PROBLEMAS DEL TRATAMIENTO COMPUTACIONAL DE LA ANÁFORA: ANÁLISIS EN UN CORPUS PARA LA ENSEÑANZA DEL ESPAÑOL LENGUA EXTRANJERA

ANDRÉS FELIPE GRAJALES RAMÍREZ

Trabajo de grado para obtener el título de:
FILÓLOGO HISPANISTA

Asesor
JORGE MAURICIO MOLINA MEJÍA
Doctor en Informática y Ciencias del Lenguaje

Evaluadora
LAURA MARCELA QUINTERO MONTOYA
Magister en Lingüística Aplicada a la
Enseñanza del Español como Lengua Extranjera

PREGRADO EN FILOLOGÍA HISPÁNICA
FACULTAD DE COMUNICACIONES
UNIVERSIDAD DE ANTIOQUIA
2019

UNIVERSIDAD DE ANTIOQUIA
1 8 0 3

Agradecimientos

Antes que nada, y en pocas palabras, quiero agradecer a algunas de las personas y entidades que, de alguna manera, participaron y apoyaron la concepción de este proyecto. Entre ellas, en primer lugar, a la Universidad de Antioquia y a los profesores de la carrera por todo lo que me han enseñado y transmitido durante estos años. Quiero dar gracias al Comité para el Desarrollo de la Investigación (CODI) por patrocinar el proyecto DICEELE en el cual me desenvolví como Joven Investigador y del cual deviene este trabajo de grado. Además, quiero agradecer al semillero *Corpus ex Machina* por permitirme trabajar en sus proyectos y a sus integrantes por el trabajo que hemos compartido en ellos. Me gustaría nombrar especialmente al profesor Jorge Mauricio Molina por ser mi guía, apoyo y dirección durante este arduo proceso. Finalmente, doy las gracias a mis amigos, Paulina y José Luis, por el apoyo y conocimiento que recibí de ellos durante este trabajo.

Resumen

La Enseñanza de Lenguas Asistida por Ordenador (ELAO) es un terreno de aplicación de la lingüística computacional que se basa en la computación para apoyar las clases de lengua. El proyecto DICEELE (Dispositivo Informático basado en Corpus para la Enseñanza del Español como Lengua Extranjera) consiste en crear un sistema de ELAO para el aprendizaje del Español como Lengua Extranjera (ELE), en donde se desarrollen las nociones fundamentales de la lingüística textual. Este trabajo de grado surge a partir de dicho proyecto, pues es necesario el procesamiento de corpus de textos para que sean interpretados y explotados por un computador. De esta manera, dentro de las múltiples tareas necesarias para la creación de dicho dispositivo se encuentra el etiquetado de la anáfora como correferente textual. Puesto que el tratamiento computacional de esta noción lingüística aún representa un reto para el Procesamiento del Lenguaje Natural (PLN), el objetivo del presente trabajo es evaluar la efectividad de un analizador de la lengua (en este caso *FreeLing 4.1*) para procesar la anáfora y, a partir de esto, describir los problemas que existen para realizar completamente esta tarea. Con esto se busca la manera en que todo este análisis pueda aplicarse al desarrollo de las actividades en el dispositivo DICEELE y a la clase de ELE.

Palabras clave: Correferencia textual, anáfora, lingüística computacional, lingüística de corpus, enseñanza del español lengua extranjera.

Abstract

Computer Assisted Languages Teaching (CALT) is a field of computational linguistics which is based on computer science to support languages courses. DICEELE project consists of create a CALT system for learning Spanish as a Foreign Language (SFL), where the notions of text linguistics be developed. This dissertation arises from DICEELE project, due to the need of process corpus of text to being interpreted and exploited by a computer. In this way, was very important the tagging of anaphora, as part of the multiple required tasks to create this CALT system. Since the computational treatment of this linguistic notion still represents a challenge for Natural Language Processing (NLP), the objective of this research is to evaluate the effectiveness of a software (in this case, *FreeLing 4.1*) to process coreferences and, based on this, describe the problems that exist to perform this task completely. With this we look the way to apply this analysis to the development of the activities in the CALT system DICEELE and in SFL classes.

Keywords: Coreference, anaphora, computational linguistics, corpus linguistics, Spanish as a Foreign Language.

Contenido

Índice de figuras

Índice de tablas

1. Introducción

El Procesamiento del Lenguaje Natural (PLN) es una de las tareas más importantes dentro de la lingüística computacional, pues sus avances permiten desarrollar y mejorar tecnologías capaces de analizar, aprender e imitar la lengua de los hombres; lo cual lo hace también de gran interés para el desarrollo de la Inteligencia Artificial (IA). Uno de los procesos más complicados para el PLN es la resolución anafórica (Saiz, 2002), la cual, en la comunicación humana, suele resolverse con sencillez, valiéndose de distintos factores como el contexto o el conocimiento del mundo. Al no poder acceder de la misma manera a este proceso cognitivo, la máquina debe ser programada con el conocimiento lingüístico suficiente para resolver estos problemas. Este trabajo le corresponde al científico de la lengua, que la estudia a profundidad con el fin de comprenderla y crear abstracciones y reglas que permitan su tratamiento computacional (Badia, 2003).

Uno de los principales problemas que se detectan en el PLN sobre la anáfora es la ambigüedad lingüística, la cual entorpece el proceso de reconocimiento automático, etiquetado y establecimiento de correferencias (Peña, 2016). Un avance en este aspecto apoyaría en gran medida, por ejemplo, el procesamiento de corpus de textos que necesitan etiquetar estas nociones. Debemos tener en cuenta que la manera en que un procesador automático del lenguaje trabaja la lengua es muy diferente a la nuestra como seres humanos. El hablante se vale del conocimiento del mundo que lo rodea, sobreentendidos y presuposiciones para comunicarse eficazmente y resolver los casos de ambigüedad lingüística que se le presenten. En cambio, estas reacciones no son posibles para una máquina que debe procesar una lengua natural. Su programación solamente le permite trabajar de manera formal y lógica el sistema lingüístico, estableciendo reglas y excepciones, pero no como lo procesa el cerebro humano. Por lo cual, sus problemas para procesar la lengua son muy diferentes a los nuestros. Esto nos motiva a indagar sobre cuáles son específicamente esas dificultades para el procesador informático con el fin de comprender su lógica y, posteriormente, buscar soluciones a ellas.

El proyecto DICEELE (Dispositivo Informático basado en Corpus para la Enseñanza del Español Lengua Extranjera[1]) se basa en el procesamiento de textos en español para crear actividades de enseñanza mediante un aplicativo informático. Este dispositivo permitirá el

[1] Proyecto avalado en 2016 por el Comité para el Desarrollo de la Investigación (CODI) de la Universidad de Antioquia.

diseño de actividades para los cursos de Español como Lengua Extranjera (ELE) con base en corpus recolectado especialmente para ello. De acuerdo con esto, se debe realizar una anotación de las nociones lingüísticas que van a trabajarse en las actividades didácticas, entre ellas, la anáfora. Aparece entonces la necesidad de métodos computacionales para agilizar en el proyecto los procesos de etiquetado e identificación de estas entidades. Por lo tanto, un estudio a profundidad del fenómeno dentro del corpus DICEELE y de las dificultades para su tratamiento computacional permitirá una mejor comprensión de la anáfora y de cómo debemos tratarla computacionalmente en el proyecto en general.

En un comienzo, este trabajo pretendía desarrollar y evaluar el procesamiento del fenómeno de la progresión temática (el cual también hace parte de los objetivos del proyecto DICEELE). Sin embargo, cuando empezamos a concebir el problema, nos dimos cuenta de que, para llegar a ella, debíamos primero resolver la cuestión de la correferencia textual y su procesamiento computacional. En este sentido, consideramos que un tratamiento basado en el etiquetado manual de las relaciones anafóricas podría desvelar sus características y generar vastas posibilidades de explotación:

> La posibilidad de corpus sin etiquetar y de corpus etiquetados con enlaces referenciales dio un fuerte impulso a la resolución de la anáfora tomando en cuenta el entrenamiento y la evaluación; los corpus (especialmente cuando están etiquetados) son un recurso de gran valor para la investigación empírica y los métodos de aprendizaje automático que animan el desarrollo de diferentes enfoques, posibilitando también medios para la evaluación de algoritmos desarrollados. Desde simples reglas de co-ocurrencia [Dagan e Itai, 1991] pasando por el entrenamiento de árboles de decisión para identificar las parejas anáfora y antecedente [Aone y Bennett, 1995], hasta algoritmos genéticos para optimizar los factores que afectan la resolución de la anáfora [Orasan et al, 2000], han sido logrados gracias a la posibilidad de contar con corpus adecuados (Morales, 2004: p. 3).

En este capítulo, expusimos la motivación y el asunto principal de nuestra investigación. En el cap. 2, enunciamos los objetivos que pretendemos alcanzar con ella. El cap. 3 se refiere a diversos trabajos que tienen relación con nuestro tema. En el cap. 4, trabajamos minuciosamente las bases teóricas centradas en la lingüística textual y el PLN y también revisamos diferentes clasificaciones sobre la anáfora y tomamos las que mejor se adapten a nuestra cuestión. Posteriormente, en el cap. 5, describimos las bases de nuestra metodología y el proceso que llevamos a cabo para desarrollar esta investigación. El cap. 6 contiene el análisis de los datos y la evaluación de los problemas identificados en el PLN para la anáfora. Finalmente, el cap. 7 expone las conclusiones y perspectivas del trabajo.

2. Objetivos

2.1. General

Identificar y analizar la problemática que actualmente existe en el tratamiento computacional de la anáfora utilizando, para ello, un analizador automático del lenguaje con capacidad de procesar correferencias y con base en un corpus especializado para la enseñanza del ELE (corpus DICEELE).

2.2.Específicos

- Crear, a partir de la literatura consultada, una tipología de la anáfora que sirva para clasificar los elementos anafóricos encontrados en el corpus DICEELE.
- Seleccionar una muestra pertinente del corpus DICEELE que respete y represente sus criterios de recolección.
- Etiquetar de manera manual las anáforas y sus respectivos antecedentes, según se encuentren en la muestra, con base en la clasificación que previamente establecimos.
- Identificar los errores que comete un procesador informático de correferencias en la identificación de anáforas presentes en nuestra muestra con base en la anotación manual que previamente realizamos.
- Evidenciar en qué tipos de anáforas suele equivocarse el sistema informático y cómo estos errores se relacionan con las características del corpus DICEELE.

3. Antecedentes

Hace pocos años, en diversos proyectos sobre el PLN, empezó a surgir una preocupación por la resolución automática de la correferencia en español, influenciada por trabajos realizados para el inglés. El Centro de Investigación en Computación del Instituto Politécnico Nacional (IPN) y el Dpto. de Lenguajes y Sistemas Informáticos de la Universidad de Alicante son muestra de múltiples publicaciones y tesis que se acercan al tratamiento computacional de este fenómeno; de la mano del Dr. Ruslan Mitkov quien adelantó su trabajo en este campo aplicado al inglés (Mitkov, 2002). Por ejemplo, podemos encontrar, desde la década pasada, trabajos como los de Martínez (2001), Saiz (2002), Morales (2004) y Olivas (2006), los cuales se enfocan en casos específicos de resolución anafórica. Los trabajos de esta época se realizaban con la finalidad de mejorar aplicaciones del PLN tales como la traducción automática, la generación de resúmenes, los sistemas de diálogos o la extracción de información (Mitkov, 2002).

En relación con los objetivos de este trabajo han sido diversos los estudios que analizan el fenómeno de la anáfora para proponer un esquema de anotación específico para una lengua. Tenemos, por ejemplo, los trabajos de Ceberio *et al*. (2008) en los cuales se identifican las características de la anáfora en el euskera para mejorar el procesamiento de anotación en una fase posterior, mediante el uso de herramientas informáticas. Otro estudio relacionado es la tesis doctoral de Navarro (2007) la cual, en parte, analiza distintas metodologías para anotar corpus con información anafórica y propone campos en los que este tipo de anotación puede ser explotada. Este autor ha trabajado, además, en el proceso de enriquecimiento del corpus Cast3LB con información correferencial.

Además de estos, son de singular importancia los aportes de Marta Recasens y M. Antònia Martí (2007; 2010), debido a su amplio conocimiento en anotación de corpus con información anafórica como lo fue el corpus CESS-ECE en español, con el cual se pretendía construir un sistema de resolución anafórica basado en aprendizaje automático a partir de información previamente anotada (Recasens *et al*., 2007). Destacamos también en estas autoras la anotación de correferencias del corpus AnCora-CO para el español y el catalán (Recasens & Martí, 2010). A partir de este trabajo, nos vimos influenciados para enfocarnos en

los casos problemáticos del tratamiento computacional de la correferencia. Esto, porque además de presentar un esquema de anotación, realizan un listado de las dificultades lingüísticas con las que se debe tratar al etiquetar relaciones anafóricas en lenguas similares como son el español y el catalán.

De igual manera, relacionados también con los objetivos de nuestra tesis, están los trabajos que no se enfocan en el corpus trabajado, sino en el programa informático capaz de procesar correferencias. Dentro de este grupo, queremos mencionar, en primer lugar, el modelo *MICE* (Arévalo *et al*., 2004) el cual se basa en el reconocimiento de entidades nombradas, como fundamentalmente también lo hacen otros sistemas. Por otra parte, se encuentra el software *RelaxCor* (Sapena *et al*., 2013), el cual está implementado en la herramienta para resolver correferencias de *FreeLing 4.1*[2]. Por último, hacemos referencia a *FunGramKB*,[3] una base de conocimiento léxico-conceptual que se basa en la gramática funcional para relacionar lo léxico con lo cognitivo y, a partir de ello, desarrollar aplicaciones de PLN (Ruíz, 2012). Cabe destacar aquí el análisis que realiza Carrión (2014) sobre el procesamiento de correferencias por medio del conocimiento cultural disponible en *FunGramKB* y de cómo, utilizando esta base de conocimiento se puede ampliar el potencial de la herramienta para resolver problemas de ambigüedad en el reconocimiento de anáforas.

[2] *FreeLing* es un sistema desarrollado por el centro de investigación TALP de la Universitat Politècnica de Catalunya bajo software libre que posibilita diversos tipos de análisis lingüísticos, entre ellos el morfológico, la detección de entidades nombradas, la desambiguación del sentido de las palabras, el análisis sintáctico y de correferencias. Se encuentra disponible en: http://nlp.lsi.upc.edu/freeling/node/1.

[3] *Functional Grammar Knowledge Base*. Para más información, consultar el siguiente enlace: http://www.fungramkb.com/conference.aspx.

4. Marco teórico

4.1. Lingüística Computacional y PLN

4.1.1. Definición del campo

La Lingüística Computacional (LC) es una disciplina científica y aplicada cuyo objetivo es alcanzar la competencia comunicativa del hombre o, por lo menos, la simulación de alguna subcompetencia de esta (Martínez, 2001; Tordera, 2011) por medio de ordenadores y máquinas. De la misma manera, como su nombre lo indica, el Procesamiento del Lenguaje Natural, al intentar simular el comportamiento lingüístico humano (Saiz, 2002), se corresponde con los objetivos de la LC y hacer una distinción entre los dos campos no tendría sentido (Tordera, 2011); incluso algunos autores los utilizan indistintamente (Moreno, 1998; Hausser, 2014). Tomemos entonces la definición de Moreno Sandoval (1998) quien afirma que ambas disciplinas tratan del "desarrollo de programas de ordenador que simulen la capacidad lingüística humana" (p.14), ya sea de manera parcial, e "independientemente de su carácter comercial o de investigación básica" (p.17).

Por otra parte, Moreno Sandoval (1998) expone la necesidad que surge de *modelizar* el lenguaje natural en este campo. Se comprende la lengua desde un enfoque comunicativo, en el que emisor y receptor procesan la información con base en un conocimiento lingüístico y un conocimiento del mundo compartido. Es así como el lingüista computacional debe reflejar todos estos conocimientos de manera estructurada en un lenguaje comprensible para un ordenador: "todas las construcciones que puedan aparecer en una aplicación real tienen que ser explícitamente codificadas en el programa, pues de otra forma no serían reconocidas por el sistema" (Moreno, 1998: p.26). Este autor nos presenta los tres tipos más comunes de modelización: modelos simbólicos, estadísticos y biológicos.

Defendemos, además, como argumenta Nerbonne (2005), que el estudio a fondo de un fenómeno lingüístico desde una perspectiva computacional puede ayudar a su mejor comprensión desde un punto de vista teórico y aplicado:[4] "Existen oportunidades de contribuir

[4] Las traducciones de citas textuales en lengua extranjera irán dentro del cuerpo del texto. La cita en su idioma original la ofrecemos en notas al pie. Todas las traducciones que aparecen en este texto son de nuestra autoría.

computacionalmente a la lingüística pura"[5] (p.2). Es necesario un estudio lingüístico muy profundo de la subcompetencia que se busca tratar en determinado contexto para lograr mejores resultados en una investigación sobre PLN.

4.1.2. Usos y aplicaciones

Comúnmente se caracteriza la LC desde una perspectiva teórica y otra aplicada, o ingenieril, (Moreno, 1998; Nerbonne, 2005; Tordera, 2011). La teórica apunta a construir abstracciones lingüísticas, por lo que su principal función consiste en proponer algoritmos que sean apropiados y eficaces frente a las necesidades y estructuras planteadas a nivel del lenguaje (Nerbonne, 2005). En relación con la aplicada, surgen múltiples campos en los que la LC crea sistemas útiles que involucran la manipulación del lenguaje. Moreno Sandoval (1998: p.29) divide en cuatro partes las aplicaciones más comunes en LC y PLN: 1) Sistemas que tratan de emular la capacidad humana de procesar lenguas naturales; por ejemplo, la traducción automática, recuperación y extracción de la información y las interfaces hombre-máquina. 2) Sistemas que ayudan en las tareas lingüísticas, como las herramientas de análisis textual, de manejo de corpus (etiquetadores) o bases de datos lexicográficos. 3) Programas de ayuda a la escritura y composición textual; p. ej., correctores ortográficos, sintácticos o de estilo. 4) Enseñanza Asistida por Ordenador, en donde destacan la Enseñanza de Lenguas Asistida por Ordenador (ELAO) y el Aprendizaje de Lenguas Asistido por Ordenador (ALAO): "muchos realizan un verdadero análisis […], ya que plantean ejercicios gramaticales y de composición corrigiendo posteriormente las respuestas" (1998: p.29). Este último, por estar relacionado directamente con este trabajo, será tratado en el apartado siguiente.

4.2. Enseñanza y Aprendizaje de Lenguas Asistidos por Ordenador

ALAO y ELAO hacen referencia a toda herramienta tecnológica utilizada en la enseñanza y el aprendizaje de lenguas, por lo que se enmarca en las Tecnologías de la Información y la Comunicación (TIC): "El ALAO busca emplear los ordenadores para mejorar el aprendizaje

[5] Traducido de: "*there are opportunities for computational contributions to 'pure' linguistics"* (Nerbonne, 2005: p.2).

de lenguas. Abarca el rango de actividades en la pedagogía de idiomas -escucha, habla, lectura y escritura- y se acerca a todas las áreas de las TIC"[6] (Nerbonne, 2003: p.631). Un sistema de ALAO no pretende remplazar la labor del profesor, sino facilitarla y fortalecerla; debe ser tomado como una herramienta[7] y nunca como un sustituto: como una innovación técnica en la enseñanza y aprendizaje de segundas lenguas y extranjeras (Nerbonne, 2003). Generalmente, lo que se busca con estos sistemas es una *enseñanza híbrida*, donde se combinen los cursos presenciales con un aprendizaje autónomo guiado: el docente guía desde la distancia (Charlier *et al*., 2006).

Así pues, frente a un enfoque transmisionista y tradicional en la enseñanza de lenguas extranjeras aparece un enfoque pedagógico cercano al constructivismo social: "Los estudios sobre los procesos mentales derivados de la psicología cognitiva y constructivista describen el aprendizaje como un proceso de construcción de conocimiento que es fruto de la interrelación entre lo que el estudiante conoce y aquello que es nuevo para él" (Cánovas, 2009: p.104). Este enfoque nos permite diseñar, por ejemplo, actividades que comiencen por un proceso inductivo lingüístico a partir de textos auténticos en la lengua meta y apoyadas por las herramientas que ofrece el computador. En relación con esto, los sistemas de ELAO también pueden plantearse como *abiertos* y *parametrables* (Antoniadis, 2010; Molina, 2015); es decir, que permitan al docente preparar y diseñar actividades de manera dinámica y a los *aprendientes*[8] realizar dichas actividades y ser evaluados. Según George Antoniadis (2010), los sistemas de ALAO deben responder a las siguientes finalidades didácticas:

- Facilitar el proceso de planificación de actividades a los docentes de lengua extranjera.
- Volver más interesante y participativo el aprendizaje de la lengua extranjera para los aprendientes.

[6] Traducido de: "*CALL seeks to employ computers in order to improve language learning. CALL spans the range of activities in language pedagogy -hearing, speaking, reading and writing- and draws from nearly all areas of ICT*" (Nerbonne, 2003: p.631):

[7] Hay quienes hablan incluso de la *muerte de la ELAO*. Se refieren a la llamada *normalización*, como nos cuenta E. Martín Monje (2012): "aquella etapa en la que la tecnología asociada a esta disciplina sea considerada algo tan ordinario y común como un libro de texto" (p.204).

[8] Se utiliza el término 'aprendiente', en lugar de los más comunes 'estudiante' (muy general) o 'aprendiz' (más relacionado con la formación técnica), puesto que este hace referencia al campo de la enseñanza y el aprendizaje de lenguas extranjeras (Cuq, 2003). El participio presente de la palabra apunta mucho mejor a la situación de aquel que está en el proceso de aprendizaje de una nueva lengua.

- Proporcionar un aprendizaje a partir de textos auténticos que den cuenta de la lengua en la que se expresan los hablantes nativos de esta.

4.2.1. Uso de corpus en ELAO y ALAO

La utilización de corpus de textos en sistemas de ALAO para la enseñanza y el aprendizaje del ELE no es muy común. Sin embargo, el uso de corpus lingüísticos como base para la enseñanza de lenguas sí se ha planteado en mayor medida (Bernal & Hincapié, 2018). Por ejemplo, Pitowsky y Vásquez (2009) exponen cómo el uso de estos en la enseñanza de la lengua permite "estudiar la lengua integrada en el contexto discursivo, a través de ejemplos reales y precisos de uso: en contraposición al empleo de la introspección" (p.31). De igual opinión se presenta Nerbonne (2003), desde una perspectiva del PLN: "Los corpus son valorados por proporcionar acceso al uso real de la lengua"[9] (p.642) y añade que su utilización resulta más efectiva con estudiantes no principiantes, el acompañamiento de instructores y buenas herramientas computacionales.

De manera que entendemos por corpus una colección de textos que se constituyen en una muestra auténtica de la lengua, como lo conciben McEnery y Hardie (2012). Además, debemos diferenciar entre corpus representativo y especializado, en donde importa más para el primero la cantidad de textos que constituyen el corpus que la calidad, proveniencia o tratamiento y precisión de este. En un corpus especializado, los textos se caracterizan por la variedad, relevancia y tratamiento preciso y revisado de la muestra (etiquetados apropiadamente según su aplicación, p. ej.), más que por la cantidad. Un corpus que se encuentre bajo estas condiciones es más adecuado para ser utilizado en la enseñanza de lenguas extranjeras (Teubert, 2009).

4.3. Lingüística Textual

Dicho todo lo anterior, la lingüística textual (el estudio sobre tipos de texto, estructuras y características de este) ha aparecido también en la enseñanza del ELE como una manera más completa de acercarse a la lengua (Lineros, 1998; Arcas, 2000). En este apartado se trabajará el nivel textual de la lengua como unidad de sentido y la anáfora como un elemento que

[9] Traducido de: "*Corpora are valued for providing access to authentic language use*" (Nerbonne, 2003: p.642).

contribuye a la cohesión de este. Fundamentamos aquí el estudio del discurso y del texto como una herramienta necesaria para la comprensión y producción en un aprendiente de lengua extranjera, lo cual le permitirá aumentar su competencia comunicativa. Es decir: describir, argumentar, dialogar, dar instrucciones, etc., cuando sea pertinente hacerlo. El texto sostendrá las emisiones lingüísticas y las funciones comunicativas con el aval del conocimiento del mundo que posea el aprendiente (Arcas, 2000: p.12).

