problems of Philosophy,* Londres. Home University Library. [Vers. Esp. *Los problemas de la filosofía*, Barcelona, Labor, 1970 (1928 1º edición].

Searle, J. Proper names and descriptions. *The encyclopedia of philosophy*, edited by Paul Edwards, The Macmillan Company & The Free Press, New York, and Collier-Macmillan Limited, London, 1967, Vol. 6, pp. 487–491.

Snow, C. (2000). Las dos culturas. Buenos Aires: Ediciones Nueva Visión. (Obra original publicada en 1959).

Solorio, T. (2005). *Taking Advantage of Existing Named Entity Taggers by Machine Learning.* Tesis doctoral. Puebla: National Institute of Astrophysics.

Vendler, Zeno. 1957. Verb and Times. *The Philosophical Review*. Vol. 66, No. 2. 143-160.
-(1967). Linguistics and Philosophy. Ithaca, New York: Cornell University Press, 1967.

Villayandre Llamazares, M. (2010). *Aproximación a la Lingüística Computacional.* Tesis doctoral. León: Universidad de León.

Wall, L., Christiansen, T., & Orwant, J. (2000). *Programming Perl.* California : O’Reilly & Associates, Inc.