4.3.1. Definición y contexto

Como su nombre lo indica, la lingüística textual (*lingüística del text*o o *textolingüística*) hace referencia a una disciplina cuyo objeto de estudio es el texto, a través de distintos enfoques lingüísticos. Un primer acercamiento a este campo surge durante los años 60, en algunas universidades de Europa Central, donde existió un interés por el estudio del texto traspasando los límites de la oración como unidad de análisis. En aquella época el objetivo de la textolingüística era dar cuenta de la cohesión y coherencia de un texto (Martín Peris, 1997). A finales de la década de los 70, Teun A. Van Dijk (1997) expresa la necesidad de una ciencia del texto desde un enfoque interdisciplinar y establece los parámetros para una *gramática del texto*. De manera similar, propone alejarse de los estudios lingüísticos centrados en el análisis de la oración como unidad autónoma. Así pues, la lingüística textual pretende explicar cómo el contexto, tanto lingüístico como extralingüístico, influye en el texto: en su sentido, producción y comprensión.

Debemos considerar, además, que el objetivo de la lingüística textual no es simplemente separarse del análisis de oraciones aisladas; no obstante, este aspecto sí es de suma importancia a la hora de convertirse en objeto de estudio de la lingüística computacional. Esto último se evidencia en el hecho de que gran cantidad de analizadores automáticos del lenguaje natural utilizan la oración como unidad de sentido y de análisis. Tanto en el campo del PLN como en la enseñanza de lenguas extranjeras la unidad de análisis de la lengua no puede ser exclusivamente la oración: “solo el texto da cuenta del sentido y de la intención comunicativa” (López, 2008: p.88). Por ejemplo, los fenómenos lingüísticos relacionados con la cohesión y la coherencia de un texto hacen referencia, frecuentemente, a elementos externos a la oración, al párrafo o, incluso, a elementos fuera del mismo texto (otros textos, contexto extralingüístico, etc.).

4.3.2. Texto y discurso

Se ha dicho anteriormente que la lingüística textual se encarga de estudiar el texto y el discurso. El uso coloquial de estos dos términos es indiscriminado unas veces y diferenciado otras: uno como producto de la oralidad (discurso oral) y otro como producción escrita (texto escrito); algunas lenguas poseen solo una palabra para referirse a ambos (alemán y holandés, p.ej.) e incluso en ámbitos académicos se usan indistintamente (Villegas, 1993: p.22). Por lo tanto, resulta pertinente realizar aquí una distinción entre ambos conceptos. Aunque existan diferencias entre los planteamientos de los autores seleccionados, intentaremos llegar a un consenso entre las definiciones que cada uno aporte. De acuerdo con el *Diccionario de términos clave de ELE* (Martín Peris, 1997), un texto es un producto verbal –oral o escrito –que corresponde a la unidad mínima con plenitud de sentido y un discurso hace referencia al uso de la lengua en las diversas actividades comunicativas.

De acuerdo con Ferrari (2014), el discurso se produce cuando se enuncia un texto con la intención de comunicarse: "Cuando efectuamos actos lingüísticos con propósitos comunicativos, producimos discursos"[10] (p.35). Este discurso puede explicarse desde tres dominios diferentes: el gramatical (la construcción lingüística), el contextual (factores sociales, extralingüísticos, paralingüísticos, etc.) y el textual (las formas en que el discurso toma una unidad desde el punto de vista formal o semántico). De esta manera, el texto hace referencia a la forma que da unidad al discurso, lo cual correspondería a otro nivel de análisis de este: el nivel textual (Ferrari, 2014), tal como lo es el fonético, morfológico o sintáctico. Para esta autora, la labor de la textolingüística es dar cuenta del dominio textual del discurso y de los recursos que el texto utiliza para darle significado en un todo coherente y cohesionado. Presenta, además, que existe una acepción más amplia por parte de otros autores influyentes en el campo, como son De Beaugrande y Dressler (1997), que conciben la lingüística textual, el texto y el discurso desde una perspectiva de análisis más dinámica (Ferrari, 2014: p.37).

Por otra parte, Villegas (1993) sostiene que el discurso hace referencia a las intenciones comunicativas de quien lo emite y de cómo se relaciona con las intenciones de quien lo recibe: "implica los procesos correlativos de producción y comprensión y es objeto de una hermenéutica interpretativa" (p .22), mas no de la textolingüística. Mientras que el texto hace

[10] Traducido de: "*Quando compiamo atti linguistici a scopo comunicativo, produciamo dei discorsi*" (Ferrari, 2014: p.35).

referencia al mensaje en sí, a la emisión "en cuanto producto sensible -oral o escrito- y se convierte en objeto del análisis textual" (p.22). Además, enfatiza Villegas (1993) la imposibilidad de utilizar ambos términos como sinónimos: "El discurso, en consecuencia, preexiste y trasciende al texto, en cuanto éste no es más que una de las infinitas actualizaciones posibles de aquél" (p.24). El discurso se rige por una organización de la macroestructura y el contexto, mientras que el texto está sujeto a las reglas del sistema lingüístico. El análisis de cada uno corresponde a materias diferentes, a pesar de su estrecha relación (Villegas, 1993).

Optamos, finalmente, por la concepción de discurso que plantea Olivas Zazueta desde el tratamiento computacional de la anáfora: "Una secuencia de oraciones producidas por una o más personas con la intención de transferir o intercambiar información" (2006: p.12), en donde dichas secuencias corresponden a la noción de texto a la cual nos aproximamos con los autores anteriores. Estas *secuencias de oraciones* que menciona Olivas ya las había tratado Van Dijk (1997: p.36) determinadas como enunciados complejos que conformarían un texto por sí mismos, a partir de una oración compuesta o de una serie de oraciones. Teniendo en cuenta esto, podemos afirmar que el texto corresponde a los elementos lingüísticos que conforman el discurso según unas reglas de construcción y el discurso sería la realización pragmática del texto.

Para Van Dijk (1997), las relaciones entre estas secuencias de oraciones son de tipo semántico, pues el significado de cada oración aislada corresponde a una proposición (p.38). Es así como se crean las conexiones entre secuencias, mediante las relaciones entre el significado de las proposiciones (coherencia semántica) y entre las referencias de frases (cohesión lineal). Además, agrega Van Dijk: "Las relaciones de conexión no tienen por qué ser continuadas, sino que también pueden existir proposiciones que no se sigan continuamente" (1997: p.46), por lo que el seguimiento de la cohesión y coherencia textual se hace más complejo. Inclusive, existen conexiones que pueden establecerse exclusivamente con el tema del texto o el contexto y no con alguna de las proposiciones dentro del mismo. Estas últimas se conocen como conexiones indirectas (Van Dijk, 1997: p.46). De todo esto se concluye que el seguimiento de las relaciones entre secuencias es de naturaleza compleja, pues cada oración debe ser entendida y agregada en un creciente banco de información textual y esto se logra gracias a "la claridad de los enlaces entre la oración actual y el discurso previo" (Olivas,

2012: p.12). Así como este proceso es complejo en la interacción verbal de un hablante nativo, lo es más para un hablante extranjero y mucho más para una computadora encargada de procesar el lenguaje natural. Pasemos entonces a los mecanismos que utiliza el texto para establecer estas relaciones entre secuencias y proposiciones.

4.3.3. Correferencia textual

La dimensión referencial del texto está dada por la necesidad de evocar diferentes referentes textuales y extralingüísticos que se pueden presentar en un acto discursivo. Como se dijo anteriormente, las conexiones entre estos elementos no tienen que ser continuadas y, de todos modos, puede establecerse una continuidad textual por medio de diversos recursos: "Una vez evocados, pueden ser caracterizados a partir de una propiedad o insertos en eventos particulares"[11] (Ferrari, 2014: p.179). A la relación que surge cuando dos elementos lingüísticos en un texto se refieren a un mismo *referente*[12] se le conoce como *correferencia*. Esta relación entre un referente y sus correferentes es nombrada *cadena de correferencia* (Sapena *et al*., 2013). Antes de continuar con la capacidad referencial del texto, vale la pena explorar dos términos que ya han sido mencionados en este trabajo y que constituyen la motivación para la conexión entre los elementos del texto. Estos son, desde De Beaugrande & Dressler (1997), las dos normas de textualidad más obvias: "la cohesión, que se manifiesta en la superficie textual, y la coherencia, que subyace en los mundos textuales. La cohesión y la coherencia indican de qué manera se integran y adquieren sentido los elementos que componen un texto" (p.169).

De acuerdo con De Beaugrande & Dressler (1997), la cohesión de un texto se basa en los elementos diafóricos (anáfora y catáfora) que, situados dentro del texto, remiten a elementos anteriores o posteriores del mismo "con los que son correferenciales y que constituyen la base de la *coherencia interna* del texto" (Villegas, 1993: p.37). Esta noción permite dotar de estabilidad a un texto, gracias a la continuidad de sus elementos. Siguiendo a De Beaugrande & Dressler, esto se basa en el supuesto de que existe una relación entre los elementos lingüísticos que configuran el texto: "o expresado en términos cognitivistas: cada

[11] Traducido de: "*una volta evocati, possono essere caratterizzati tramite una proprietà o inseriti in particolari eventi*" (Ferrari, 2014: p.179).
[12] Se entiende *referente* como entidad extralingüística o hecho al que se remite por medio de las palabras. No debe confundirse con el *referente* que utilizamos más adelante para denominar un antecedente de anáfora. Aquí, como en Pena (2006), utilizamos los términos *antecedente* y *referente* indistintamente, como sinónimos.

elemento lingüístico es un instrumento eficaz para acceder a otros elementos lingüísticos" (1997: p.89).

Por otra parte, la coherencia se aleja más del plano textual y apunta al ámbito discursivo. Entonces, a diferencia de la cohesión, la coherencia hace referencia a la continuidad y unidad entre significados y no a la conexión entre referentes textuales explícitamente: "La continuidad del sentido está en la base de la coherencia, entendida como la regulación de la posibilidad de que los conceptos y las relaciones que subyacen bajo la superficie textual sean accesibles entre sí e interactúen de un modo relevante" (De Beaugrande & Dressler, 1997: p.135). Villegas (1993) afirma que la coherencia no es una propiedad intrínseca de los textos: los discursos serán coherentes en la medida en que sean interpretables y depende de las condiciones definidas por los estadios de conocimiento y expectativas de los participantes (p.42). Asimismo, ubica la cohesión como componente textual y la coherencia como uno discursivo (p.25). Por lo tanto, nos interesamos por la cohesión en cuanto se relaciona directamente con la correferencia textual.

Dicho esto, se habla de correferencia textual cuando una expresión lingüística se refiere a un elemento que ha sido evocado en otro lugar al interior del mismo texto (Ferrari, 2014). Tales conexiones referenciales se pueden concebir como conexiones directas o asociativas:

a) ***La película*** [13] *está en proceso de filmación.* ***La*** *veremos en cartelera para el próximo año*.

b) *La película está en proceso de filmación.* ***Los espectadores*** *serán testigos de algo único*.

Las conexiones directas son aquellas que toman un referente específico del texto. En **a)**, el pronombre personal '*La*' remite a un elemento que ya ha sido mencionado en el texto y es identificable en la oración anterior: '*La película*'. Por otro lado, en las conexiones asociativas el referente es generado por el contexto, sea por asociación semántica del léxico o por el conocimiento del mundo (Ferrari, 2014: p.183). En el ejemplo **b)**, el sintagma nominal *Los espectadores* aparece como un nuevo referente, del que se podría decir que no hay un

[13] Son nuestras todas las negritas utilizadas tanto en los ejemplos como en las citas textuales con la intención de enfatizar, sobre todo, las expresiones anafóricas y sus antecedentes. También son de nuestra autoría estos dos ejemplos.

antecedente en el texto. Sin embargo, aunque no haya un antecedente directo, la aparición de este sintagma es posible gracias a las relaciones que se establece con el contexto y el significado de las proposiciones (Peña Martínez, 2006). *Los espectadores* hacen parte del campo semántico de *La película* y su contexto, por lo que esta relación se considera una conexión de tipo asociativa (aportando a la coherencia del texto). En este trabajo solo nos enfocamos en las relaciones entre conexiones directas, las cuales se acercan más adecuadamente a la cohesión textual. Uno de los elementos lingüísticos referenciales utilizados para responder al mecanismo de cohesión es la *anáfora*, la cual revisaremos a continuación.

4.3.4. Anáfora

Como comenta Saiz (2002), la anáfora es uno de los fenómenos que, junto con la elipsis y la deixis, responden al mecanismo de *economía lingüística* (p.2). Este mecanismo permite agilizar la comunicación natural y se basa en que el intérprete está capacitado para entender aquello a lo que se hace referencia o se está omitiendo. La capacidad referencial del lenguaje permite no solo establecer conexiones con otros referentes lingüísticos dentro del texto, sino que puede salirse de este y cubrir otras relaciones de tipo palabra-mundo. Por lo tanto, existen dos tipos de relaciones anafóricas: la *exófora*, como aquella que hace referencia al mundo externo del mensaje lingüístico y la *endófora*, que se presenta cuando la referencia está dentro del contexto lingüístico. De esta manera: "Tanto la anáfora como la catáfora se consideran categorías de endófora, la cual viene definida por su dependencia del contexto lingüístico, en oposición a la exófora, que se desarrolla en el contexto situacional" (Saiz, 2002: p.14). Dicho esto, nos ocuparemos en este texto de las relaciones endofóricas.

En ocasiones, el concepto de anáfora es entendido, de una manera más general, como todos aquellos elementos lingüísticos que establecen una relación de conexión; es decir, todas las relaciones endofóricas. Entendida así, vale la pena diferenciar entre dos tipos de relaciones endofóricas: cuando el elemento fórico se conecta con un elemento del contexto anterior (antecedente) se denomina anáfora, pero cuando lo hace con el texto siguiente se denomina *catáfora* (Martínez, 2001; Peña Martínez, 2006):

c) *Si necesitas **una**, hay **toallas** en el ropero*.[14]

[14] Ejemplo tomado de Morales (2004: p.42).

En el ejemplo anterior, a diferencia de **a)** y **b)** (dos tipos de anáfora), el elemento específico al que se hace referencia (*toallas*) aparece después de ser aludido deícticamente (*una*). Este tipo de referencias catafóricas no serán tratadas en el desarrollo de este marco ni tomadas en cuenta para el análisis; por lo cual, nos enfocaremos en definir y delimitar el campo de la anáfora, entendida como la conexión *antecedente > elemento endofórico*.

Etimológicamente, el término anáfora proviene del griego αναφορα, donde ανα- denota 'atrás' (o hacia atrás) y φορα 'llevar'; el sentido de la expresión es equivalente a 'recordar', 'repetir' (Olivas, 2006; Peña Martínez, 2006). Por consiguiente, entendemos como anáfora un elemento lingüístico que, en una relación endofórica, cumple la función de referirse a otro elemento previamente enunciado. Los diferentes mecanismos anafóricos para establecer correferencias dan lugar a diversas clasificaciones sobre la anáfora. Por ejemplo, vimos anteriormente, en **a)** y **b)**, la diferencia entre conexiones directas y conexiones asociativas. De dicha diferenciación surge también la anáfora directa y la indirecta (asociativa), aunque, como mencionamos ya, nos centramos solamente en el tratamiento de las relaciones directas[15]. Pasemos entonces a los demás tipos de clasificación sobre la anáfora.

4.3.4.1. Tipos de anáfora

Ahora bien, dentro de la anáfora pueden encontrarse diferentes tipologías, las cuales varían según la perspectiva que se tome: morfosintáctica, semántica o pragmática, por ejemplo. Para este trabajo se han seleccionado clasificaciones de diferentes autores, las cuales se complementan en ciertos puntos, como veremos más adelante. La selección de dichos autores (Saiz, 2002; Olivas, 2006; Ferrari, 2014) se debe a que sus criterios de clasificación corresponden con los objetivos de análisis que aquí planteamos y con las nociones de texto y correferencia textual que expusimos en los apartados anteriores.[16]

Por ejemplo, Olivas propone, primeramente, una clasificación desde "el marco en que sucede la anáfora" (2002: p.21). Esto es: si la relación anafórica se establece dentro de la

[15] En primera instancia, tomamos este tipo de decisiones con el fin de delimitar el objeto de estudio de este trabajo. Sin embargo, puesto que el objetivo de este es evaluar la efectividad de los analizadores del lenguaje, descartamos de antemano aquellos elementos que precisan de conocimiento del mundo o que implican otros procesos de análisis para la máquina y, por lo tanto, para nuestro análisis. Sobre la anáfora indirecta puede consultarse la tesis doctoral de Morales Carrasco (2004).

[16] Descartamos, por ejemplo, tipologías que obedecen a criterios pragmáticos como la que propone Peña Martínez (2006) para el análisis de las marcas anafóricas: "en la que determinados procesos cognitivos de carácter pragmático permiten la asignación de un referente" (Peña Martínez: p.50).

oración, es *intraoracional* y si la conexión entre elementos se da fuera de una sola oración, es *interoracional*. Lo anterior coincide con la visión que ya expusimos de texto y de cohesión textual, pues "actúa como un instrumento lingüístico el cual ayuda a mantener el discurso como una unidad de sentido, debido a la creación de relaciones de unión entre las distintas partes del texto" (Olivas, 2002: p.21). Dicho esto, pasemos a revisar otros criterios de clasificación propuestos por los autores y sus relaciones entre sí. Mencionamos, además, que estos criterios están basados en la clasificación de anáfora de Ruslan Mitkov (2002).

Criterios basados en el antecedente

Algunos autores sugieren analizar la función del antecedente como criterio de clasificación de la anáfora. Uno de los tres criterios que propone Saiz (2002) se basa en la categoría gramatical del antecedente. Es decir, a qué tipo de elementos gramaticales se puede hacer referencia con una anáfora dentro de un texto; por lo cual se presentan las siguientes categorías:

- Sintagma nominal: En el que el núcleo del antecedente es un nombre común o propio.
- Sintagma verbal: El núcleo es un verbo; para referirse a este se debe utilizar también un verbo (2002: p.15).
- Sintagma adverbial: El antecedente anafórico está representado por un adverbio, como en el siguiente ejemplo: "María está ***arriba***. *Allí* se trabaja mejor" (Saiz, 2002: p.16).
- Oración completa: Donde la anáfora hace alusión a un hecho o idea mencionados anteriormente en una oración: "***María está embarazada***. Su marido no *lo* sabe" (2002: p.16).

Es similar, en este sentido, el segundo criterio expuesto por Olivas (2006): "en función de la accesibilidad del antecedente" (p.22). Es decir, la anáfora puede clasificarse también según la facilidad con la que se puede hallar la conexión correferencial, o bien, la dificultad para establecer el antecedente de una expresión anafórica. Este autor plantea tres niveles de accesibilidad de la anáfora, de mayor a menor accesibilidad:

- *Anáfora morfosintáctica*

La de máxima accesibilidad, pues se hace referencia a un antecedente que corresponde con un sintagma nominal, sintagma verbal o una oración y, comúnmente, la expresión anafórica es un pronombre.

- *Anáfora semántica*

De una accesibilidad media, esta anáfora se establece con su antecedente por medio de relaciones de significado: sea por sinonimia, hiperonimia o por contexto, pues el antecedente puede ser una entidad o idea expuesta anteriormente. Esta última concuerda con la conexión asociativa mencionada en este trabajo, pues se requiere de información semántica para establecer dicha relación.

- *Anáfora pragmática*

La de menor accesibilidad, puesto que el hallazgo de su antecedente requiere de conocimiento enciclopédico y del mundo.

Criterios basados en la relación entre anáfora y antecedente

Con respecto a este criterio, Saiz (2002) distingue dos tipos básicos de anáfora, que se basan en la relación que anáfora y antecedente poseen con un referente, entendido este como entidad del mundo, hecho o idea al que remite el antecedente (p.15):

- *Anáfora de referencia (profunda)*

En este tipo de anáfora se comparte el mismo referente. Por ejemplo, en **d)** el cigarrillo que se prende es el mismo que se apaga; la correferencia se da también con el referente.

d) *El policía prendió* ***un cigarrillo****, pero su compañero se* ***lo*** *apagó.*

- *Anáfora de sentido (superficial)*

Aunque los elementos que sostienen la correferencia comparten el mismo significado, no comparten el mismo referente. Así, en **e)** existe correferencia entre el antecedente y el elemento anafórico, pero se trata de dos referentes distintos:

e) *El policía había perdido* ***su encendedor*** *y a su compañero se* ***lo*** *habían robado*.

También Olivas (2006) hace esta distinción entre anáforas profundas y superficiales. No obstante, la anáfora superficial se entiende como una referencia parcial, en la que no existe correferencia, y que puede introducir nuevos elementos como vimos con la anáfora asociativa (Ferrari, 2014).

Por otra parte, Ferrari (2014), atendiendo también a la relación entre antecedente y anáfora, presenta tres maneras en que estas se pueden dar: por repetición, por sustitución y por continuidad de significado. La anáfora por repetición se da cuando hay coincidencia lexical entre el antecedente y la expresión anafórica y puede ser exhaustiva o parcial (p.186):

f) *Hoy en día,* ***el internet*** *es una gran herramienta. Sin embargo, existen muchos problemas sociales relacionados con* ***el internet****.*

g) **El Ingenioso Hidalgo Don Quijote de la Mancha** *es un libro del escritor Miguel de Cervantes.* […] **Don Quijote de la Mancha** *fue publicado en 1615*.[17]

En ambos ejemplos hay una relación anafórica por repetición. Es exhaustiva en **f)** porque se utiliza exactamente la misma expresión y parcial en **g)** puesto que se repite solo una parte del antecedente para hacer referencia a este.

En cuanto a la forma de relacionarse por continuidad de significado, no es más que la anáfora indirecta o asociativa que ya hemos visto, que hace referencia a la generación de nuevos referentes a partir de campos semánticos y otras relaciones semánticas (Ferrari, 2014: p.194). Por último, dentro de este criterio de clasificación, Ferrari explica la relación por sustitución, la cual se da cuando "la anáfora indica el referente designado para el antecedente con una expresión lingüística diferente de la anterior"[18] (p.187). Lo cual abre la posibilidad de una amplia categorización de la anáfora según las diversas categorías y mecanismos lingüísticos que puede tomar la expresión anafórica, como veremos en el siguiente apartado.

Criterios basados en la expresión anafórica

Este último criterio de clasificación es tomado en cuenta por los tres autores que seleccionamos y es trabajado entre ellos con algunas diferencias. Consiste, la mayoría de las veces, en

[17] Los últimos cuatro ejemplos son de nuestra autoría.

[18] Traducido de: "*l'anafora indica il referente designato dall'antecedente con un'espressione linguistica diversa dalla precedente*" (Ferrari, 2014: p.187).

agrupar las anáforas de acuerdo con la categoría gramatical a la que esta se asocie. A continuación, observaremos la clasificación de la anáfora según los distintos elementos lingüísticos que pueden constituir una expresión anafórica para estos autores.

- *Anáfora pronominal*

Para Saiz (2002), esta es la más frecuente de las expresiones anafóricas y responde a todos los tipos de pronombres establecidos. Es decir, una anáfora puede ser: un pronombre personal (tanto las formas de sujeto: *yo, tú, nosotros, ellos*, etc., como las formas tónicas y átonas de estos: *mí, me, ti, te, nos*, etc.), un pronombre omitido (sujeto implícito[19] en el verbo, muy común en lenguas como el español o el italiano), un pronombre demostrativo (*este, aquel, esa*, entre otros). O bien, pueden ser también expresiones anafóricas los pronombres posesivos (*su, mi, suyo, mío*, etc.), pronombres reflexivos –los cuales correfieren por definición con el sujeto del verbo del que depende (Saiz, 2002) –pronombres relativos (expresados por *que* como pronombre) y el tipo *one anaphora* que se da "exclusivamente en el caso del inglés" (p.22).

Olivas (2006) agrega a este tipo de anáfora el hecho de que permiten mayor accesibilidad al antecedente, pues se suelen presentar a una distancia corta de este ya que poseen poca carga léxica (p.31). Por otra parte, Ferrari (2014) considera que la anáfora pronominal no constituye el principal número de casos de anáfora y añade a este tipo, además de los ya mencionados por Saiz (2002), los pronombres indefinidos con ciertos usos, como *ambos, el uno y el otro* o el pronombre *todos* cuando se refiere a la totalidad de elementos que denotan el antecedente (Ferrari, 2014: p.190). Así mismo, toma en cuenta también la anáfora de sujeto implícito, o tácito, pero no dentro de la pronominal, sino como una categoría aparte.

- *Anáfora nominal*

Esta anáfora también es llamada *descripción definida* (Saiz, 2002; Olivas, 2006) y consiste en que el antecedente puede ser evocado o sustituido por un sintagma nominal. Saiz la subdivide según la formación de dicho sintagma: si se forma con un artículo definido (determinado: p. ej., ***El*** *obrero*), con un determinante demostrativo (***Este*** *obrero*) o si el sintagma

[19] Aunque este tipo de anáfora podría considerarse también como un tipo de elipsis, se opta por trabajarla como un tipo de anáfora pronominal "bajo la suposición de que el sujeto elidido es un pronombre que concuerda morfológicamente con el verbo que acompaña", también se conoce como anáfora cero (Saiz, 2002: p.20).

se forma con determinante posesivo: "La casa de ***María*** es enorme. ***Su salón*** tiene 30 metros cuadrados" (Saiz, 2002: p.24). Olivas (2006), en cambio, entiende la anáfora nominal dentro de la categoría que se da por repetición, a diferencia de Ferrari que solo la admite en la sustitución (2014).

Según Ferrari (2014), el principal tipo de sustitución se da por medio de la anáfora nominal, la cual considera más frecuente. La autora también realiza una subdivisión de esta, pero atendiendo, además, a criterios de tipo semántico. Concibe primero, la sustitución nominal a través de una relación de sinonimia, en la que la anáfora nominal la constituye un sinónimo del antecedente. Esta relación también puede existir únicamente dentro del texto mismo: la correferencia solo tiene sentido en el mundo evocado del texto, lo que se conoce como *sinónimos contextuales* (p.187). Finalmente, también puede darse esta anáfora a través del uso de hiperónimos e hipónimos, de nombres propios, perífrasis o epítetos.

- *Anáfora verbal*

Este caso se presenta, en español, cuando la forma pronominal "lo" se refiere a un verbo o a un sintagma verbal (Saiz, 2002). Este elemento lingüístico va acompañado siempre de un verbo auxiliar o *pro-verbo*, el cual, en la mayoría de los casos, es el verbo "hacer" (es decir: 'hacerlo'): "Este verbo anafórico no proporciona rasgos semánticos específicos, por lo que la aplicación de esta fuente de información no resulta especialmente útil para su resolución" (Saiz, 2002: p.26). A lo anterior, podemos añadir, desde Ferrari (2014), que la anáfora verbal (además de la estructura antes mencionada de pronombre y pro-verbo) también puede realizarse por medio de otros elementos:

h) *Debería castigarte, pero no* ***lo haré*** / *pero no* ***actuaré así*** / *pero no* ***tendré tal comportamiento***.[20]

- *Anáfora adverbial*

Saiz (2002) divide la anáfora adverbial en dos tipos: temporales y locativas, dependiendo de si el antecedente describe una circunstancia temporal o espacial (p.26). Los elementos lingüísticos utilizados como anáfora adverbial pueden ser muy variados: *Allí, ahí,*

[20] Ejemplo tomado y traducido de Ferrari: "*Dovrei punirti, ma non lo farò così/ ma non agirò così/ non terrò tale comportamento*" (2014: p.192).

aquí, *entonces*, *cuando*, entre otros. Olivas (2006) acota que estas anáforas encuentran siempre su antecedente en la referencia local o temporal más cercana, más reciente, en el texto. Además, añade, que los complementos circunstanciales también pueden formar parte de esta relación anafórica (p.37). En este tipo de anáfora, la información semántica contenida por la expresión anafórica es muy general y esto dificulta su resolución. A diferencia de estos dos autores, aunque también contempla los adverbios como posibles anáforas, Ferrari (2014: p.190) sitúa estos elementos dentro de la anáfora pronominal y los nombra *avverbi pronominali* (adverbios pronominales).

- *Anáfora adjetival y superficial numérica*

De la anáfora tipo *one*, que Saiz (2002) ubicaba dentro de la pronominal y limitaba al inglés, Olivas (2006) desprende para el español dos tipos de anáfora: adjetival y superficial numérica. En el siguiente ejemplo, la expresión anafórica *one* permite elidir elementos del antecedente e introducir un nuevo referente por medio de la correferencia: "*Wendy didn't give either boy* ***a green tie-dyed T-shirt****, but she gave Sue* ***a red one***" (p.34). De esta manera, en español, se denominaría a esta anáfora de tipo adjetivo, pues al sintagma nominal que introduce la expresión anafórica le hace falta el núcleo nominal y el adjetivo cumple su función: "Wendy no le dio a ningún niño **una camisa verde de manga corta**, pero a Sue le dio **una roja**" (p.34).

Del mismo modo, con una estructura similar, la anáfora superficial numérica se especializa (para distinguirla de la anáfora adjetival) en los ordinales, numerales y todas las expresiones anafóricas que signifiquen cierto grado de orden. En otras palabras: "alude al orden establecido por sus antecedentes" (Saiz, 2002: p.25), esto es, *el primero*, *el último*, *el tercero*, etc., lo cual aporta "información que será especialmente útil para determinar su antecedente" (Olivas, 2006: p.36). Saiz (2002) también admite entre este tipo de anáfora algunos pronombres distributivos, en ciertos casos: *Los unos* […] *los otros*; *Estos* […] *aquellos*, en los que se alude a un orden de aparición de varios antecedentes, igual que Ferrari (2014) aludía a estos simplemente como pronombres indefinidos.

4.4. Resolución y desambiguación anafórica

En el PLN la desambiguación hace referencia al proceso que debe realizar la computadora para resolver problemas de ambigüedad en el análisis de una lengua natural: "En un cierto momento, el sistema de procesamiento del lenguaje se encuentra con un fragmento de información para el cual tiene múltiples interpretaciones y debe decidir cuál es la interpretación correcta en el contexto preciso donde se ha producido" (Márquez, 2002: p.136). Es, además, nuestro trabajo como investigadores y lingüistas precisar la información necesaria para que la máquina resuelva por sus propios medios estas ambigüedades.

4.4.1. Ambigüedad lingüística

Un enunciado ambiguo es aquel que es susceptible de dos o más interpretaciones (Alcaraz & Martínez, 1997). Similar concepción poseen López (2008), Anjali & Babu (2014) y Peña (2016), para quienes un mensaje ambiguo (fonema, morfema, palabra, oración, enunciado, etc.) puede definirse como la secuencia de signos a la que es posible asignar más de un significado y de la cual surgen dos o más interpretaciones distintas. La ambigüedad es particularidad propia de todas las lenguas naturales (Peña, 2016: p.41).

A pesar de que los autores que trabajan el fenómeno coinciden generalmente en una definición de ambigüedad, al momento de realizar una tipología de esta difieren bastante entre ellos. Esto se debe a que la ambigüedad lingüística se da desde los estadios más pequeños de la lengua, lo cual genera ejemplos numerosos, pero, sobre todo, bastante diversos, complicando así una generalización y una formalización de esta. López (2008: p.28) señala varios casos que permiten la aparición de la ambigüedad en la comunicación humana. Las relaciones que establecen los signos entre sí son tomadas en cuenta inconscientemente por el hablante y hay algunas que no pueden ignorarse al momento de procesar el lenguaje natural.

4.4.1.1. Tipología de la ambigüedad

López concibe cuatro tipos de ambigüedad lingüística (2008: p.32):

- **Fónica**: Se da en el habla y desaparece en la manifestación escrita: *gente de mente*.
- **Sintagmático-sintáctica**: Sintagmas que pueden contraer más de una función sintáctica: *la perra de tu hermana*.

- **Sintagmático-semántica**: El enunciado es analizable desde dos relaciones semánticas entre sintagmas: *la condena del juez* (él realiza la condena o él es condenado).
- **Léxica**: Homonimia: *Todos sabemos que a George le gusta la **heroína***. Este último caso se resuelve por la comprensión del contexto.

Por otra parte, Anjali & Babu (2014) conciben su tipología desde los problemas que la ambigüedad puede producir en el PLN (p.392):

- **Léxica**: Ambigüedad que corresponde solamente a una palabra y a sus connotaciones, según pueda funcionar como verbo, adjetivo o sustantivo.
- **Sintáctica**: Ambigüedad de *scope* (operadores y cuantificadores): en 'todo hombre ama a una mujer' se generan dos sentidos: a) por cada hombre hay una mujer para amar, b) todos los hombres aman a la misma mujer. Ambigüedad de *attatchment*: sintagma que puede cumplir más de una función en un árbol de análisis sintáctico.
- **Semántica**: Las ambigüedades semánticas en PLN nacen del hecho de que una computadora no está en condiciones de distinguir lo que es lógico de lo que no lo es en una lengua natural. A partir de este nivel de la lengua empieza a complicarse más el procesamiento ante ambigüedad.
- **Discursiva**: Se necesita de un conocimiento compartido del mundo y la interpretación se lleva a cabo utilizando este contexto. Introducen aquí la **ambigüedad anafórica** pues pertenece al nivel del discurso (Anjali & Babu, 2014: p.393).
- **Pragmática**: La más complicada para el PLN ya que involucra el procesamiento de las ideas, la intención del hablante, las creencias, los modales, etc.

4.4.2. Métodos de desambiguación

Uno de los métodos más utilizados de desambiguación se ejecuta en el *Part of Speech Tagging* (*PoS-tagging*); el cual se utiliza para identificar la correcta categoría morfosintáctica de las palabras. Lo que denominamos ambigüedad léxica, según Anjali & Babu (2014) "puede resolverse por desambiguación de la categoría gramatical, p. ej., *parts-of-speech tagging*"[21] (p.392). Otros métodos utilizados son el *Word Sense Disambiguation* (*WSD*) para

[21] Traducido de: "*Can be resolved by Lexical category disambiguation i.e, parts-of-speech tagging*" (Anjali & Babu, 2014: p.392).

la identificación y selección del sentido de una palabra polisémica o el *PP- attatchment disambiguation*: para la resolución de la ambigüedad estructural o sintáctica introducida por los sintagmas preposicionales.

Para cada nivel de ambigüedad en la lengua se requiere un método de desambiguación diferente (Moreno, 1998: p.33). Como dijimos, en el nivel discursivo se hace aún más complicado el desarrollo de sistemas para la identificación correcta en casos ambiguos. La desambiguación en niveles superiores al análisis oracional resulta muy costosa para una máquina, en términos computacionales, y requieren de la utilización de un volumen desproporcionado de conocimiento, afirma Rodríguez Hontoria (2003). Este autor desarrolla un ejemplo con los problemas que las *interfaces de lenguaje natural* deben superar, entre ellos la ambigüedad. El procesamiento en los niveles más básicos de la lengua ha avanzado en gran medida, pero lograr avances en un *nivel ilocutivo* resulta más complejo, ya que "está a caballo entre lo que es lingüístico y lo que es cognitivo" (Rodríguez, 2003: p.139). La mayoría de las técnicas de PLN para resolver estos problemas de ambigüedad son inaplicables en este nivel de sentido de la lengua: tanto el *PoS-tagging*, el *WSD*, el establecimiento de hipótesis para palabras desconocidas, como el análisis sintáctico se quedan cortos ante casos de ambigüedad discursiva, anafórica o pragmática, aunque pueden ser útiles para su resolución.

La dificultad del tratamiento de la ambigüedad radica en su relación con factores externos, con el conocimiento del mundo (Badia, 2003: p.247). El proceso de desambiguación no es un fin en sí mismo, sino que debe tomarse como una tarea intermedia en el PLN. La desambiguación es transversal a todos los procesos de lenguaje natural, pues la ambigüedad de las lenguas naturales impide un tratamiento computacional completo de estas. Por lo tanto, el mejoramiento de este proceso favorecerá las diferentes tareas de PLN, por ejemplo, la identificación de correferentes, de relaciones anafóricas o de una progresión temática (tema y rema), relacionadas con este proyecto; tanto de otras tareas como las interacciones humano-máquina, interfaces de lenguaje natural o el desarrollo de Inteligencia Artificial, en un ámbito más general. La ambigüedad y la imprecisión son algunos de los principales problemas para formalizar el lenguaje: "La posibilidad de procesar el lenguaje depende del grado de formalización que podamos alcanzar en su descripción" (Badia, 2003: p.200).

4.4.3. El caso de la anáfora

Basados en la definición y tipología de la anáfora vistas en apartados anteriores y en la categorización de la ambigüedad, veamos ahora lo que se entiende por ambigüedad anafórica y resolución de la anáfora.

i) ***El caballo*** *arrolló al* ***perro****, pues* ***este*** *no lo vio.*

En el ejemplo **i)** existe una ambigüedad en cuanto al pronombre demostrativo *este*, pues no queda claro si hace referencia a la entidad introducida por *caballo* o por *perro*. La ambigüedad anafórica (Anjiali & Babu, 2014) es nombrada también por Saiz (2002), cuando introduce el fenómeno de la anáfora como: "elemento fundamental de la cohesión entre oraciones y marcador de la coherencia del texto, se enmarca en el ámbito de la ambigüedad referencial" (p.2).

La dificultad de la desambiguación anafórica radica en que muchas veces se debe acudir a factores externos a la oración, que, a su vez, se relacionan con otros elementos anafóricamente (cadenas de correferencia) y, otras veces, con elementos fuera del discurso. Debemos tener en cuenta también que no siempre los casos de ambigüedad son los mismos que se presentan en la comunicación humana, como en el ejemplo anterior. Algunos que son comprensibles y solucionables para nosotros sin mayor esfuerzo, no lo son para el ordenador:

j) ***Susana*** *estaba sentada en* ***la silla****.* **Su** ***hermana*** le *rompió las patas a* esta *y, por culpa de* ello*, cayó de espaldas*.

En **j)** podemos identificar varios referentes que se introducen en el texto a los que luego se hace referencia por medio de una anáfora. El primero de ellos es ***Susana***, el segundo es ***la silla*** y el tercero ***su hermana***. La interpretación de esta secuencia de oraciones no representaría mayores dificultades para un hablante nativo competente en la lengua. En cambio, en el análisis por ordenador, el establecimiento de las relaciones anafóricas presenta problemas de ambigüedad si solo se vale de información morfosintáctica para su desambiguación. Es decir, la primera conexión interoracional que se presenta obvia para nosotros (entre el pronombre posesivo de tercera persona *su* y el sintagma nominal *Susana*) podría representar un caso de ambigüedad, ya que el pronombre *su* también concuerda morfosintácticamente con *la silla* (p. ej., en vez de ***Su*** *hermana*: ***su*** *espaldar*) en cuanto a género y número.

Para resolver las ambigüedades de **j)** se requieren más datos que solamente morfológicos y sintácticos. Por ejemplo, en la correferencia anterior se necesita de información semántica, pues se sabe que la categoría *hermana* normalmente no aplica para *silla* y en este contexto se establece con ***Susana***.[22] Tenemos también la anáfora introducida por los pronombres ***le*** y *a* ***esta***, en donde el antecedente de estos podría ser de nuevo ***Susana*** o ***la silla***. Aquí el contexto es el que nos permite desambiguar, pues el verbo *romper* y su complemento *las patas* proporcionan la información semántica necesaria para establecer la correferencia de los pronombres con ***la silla***, gracias a que *patas* se relaciona metonímicamente con *silla*. Sin embargo, en otros contextos *patas* también podría ser parte de *Susana*, por lo que la información que nos permite como hablantes interpretar correctamente la secuencia también es pragmática (Van Dijk, 1997). Además de los otros casos de ambigüedad que puedan presentarse en **j)**, podemos hablar también de otra anáfora producida por el pronombre personal neutro *ello*, que se refiere al acto de romperle las patas a la silla, aunque su relación se resuelve por proximidad (el acto *estaba sentada en la silla* está más alejado que *le rompió las patas*). Este último ayuda a establecer el correferente del último verbo con pronombre implícito *cayó*, pues se presenta como consecuencia (*por culpa de* ***ello***) de romperle ***las patas*** a ***la silla*** en la que estaba sentada ***Susana***.

La Resolución de la Anáfora –del inglés *Anaphora Resolution* (Mitkov, 2002) –consiste entonces en localizar el antecedente que introdujo una nueva entidad en el texto para colocarlo en el lugar del material anafórico que se refiere a él y así completar el contenido de los enunciados: "A cada entidad generada se le asigna un nombre, una constante a la que se puede hacer referencia en el proceso posterior" (Moreno, 1998: p.124). Lo anterior hace referencia a la necesidad de etiquetar estas nociones para su procesamiento e identificación: "A medida que se procesan las oraciones, se va elaborando una lista en la que figuran las entidades del discurso" (Moreno, 1998: p.124). Como lo hemos descrito, este proceso es bastante complejo y depende de muchos factores que lo constituyen todavía como un reto para el PLN.

[22] Para más información sobre estas relaciones semánticas introducidas por los pronombres posesivos, *véase* Van Dijk (1997: p.46).

5. Metodología de investigación

Antes de pasar a describir los procesos y bases metodológicas que utilizamos para el desarrollo de este trabajo, definiremos el enfoque de investigación y el alcance bajo los cuales este ha sido planteado. En primera instancia, nuestro trabajo se enmarca en un enfoque metodológico de tipo mixto y un alcance de carácter exploratorio-descriptivo (Hernández et al., 2014). Lo primero se debe a que, por una parte, empleamos datos cuantitativos con el fin de caracterizar la anáfora en el corpus y de evaluar los errores en su tratamiento y, por otra parte, analizamos de manera cualitativa los casos problemáticos para establecer hipótesis sobre la dificultad de su tratamiento computacional. En este sentido, el alcance de nuestro trabajo tiende a ser, en primer lugar, exploratorio, pues basados en la lingüística de corpus como metodología de trabajo buscamos indagar acerca del comportamiento de la anáfora dentro de un corpus y los problemas que implica su procesamiento por intermedio de un software especializado para este fin. En segundo lugar, es descriptivo puesto que describimos cuáles son dichos problemas y analizamos, con base en la teoría lingüística, las causas que provocan este fenómeno.

En la sección que se encuentra a continuación, definimos las bases metodológicas que nos aportó la lingüística de corpus para nuestro trabajo y, posteriormente, describimos todo el proceso que seguimos para llevarlo a cabo.

5.1. Lingüística de corpus como herramienta metodológica

Como lo hemos mencionado anteriormente, la lingüística de corpus resulta de gran utilidad en el área de la enseñanza de lenguas y en el desarrollo de sistemas de ELAO y ALAO. Para esta investigación, además, la lingüística de corpus es esencial como base metodológica, a pesar de la controversia sobre si su naturaleza es teórica o metodológica (Tognini-Bonelli, 2001; Parodi, 2010; McEnery & Hardie, 2012). De esta manera, aceptamos la concepción de Giovanni Parodi (2010) sobre este campo como “un método de investigación que puede ser empleado en todas las ramas o áreas de la lingüística y desde enfoques teóricos diferentes” (p.15). Lo relevante de este método empírico es el uso de datos observables a modo de evidencia científica, es decir, de material auténtico de la lengua que permite estudiar una reali-

dad lingüística. Por otra parte, añade el propio Parodi (2010), dichos datos deben estar almacenados como corpus electrónicos para que puedan ser explotados, más adelante, por computadora.

De igual manera, Bernal e Hincapié (2018) muestran a la lingüística de corpus como metodología, la cual se encarga de sistematizar y analizar conjuntos de datos textuales ordenados con criterios lingüísticos, sociales, culturales o literarios. Estos autores presentan también las bases bajo las cuales puede aplicarse la lingüística de corpus: principalmente, esta debe centrarse en la lengua en uso como principal insumo, esto es, corpus conformados por muestras reales de una determinada lengua. Por otra parte, se basa en el análisis sistemático de la lengua sustentado por un conjunto de reglas de recolección, almacenamiento y anotación y por la posibilidad de trabajar desde un enfoque cualitativo o cuantitativo (Bernal & Hincapié, 2018: p.11).

Es bien sabido, cuando se habla de lingüística de corpus, que suele hacerse una distinción según el enfoque que se sigue: *corpus-based* o *corpus-driven*. Comúnmente, esta dicotomía se encuentra apoyada y repetida entre los estudiosos del campo, como veremos en la siguiente tabla que resume las principales características de cada enfoque según Tognini-Bonelli (2001) y Parodi (2010):

Tabla 1. *Enfoques de la lingüística de corpus basados en Tognini-Bonelli (2001) y Parodi (2010).*

Corpus-based	***Corpus-driven***
Las categorías de análisis están previamente determinadas y enmarcadas en una opción teórica.	Las categorías emergen del análisis y dan sustento a la construcción de una teoría guiada por los datos.
El corpus y las herramientas actúan como un método de investigación, indagación y corroboración de ideas preexistentes.	El corpus dará lugar a un nuevo conocimiento y es parte integral de la investigación.
Pone a prueba o ejemplifica teorías ya formuladas.	La teoría no existe de manera independiente de la evidencia.
Buscan sustento en grandes muestras de textos disponibles digitalmente.	"La observación conduce a las hipótesis, las que conducen a las generalizaciones, las que conducen a la unificación de una afirmación teórica" (Parodi, 2010: p.47).

Según esto, en esta investigación específicamente podrían aplicarse las características del enfoque *basado en corpus*, puesto que trata de identificar los aspectos teóricos sobre la anáfora, explicados anteriormente, en un corpus dado; pero, además, busca también observar los problemas de su tratamiento computacional para categorizarlos como nueva información

que solo el corpus puede ofrecer, es decir, *guiado por el corpus*. Por lo tanto, siendo consecuentes con la concepción adoptada sobre la lingüística de corpus como metodología, aceptamos la posición de McEnery y Hardie (2012) al negarle una dimensión exclusivamente teórica al corpus, la cual rechaza la distinción binaria entre *corpus-based* y *corpus-driven*: "Desde este punto de vista, toda la lingüística de corpus solo puede ser descrita como *corpus-based*."[23] (McEnery & Hardie, 2012: p.6).

Dicho lo anterior, para trabajar a través de la lingüística de corpus, debemos aclarar a continuación los criterios que se adoptan en esta investigación acerca de la noción de corpus. Partamos de las palabras con las que Parodi (2010) resume al máximo el término: "Un corpus es una colección o conjunto de textos que está formado por al menos dos o más textos" (p.25). La razón para utilizar aquí esta definición minimalista es que nos permite entender que un corpus útil para la investigación no está limitado por restricciones generales sobre su constitución, sino que a cada caso particular se adapta un tipo de corpus con diferentes características.

Así las cosas, destacamos en primera instancia la naturaleza de la información lingüística que debe conformar un corpus: original y completa, con unidad de sentido y con un propósito comunicativo específico. Además, esta información debe ser representativa de la lengua, de este modo, los resultados generados podrán ser aplicados y generalizados a la misma (Parodi, 2010). En este caso, la representatividad de un corpus no apunta hacia la cantidad de textos o palabras que pueda contener este, sino más hacia la variedad y las diferentes manifestaciones de la lengua que pueda incorporar en los textos que recoge. La extensión se decide según los objetivos del corpus (Bernal & Hincapié, 2018). Otros autores agregan también que sean legibles por computadora y que sean escogidos por muestreo (Bolaños, 2015). Por todas estas razones, acordamos con las características para definir un corpus de Bernal e Hincapié (2018):

> Un corpus debe constituirse como una muestra de lengua real con diferentes posibilidades de composición que relaciona la teoría y los datos, brinda información adicional a la explícita en los textos, facilita la extracción de datos homogéneos y cuantificables, no se rige por un tamaño estándar establecido, es representativo y diverso, tiende al equilibrio, es digital y de fácil acceso. Estas características hacen de los corpus fuentes de datos aptas para investigaciones lingüísticas (p.24).

[23] Traducido de: "*From this point of view, all corpus linguistics can justly be described as corpus-based.*" (McEnery & Hardie, 2012: p.6).

5.2. Desarrollo metodológico

5.2.1. Descripción del corpus DICEELE

El corpus DICEELE es la base textual con la que se desarrollarán las actividades de lingüística para el proyecto DICEELE, el cual apunta hacia el ALAO y la ELAO para aprendientes de ELE. El corpus está constituido por 150 textos cortos clasificados en los niveles B1, B2 y C1 del Marco Común Europeo de Referencia (MCER) para la lengua española (50 textos por cada nivel). La recopilación de los textos estuvo guiada por ciertos criterios específicos, teniendo en cuenta que se trata de un corpus especializado para la enseñanza del ELE en el cual se trabajarán actividades de lingüística textual. Se definió que las características que debían tener los textos serían las siguientes:

Textos auténticos: Además de seguir los lineamientos de la lingüística de corpus sobre textos originales, el corpus DICEELE se rige por brindar material para la enseñanza y el aprendizaje de la lengua que represente el uso real de esta. No utiliza, por lo tanto, textos adaptados, didactizados o diseñados por profesores de ELE con intención explícita de trabajar algún aspecto lingüístico.

Textos clasificados según el MCER dentro de los niveles B1, B2 y C1: Es necesario conocer el nivel de dificultad que representa un texto que va dirigido a un aprendiente. La selección de los niveles B1, B2 y C1 estuvo sujeta a criterios de pertinencia, pues a partir del nivel umbral (B1) los textos dejan de estar conectados directamente con situaciones inmediatas y campos semánticos desconectados (familia, ocupaciones, etc.) y comienzan a hacer referencia a cuestiones abstractas y a discursos relativamente elaborados. Se procuró buscar textos que ya hubiesen pasado por un proceso de clasificación y que fueran utilizados para el aprendizaje del ELE en instituciones o portales en línea para la enseñanza.[24]

Textos cortos de diferentes géneros: Con el fin de privilegiar la variedad y la representatividad y, así, favorecer el desarrollo de las actividades con los fenómenos lingüísticos, se decidió que hubiese diversidad en los géneros discursivos y en los tipos de texto, además de una consistencia en el número de estos por género y en su extensión. En otras palabras,

[24] Sin embargo, también se utilizaron textos de otras fuentes que no poseían clasificación alguna, por lo que esta tuvo que realizarse con base en las especificaciones del *Plan curricular del Instituto Cervantes* (Instituto Cervantes, 2007) y con la ayuda de la aplicación web de Dimitris Lamprinos: *Programa informático para calcular el nivel de dificultad de textos en español*, disponible en: http://www.ajugar.gr/es/clasificar.

los textos debían ser breves (entre 300 y 1000 palabras) y cada nivel del MCER trabajado debía poseer textos informativos, narrativos, argumentativos y explicativos en igual medida. A su vez, dentro de cada género también debía existir variedad, es decir, los textos narrativos del nivel C1, por ejemplo, no debían ser exclusivamente cuentos sino también fábulas, mitos, relatos, etc. Todo esto con base en los planteamientos del MCER sobre tipos de texto y discurso (Consejo de Europa, 2002: p.91).

Textos de distintas fuentes: Siguiendo los lineamientos de variedad y representatividad, se buscó que los textos no fueran extraídos únicamente de una misma página web o de un solo libro de texto, por ejemplo. Así mismo, se recolectó material de distintos países hispanohablantes con el fin de no centralizar una sola variedad del español.

Debemos también mencionar que este corpus se encuentra almacenado en formato .txt, .docx y, finalmente, en formato .xml.[25] Dentro de este último, en los documentos XML, se encuentra consignada toda la información sobre las características de cada texto, incluidas las que acabamos de definir arriba. A esta sección se le denomina *metadatos*: "los metadatos, es decir, la información que acompaña a cada una de las expresiones o los textos incluidos en el corpus, ayuda fundamentalmente a contextualizar aspectos relevantes de dicho corpus" (Bolaños, 2015: p.36). En ellos puede asignarse la información sobre el contexto que rodea al texto o características sobre este, por ejemplo, la fecha de producción y publicación, el lugar de origen, el autor, la cantidad de palabras, etc. Estos datos son útiles al momento del análisis puesto que facilitan la posibilidad de identificar qué tanto influyen las variables tales como nivel de dificultad, tipo de texto, país, entre otras, en el tratamiento de un fenómeno lingüístico específico, como es la anáfora en este caso. En la siguiente figura se ilustra la manera en que estos datos fueron asignados entre las etiquetas <metadata></metadata>:

[25] XML (siglas de *eXtensible Markup Languague*) hace referencia a un tipo de lenguaje de marcado o metalenguaje (similar a HTML) que se utiliza para etiquetar y organizar documentos con el fin de posibilitar nuevos usos informáticos a partir de la información que se le agregue.

Figura 1. Metadatos de un documento del corpus etiquetado en XML.

5.2.1.1. Selección de la muestra para este trabajo

Debido a la extensión del corpus y según los objetivos de esta investigación monográfica se decidió que era improcedente trabajar con la totalidad de este (150 textos) y se debía optar por una muestra menor. En primer lugar, de manera cualitativa, la muestra seleccionada debía representar la variedad tanto de niveles del MCER como de géneros discursivos y tipos de textos, en igual medida que estos se presentaban en el corpus DICEELE, con el fin de mantener una coherencia con los criterios de recolección de este. Es decir, por cada nivel (B1, B2 y C1) tendríamos que seleccionar la misma cantidad de textos y dentro de cada cantidad debía existir variedad en cuanto a géneros y tipos textuales. Por otro lado, otras variables como el país de origen del texto no fueron controladas para esta selección. Finalmente, se decidió que, de manera cuantitativa, era prudente elegir una muestra que correspondiera al 10% de la totalidad del corpus.

Dicho esto, el corpus que se conformó para esta investigación consta de 15 textos clasificados en los niveles B1, B2 y C1 del MCER (5 textos por cada nivel), los cuales representan todos los géneros discursivos presentes en el corpus DICEELE. Así mismo, entre los tipos de textos podemos encontrar el relato cotidiano, el cuento, la columna de opinión, la reseña, la entrevista, la noticia y la receta. La extensión de los textos varía entre 300 y 800 palabras y, en total, la muestra se compone de 6094 palabras y 120 párrafos. Para un mejor manejo de los datos, los archivos que contienen los textos fueron nombrados con información codificada sobre el nivel, género y tipo de texto, país de origen y ubicación dentro del corpus

DICEELE. De esta manera, por ejemplo, una columna de opinión chilena del nivel B1 sería nombrada así: B1_ACOL_chi_015.

5.2.2. Etiquetado morfosintáctico del corpus DICEELE

Como lo mencionamos anteriormente, con el fin de que un corpus sea más apropiado para la investigación lingüística es conveniente que se encuentre anotado con información adicional. Además de la asignada en los metadatos, también puede suministrarse la información metalingüística sobre el contenido del texto por medio de etiquetas informáticas. De esta manera, el corpus no solo está constituido por la materia real de la lengua, sino que está enriquecido con información sobre ella que permite su explotación (McEnery & Hardie, 2012). Estos datos pueden referir al carácter morfológico, sintáctico, semántico, pragmático, etc., de cada una de las partes del texto que deseemos etiquetar: "En términos generales, cada palabra de un corpus anotado tiene una o varias etiquetas que indican sus características" (Bernal & Hincapié, 2018: p.59).

En este proyecto, el primer etiquetado que se realizó fue el morfosintáctico. Este consiste, básicamente, en la asignación de una etiqueta con información sobre la categoría gramatical, las flexiones, conjugaciones, etc., de cada palabra del texto. Este tipo de etiquetado es conocido como *PoS-tagging*, como mencionamos más arriba. Para ser más precisos, inicialmente, el *PoS-tagging* divide el texto en unidades de análisis llamadas *tokens*: "El token […] corresponde, en lenguaje computacional, a cada una de las cadenas de caracteres dividida por espacios" (Bernal & Hincapié, 2018: p.60); además de las palabras, también se incluyen los signos de puntuación. En términos computacionales, este proceso de *tokenización* facilita en cierta medida el PLN puesto que separa cada uno de los elementos de un texto. Posteriormente, el sistema se encarga de asignar a cada token una etiqueta morfosintáctica.

Para el etiquetado morfosintáctico del corpus DICEELE se utilizó el analizador del lenguaje *FreeLing* y su herramienta de *PoS-tagging*. Este etiquetador realiza la tokenización, lematización[26] y asignación de categorías gramaticales a cada token. Además, permite la salida de estos datos en diferentes formatos como CONLL, XML, JSON, entre otros. A continuación, veremos un ejemplo gráfico sobre la información que este etiquetador ofrece:

[26] La lematización consiste en asignarle a cada palabra su lema (Bernal & Hincapié, 2018). El lema hace referencia a la forma en que una palabra aparecería en el diccionario, es decir, sin ninguna flexión o conjugación. Esto permite agrupar todas las formas que puede obtener cada palabra de un corpus.

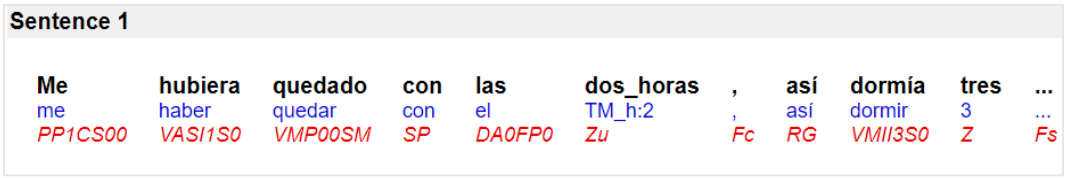


Figura 2. Ejemplo de una salida del *PoS-tagging* de *FreeLing*.

Como podemos observar en esta figura, la información metalingüística que proporciona *FreeLing* viene codificada por una serie de números y letras (p. ej., *PP1CS00*, *VASI1S0*, *SP*, *RG*). Se trata, en este caso, del sistema de etiquetas –o *tagset* –diseñado por el *Expert Advisory Group on Language Ingineering Standards* (EAGLES), el cual propone una normalización para el PLN en las lenguas europeas, por lo que ofrece este *tagset* que *FreeLing* y otros etiquetadores utilizan.[27] En este trabajo se utilizarán también las etiquetas EAGLES al momento de referirnos a aspectos morfosintácticos.

Así pues, como en los metadatos, la anotación morfosintáctica del corpus DICEELE también se realizó en formato XML y fueron estos documentos los que sirvieron como base para la investigación. En la figura 3 se observa un pequeño fragmento del corpus tokenizado y etiquetado en XML; notemos que, además de las EAGLES, esta salida ofrece explícitamente los atributos y valores que vienen codificados en ellas:

```
<token begin="40" ctag="SP" end="42" form="en" id="t3.7" lemma="en" pos="adposition" tag="SP" type="preposition">
  <morpho>
    <analysis ctag="SP" lemma="en" pos="adposition" selected="1" tag="SP" type="preposition"/>
  </morpho>
</token>
<token begin="1" ctag="NP" end="5" form="Alemania" id="t3.8" lemma="alemania" pos="noun" tag="NP00000" type="proper">
  <morpho>
    <analysis ctag="AQ" gen="feminine" lemma="alemania" num="singular" pos="adjective" selected="1" tag="AQ0FS00" type="qualificative"/>
  </morpho>
</token>
```

Figura 3. Ejemplo de dos tokens del corpus (“en” / “Alemania”) etiquetados en XML.

Al mismo tiempo, consideramos que era importante preguntarnos por la relevancia del etiquetado morfosintáctico para los fines de este trabajo de investigación. En primer lugar, resultó de gran utilidad la estructura de tokenizado que realiza *FreeLing* para el proceso de establecer las cadenas de correferencia entre los elementos de un texto. Dicho proceso,

[27] Las EAGLES para el español pueden encontrarse en el manual de *FreeLing*: https://talp-upc.gitbook.io/freeling-4-0-user-manual/tagsets/tagset-es.

además de dividir el texto en tokens para su análisis individual, le provee a cada token un identificador único a través de la etiqueta *id*; en la figura 3 podemos observar, por ejemplo, ambos tokens identificados con los *id* "t3.7" y "t3.8" respectivamente. De esta manera, para anotar las relaciones de correferencia de cada texto, bastaba con referirnos a los *id* de cada elemento, lo cual facilitó el etiquetado de anáforas, realizado de forma manual.

En segundo lugar, una vez realizado el etiquetado de las relaciones de correferencia, la información morfosintáctica de cada token anotado nos sirvió para establecer diferentes análisis estadísticos sobre la frecuencia con la que ciertos elementos gramaticales suelen referirse a otros. Por último, todos estos datos se pueden utilizar con miras al diseño de las actividades planteadas para la plataforma de enseñanza del español al ofrecerle información lingüística explícita al aprendiente de ELE.[28]

Finalmente, debido a que en la actualidad, los textos del proyecto DICEELE se encuentran en proceso de revisión del *PoS-tagged* y las herramientas que lo ejecutan presentan siempre un cierto margen de error, nos vimos forzados a realizar un proceso de revisión y corrección de las etiquetas: "La anotación se puede llevar a cabo de manera automática o de manera manual; sin importar el método que se utilice, siempre debe existir una fase de revisión del material anotado" (Bernal & Hincapié, 2018: p.60). Dicha revisión debería permitirnos poder confiar más en los resultados relacionados con el análisis morfosintáctico, mencionado en el párrafo anterior. Por ejemplo, fue común encontrarnos con errores en el anotado de la palabra *que*, pues funciona como conjunción o pronombre relativo. Era importante corregir este error puesto que en varias ocasiones el pronombre *que* funciona como elemento anafórico. De igual manera y por la misma razón, debimos corregir nombres propios o verbos marcados erróneamente (por ejemplo: *cogía*, que puede ser primera o tercera persona), entre otros errores menos frecuentes.

[28] Un ejemplo de esto serían las actividades en línea para el aprendizaje del latín que ofrece el *School Classic Project* de la Universidad de Cambridge (*Cambridge latin course*). En esta plataforma, al colocar el cursor sobre cada palabra, se le muestra al estudiante la información metalingüística sobre ella. Puede encontrarse más información sobre esto en la siguiente dirección: https://www.clc.cambridgescp.com/online-activities.

5.2.3. Etiquetado anafórico manual de la muestra

Desde el comienzo de este trabajo, hemos mencionado la complejidad que representa la correferencia para el PLN, tanto su identificación como su resolución. Al día de hoy, hemos podido constatar que los sistemas informáticos encargados de procesar estas nociones lingüísticas no están tan desarrollados como para igualar, por ejemplo, a un analizador morfosintáctico en cuanto a su nivel de precisión. Aunque cada vez más los sistemas se perfeccionan en este proceso, el etiquetado de fenómenos como la anáfora es inclusive complejo para un anotador humano que lo trabaje manualmente (Navarro, 2007: p.133). El proceso de anotación anafórica consiste en poder identificar y seleccionar, en un corpus, todos los casos de anáfora para indicar, de manera precisa, a qué entidad del texto precedente se están refiriendo, es decir, indicar su antecedente (McEnery & Hardie, 2012). Según Mitkov (2002: p.130), este tipo de anotación resulta muy útil para desarrollar nuevas aproximaciones a la resolución de la anáfora, pero, además, también para la evaluación objetiva de sistemas ya desarrollados, como es el caso del presente trabajo.

Así pues, debido a la tecnología disponible y a nuestras intenciones investigativas, la anotación anafórica para este trabajo fue realizada de manera manual. Para ello, nos basamos en los principios que presenta Navarro (2007) los cuales permiten realizar un etiquetado efectivo: rapidez, consistencia y profundidad. A partir de estos, el proceso de anotación debe ser simple y apoyado por sistemas, debe haber un acuerdo entre los anotadores (en caso de ser varios) y debe reflejar datos relevantes de la lengua. Además, se debe prestar atención a la teoría "con el objetivo de desarrollar una representación de la información lingüística fundamentada en los conocimientos científicos actuales sobre las lenguas" (Navarro, 2007: p.4). De esta manera, con base en lo investigado sobre la anáfora en este trabajo, pasamos a la identificación y selección de esta noción dentro del corpus. Utilizamos, especialmente, la tipología desarrollada en el apartado *4.3.4.1*, la cual fue de gran utilidad tanto para la selección de anáforas a etiquetar como para la clasificación de estas:

Tabla 2. *Tipología seleccionada para etiquetas e identificación, con base en el apartado 4.3.4.1.*

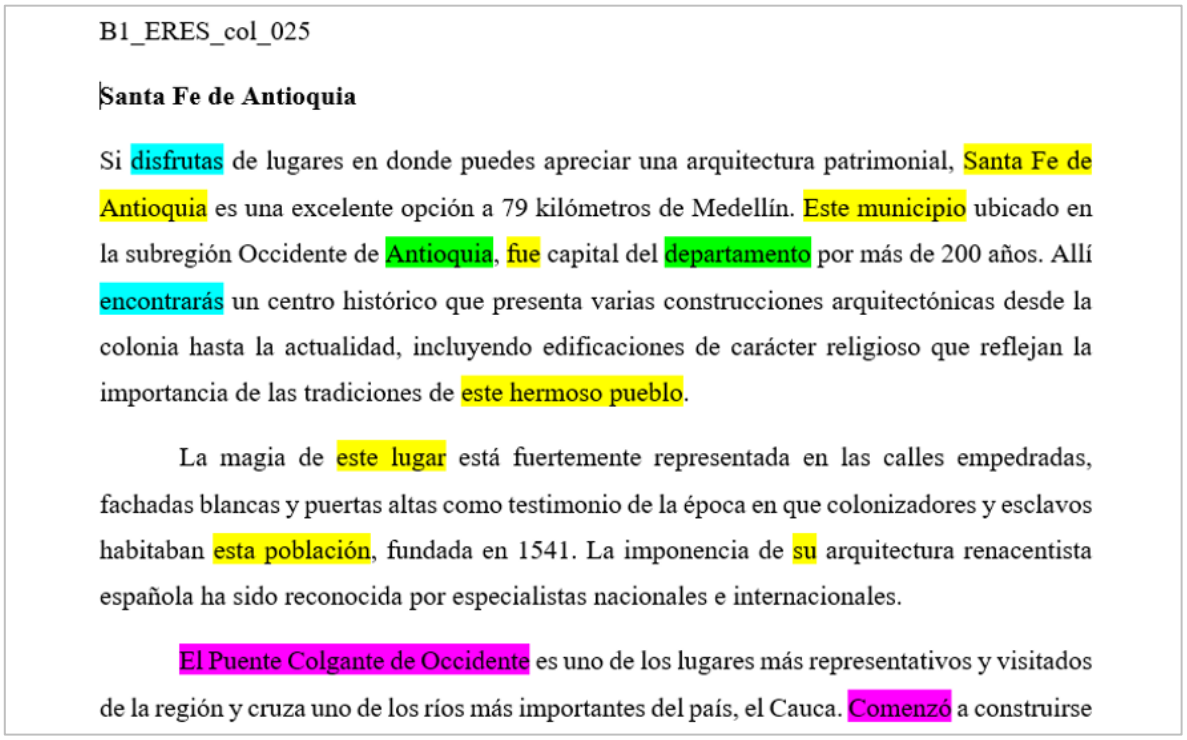
<table>
<tr><th colspan="4">Tipología anafórica</th></tr>
<tr><th colspan="2">Según el antecedente</th><th colspan="2">Según la expresión anafórica</th></tr>
<tr><td>Sintagma nominal</td><td>Sintagma verbal</td><td>Anáfora pronominal</td><td>Anáfora nominal</td></tr>
<tr><td rowspan="2">Sintagma adverbial</td><td rowspan="2">Oración completa</td><td>Anáfora verbal</td><td>Anáfora adverbial</td></tr>
<tr><td colspan="2">Anáfora adjetival y superficial numérica</td></tr>
<tr><th colspan="4">Tipos de relaciones anafóricas</th></tr>
<tr><th colspan="2">Según el referente</th><th colspan="2">Según la relación anáfora/antecedente</th></tr>
<tr><td rowspan="2">Anáfora profunda</td><td rowspan="2">Anáfora superficial</td><td>Por repetición</td><td>Por sustitución</td></tr>
<tr><td colspan="2">Por continuidad de sentido</td></tr>
</table>

A partir de la información descrita en la tabla 2, solo tomamos en cuenta los criterios basados en el tipo de antecedente y expresión anafórica para la clasificación que debían llevar las etiquetas. Por su parte, los criterios basados en el tipo de relación que sostienen entre sí fueron utilizados para la identificación y selección de anáforas y antecedentes, o bien, para su descarte, mas no se incluyeron en las etiquetas. Todo esto, con excepción de la anáfora por repetición que sí se incluyó para el estudio, como lo veremos más adelante.

Teniendo en cuento lo aconsejado por Navarro (2007: p.115), lo más indicado para este tipo de etiquetado es utilizar sistemas que ayuden al anotador a identificar posibles anáforas y antecedentes, basándose en un etiquetado previo y, de esta manera, facilitar el fatigoso proceso de la lectura metalingüística. Infortunadamente, el etiquetado morfosintáctico que posee el corpus DICEELE no permitió desarrollar un sistema similar al que propone dicho autor, pues se necesitaba, además, poder contar con información sobre sintagmas y oraciones, los cuales no fueron anotados para este fin. Por lo cual, el proceso de identificar y marcar posibles anáforas y antecedentes debió hacerse de manera manual; esto implicó un mayor esfuerzo y concentración.

5.2.3.1. Selección de anáforas y antecedentes[29]

En relación con lo anterior, esta fase consistió en hallar todas las posibles expresiones anafóricas, a partir de la tipología seleccionada, con el fin de buscar a qué entidad o enunciado podrían hacer referencia en el texto anterior: "El principal criterio de anotación propuesto para anotar las relaciones anafóricas es marcar siempre el antecedente nominal expreso más cercano a la expresión anafórica semánticamente plena" (Navarro, 2007: p.111), esto requería devolverse en el texto cada vez que se hallaba una expresión anafórica. Una vez ubicado el antecedente que le correspondía, verificábamos que las expresiones anafóricas siguientes también se refirieran al mismo; en caso de no hacerlo, debíamos buscar otro antecedente. Cada vez que hallábamos una nueva relación entre anáfora y antecedente, la marcábamos con un nuevo color para diferenciarla de las demás. De este modo, constituimos las cadenas de correferencia de cada texto. Recordemos que estas cadenas son el conjunto de anáforas que correfieren con un mismo referente (Sapena *et al*., 2013). La siguiente figura refleja dicho proceso:



B1_ERES_col_025

Santa Fe de Antioquia

Si disfrutas de lugares en donde puedes apreciar una arquitectura patrimonial, Santa Fe de Antioquia es una excelente opción a 79 kilómetros de Medellín. Este municipio ubicado en la subregión Occidente de Antioquia, fue capital del departamento por más de 200 años. Allí encontrarás un centro histórico que presenta varias construcciones arquitectónicas desde la colonia hasta la actualidad, incluyendo edificaciones de carácter religioso que reflejan la importancia de las tradiciones de este hermoso pueblo.

La magia de este lugar está fuertemente representada en las calles empedradas, fachadas blancas y puertas altas como testimonio de la época en que colonizadores y esclavos habitaban esta población, fundada en 1541. La imponencia de su arquitectura renacentista española ha sido reconocida por especialistas nacionales e internacionales.

El Puente Colgante de Occidente es uno de los lugares más representativos y visitados de la región y cruza uno de los ríos más importantes del país, el Cauca. Comenzó a construirse

Figura 4. Selección manual de cadenas de correferencia en un el documento B1_ERES_col_025.

[29] Este proceso fue validado por el coordinador del proyecto DICEELE, el profesor Jorge Molina (2015), quien trabajó la correferencia textual en su tesis doctoral (Molina, 2015).

Por otra parte, este ejemplo nos sirve como referencia para explicar los criterios de selección que tuvimos en cuenta para este ejercicio. Lo primero que debemos mencionar es la exclusión del título como elemento de análisis. Como se observa en la Figura 4, el título del texto '*Santa Fe de Antioquia*' podría correferir con la cadena de color amarillo, pues se refiere a la misma entidad. Sin embargo, consideramos que el título no hace parte del texto como unidad, sino que pertenece a un grupo de elementos llamado *paratexto*, el cual, según Gerard Genette (1997: p.7), se considera un texto en sí mismo, independiente del texto al que hace referencia. Por lo cual, este elemento no fue etiquetado lingüísticamente.

En el anterior ejemplo se evidencia que no hay relación endofórica entre el título y los elementos del texto, pues se vuelve a introducir el sintagma nominal 'Santa Fe de Antioquia' y este es tratado como antecedente de las subsiguientes anáforas resaltadas en amarillo. No se trata, por lo tanto, de un caso de anáfora por repetición, la cual sí incluíamos en nuestra selección, a pesar de que su resolución pueda parecer un poco obvia (una anáfora que usa exactamente las mismas palabras del antecedente). En realidad, la anáfora por repetición, sobre todo cuando es parcial, puede generar problemas de ambigüedad. Por lo tanto, decidimos marcarla con el fin de analizar los problemas que el sistema pudiese tener para detectarla.

Además de las evidentes anáforas nominales –según el apartado 4.3.4.1 –presentes en la figura 4, las cuales siempre fueron marcadas, debemos mencionar en qué casos consideramos que estábamos ante una anáfora pronominal. Con respecto a esta, seleccionamos todos los casos de anáfora que fueran pronombres demostrativos, personales o posesivos, puesto que, según la teoría revisada para este trabajo, son los casos más representativos y notorios de anáfora pronominal. Por otra parte, para las anáforas de pronombre relativo y de sujeto implícito tuvimos en cuenta otras consideraciones. De esta manera, si el elemento estaba yuxtapuesto al antecedente, no se marcaba como anáfora; solamente cuando estuviera separado por otros elementos que pudiesen interferir con su resolución, separado por puntuación o si se encontraban en otra oración o párrafo.

Observemos un ejemplo de esto último a partir de la figura 4: '*El puente Colgante de Occidente*', como antecedente de la cadena de color fucsia, va seguido del verbo '*es'* y, más adelante, del verbo '*cruza'*. Dentro de estos verbos no existe anáfora pronominal de sujeto

implícito, pues el sujeto está ligado a los verbos en la sintaxis de esta oración;[30] sin embargo, el tercer verbo '*Comenzó'* sí hace correferencia interoracional con el antecedente y, además, presenta elementos que interfieren con su resolución. Esto es, entre 'cruza' y el antecedente sí hay otros elementos, pero ninguno de estos podría ser ambiguo para la máquina. En cambio, desde un punto de vista computacional, '*Comenzó'* podría correferir con 'el *Cauca'*, si no se posee la información lingüística y referencial necesaria para obviar este caso de ambigüedad.

Otros criterios de selección más específicos se verán evidenciados uno por uno en la sección de análisis correspondiente. Dicho esto, pasamos ahora a describir cómo se formalizó toda esta información lingüística en las etiquetas.

5.2.3.2. Formalización en XML

Con el fin de representar las relaciones anafóricas presentes en el corpus, utilizamos también el lenguaje de marcado XML. El lenguaje XML resulta muy útil para estas tareas pues permite anotar cualquier noción y convertirla en información explotable para una computadora. Además, existen varios programas para marcar y editar textos en este formato y representa un modo de intercambio de corpus entre la comunidad científica (Navarro, 2007: p.133).

El primer paso para etiquetar las anáforas en XML era localizar los *ids* que identificaban a los elementos marcados como correferentes en la fase anterior. De esta manera, se podía utilizar la información ya anotada para apoyar la construcción de las nuevas etiquetas correferenciales. Esto es, buscamos los atributos *tag* y *form* del etiquetado morfosintáctico previo (ver figura 3) para extraer las etiquetas EAGLES y las palabras correspondientes a cada correferencia, con el fin de ahorrar tiempo en la anotación manual mediante un algoritmo.

El diseño del etiquetado anafórico en XML se desarrolló con base en los trabajos de Molina (2015) y Navarro (2007), los cuales utilizan un esquema de anotación similar al usado en la *Message Understanding Conference* (MUC), como lo muestra Mitkov (2002: p.132).

[30] Una explicación más detallada puede encontrarse en los ejemplos de Van Dijk (1997) sobre *secuencias de oraciones* (p.36).

Estas etiquetas[31] usan una estructura en la cual **<REF>** se corresponde con los antecedentes y **<COREF>** con las anáforas. Además, estos autores etiquetan las correferencias en el cuerpo del texto. Por otra parte, también nos basamos en el esquema de anotación de *Free-Ling* que establece otro tipo de etiquetado agrupando las cadenas de correferencia y reescribiendo la información a partir de los *ids*. Esto facilitó en gran medida el trabajo de comparación de etiquetado.

Además de esta información, introducimos en las etiquetas la clasificación de cada correferencia, según la tipología dispuesta en el apartado *4.3.4.1*. Sin embargo, durante el proceso de etiquetado nos dimos cuenta de que dicha tipología no era suficiente para cubrir todos los casos de antecedentes y anáforas que habíamos seleccionado del corpus. Por lo tanto, agregamos nuevas categorías, como se observa en la siguiente tabla, donde también incluimos las abreviaciones que usamos para las etiquetas:

Tabla 3. *Clasificación para etiquetas de antecedentes y anáforas según la información del corpus*.

Tipos de antecedente	**Subtipo**	**Tipos de anáfora**	**Subtipo**
Nominal <**nom**>	Sintagma nominal <**SN**> Nombre propio <**NP**>	Pronominal <**pron**>	Sujeto implícito <**imp**>, pronombre personal <**PP**>, posesivo <**pos**>, relativo <**rel**>, demostrativo <**dem**>, indefinido <**ind**>
Verbal <**verb**>	Sintagma verbal <**SV**>	Nominal <**nom**>	Nombre propio <**NP**>, nombre común <**NC**>, sintagma nominal <**SN**>
Enunciado <**enun**>	Frase <**fra**>, oración <**orac**>, Secuencia de oraciones <**sec-orac**>	Verbal <**verb**>	Pro-verbo <**estar**>, <**hacer**>, etc.
		Adverbial <**adv**>	Adverbio temporal <**temp**>, locativo <**loc**>, de modo <**mod**>
		Adjetival <**adj**>	Adjetivo <**adj**>, superficial numérica <**num**>, posesivo <**pos**>, elipsis <**elip**>
			Encapsuladora <**resumativa**>

[31] Las conferencias MUC surgen en 1991 como competencias para desarrollar y mejorar tecnologías para la extracción de información en el ámbito militar (Mitkov, 2002). En el siguiente enlace puede consultarse el esquema de anotación y los criterios que se establecieron en la MUC-7 (1997) sobre la resolución correferencial: *MUC-7 Coreference Task Definition:*
https://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

El primer ajuste que se puede observar en la anterior tabla, con respecto a la tabla 2, es la exclusión del tipo de antecedente 'Oración completa' que proponían los autores Saiz (2002) y Olivas (2006); si bien es cierto que una oración completa puede ser objeto de referencia de una anáfora, en nuestro corpus encontramos antecedentes mucho más extensos, los cuales no podían ser abarcados por tal categoría. Por esta razón, optamos por un término más amplio para el etiquetado: el enunciado, como hace también Molina (2015). De esta manera, podemos abarcar tanto pequeñas proposiciones como numerosas secuencias de oraciones (Van Dijk, 1997).

En relación con lo anterior, agregamos un nuevo subtipo de anáfora que, ante un antecedente relacionado con hechos o ideas generales, generalmente extenso, cumplía una función *resumativa* (Gaspar, 2009). Al respecto, fue muy útil la categoría de anáfora que menciona Ferrari (2014) a la cual nombra *incapsulatori*, para el caso italiano (2014: p.192). En español, la anáfora *encapsuladora* consistiría en un sintagma nominal (*esta tragedia*, *el hecho*, *estos pensamientos*, etc.) o un pronombre, generalmente demostrativo, (*eso*, *aquello*) que cumple la función de retomar por completo el contenido de un enunciado extenso. No las consideramos como un tipo diferente de anáfora, pues encajaban en las ya establecidas. Sin embargo, debido a su particularidad y función específica, las añadimos como un nuevo subtipo que puede pertenecer tanto a una anáfora nominal como pronominal.

De igual forma, complementamos las demás categorías con otros tipos pronominales (indefinido), adverbiales (de modo) y adjetivales (posesivo y elipsis), según se iban hallando paulatinamente dentro del corpus. Con respecto a la anáfora adjetival a la que nombramos *elipsis*, surge a partir de ciertos casos en los que se aplicaba la definición de anáfora adjetival (sustituían al núcleo nominal), pero no mediante un adjetivo –ni ningún tipo de número. En el siguiente ejemplo (figura 5) se ilustra cómo el primer caso de anáfora adjetival 'la tercera de Colombia' se ajusta con la teoría sobre anáfora superficial numérica, mientras que 'la del Once Caldas' no aplica para ninguna categoría, a pesar de cumplir la misma función. Por lo tanto, decidimos nombrar este caso como elipsis, puesto que, por medio de este mecanismo, el núcleo nominal al que se refiere anafóricamente queda omitido: 'Copa Libertadores':

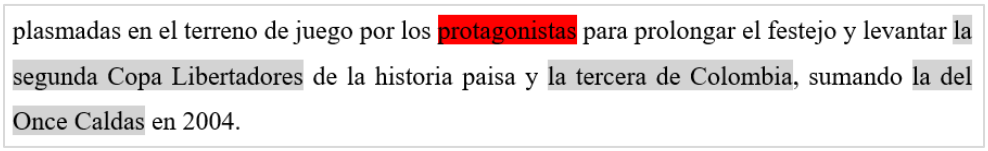


Figura 5. Ejemplos de anáfora adjetival (con gris) en el documento B1_INOT_col_013.

Dicho esto, después de seleccionar, clasificar y etiquetar manualmente en XML cada caso de anáfora y antecedente, ejecutamos un *script* en Python (Anexo 1) que completaba la información de las etiquetas generando el siguiente resultado, similar a las MUC-7 mencionadas anteriormente:

```
3246 <COREF ANA="adj" type="esp" id="64" from="390" to="394" chain="6">
3247
3248   <element morpho="DA0FS0" id="390">la</element>
3249   <element morpho="SP" id="391">de</element>
3250   <element morpho="DA0MS0" id="392">el</element>
3251   <element morpho="NP00000" id="393">Once</element>
3252   <element morpho="NP00000" id="394">Caldas</element>
3253 </COREF>
```

Figura 6. Ejemplo de anotación de una anáfora en XML en el documento B1_INOT_col_013.

En la figura 6 exponemos el mismo ejemplo de anáfora adjetival 'la del Once Caldas', donde podemos observar su clasificación, identificación y cadena de correferencia en XML. Cada token que conforma el elemento anafórico va descrito dentro de la baliza *element*, como en Molina (2015: p.284), y morfológicamente en el atributo *morpho*. Se realizó, además, un segundo etiquetado en el que se organizaban los mismos datos de otra manera:

```
3301 <chain id="6">
3302 <REF ANT="nom" type="NP" id="4" words="Copa Libertadores de América "/>
3303 <COREF ANA="nom" type="NP" id="24" words="Copa Libertadores "/>
3304 <COREF ANA="nom" type="SN" id="37" words="dos Libertadores "/>
3305 <COREF ANA="nom" type="NP" id="47" words="Copa Libertadores de América "/>
3306 <COREF ANA="nom" type="NP" id="54" words="Libertadores "/>
3307 <COREF ANA="nom" type="NP" id="61" words="Copa Libertadores "/>
3308 <COREF ANA="adj" type="num" id="62" words="la tercera de Colombia "/>
3309 <COREF ANA="adj" type="esp" id="64" words="la de el Once Caldas "/>
3310 </chain>
```

Figura 7. Ejemplo de una cadena de correferencia etiquetada en XML con sus anáforas y su respectivo antecedente.

En este caso (figura 7), el script mencionado anteriormente (Anexo 1) recurría al etiquetado de la figura 6 para crear otro tipo de etiquetas similares a las que realiza FreeLing, con base en las cadenas de correferencia. Observemos cómo estas etiquetas agrupan la información según el grupo de correferentes <COREF> que comparten un mismo referente <REF> en la cadena 6, el cual, en este caso, corresponde a 'Copa Libertadores de América'.

5.2.4. Sistematización de datos

Una vez etiquetadas todas las anáforas, referentes y cadenas de correferencia, nos dedicamos a realizar un conteo de la información anafórica que poseíamos en nuestro corpus. Para esta tarea utilizamos el software *AntConc*[32] (Anthony, 2019) y la herramienta *concordances* que permite buscar términos en diferentes textos a la vez, observar su contexto y conocer el número de apariciones o de ocurrencias. Esta herramienta resultó de gran utilidad al momento de establecer la cantidad de anáforas, de tipos y subtipos y la posición que estas ocupaban en la muestra según diferentes criterios de análisis.

En primer lugar, realizamos una búsqueda para establecer la cantidad total de cadenas de correferencia presentes en el corpus, organizadas, además, por nivel de dificultad y tipo de texto. Luego, realizamos un conteo general de anáforas y referentes bajo los mismos criterios y, posteriormente, un conteo de los tipos de anáfora y subtipos con los que contaríamos para el análisis. También organizamos esta información por niveles del MCER y tipos de texto. Esta búsqueda nos permitió caracterizar el corpus cuantitativamente desde el punto de vista de la anáfora, pues dio una primera impresión de la importancia de este elemento en la constitución de los textos. Por otra parte, este proceso facilitó el seguimiento de la distribución de las anáforas y sus diversos tipos según las variables que seleccionamos, lo cual nos llevó a establecer diferentes hipótesis.

Los datos recogidos por medio de *AntConc* fueron consignados en tablas de Excel pues simplificaban el proceso de sumar los resultados según se iban hallando. El cruce de estas variables para la autosuma fue el que nos indicó resultados evidentes sobre las características de nuestro corpus y del desarrollo de la anáfora en textos de diferentes características. Además, de esta manera, organizamos los datos para la sección de análisis, como enseñaremos más adelante.

5.2.5. Evaluación del etiquetado anafórico automático

De acuerdo con lo planteado desde un comienzo, esta fase consistió en la revisión de la capacidad de un programa informático para reconocer relaciones anafóricas. Recordemos

[32] *AntConc* es una herramienta informática de uso libre para el procesamiento de corpus desarrollada por Lawrence Anthony (2019), la cual permite la extracción de información mediante funciones como búsqueda de concordancias, colocaciones, expresiones comunes, lista de palabras, etc. Para más información al respecto consultar: http://www.laurenceanthony.net/software/antconc/

que el objetivo principal de este trabajo es identificar los problemas que en la actualidad están presentes en el tratamiento computacional de la anáfora. Una vez etiquetado nuestro corpus anafóricamente de manera manual, consideramos que no podíamos partir de suposiciones para definir las dificultades que representa el procesamiento de la anáfora para un ordenador. Por esta razón, debíamos encontrar un analizador del lenguaje natural capaz de procesar relaciones anafóricas en español. Como hemos visto hasta el momento, muchos son los trabajos relacionados con la resolución de la anáfora, pero son pocos los que desembocan en el desarrollo de software, lo cual dificultó nuestra búsqueda. Además, la problemática que surgiese de nuestro trabajo podría variar según el software que escogiéramos, pues no todos analizan la información con los mismos recursos (Ruiz, 2012: p.15).

Teniendo esto en cuenta, para este análisis seleccionamos de nuevo *FreeLing* y su herramienta para analizar correferencias. Este software es una librería de uso libre –*open source* –que se puede ejecutar fácilmente desde diferentes sistemas operativos y es de los pocos que logramos encontrar que procesa correferencias para el español. Además de esto, *FreeLing 4.1* funciona con un sistema llamado *RelaxCor* (Sapena *et al.*, 2013) el cual fue implementado en su estructura para mejorar el análisis correferencial. Este programa, desarrollado también por el grupo TALP, ha participado en competencias donde se evalúa la efectividad para el procesamiento de correferencias y ha obtenido muy buenos resultados.[33] Por tales razones decidimos que estas herramientas eran las más adecuadas para este trabajo y pasamos a procesar todo nuestro corpus. Igualmente, seleccionamos el formato de salida en XML para facilitar el proceso de comparación. A continuación (figura 8), damos un ejemplo de este etiquetado, similar al que enseñamos en la figura 7:

[33] Similares a la MUC-7 que mencionamos anteriormente, RelaxCor participó en conferencias orientadas a la resolución de la anáfora como la SemEval-2010 (*International Workshop on Semantic Evaluation*) o la CoNLL-2011 (*Conference on Computational Natural Language Learning*); en esta última ocupó el segundo lugar (Sapena *et al.*, 2013).

```xml
<coreferences>
 <coref id="co4">
  <mention from="t1.13" id="m4.1" to="t1.13" words="Santa_Fe_de_Antioquia"/>
  <mention from="t1.15" id="m4.2" to="t1.21" words="una excelente opción a 79_kilómetros de Medellín"/>
  <mention from="t2.7" id="m4.3" to="t2.7" words="Occidente_de_Antioquia"/>
  <mention from="t6.1" id="m4.4" to="t6.2" words="El Puente_Colgante_de_Occidente"/>
  <mention from="t8.1" id="m4.5" to="t8.1" words="Santa_Fe_de_Antioquia"/>
 </coref>
 <coref id="co13">
  <mention from="t3.8" id="m13.1" to="t3.10" words="este hermoso pueblo"/>
  <mention from="t9.1" id="m13.2" to="t9.2" words="El pueblo"/>
 </coref>
```

Figura 8. Resultado en formato XML de un etiquetado automático de correferencias en *FreeLing 4.1*.

El proceso para evaluar estos resultados consistió simplemente en tomar como referencia nuestro corpus etiquetado previamente de manera manual: "comparar las anáforas detectadas y anotadas automáticamente con las anáforas detectadas, anotadas y validadas por humanos" (Navarro, 2007: p.166). El primer paso para realizar esto fue procesar toda nuestra muestra por el sistema *FreeLing 4.1* y su análisis de correferencias. Después, con base en nuestro etiquetado previo de anáforas, identificamos cada uno de los errores y aciertos que cometía la máquina. Por último, utilizamos los documentos en XML, donde poseíamos las anáforas clasificadas, para consignar esta información en nuevas etiquetas que contenían la valoración sobre si la unidad había sido reconocida o no, con el fin de reunir estos datos por medio del software *AntConc* (Anthony, 2019).

Además, para esta revisión, utilizamos los lineamientos y medidas MUC[34] que suelen utilizarse en las competencias de correferencias como la CoNLL-2011 (Pradhan *et al*., 2011). De estos lineamientos, puesto que nuestro objetivo no es calificar a fondo el procesamiento del sistema, como en dichas competencias, solo tomamos en cuenta *recall* como medida estadística. Esta establece el porcentaje de correferencias que reconoció el sistema con respecto al total que efectivamente era.[35] En otras palabras, nos basamos en *recall* para medir el porcentaje de anáforas que *FreeLing* logró y no logró reconocer.

Para el análisis que aquí realizamos, podríamos haber utilizado un programa que comparara ambos etiquetados y determinara sus niveles estadísticos, pues estos poseen una estructura similar en XML. Sin embargo, muchos de estos casos merecen una revisión más

[34]Disponibles en el enlace que mencionamos antes: https://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

[35] Para comprender mejor cómo funcionan tales medidas, revisar el apartado sobre las métricas utilizadas en la CoNLL-2011, que son descritas en Pradhan *et al*. (2011: p.14).

detenida. Por ejemplo, se encuentran casos en los que anáforas por repetición solo son reconocidas parcialmente, a pesar de ser exhaustivas, por lo que no puede considerarse como un error, pero tampoco como acierto. En este tipo de casos optamos por calificarlas con medio punto. Este conteo nos permitió, gracias al etiquetado previo de tipos y subtipos de correferencia, determinar qué tipos de anáfora y antecedentes representaban más problema para su identificación y en qué medida. Finalmente, a partir de esto realizamos el análisis sobre la problemática del procesamiento de la anáfora y sobre las posibles razones que la generan. La identificación de las falencias que se detectaron en el programa nos permitió una mayor comprensión del fenómeno y de su tratamiento computacional, como lo veremos en el siguiente capítulo.

6. Análisis de resultados

En este capítulo expondremos, por una parte, los datos recolectados a partir de las herramientas mencionadas en el capítulo anterior, analizando, para ello, de forma estadística el fenómeno de la anáfora en nuestro corpus. Por otra parte, examinaremos los resultados del tratamiento computacional de la correferencia realizado por medio del software *FreeLing*, describiendo, de esta manera, tanto los problemas más representativos como los más particulares que representan la imposibilidad de procesar la anáfora de una manera completamente correcta, con la tecnología actual.

6.1. Datos sobre la anáfora en nuestro corpus

Como lo describimos en el apartado 5.3.4, para este análisis realizamos un conteo, por medio de la herramienta *AntConc*, de todos los elementos correferenciales que poseíamos. Recordemos el tamaño de la muestra seleccionada para este trabajo: 15 textos escritos, de entre 300 y 800 palabras, y un total de 6094 palabras y 120 párrafos. Debemos tener en cuenta que, en nuestro caso, la extensión de los textos varía según el nivel de dificultad y el género o tipo textual al que pertenezca, por lo que no hay una equivalencia en este aspecto. Esto es, una receta de cocina del nivel B1, generalmente, va a ser más corta que una columna de opinión del C1 y, en consecuencia, la cantidad de anáforas, referentes y cadenas de correferencia cambiará dependiendo de estos factores.

En términos generales, encontramos **788 expresiones anafóricas** que hacían referencia a **202 referentes**, conformando así **198 cadenas de correferencia**. En suma, la cantidad de elementos correferenciales representaría, aproximadamente, un 20% del total de palabras, frente a las más de seis mil de nuestra muestra; teniendo en cuenta que muchas de nuestras unidades están conformadas por más de una palabra (locuciones, sintagmas, enunciados, párrafos, etc.). Este primer acercamiento revela superficialmente la importancia de este elemento en la constitución de los textos y la necesidad de revisar detalladamente este fenómeno para su comprensión y aplicación al PLN o a los cursos de ELE.

Comentaremos, primero, los datos sobre los antecedentes y las anáforas del corpus en general y, posteriormente, los datos clasificados por nivel de dificultad y tipo de texto.

6.1.1. Número y tipos de antecedentes

Con base en la información consignada en la tabla 4, describiremos las características de los antecedentes presentes en nuestro corpus. Decidimos incluir el número de cadenas de correferencia, pues estas poseen una estrecha relación con el número de referentes.

Tabla 4. *Cantidad en Excel de antecedentes (referentes) y de tipos encontrados en el corpus.*[36]

Tipo de antecedente	Subtipo	Cantidad
Enunciado	Orac	5
	Sec. Orac	8
	Total	13
Verbal	SV	4
	Total	4
Nominal	NC	6
	NP	37
	SN	142
	Total	185
Referentes		202
Cadenas de correferencia		198

Lo primero que resaltaremos de tabla 4 es la relación entre referentes y cadenas de correferencia. Según hemos visto hasta el momento, un referente implica la apertura de una nueva cadena y es el elemento al que, en principio, las anáforas subsiguientes harán referencia. Por lo tanto, cada cadena debería comenzar con un solo referente y la cantidad de estos equivaldría al número de cadenas. La razón por la cual en nuestro corpus el número de referentes (202) es un poco mayor que el de cadenas de correferencia (198) es porque algunas relaciones poseían más de un referente. La siguiente cadena ejemplifica lo anterior:

a) <chain id="4">
<REF ANT="nom" type="SN" id="3" words="**yo** "/>
<REF ANT="nom" type="NP" id="15" words="**María** "/>
<COREF ANA="pron" type="PP" id="16" words="**nos** "/>
</chain>[37]

En **a)**, el pronombre anafórico '*nos'* hace referencia a '*yo'* y a '*María'* al mismo tiempo. De esta manera, agrupa en un elemento anafórico dos referentes que, incluso, pueden pertenecer a otras cadenas de correferencia. Esta información nos permitió evidenciar este

[36] Utilizaremos en este capítulo las mismas abreviaciones que aparecen en la tabla 3 del capítulo anterior.
[37] Cadena tomada del documento C1_NCUE_ecu_031.

tipo de casos, los cuales resultan problemáticos para la anotación anafórica manual, por lo cual, gracias a esto, se tendrán en cuenta para el acuerdo entre anotadores de la anáfora del proyecto DICEELE. El ejemplo **a)** también nos sirve para exponer que una cadena de correferencia puede formarse solamente con un elemento anafórico, por lo que, entre las 198 cadenas, podemos encontrar relaciones de tan solo dos elementos –un REF y un COREF– o de más de cincuenta –una cadena perteneciente al personaje principal de un relato.

Ahora bien, lo primero que observamos al examinar los datos sobre los tipos de referentes en la tabla 4 es el predominio del tipo nominal sobre los referentes verbales y los enunciados, como lo grafica la siguiente figura:

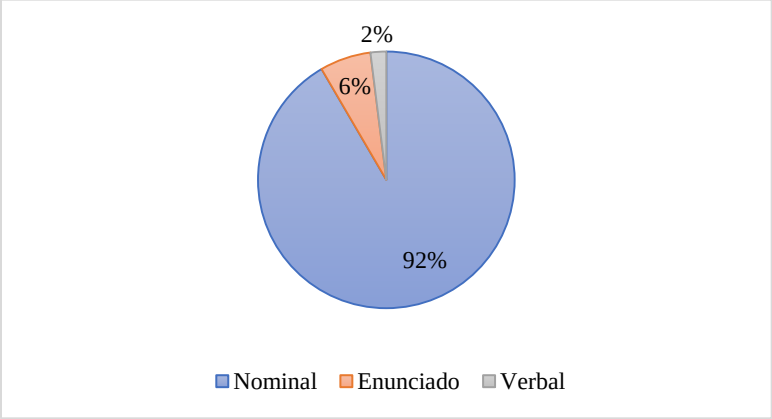


Figura 9. Porcentaje de cada tipo de referente en el corpus.

Lo anterior concuerda con las definiciones de antecedente que proporcionaba Pena (2006), en las que este se componía mayormente por entidades o elementos lingüísticos altamente referenciales. Esto significa, a partir de nuestro corpus, que la mayoría de las anáforas terminan haciendo correferencia con elementos nominales que designan entidades del mundo. Evidentemente, los referentes no solo se conforman por tales entidades, como lo vimos en la clasificación de Saiz (2002), sino también por hechos a los que puede hacerse referencia anafóricamente. Estos hechos, según la figura 9, se ven representados en el 8% de antecedentes, constituido por los enunciados (6%) y los sintagmas verbales (2%); pero, sin ninguna duda, son mucho menos frecuentes que las entidades nominales.

Con respecto a los subtipos de antecedentes, solo consideramos relevante comentar la variedad en los referentes de tipo nominal. Dentro de este tipo encontramos que predomina el sintagma nominal (77%); sin embargo, esto solo se debe a razones de procedimiento. En realidad, todos los referentes de este tipo pertenecen al subtipo SN; no obstante, decidimos

poner aparte en este conteo los NP y los NC, con lo cual hallamos un alto número de NP (20%) y pocos NC individuales (3%), como observamos en la tabla 4. De todas maneras, el tipo de antecedente más común resultó ser aquel que posee una estructura un poco más compleja, introducida por determinantes, posesivos o acompañado de preposiciones, adjetivos, etc., por ejemplo: '*25 vueltas Olímpicas*', '*el redondel*', '*los abogados de oficio'*, etc.

Finalmente, cabe mencionar que no mostraremos los datos de los referentes clasificados por variables, como sí lo haremos con la anáfora más adelante. Esto se debe, principalmente, a que dicho análisis no nos proporcionó información nueva sobre nuestro corpus debido a que la mayoría de los antecedentes era de tipo nominal y bastaba con describir lo que ya se ha descrito más arriba. Por lo demás, este trabajo se enfoca en la anáfora y, aunque el referente esté directamente relacionado con ella, no es nuestro principal enfoque. Por consiguiente, en adelante describiremos los datos relacionados con las anáforas halladas en el corpus.

6.1.2. Número y tipos de anáforas

Tabla 5. *Número de anáforas (correferentes) y sus tipos en Excel según lo hallado en el corpus.*

Tipo de anáfora	**Subtipo**	**Cantidad**			
Adjetival	Adj	4	**Pronominal**	Dem	5
	Elip	6		Imp	178
	Num	12		Ind	5
	Pos	4		Pos	66
	Total	26		PP	205
Adverbial	Loc	7		Rel	20
	Mod	1		Resum	12
	Temp	0		Total	491
	Total	8	**Nominal**	NC	11
Verbal	Estar	1		NP	70
	Hacer	3		SN	174
	Total	4		Resum	4
				Total	259
					788

Con base en la tabla 5, coincidimos con la afirmación de Saiz (2002) sobre la anáfora pronominal como la más frecuente, en oposición a Ferrari (2014) que sostenía que la más frecuente era la nominal. En nuestra muestra, la anáfora pronominal (62%) casi dobla la cantidad de anáforas nominales (33%) y el resto de los tipos, adjetival (3%), verbal (1%) y adverbial (1%), representa un porcentaje bastante mínimo, como se observa en el siguiente gráfico:

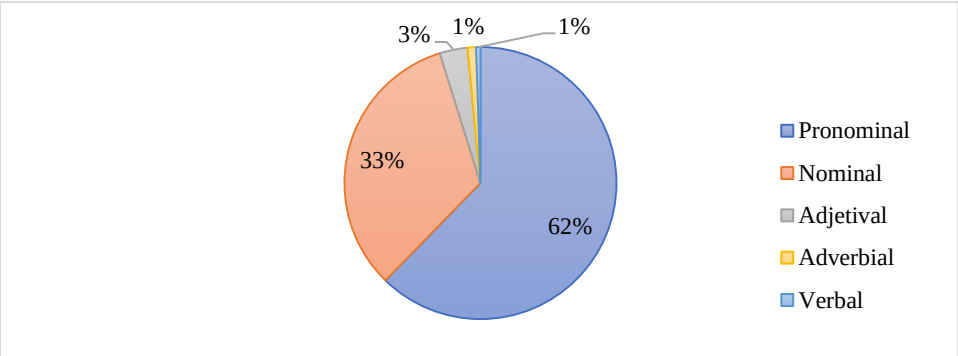


Figura 10. Porcentaje de cada tipo de anáfora presente en el corpus.

En relación con los subtipos de anáfora, la tabla 5 también nos permite observar que, dentro de la anáfora pronominal, la más frecuente es la de pronombre personal (42%) –lo que también coincide con Saiz (2002) –seguida de la que hemos nombrado anáfora de sujeto implícito (36%) y de pronombre posesivo (13%). Incluso, la anáfora pronominal encapsuladora (*resumativa*) supera en número a la resumativa nominal. Con esto queremos mostrar, además, la variedad que presenta la anáfora pronominal como una razón para considerarla ejemplo prototípico de anáfora, como ya lo habíamos mencionado anteriormente.

En cuanto a la anáfora nominal, sucede lo mismo que hemos mencionado a propósito de los referentes con sintagma nominal. Resulta interesante comparar ambos porcentajes de anáforas y antecedentes nominales, pues los resultados son bastantes similares: anáforas como NP: 27% y antecedentes como NP: 20%; anáforas como NC: 4% y antecedentes como NC: 3%; también sigue siendo mayoritario el número de otras construcciones nominales que no encajan en las anteriores: anáforas como SN: 68% y referentes como SN: 77%. Esto permite advertir que hay una consistencia en el uso de estos elementos independientemente de que la función que cumplan sea anafórica o referencial. A excepción de la anáfora encapsuladora, la cual posee una función anafórica muy específica, aunque solo representa un 1% del tipo nominal.

Por otra parte, dentro de la anáfora adverbial no pudimos encontrar del tipo temporal (*temp*) que menciona Saiz (2002) y en su mayoría fueron anáforas de lugar (*loc*). Asimismo, de los pocos casos registrados, la anáfora verbal más común resultó ser la que utiliza como pro-verbo el verbo ‘hacer’ –como también asegura Saiz (2002)– y solo encontramos un caso de anáfora verbal que usaba uno diferente.

En último lugar, debemos resaltar que el subtipo de anáfora superficial numérica (*num*) aparece como el más frecuente dentro de las adjetivales, pero que esto solo se debe al uso desmedido de este recurso por parte del autor de uno de los documentos de nuestro corpus, como se ve en el ejemplo **b):**

> **b)** *los verdolagas han jugado **18 finales**, **10** de la Liga local con ocho títulos; **dos** de Copa Colombia y **las dos** con triunfo; **cuatro** de la Superliga, con dos festejos; **una** de la Copa Sudamericana*[38]

En este ejemplo, los números funcionan como anáfora adjetival al sustituir el núcleo del antecedente, que corresponde en este caso a 'finales'. El texto del anterior fragmento presenta diez casos de este subtipo y, solo por esta razón, la anáfora superficial numérica es más frecuente que los demás subtipos adjetivales en nuestra muestra.

A continuación, analizaremos los datos sobre la anáfora según los niveles del MCER y el género discursivo al que pertenezcan los textos. Esto nos permitirá observar características que pueden aplicarse al corpus en su totalidad y, conjuntamente, generar hipótesis sobre la lengua en general.

6.1.3. Anáforas por nivel de dificultad del MCER

Teniendo en cuenta que nuestros textos hacen parte de un corpus especializado para la enseñanza del ELE, pasamos ahora a considerar la información analizada desde el punto de vista que tiene relación con el nivel de dificultad. El hecho de que, en el momento de seleccionar la muestra para este trabajo, se haya decidido utilizar un igual número de textos por nivel de dificultad del MCER nos permite desvelar más fácilmente el comportamiento de la anáfora dependiendo del nivel en el que se encuentre. Agregamos también en esta información el número de cadenas de correferencia, puesto que, al parecer, también está directamente relacionado con el número de correferentes. De esta manera, podemos formular algunas conjeturas mediante la comparación de los datos sobre los tipos y subtipos de anáfora que aparecen en los niveles B1, B2 y C1, organizados en la siguiente tabla:

[38] Fragmento tomado del documento B1_INOT_col_013.

Tabla 6. *Cantidad en Excel de tipos y subtipos de anáfora presentes en cada nivel del MCER.*

Tipo de anáfora	Subtipo	Nivel del MCER			
		B1	B2	C1	
Pronominal	Imp	23	64	91	
	PP	17	69	119	
	Pos	9	24	33	
	Rel	6	5	9	
	Resum	0	9	3	
	Ind	0	5	0	
	Dem	0	0	5	
	Total	55	176	260	491
Nominal	NP	21	27	22	
	SN	70	60	44	
	NC	7	3	1	
	Resum	2	0	2	
	Total	100	90	69	259
Adjetival	Elip	3	3	0	
	Num	11	0	1	
	Pos	0	0	4	
	Adj	0	2	2	
	Total	14	5	7	26
Adverbial	Loc	1	5	1	
	Mod	0	1	0	
	Total	1	6	1	8
Verbal	Hacer	0	1	2	
	Estar	1	0	0	
	Total	1	1	2	4
Correferentes		171	278	339	788
Cadenas de correferencia		53	66	79	198

Si observamos el final de la tabla 6, notamos que la cantidad de correferentes y cadenas de correferencia crece según se aumenta el nivel de dificultad. Así, en el nivel C1 el número de correferentes y cadenas representa el 43% y 40% de sus totales, respectivamente. Esto puede significar, en primer lugar, que el uso de anáforas se intensifica en los niveles avanzados de la lengua española y que, por ello, los textos de niveles superiores exigen una mayor capacidad para retener entidades y hechos del discurso, lo cual tiene sentido que suceda en estos niveles pues es una característica que los hace más complejos.[39] Del mismo modo, en segundo lugar, los textos de estos niveles requieren de más esfuerzo para resolver mayor cantidad de anáforas, tanto para un aprendiente de ELE como para un procesador informático de la lengua.

[39] Según lo expuesto en Van Dijk (1997), cada referente del texto se iría acumulando en lo que se denomina *memoria semántica a corto plazo* (p.189). Por lo que, entre más cadenas de correferencia tenga que retener en su memoria un aprendiente de ELE para determinar a qué se refiere una anáfora, más dificultad representará para él la comprensión de un texto.

Ahora bien, queremos resaltar también de la tabla anterior lo que ocurre con las anáforas nominales y pronominales. Habíamos notado anteriormente que el tipo de anáfora más común en nuestro corpus era la pronominal. Si nos fijamos en ella en la tabla 6, observamos como el número de estas avanza según aumenta el nivel de dificultad, lo cual tiene sentido ya que el nivel C1 posee más anáforas que los otros dos. Lo que llama la atención sobre esto, es que la anáfora pronominal resulta no ser la más frecuente para el nivel B1. En este caso, el número de pronominales (55) se ve casi doblado por las nominales (100). También observamos como la anáfora nominal disminuyen al aumentar el nivel de dificultad (100/90/69), a pesar de que el nivel C1 posee más cantidad de anáforas.

Lo anterior puede explicarse, si tomamos en cuenta la consideración que realiza Saiz sobre los pronombres como "elementos textuales sin carga semántica" (2002: p.2). Con esto queremos señalar, que, en términos de referencialidad, es mucho más claro semánticamente un sintagma nominal que un pronombre. Por esto, es mucho más simple, generalmente, encontrar la correferencia de una anáfora nominal que la de una pronominal. Esto se relaciona directamente con los niveles de dificultad del MCER que empleamos para nuestro estudio, puesto que, por tal motivo, tiene sentido que en los primeros niveles se utilice con menos frecuencia el tipo de anáfora pronominal y se opte por el tipo nominal, el cual da más pistas sobre su antecedente.

Por otra parte, con respecto a los subtipos de anáfora, volvemos a encontrar una diferencia en el nivel B1 en el que predomina la anáfora de pronombre implícito (42%) sobre la de pronombre personal (31%). En los otros dos niveles es más común encontrar el PP como elemento anafórico, aunque los porcentaje son muy similares. Otro elemento que varía es el aumento de anáforas encapsuladoras en los niveles avanzados. La dificultad en este tipo de anáforas podría residir en la información necesaria para encontrar su respectivo antecedente, por lo cual, es menos probable que se encuentren en textos de niveles más básicos. El resto de los porcentajes de subtipos pronominales no varía sustancialmente entre niveles.

Finalmente, los otros tipos y subtipos de anáfora presentan tan pocos casos que resulta muy difícil realizar declaraciones relevantes sobre ellos con base en esa información. Sin embargo, estos pueden tenerse en cuenta para realizar un análisis posterior en el proyecto DICEELE completo.

6.1.4. Anáforas por tipo de texto

En este caso, no hemos realizado el análisis de la misma manera que se hizo en la fase anterior, puesto que el género discursivo y tipo de texto no poseían una cantidad equivalente dentro del corpus DICEELE, como sí sucedía en el caso del nivel de dificultad. La mayoría de estos eran textos narrativos, seguidos de textos argumentativos, informativos y, en menor medida, expositivos. Por consiguiente, nuestra muestra estaba compuesta por un mayor número de narraciones, por lo que la cantidad de anáforas aumentaría para este grupo y disminuiría para los de tipo expositivo. Sin embargo, comparando proporcionalmente los textos sí es clara la diferencia en número de anáforas. Esto pudimos evidenciarlo con base en la información de la siguiente tabla:

Tabla 7. *Cantidad de tipos y subtipos de anáfora en Excel según el tipo de texto en el que aparecen.*

Tipo de anáfora	Subtipo	Tipos de texto				
		Narrativo	Informativo	Argumentativo	Expositivo	
Pronominal	Imp	104	29	37	8	
	PP	136	13	52	4	
	Pos	27	10	27	2	
	Rel	3	6	11	0	
	Resum	7	2	3	0	
	Ind	1	3	0	1	
	Dem	4	0	1	0	
	Total	282	63	131	15	491
Nominal	NP	28	22	18	2	
	SN	76	54	34	10	
	NC	4	1	5	1	
	Resum	2	2	0	0	
	Total	110	79	57	13	259
Adjetival	Elip	1	5	0	0	
	Num	2	10	0	0	
	Pos	1	0	3	0	
	Adj	0	0	3	1	
	Total	4	15	6	1	26
Adverbial	Loc	4	1	1	1	
	Mod	1	0	0	0	
	Total	5	1	1	1	8
Verbal	Hacer	1	1	1	0	
	Estar	0	0	1	0	
	Total	1	1	2	0	4
Correferentes		402	159	197	30	788
Cadenas de correferencia		79	48	60	11	198

Como decíamos, anteriormente, el número de correferentes para el tipo narrativo representa más de la mitad del total de anáforas, debido a que había más textos de este tipo en la muestra. De todas formas, calculamos la media de correferentes por cada texto y los resultados coinciden con la proporción de la muestra. Esto es, en promedio, un texto narrativo

posee 80 anáforas, uno argumentativo: 49, uno informativo: 40 y uno expositivo tan solo 15. De igual manera, se mantiene una consistencia porcentual entre el número de correferentes y cadenas de correferencia. Es probable que esto ocurra por las características de cada tipo de texto. En un texto de tipo cuento se da una mayor posibilidad de tener que remitirse a distintos referentes como personajes u objetos que pertenecen al mundo de la narración, lo cual es menos frecuente en textos como instrucciones, manuales, recetas de cocina o reseñas.

Por otra parte, la tabla 7 también nos permite hacer comparaciones entre las anáforas que aparecen dentro de un mismo tipo de texto. De esta forma, notamos que se sigue el paradigma sobre la anáfora pronominal como la más común en todos los tipos de texto, excepto en el caso de los textos informativos, donde predomina la anáfora nominal. Si relacionamos esto con el punto anterior, podemos notar que este tipo de textos resultaría más conveniente para la enseñanza del ELE en niveles básicos como el B1; puesto que se inclina por el uso de anáforas nominales, las cuales, probablemente, sean de más fácil resolución. Además, encontramos que, en los textos informativos, es mayor el uso de la anáfora de pronombre implícito (46%) sobre la de pronombre personal (21%), mientras que en los demás tipos de texto se mantiene el paradigma del PP. Esto se relaciona, de nuevo, con lo que ocurría en los textos pertenecientes al nivel B1.

6.2. Problemas en el procesamiento de la anáfora mediante el análisis del sistema *FreeLing 4.1*

Con el fin de completar los objetivos de este trabajo, dedicamos esta sección a las dificultades del sistema *FreeLing 4.1* para detectar correctamente los elementos anafóricos por medio de su herramienta de análisis de correferencias. Nos referimos aquí a las causas que generan el error en el procesamiento de la anáfora: por ejemplo, la ambigüedad lingüística, el desconocimiento de información contextual o, incluso, la limitación provocada por la programación misma del sistema. Para esta etapa utilizamos la metodología descrita en el apartado 5.2.5 donde también explicamos la razón que tuvimos para seleccionar el software y la manera de calificarlo. En este sentido, presentaremos, a continuación, los resultados que, en total, obtuvo *FreeLing* en el etiquetado de nuestra muestra seguidos de una pequeña explicación y, en los apartados posteriores, revisaremos los casos problemáticos organizados por tipos de

anáfora, en donde comentaremos algunos de ellos. Así pues, observemos la siguiente tabla que resume el desempeño del sistema para nuestro análisis:

Tabla 8. *Resultados generales del proceso de etiquetado automático por* FreeLing 4.1.

	Unidades reconocidas	**Porcentaje de acierto**	**Porcentaje de error**
Referentes	123/202	60.9%	39.1%
Correferentes	395/788	50.1%	49.9%
Unidades en total	518/990	52.3%	47.7%

La tabla 8 nos enseña, en principio, que hay un alto grado de error en la anotación de nuestra muestra. En el caso de la anáfora, la unidad que más nos interesa, el porcentaje de unidades reconocidas (50.1%) es casi el mismo que el de las no reconocidas (49.9%). Esto nos indica que debe haber diversas dificultades en el tratamiento de esta noción lingüística y nos da pie para indagar a profundidad sobre las razones que las causan. No obstante, antes de profundizar en los casos problemáticos, exploraremos un par de cuestiones que influyeron en el anterior resultado.

Si observamos de nuevo la tabla 8, nos damos cuenta de que los resultados son mejores para los referentes que para los correferentes. Debemos recordar el análisis anterior sobre los datos de la muestra en donde mostrábamos el predominio del antecedente de tipo nominal frente a los demás tipos y en donde, además, hipotetizábamos sobre la facilidad para resolver las unidades nominales pues eran semánticamente más plenas. Lo anterior se podría justificar con base en la información consignada en la siguiente tabla:

Tabla 9. *Porcentaje de antecedentes reconocidos según su tipo y subtipo*.

Tipo de antecedente	**Subtipo**	**Unidades reconocidas**	**Porcentaje de acierto**
Enunciado	Orac	0/5	0%
	Sec-orac	0/8	0%
	Total	0/13	0%
Verbal	SV	0/4	0%
	Total	0/4	0%
Nominal	NC	4/6	67%
	NP	28/37	75%
	SN	87/142	61%
	Total	119/185	64%

Es evidente que en la tabla 9 los porcentajes de acierto son mucho más favorables para el tipo de antecedente nominal que para los otros dos. Esto tiene su explicación en la

construcción del sistema *RelaxCor* el cual está implementado en *FreeLing*. Como lo mencionamos anteriormente, las anáforas siempre terminan refiriéndose a una entidad o hecho, las cuales están representadas en los tipos de antecedentes. Los referentes constituidos por hechos –que se corresponden con los enunciados y antecedentes verbales –son unidades que no poseen una estructura sintáctica fija y que suelen delimitarse por la comprensión del contexto. Por esta razón, estas unidades son mucho más difíciles de identificar que las entidades nominales. Dicho esto, parece ser que *RelaxCor* no está programado para buscar este tipo de referentes, dado que, en un artículo que explica el funcionamiento del sistema, indican que estos no serán tratados: "no nos ocupamos de la correferencia que involucra hechos y solo nos enfocamos en la correferencia entre entidades"[40] (Sapena *et al.*, 2013: p.848). Por lo tanto, creemos que este sistema se basa solamente en el reconocimiento de entidades nombradas.[41]

Nuevamente, consideramos que, aunque no debemos centrarnos en el análisis del antecedente, no podemos dejarlo a un lado ni negar que de su correcta identificación depende buena parte del acierto en la resolución de la anáfora. Prueba de ello, adelantándonos un poco al análisis que sigue, es el gran porcentaje de anáforas por repetición reconocidas por el sistema (aproximadamente 99%), puesto que se basan en utilizar las mismas palabras que el antecedente y, por tanto, el analizador informático las relaciona directamente.

6.2.1. Problemas con la correferencia pronominal

Comenzaremos por analizar el comportamiento de *FreeLing* ante el caso más común de anáfora y el que hemos considerado complejo por su poca carga semántica (Saiz, 2002).

Tabla 10. *Porcentaje de anáforas pronominales identificadas según su subtipo.*

Subtipo de anáfora pronominal	Porcentaje de acierto	Porcentaje de error
Demostrativa	30%	70%
Implícita	36%	64%
Indirecta	30%	70%
Posesiva	28%	72%
Pronombre Personal	57%	43%
Relativa	55%	45%

[40] Traducido de: "*In this article, we do not deal with coreference involving events, and focus only on entity coreference*." (Sapena *et al.*, 2013: p.848).
[41] NER, por sus siglas en inglés de *Named Entity Recognition*.

De la tabla 10 podemos comentar varios aspectos. Lo primero que sobresale es el bajo porcentaje de acierto, en general, en este tipo de anáfora para la mayoría de las categorías. Sin embargo, si nos basamos en la cantidad de anáforas pronominales en total que sí logró reconocer el sistema (218/491), el panorama no parece tan desalentador. Por otra parte, notamos al final de la tabla un par de excepciones. Las anáforas de pronombre personal y relativo presentan valores más altos que la media general. En el caso de la primera, suele ser más común que la encontremos muy cerca del referente o del correferente anterior en la cadena, como en el ejemplo siguiente:

c) ***El amigo*** *se murió. Niño, no pienses más en* ***él****…*[42]

De esta manera, el sistema asigna el pronombre al posible antecedente más cercano, según su sistema de reglas. Otro motivo por el cual este porcentaje resultó ser más alto que los demás fue porque muchos de los PP se utilizaban varias veces con las mismas palabras. Es decir, en **c)**, el pronombre '*él*' continuaba repitiéndose durante la cadena, de manera que funcionaba como una anáfora por repetición y el sistema, por esa razón, lo detectaba como parte de la cadena, sin importar qué tan alejado pudiese estar de los demás elementos. Por otro lado, lo que ocurre con el pronombre relativo '*que*' puede atribuirse a que la mayoría de las veces aparece lo más cerca posible de su antecedente. Si está muy alejado, el sistema se lo asigna a otra unidad.

Ahora bien, revisemos algunos de los ejemplos pronominales que generaron dificultades en su identificación por medio del software. Los siguientes casos fueron seleccionados del corpus porque representan los problemas más comunes de identificación de anáfora o porque su particularidad y complejidad demuestran por sí mismas que su resolución informática no ha sido posible:

d) ***El Puente Colgante de Occidente*** *es uno de los lugares más representativos y visitados de la región y cruza uno de los ríos más importantes del país, el Cauca.* ***Comenzó*** *a construirse en 1887…*[43]

La razón por la cual la relación señalada en **d)** no fue reconocida por el sistema es porque el sujeto implícito se encuentra en otra oración. En este caso, el problema es también

[42] Fragmento extraído del documento B1_NCUE_esp_021.
[43] Fragmento extraído del documento B1_ERES_col_016.

generado por la programación inicial del sistema puesto que, al parecer, este no ha sido diseñado para identificar los sujetos implícitos como relaciones anafóricas. En cambio, este tipo de correferencias sí son reconocidas en el análisis oracional la mayoría de las veces. Es decir, si la anáfora de pronombre implícito se encuentra en la misma oración del referente, *FreeLing* articula ambos elementos en su análisis sintáctico y acierta, generalmente, en la relación, pero si estos están separados por un punto seguido o aparte, este análisis no es posible. Esto nos indica que, en la programación del sistema, este subtipo de anáfora pronominal no está concebido como una relación correferencial.

Sin embargo, nosotros nos preguntamos, a partir de estos ejemplos, si un lector humano, cuando se encuentra con uno de estos casos, no realiza un proceso muy similar al que sucede cuando se detiene ante un PP para buscar su antecedente; sobre todo cuando este está en otra oración o párrafo. Nuestra opinión es que las características de la anáfora de pronombre implícito en español la hacen meritorias para considerarla como una relación correferencial (Saiz, 2002; Recasens & Martí, 2010). Por otra parte, es también válido considerarla como un tipo de elipsis (que también es mecanismo de economía lingüística) y es probable que así lo hayan hecho los programadores de este sistema. Este tipo de decisiones dependen de la teoría lingüística utilizada para conformar las propias clasificaciones y, por el momento, no hay resoluciones definitivas al respecto.

e) *Había empezado a leer* ***la novela*** *unos días antes.* ***La*** *abandonó por negocios urgentes, volvió a abrir****la*** *cuando regresaba en tren a la finca*…[44]

El fragmento expuesto en el ejemplo **e)** nos muestra las relaciones anafóricas del PP átono singular femenino '*la*' con el SN '*la novela*'. Este tipo de correferencia fue identificada satisfactoriamente en la mayoría de los casos, como lo vimos en la tabla 10. Incluso, hallamos relaciones bien etiquetadas que utilizaban el mismo tipo de pronombre y referente. Lo que parece problemático con este caso es la posición del primer PP al comienzo de una oración. Por esto, *FreeLing* en el etiquetado morfosintáctico analiza '*la*' como un DA0FS0 y no como un PP. Esto genera una de las causas más comunes para el error en la resolución de la anáfora: un mal etiquetado previo. Dado que el sistema identifica la anáfora como un artículo, el cual no es anafórico por sí solo, no lo propone como un posible correferente. De esta forma, la

[44] Fragmento extraído del documento C1_NCUE_arg_32.

unidad señalada siguiente, aunque sí es reconocida como PP por su posición enclítica, tampoco es reconocida puesto que se ha roto la cadena de correferencias. Esto lo comprobamos al volver a analizar el fragmento sin la frase '*La abandonó por negocios urgentes*' y, efectivamente, la relación anafórica entre '*abrirla*' y '*la novela*' se detectaba acertadamente.

f) *Nos daría igual que los presupuestos nos los hiciera* ***Merkel*** *si* ***ella*** *pusiera también la pasta…*[45]

El fragmento de **f)** parecería un caso de sencilla resolución, pues el referente y el correferente están prácticamente contiguos y el PP '*ella*' indica, la mayoría de las veces, la presencia de una anáfora. En realidad, si etiquetamos solamente esta oración, la resolución es efectiva. Por el contrario, cuando el sistema analiza el texto completo, asigna como antecedente de '*ella*' otra unidad anterior con la cual es más probable que se dé la concordancia: '*la ropa*'. Para resolver este caso es necesario saber que '*Merkel*' es el apellido de la canciller alemana Angela Merkel, por lo cual es un nombre femenino y puede correferir con el PP femenino. Creemos que la resolución es efectiva al analizarse la oración aislada porque el sistema no encuentra ningún otro referente que coincida con '*ella*' y elige '*Merkel*' por descarte. Es probable que *FreeLing* no posea exactamente esta información referencial de tipo cultural en su base de datos y, por lo tanto, no tenga los recursos para resolver la correferencia.[46]

g) *Habría que llevar a* ***esa muy granada recua de imbéciles*** *hasta Siria, porque desde que se firmó el acuerdo de* paz *por aquí no se han vuelto a repetir las escenas del horror, y porque si* ***ellos*** *no sienten el zumbido de las balas sobre* ***sus*** *cabezas ni el detonar de los* morteros *derribando las puertas de* ***sus*** *casas…*[47]

Como hemos aludido anteriormente, la identificación correcta de la anáfora pronominal depende mucho de la cercanía con el antecedente. Uno de los casos en el que el análisis computacional más falla es en la anáfora de pronombre posesivo (28% de aciertos), pues casi siempre hay otros elementos de por medio que interfieren en su resolución. Observemos en

[45] Fragmento extraído del documento C1_ACOL_esp_023.
[46] Sin embargo, cabe destacar que *RelaxCor* se permite el uso de conocimiento del mundo para apoyar la resolución de la correferencia a partir de artículos de Wikipedia con el fin de solucionar este tipo de casos (Sapena *et al*., 2013).
[47] Fragmento extraído del documento B2_ACOL_col_016.

el ejemplo **g)** la manera correcta en la que los posesivos debieron de ser asignados. Los elementos señalados con negrilla pertenecen a una cadena de correferencia que el sistema no logró reconocer. De todas maneras, este ejemplo nos sirve para enseñar el comportamiento del analizador informático ante este tipo de pronombres. Puesto que el sistema no logra identificar el referente de los posesivos, debe buscar los candidatos que considere más lógicos. En este caso, asignó '*paz'* como antecedente para el primer '*sus'* y '*los morteros'* para el segundo posesivo. Otro caso de pronominal en el que el sistema se confunde por interferencia con otros elementos suele darse con los PP, como en el ejemplo **e)**, y se produce un error:

h) *Aunque las autoridades no han revelado el nombre d**el atacante**, muerto tras* el ataque, *varios medios estadounidenses **lo** han identificado como **Devin Kelley**…*[48]

En **h)**, entendemos que el PP '*lo'* se refiere a '*el atacante*', principalmente, porque el contexto posterior nos muestra un NP que también correfiere con este. El proceso informático de *FreeLing* se confunde con el SN anterior '*el ataque*' y lo liga al PP3MSA0 debido a que también concuerdan morfosintácticamente. De estos dos ejemplos, podemos percibir que una de las condiciones para que sea posible la resolución pronominal es que el elemento anafórico esté lo más cerca posible del elemento con el que correfiere.

Por último, mencionaremos un tipo de problema que es más común encontrarlo en textos de tipo narrativo e informativo. Observemos el siguiente ejemplo:

i) ***Volví** a cerrar los ojos, **me** aletargué, y desperté de nuevo en la consulta de **mi** psicoanalista. No se creerá lo que **me** acaba de ocurrir, le **dije** para prepararla. Me lo **tendrá** que contar el martes próximo, dijo ella, por hoy hemos terminado*.[49]

Resaltamos en **i)** dos cadenas de correferencia diferentes que provocan confusión entre ellas para el sistema. Como podemos observar, los elementos marcados en negrilla se refieren al narrador principal y la cadena resaltada en color gris se refiere a su psicoanalista. El gran problema que generan estos tipos de texto para el establecimiento de cadenas de correferencia es el cambio de la voz enunciante. Este cambio produce ambigüedad para la máquina que no comprende el significado de lo que está sucediendo en el texto, especialmente si no se marca, por ejemplo, con elementos tipográficos como el guion de diálogo

[48] Fragmento extraído del documento B2_INOT_esp_026.
[49] Fragmento extraído del documento C1_NREL_esp_024.

(Martínez, 2001). En el ejemplo anterior, *FreeLing* marcó correferencia entre todos los PP '*me*' (deducido por repetición), ignorando que el tercero de ellos, en realidad, es correferente de la psicoanalista. Igualmente, es problemático el subrayado final en el que se agrupan ambos referentes con la primera persona del plural '*hemos terminado*', pero esto ya lo hemos mencionado con anterioridad.

Por ejemplo, dentro de un texto de tipo informativo encontramos la siguiente cadena de correferencia en la cual, en el momento en el que el autor cita las palabras de la entrevistada, el sistema rompe la cadena y deja de reconocer los siguientes elementos como anáforas de '*Rocío Molina*', lo cual se evidencia en las etiquetas MUC marcadas con color rojo. Se pierden entonces los elementos anafóricos '*hago'*, '*mi'*, '*mí'* y '*me'* ocasionados por el cambio que realiza el autor de estilo indirecto a estilo directo al citar la voz de Rocío Molina.

j) <chain id="4">
<REF ANT="nom" MUC="y" type="SN" id="8" words="**Rocío Molina**"/>
<COREF ANA="nom" MUC="y" type="SN" id="9" words="**una chica de Málaga que comenzó a bailar desde pequeñita"/>** <COREF ANA="pron" MUC="n" type="pos" id="10" words="**su**"/> <COREF ANA="pron" MUC="y" type="pos" id="11" words="**Su**"/> <COREF ANA="pron" MUC="n" type="imp" id="13" words="**reconoce**"/> <COREF ANA="pron" MUC="n" type="imp" id="16" words="**hago**"/> <COREF ANA="pron" MUC="n" type="pos" id="17" words="**mi**"/> <COREF ANA="pron" MUC="n" type="PP" id="20" words="**mí**"/> <COREF ANA="pron" MUC="n" type="PP" id="21" words="**me**"/> <COREF ANA="pron" MUC="n" type="pos" id="23" words="**su**"/> <COREF ANA="nom" MUC="n" type="SN" id="24" words="esta flamenca"/> […][50]

6.2.2. Problemas con la correferencia nominal

Analicemos ahora el tratamiento de *FreeLing* para los correferentes de tipo nominal.

Tabla 11. *Porcentaje de efectividad de anotación en anáforas nominales según su subtipo.*

Subtipo de anáfora nominal	Porcentaje de acierto	Porcentaje de error
Nombre común	27%	73%
Nombre propio	91%	9%
Sintagma nominal	60%	40%
Nominales en Total	66%	34%

[50] Cadena de correferencia tomada del documento B2_IENT_esp_008.

De los resultados expuestos en la tabla 11, podemos decir, si los comparamos con la tabla anterior sobre la anáfora pronominal, que el tratamiento del tipo nominal por medio de *FreeLing* fue mucho más acertado. En primer lugar, el porcentaje de NP identificados correctamente es el más alto de entre todos los subtipos presentes en nuestra muestra. Esto se debe a que la mayor parte de correferentes como NP están constituidos por anáforas por repetición (parcial y exhaustiva). En consecuencia, cuando nos encontramos ante una anáfora con nombre propio muy probablemente sea el mismo NP utilizado en el referente, o una parte de este. Encontramos muy pocos casos como el del ejemplo **h)**, en donde el NP '*David Kelley*' se refiere al SN '*el atacante'*. Además, en segundo lugar, es también alto el porcentaje de SN identificados por *FreeLing*; teniendo en cuenta la media general para las correferencias. Aquí, nos aventuramos a suponer que se debe a razones similares a las del caso anterior, puesto que gran parte de los elementos nominales de nuestra muestra utilizaban palabras del referente; por ejemplo: entre '*la búsqueda de respuestas*' y '*esta búsqueda*'[51] la palabra *búsqueda* permite relacionarlos directamente como pasa con las repeticiones.

A continuación, revisaremos y analizaremos los diferentes problemas encontrados en el procesamiento de la anáfora de tipo nominal representados en los siguientes ejemplos:

k) <chain id="1">

<REF ANT="nom" MUC="y" type="NP" id="2" words="Santa Fe de Antioquia"/>
<COREF ANA="nom" MUC="n" type="SN" id="4" words="**Este municipio**"/>
<COREF ANA="nom" MUC="n" type="SN" id="9words="**este hermoso pueblo**"/>
<COREF ANA="nom" MUC="n" type="SN" id="10" words="**este lugar**"/>
<COREF ANA="nom" MUC="n" type="SN" id="11" words="**esta población**"/>
<COREFANA="nom"MUC="y"R="E"type="NP"words="Santa Fe de Antioquia"/>
<COREF ANA="nom" MUC="n" type="SN" id="17" words="El pueblo"/>
</chain>[52]

En el ejemplo **k)**, como es evidente, los elementos resaltados en color rojo no fueron reconocidos por el analizador computacional. Si observamos los que están marcados con negrilla, notamos que poseen una estructura similar a la unidad que mencionamos más arriba: '*esta búsqueda*', introducida por un demostrativo. Sin embargo, los elementos que tratamos en este ejemplo no poseen palabras iguales a las del referente '*Santa Fe de Antioquia*', como sí lo hace el correferente de repetición exhaustiva. Para resolver este caso de correferencia,

[51] Elementos tomados de la cadena de correferencia 7 del documento B1_ACOL_chi_015.
[52] Cadena de correferencia tomada del documento B1_ERES_col_025.

se necesitaría la información semántica y cultural que describe a '*Santa Fe de Antioquia*' como un *municipio*, o *pueblo* de Antioquia, el cual es hipónimo de *lugar* y se relaciona sinonímicamente con palabras como *población*, *poblado* o *localidad*. Esta es la información que nos permite, como lectores humanos, identificar sin mayor problema el referente de tales anáforas. Para el sistema, por el contrario, representa un obstáculo.

Lo que acabamos de describir arriba es el principal motivo por el cual el sistema que analizamos dejaba de identificar elementos de correferencia nominales. Encontramos, dentro de toda la muestra anotada, todo tipo de anáforas que no fueron etiquetados debidamente: '*el rey de copas*' y '*el verde paisa*' como epítetos para '*Atlético Nacional de Medellín*'[53], los cuales precisarían de conocimiento cultural para su resolución. De igual manera, con '*el jefe de estado*' para '*Nicolas Maduro*', '*el país*' para '*Venezuela*'[54], '*la novela*' para '*La Metamorfosis*' y '*el libro*', usado como sinónimo para '*la novela*'[55] y '*esos juguetes*' como hiperónimo para '*las canicas, el camión y la pistola de hojalata*'.[56] Asimismo, detectamos una dificultad para el tratamiento de siglas como '*EE.UU.*' para '*Estados Unidos*'[57]. Ninguno de los anteriores elementos pudo ser reconocido por *FreeLing* debido a las limitaciones que hemos mencionado.

l) *Un total de* 26 personas *murieron este domingo en un tiroteo ocurrido en una iglesia de Texas (EE. UU.), en el que se ha señalado ya como "la peor matanza" en la historia de este estado* […] *En una rueda de prensa, las autoridades señalaron que las edades de* las 26 víctimas *oscilan entre los 5 y los 72 años* […] *Las autoridades no detallaron si* **el autor de la matanza** *murió por un disparo de los agentes…*[58]

Lo que sucede en el ejemplo **l)**, de manera similar a los casos anteriores, es que la resolución de la correferencia depende de la información atribuida a los elementos anafóricos. La diferencia es que, en estos, la información no es proporcionada por el conocimiento de la cultura o de la semántica de la lengua, sino por el mismo texto. Aclaremos primero que

[53] Elementos tomados de cadena de correferencias del documento B1_INOT_col_013.
[54] Elementos tomados del documento B1_INOT_ven_024.
[55] Tomados del documento C1_ARES_esp_006.
[56] Tomados del documento B1_NCUE_esp_021.
[57] Tomados del documento B2_INOT_esp_026.
[58] Fragmento extraído del documento B2_INOT_esp_026.

no se tratan de casos de anáfora asociativa, puesto que no se generan nuevos referentes. Más bien, son nuevos correferentes de un referente ya establecido. Hemos señalado en color gris claro la información del texto que permite detectar las nuevas correferencias. Por ejemplo, en **l)**, '*el autor de la matanza*' correfiere con '*el atacante*'; solo que esta relación es posible, en parte, por la comprensión del contexto el cual nos señaló que el atacante fue quien realizó la masacre. Otro ejemplo está presente en la anáfora nominal '*las 26 víctimas*', la cual se relaciona con el referente '*un total de 26 personas*' (no solamente por medio del número 26), a través de la información sucesiva la cual explica que esas personas murieron en un tiroteo y, por lo tanto, son las víctimas.

m) <chain id="6">

<REF ANT="nom" MUC="n" type="SN" id="12" words="**un cuerpo extraño** "/>
<COREF ANA="pron" MUC="n" type="dem" id="27" words="aquello "/>
<COREF ANA="pron" MUC="y" type="PP" id="28" words="lo "/>
<COREF ANA="nom" MUC="n" type="SN" id="31" words="**el objeto** "/>
<COREF ANA="pron" MUC="n" type="PP" id="32" words="Lo "/>
<COREF ANA="pron" MUC="n" type="dem" id="36" words="eso "/>
<COREF ANA="pron" MUC="y" type="dem" id="41" words="aquello "/>
<COREF ANA="pron" MUC="y" type="PP" id="42" words="lo "/>
<COREF ANA="pron" MUC="n" type="dem" id="55" words="eso "/>
<COREF ANA="nom" MUC="n" type="SN"id="72" words="**el cuerpo extraño** "/>
<REF ANT="nom" MUC="n" type="SN" id="73" words="**un zapato** "/>
<COREF ANA="pron" MUC="y" type="PP" id="74" words="lo "/>
<COREF ANA="pron" MUC="y" type="imp" id="75" words="tiré "/>
<COREF ANA="nom" MUC="y" type="SN" id="88" words="**el zapato** "/>
<COREF ANA="nom" MUC="n" type="SN" id="106" words="**su zapato** "/>[59]
</chain>

En este último ejemplo, queremos presentar un caso particular. Este es similar a los ejemplos anteriores dado que su correcta identificación depende de la comprensión del significado del texto. El texto del ejemplo **m)** es un cuento en el cual el protagonista, siente '*un cuerpo extraño*' en el suelo de su auto mientras conduce. A medida que transcurre la narración, el personaje se refiere a este cuerpo como '*el objeto*', '*el cuerpo extraño*', entre otras anáforas pronominales como '*aquello*' y '*eso*'. Sin embargo, cuando, dentro del mundo de la narración, el personaje descubre exactamente qué cosa es dicho objeto pasa a referirse a él por su nombre: '*un zapato*', '*el zapato*', etc. La mejor manera de comprender que el zapato

[59] Cadena de correferencias tomada del documento C1_NCUE_ecu_031.

hace referencia al objeto del que se habla al comienzo es conociendo y entendiendo el significad textual del texto. Por esta razón, es muy poco probable que el sistema los relacione y, quizás, para el análisis computacional, es mejor considerar '*un zapato*' como un nuevo referente.

6.2.3. Problemas con otros tipos de correferencia

Hasta este punto del trabajo, hemos hecho notar que la cantidad de casos de anáforas nominales y pronominales no es comparable a la de los demás tipos de correferencia y no es razonable tratarlos cada uno en apartados diferentes. Por lo tanto, en esta sección, comentaremos las razones por las cuales las **anáforas adjetivales, verbales y adverbiales** presentan dificultades en su tratamiento.

Tabla 12. *Porcentaje de efectividad para los tipos y subtipos de anáforas adjetivales, verbales y adverbiales.*

Tipo de anáfora	Subtipo	Porcentaje de acierto	Porcentaje de error
Adjetivales	Numérica	16%	84%
	Adjetiva	0%	100%
	Elipsis	16%	84%
	Posesiva	0%	100%
	Adjetivales en total	22%	78%
Verbal	Con verbo hacer	33%	66%
	Con verbo estar	0%	100%
	Verbales en total	25%	75%
Adverbiales	Locativas	0%	100%
	De modo	0%	100%
	Adverbiales en total	0%	100%

Este grupo de anáforas representa un alto grado de dificultad para su procesamiento, según observamos la tabla 12. Es probable que, debido a su poca frecuencia, los programadores de sistemas para la resolución de la anáfora no le presten especial atención. Como lo mencionamos anteriormente, a lo mejor, el diseño de *RelaxCor* solo está enfocado en resolver anáforas que se refieran a entidades nominales. Recordemos que, por lo menos, la anáfora verbal se relaciona con hechos, por lo que el sistema no estaría capacitado para resolverlas. Por otra parte, ninguna de las anáforas adverbiales pudo ser reconocida. Esto nos indica que, muy probablemente, en el diseño del sistema tampoco se incluyeron este tipo de relaciones. Por último, la anáfora adjetival también suele considerarse como un tipo de elipsis, por lo cual el desempeño y el interés en la resolución de este tipo de anáfora tampoco debió haber

sido muy alto. Miremos qué sucede con estas anáforas al revisar un par de ejemplos que demuestran los problemas para su correcta identificación:

n) <chain id="6">
<REF ANT="nom" MUC="y" type="SN" id="18" words="un público "/>
<COREF ANA="adj" MUC="n" type="elip" id="19" words="**parte** "/>
<COREF ANA="adj" MUC="y" type="elip" id="22" words="**otra parte** "/>
</chain>[60]

Observamos en el ejemplo **n)** que la anáfora de elipsis '*otra parte*' se marca como positiva en la resolución de la correferencia. En realidad, no se trata de un acierto completamente, pues no ha detectado el sintagma elidido '*un público*' (tampoco tenía por qué hacerlo), sino que ha marcado la correferencia con la unidad anterior de la cadena por repetición de palabra. No obstante, dicha correferencia '*parte*' (del mismo tipo elidido) no ha sido reconocida para su referente, por lo cual la cadena de correferencia no tiene sentido. En conclusión, la naturaleza de la anáfora adjetival no permite su resolución por medio de *FreeLing*, al parecer, porque no fue considerada como anáfora en su programación.

En cuanto a los ocho casos de anáfora adverbial, creemos que ninguno de ellos fue identificado porque tampoco, en la programación inicial del sistema, se concibió a los adverbios como candidatos a elementos anafóricos. Por lo tanto, nuestros casos de *allí*, *allá*, *ahí*, *donde* y *así* no significan para *RelaxCor* una posible relación correferencial. Es muy probable, que esto se deba a que dichas anáforas no se refieren específicamente a entidades nominales y lo hacen más bien a complementos circunstanciales o a otros adverbios.

o) <chain id="9">
<REF ANT="enun" MUC="n" type="orac" id="31" words="llevar a esa muy granada recua de imbéciles hasta Siria "/>
<COREF ANA="pron" MUC="n" type="resumativa" id="32" words="**Eso** "/>
<COREF ANA="verb" MUC="y" type="hacer" id="33" words="**lo haremos** "/>
</chain>[61]

Por último, presentamos el único correferente verbal que consideramos como correcto. Sin embargo, como mencionamos anteriormente, la anáfora verbal se refiere a hechos y no a entidades, por lo cual no es posible su resolución a partir de *FreeLing*. Lo que ocurre

[60] Cadena de correferencias tomada del documento B2_IENT_esp_008.
[61] Cadena de correferencias tomada del documento B2_ACOL_col_016.

en **o)**, de hecho, es lo mismo que en el ejemplo **n)**. Sucedió que el sistema reconoció el pronombre '*lo*' de la anáfora verbal y lo identificó con el correferente anterior '*Eso*' de la cadena a la que pertenecen, la cual hacía referencia a un hecho descrito en un enunciado. Por lo tanto, tampoco se trata de un acierto en sí debido a la misma programación del sistema. Aunque sabemos que las anáforas que se refieren a hechos y enunciados no tienen resolución por parte de este sistema, consideramos importante hacer una pequeña mención a ellas, pues representan un impedimento a la hora de pretender realizar un análisis completamente correcto de las unidades anafóricas.

6.2.4. El caso de la anáfora encapsuladora

Como lo habíamos dispuesto anteriormente desde la clasificación, la anáfora encapsuladora (*resumativa*) es un subtipo de anáfora que pertenece a los tipos nominal y pronominal, puesto que se vale de estas categorías para referirse a distintos tipos de enunciados. Sin embargo, su peculiaridad y función específica merece una mención aparte en el análisis. Ahora sabemos que el problema de este subtipo de anáfora no está en la expresión misma, porque esta es evidentemente de naturaleza anafórica, sino en el antecedente con el que se relaciona, formado por oraciones simples y secuencias de oraciones complejas. Recordemos que, en la tabla 9, el porcentaje de estos enunciados identificados fue del 0%. En consecuencia, el porcentaje de anáforas encapsuladoras reconocidas fue igualmente del 0%.

p) *el verde paisa acumula 19 títulos en Colombia (15 ligas, dos Copas y dos Superligas) y seis coronas internacionales (dos Libertadores, dos Merconorte y dos Interamericanas) para 25 vueltas olímpicas oficiales, más otras seis amistosas y cuatro aficionadas. Y* ***todo este palmarés*** *lo consolidó por lo alto este miércoles al vencer en la final de la edición 2016 de la Copa Libertadores de América…*[62]

El ejemplo **p)** muestra un poco el panorama de las dificultades que representa este tipo de relaciones anafóricas. Sin embargo, para el lector humano, el elemento anafórico nominal encapsulador '*todo este palmarés*' basta para advertir que se está refiriendo al listado de triunfos que acaba de mencionar. En nuestra opinión, este subtipo de anáfora resulta muy

[62] Fragmento extraído del documento B1_INOT_col_013.

llamativo para la resolución de la correferencia pues representa un reto que todavía está completamente por desarrollar en el PLN. Por otra parte, consideramos que también puede resultar interesante enfocarse en anáforas de este tipo para el diseño de actividades de ELE, pues requieren de una alta capacidad para comprender el lenguaje humano y la comunicación.

7. Conclusiones y perspectivas

7.1. Conclusiones

Durante la realización de este trabajo, pretendimos alcanzar los objetivos planteados desde un comienzo, con el fin de generar nuevo conocimiento acerca del tratamiento computacional de la anáfora. Por otra parte, uno de los principales propósitos que buscábamos con esta investigación era apoyar el desarrollo del proyecto DICEELE a partir de nuestra metodología y de los resultados aquí encontrados. Estos deberían guiar, por una parte, a los integrantes del proyecto en cuanto al tratamiento manual de las relaciones anafóricas para la construcción del dispositivo y, por otra, al diseño de las actividades didácticas a partir del conocimiento de las características de la anáfora y de los problemas para su tratamiento. Consideramos que los objetivos que establecimos en el comienzo están satisfechos: hemos anotado toda la información lingüística que poseíamos sobre la anáfora en una muestra del corpus DICEELE y, con base en ello, hemos analizado el procesamiento del software *FreeLing* con el fin de identificar los problemas recurrentes que impiden el tratamiento eficaz de este elemento lingüístico.

En primera instancia, queremos destacar que este trabajo, como muchos otros que se realizan en el área, ratifican la necesidad de utilizar herramientas informáticas en el procesamiento de corpus, pues representan un gran ahorro de tiempo y, en ocasiones, de efectividad para trabajar con ingentes cantidades de información y analizarlas desde diferentes perspectivas. Esto concuerda con lo expuesto por autores como Giovanni Parodi (2014) o Bernal e Hincapié (2018) en relación con la metodología de la lingüística de corpus. En nuestro caso, fueron de gran utilidad los sistemas *AntConc* (Anthony, 2019), *FreeLing* y *RelaxCor* (Sapena *et al.*, 2013) para evaluar la información anotada del corpus DICEELE. Así como para el proyecto ha sido de vital importancia el etiquetado morfosintáctico automático de *FreeLing* o la herramienta de Dimitri Lamprinos para la clasificación de los textos según el nivel de dificultad del MCER. Es importante mencionar que estos programas informáticos son de libre uso y accesibles de manera gratuita en internet. Esperamos que el uso que aquí hacemos de estas herramientas sea de ayuda tanto para los integrantes del proyecto como para los futuros investigadores del área.

Además, queremos también resaltar, basándonos, de nuevo, en la lingüística de corpus, la importancia de usar documentos auténticos para el análisis de las lenguas naturales, sobre todo en trabajos sobre el PLN. De esta forma, podemos partir de información real para hipotetizar sobre la lengua y proponer soluciones a los conflictos que se encuentren en el manejo de los datos y en las concepciones que se tengan sobre aspectos de la lengua. En nuestro caso, partir de documentos auténticos nos permitió reflexionar sobre las características de la anáfora desde distintos tipos de producción lingüística. Por otra parte, el hecho de que se tratara de un corpus recolectado con la intención de servir para la enseñanza del ELE abre la posibilidad de utilizar para este fin el conocimiento recogido a lo largo de este trabajo.

Ahora bien, sobre el análisis que hemos realizado aquí para la correferencia textual, concluimos que, en primer lugar, fue pertinente utilizar una tipología extraída de varios autores, adaptarla e, incluso, agregarle nueva información según las características del corpus. Esto fue necesario porque muchos de los elementos encontrados no encajaban en todas las categorías, lo cual facilitó el descubrimiento de otros mecanismos de cohesión textual relacionados con la anáfora. En segundo lugar, también consideramos acertada la elección del programa *FreeLing 4.1* para el procesamiento de la correferencia, puesto que, a pesar de los casos adversos, presentó resultados bastante favorables que permitieron las relaciones establecidas entre aciertos y errores.

También, con relación a los datos que recogimos del análisis, recordamos la influencia que efectivamente poseen factores como el nivel de dificultad y el tipo de texto en la presencia de la anáfora como mecanismo de cohesión textual. Esto nos indicó, por una parte, que los niveles avanzados del MCER exigen una mayor capacidad de comprensión e interpretación de la lengua al requerir que el aprendiente de ELE resuelva mayor número de relaciones correferenciales y retenga más entidades y referentes en la lectura de un texto. Por otra parte, definimos que ciertos tipos de textos se adecuan de mejor manera a ciertos niveles del MCER al compartir las mismas características de la anáfora. De esta manera, esta información puede ser explotada en la construcción del dispositivo DICEELE con respecto a la selección de los textos para el diseño de las actividades.

En este mismo sentido, encontramos que es más común el uso de anáforas de tipo pronominal que de otro tipo para el caso del español. Sin embargo, también descubrimos que sería más conveniente, para los niveles más básicos, utilizar textos que contengan menos

cantidades de relaciones anafóricas y, preferentemente, de tipo nominal, dado que se muestran semánticamente más claros. La anáfora pronominal se relaciona más con la dificultad de la resolución y requiere de un mayor esfuerzo. También recordamos, que, dentro de las anáforas pronominales, el tipo más común para el español son las de pronombre personal, sobre todo, las formas átonas de objeto directo.

En último lugar, para recapitular la información que se ha obtenido sobre los problemas en el tratamiento computacional de la anáfora, el cual era el objetivo principal de esta investigación, presentamos un listado que resume la *problemática* según las dificultades lingüísticas que encontramos para el procesamiento correferencial. Estas son:

1) La programación inicial del sistema define, desde sus propias reglas, lo que puede ser o no un referente o un correferente. De este modo, el sistema se limita solamente a procesar los casos más comunes y descarta, desde un principio, los casos excepcionales, poco comunes o de difícil procesamiento, como el reconocimiento de anáforas adverbiales o de antecedentes como hechos y enunciados.

2) Los sistemas de resolución correferencial necesitan de información cultural sobre el mundo y aunque la tengan, esta no abarca completamente toda la información que pueda definir a una entidad; por ejemplo, como vimos con los epítetos o las siglas.

3) Puede haberse hecho un etiquetado previo incorrecto. Si hay errores en el etiquetado morfosintáctico, semántico u oracional, se van a ver reflejados en el análisis correferencial, pues este se basa en dicha información. Claramente, también es un problema que no exista tal información etiquetada, como en el caso del conocimiento del mundo.

4) La no detección de un referente provoca que las anáforas de dicha cadena de correferencia no tengan a qué elemento dirigirse y se rompan, por lo tanto, las relaciones subsiguientes.

5) Para el caso de los pronombres, deben estar muy cerca de su antecedente o de la anáfora anterior en la cadena de correferencias para ser asignados correctamente a su referente.

6) Los casos de anáfora que agrupan varios referentes en una sola unidad no logran ser reconocidos.

7) Los textos en donde existe un cambio de voces entre las entidades de un texto por medio del intercambio entre estilos directo e indirecto presentan ambigüedades complejas para la máquina.

8) El analizador automático no logra obtener una comprensión a nivel textual de la información nueva que da el texto sobre sus referentes.

9) Las ambigüedades que se generan en lengua natural sobre la asignación de referentes para las anáforas son también un problema para el PLN.

7.2. Perspectivas

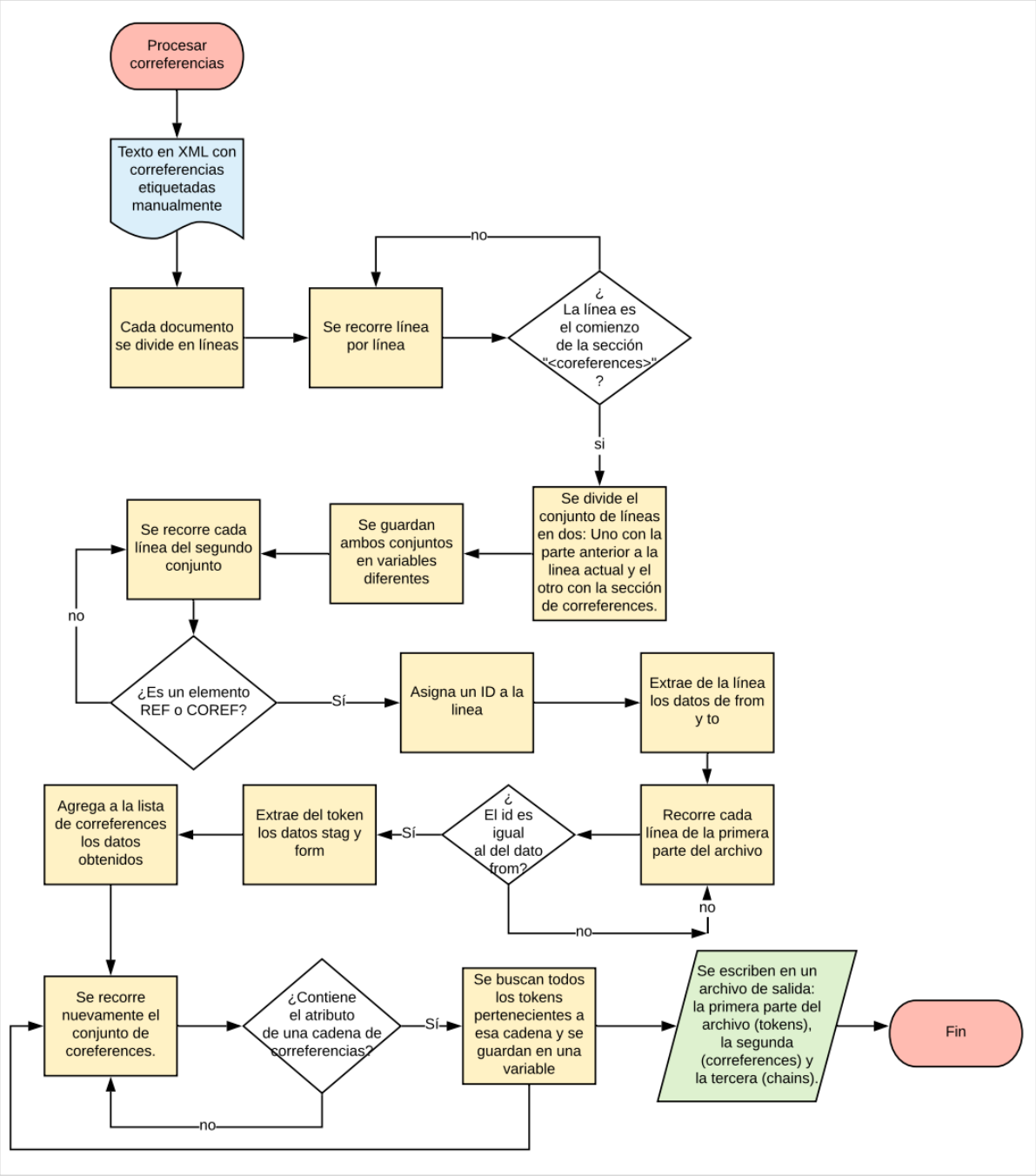
Por último, mencionamos las perspectivas que tenemos a partir de esta investigación. La primera se desprende de nuestra metodología de etiquetado, pues pretendemos completar el *script* de Python para automatizar en gran medida el etiquetado anafórico del corpus DICEELE. Este consistiría, básicamente, en relacionar los elementos marcados en un documento de Word con los tokens de los documentos etiquetados morfosintácticamente para crear un anotador de correferencias, similar a los que mencionan Navarro (2007) y Recasens & Martí (2010). De esta manera, ahorraríamos mucho tiempo en la tarea laboriosa de anotar documentos de manera manual.

Por otro lado, esperamos que las características que en este trabajo se definieron para los elementos anafóricos permitan elaborar reglas para el diseño de las actividades del proyecto DICEELE sobre correferencias. Es decir, con base en los niveles de dificultad y en las características de cada tipo de texto, hemos contemplado que sea posible generar automáticamente las actividades, sin que se conviertan en un crucigrama de correferencias imposible de solucionar, y que sean más adecuadas y consecuentes con el mecanismo de cohesión textual que representa este fenómeno. También, que sirva de ejemplo la dificultad para procesar ciertos tipos de anáforas y verificar si tales problemas también se presentan en la resolución humana por parte de aprendientes de ELE.

Finalmente, esperamos, más adelante, poder aportar a la resolución del subtipo de anáfora encapsuladora, sobre el cual se dispone de muy poco material y representa dificultades serias, aún en la actualidad, para ser procesadas automáticamente. Esto significaría un gran avance en el PLN y nos acercaría un poco más al tratamiento computacional del lenguaje sin inconvenientes.

8. Anexos

1) Diagrama de flujo del *script* en Python que facilitó la anotación manual de anáforas



Realizado por José Luis Pemberty Tamayo, integrante del semillero de investigación *Corpus ex Machina*.

9. Bibliografía

Alcaraz Varó, E. & Martínez Linares, M. A. (1997). *Diccionario de lingüística moderna*. Barcelona: Ariel S.A.

Anjali, M. K. & Babu Anto, P. (2014). Ambiguities in Natural Language Processing. *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, Special Issue 5, October 2014. pp. 392 – 394.

Anthony, L. (2019). AntConc (Version 3.5.8) [Software]. Tokyo: Waseda University. URL: http://www.laurenceanthony.net/software/antconc/

Antoniadis, G. (2010). De l'apport pertinent du TAL pour les systèmes d'ALAO. L'exemple du projet MIRTO. En: *2eCongrès Mondial de Linguistique Française (CMLF-2010)*. (pp. 2211 – 2223). La Nouvelle Orléans: USA.

Arcas, Y. (2000). Lingüística textual y la enseñanza del español como lengua extranjera. *NUCLEO*, n.17, pp. 5 – 16.

Arévalo, M., Civit, M. & M. A. Martí. (2004). MICE: A Module for Named-Entities Recognition and Classification. *International Journal of Corpus Linguistics*, Vol.9, n.1, pp. 53 – 68. Amsterdam: John Benjamins.

Badia Cardús, T. (2003). Técnicas de procesamiento del lenguaje. En: Martín, M. A. (Ed.), *Las tecnologías del lenguaje* (pp. 193 – 248). Barcelona: UOC.

Bernal Chávez, J. A. & Hincapié Moreno, D. A. (2018). *Lingüística de corpus*. Bogotá: Instituto Caro y Cuervo.

Bolaños Cuéllar, S. (2015). La lingüística de corpus: Perspectivas para la investigación lingüística contemporánea. *Forma y Función*, Vol. 28, n.1, pp. 31 – 54.

Cánovas Méndez, M. (2009). Interacción de recursos digitales en las tareas de aprendizaje de lenguas: portafolios electrónicos y traducción asistida por ordenador. *Campo abierto*, Vol. 28, No. 2, pp. 103 – 119.

Carrión, M. (2014). Resolución de anáforas que requieren conocimiento cultural con la herramienta FunGramKB. *Revista de Lingüística y Lenguas Aplicadas*, Vol 9, pp. 1-13.

Ceberio, K., Aduriz, I., Diaz De Ilarraza M.A. y Garcia-Azkoaga, I.M. (2008). Análisis de la correferencia para su anotación en un corpus en euskara. En: Moreno Sandoval, A: *Actas del VIII Congreso de Lingüística General*. (pp. 496 - 512). Madrid.

Charlier, B., Deschryver, N. & Peraya, D. (2006). Apprendre en présence et à distance: une définition des dispositifs hybrides. *Distances et savoir*s, Vol. 4, pp. 469 – 496.

Consejo de Europa. (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. Madrid: Instituto Cervantes / MECyD / Anaya.

Cuq, J. P. (2003). *Dictionnaire de didactique du français Langue étrangère et seconde*. París: CLE International.

De Beaugrande, R. A., & Dressler, W. U. (1997). *Introducción a la Lingüística del texto*. Barcelona: Ariel, S.A.

Ferrari, A. (2014). *Linguistica del testo. Principi, fenomeni, strutture.* Roma: Carocci.

Gaspar Marques, I. (2009). *Anáfora associativa - propostas de abordagem em contexto escolar*. (tesis de maestría) Faculdade de Letras da Universidade de Coimbra.

Genette, G. (1997). *Paratexts: Thresholds of Interpretation* (J. Lewin, Trad.). Cambridge: Cambridge University Press.

Hausser, R. (2014). *Foundations of Computational Linguistics*. Berlin: Springer.

Hernández, R., Fernández, C. & Baptista, P. (2014). *Metodología de la investigación*. Ciudad de México: McGraw-Hill.

Instituto Cervantes. (2007). *Plan curricular del Instituto Cervantes*. Madrid: Cervantes/Edelsa. URL:http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/

Lineros Quintero, R. (1998). Aportaciones teóricas de la lingüística textual a la enseñanza/aprendizaje del español como lengua extranjera. En: *II Congreso Nacional sobre Metodología y Didáctica del español como L2*, (pp. 1 – 11). Universidad de Cádiz.

López, A. M. (2008). *Semántica y sintaxis.* Bogotá: Universidad Santo Tomás.

Márquez, L. (2002). Aprendizaje automático y procesamiento del lenguaje natural. En: Martín, M. A. & Listerri, J. (Ed.), *Tratamiento del lenguaje natural* (pp. 133 – 188). Barcelona: Edicions de la Universitat de Barcelona.

Martín Monje, E. (2012). Presente y futuro de la Enseñanza de Lenguas Asistida por Ordenador: ¿El final de una era? *Revista de Lingüística y Lenguas Aplicadas.* Vol. 7, pp. 203 – 212.

Martín Peris, E. (1997). *Diccionario de términos clave de ELE*. Madrid: Instituto Cervantes.

Martínez Barco, M. P. (2001). *Resolución computacional de la anáfora en diálogos: estructura del discurso y conocimiento lingüístico*. (tesis doctoral) Universidad de Alicante. Depto. de Lenguajes y Sistemas informáticos.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. New York: Cambridge University Press.

Mitkov, R. (2002). *Anaphora resolution*. London: Longman.

Molina Mejía, J. M. (2015*). ELiTe-[FLE]² : Un environnement d'ALAO fondé sur la linguistique textuelle, pour la formation linguistique des futurs enseignants de FLE en Colombie.* (tesis doctoral) Université Grenoble Alpes.

Morales Carrasco, R. (2004). *Resolución automática de la anáfora indirecta en español.* (tesis doctoral) Instituto Politécnico Nacional.

Moreno Sandoval, A. (1998). *Lingüística computacional. Introducción a los modelos simbólicos, estadísticos y biológicos*. Madrid: Síntesis, S.A.

Navarro Colorado, F. (2007). *Metodología, construcción y explotación de corpus anotados semántica y anafóricamente.* (tesis doctoral) Dpto. de Lenguas y Sistemas Informáticos. Universidad de Alicante.

Nerbonne, J. A. (2003). Computer-assisted language learning and natural language processing. En: Mitkov, R. (Ed.), *The Oxford handbook of computational linguistics* (pp. 670 – 698). Oxford: Oxford University Press.

Nerbonne, J. A. (2005). Linguistics challenges for computationalists. En: Nicolas Nicolov *et al.* (Ed.), *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005* (pp. 1 – 16). Borovets: John Benjamins, B.V.

Olivas Zazueta, O. A. (2006). *Sistema de construcción de redes semánticas con detección de anáfora.* (tesis de maestría) Instituto Politécnico Nacional. Centro de Investigación en Computación.

Parodi Sweis, G. (2010). *Lingüística de corpus: de la teoría a la empiria*. Madrid: Iberoamericana Vervuert.

Peña, H. (2016). La ambigüedad. *Revista Documentos Lingüísticos y Literarios UACh*, 2016, n.8, pp. 41 – 45.

Peña Martínez, G. (2006). *La anáfora y su funcionamiento discursivo: Una aproximación contrastiva*. Valencia: Servei de Publicacions, Universitat de València.

Pitkowski, E. F. & Vásquez, J. (2009). El uso de los corpus lingüísticos como herramienta pedagógica para la enseñanza y aprendizaje de ELE. *TINKUY*, n.11, pp. 31 – 51.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., & Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. En: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 1 –27). Association for Computational Linguistics.

Recasens, M., Marti, M. A., & Taulé, M. (2007). Where anaphora and coreference meet. Annotation in the Spanish CESS-ECE corpus. *Proceedings of RANLP*, pp. 504-509.

Recasens, M., & Martí, M. A. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language resources and evaluation*, n.44 (4), pp. 315 - 345.

Rodríguez Hontoria, H. (2003). Interfaces en lenguaje natural. En: Martí, M. A. (Ed.), *Las tecnologías del lenguaje* (pp. 130 – 155). Barcelona UOC.

Ruiz, M. (2012). *Resolución de la anáfora correferencial con FunGramKB.* (tesis de maestría) Universidad Nacional de Educación a Distancia: Madrid.

Saiz Noeda, M. (2002). *Influencia y aplicación de papeles sintácticos e información semántica en la resolución de la anáfora pronominal en español*. (tesis doctoral) Universidad de Alicante. Depto. de Lenguajes y Sistemas informáticos.

Sapena, E., Padró, L. & Turmo, J. (2013) A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution. *Computational Linguistics*, Vol. 39, n.4, pp. 847 – 884.

Teubert, W. (2009). La linguistique de corpus : une alternative. *Semen. Revue de sémio-linguistique des textes et discours*, n.27, pp. 185 – 211.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: Benjamins.

Tordera Yllegas, J. C. (2011). *Lingüística computacional. Tecnologías del habla*. Valencia: Universitat de Valéncia.

Van Dijk, T. (1997). *La ciencia del texto*. Barcelona: Paidós.

Villegas, M. (1993). Las disciplinas del discurso: hermenéutica, semiótica y análisis textual. *Anuario de Psicología*, n.59, pp. 19 – 60.

Referencias de la muestra seleccionada

B1_ACOL_chi_015: http://www.mifuturo.cl/index.php/2013-03-06-18-20-53/columnas-de-opinion/355-maxbanados

B1_ERES_col_025:http://medellin.travel/MedellinTravelWeb/locations/117/santa-fe-de-antioquia

B1_INOT_col_013:http://www.elcolombiano.com/deportes/futbol/atletico-nacional-vs-independiente-del-valle-minuto-a-minuto-ME4655187

B1_INOT_ven_024:http://www.ultimasnoticias.com.ve/noticias/politica/maduro-despues-5-anos-me-siento-orgulloso-he-leal-chavez/

B1_NCUE_esp_021: http://ciudadseva.com/texto/el-nino-al-que-se-le-murio-el-amigo/

B2_ACOL_col_016: https://www.elheraldo.co/columnas-de-opinion/una-dosis-de-realidad-468254.%20Consultado%20el%2010/03/2018

B2_EREC_esp_023:
https://www.abc.es/recetas-cocina/receta cocina.asp?cocina=5906&receta=Leche+frita

B2_IENT_esp_008:https://hablacultura.com/cultura-textos-aprender-espanol/artes-escenicas/rocio-molina/

B2_INOT_esp_026:https://www.practicaespanol.com/las-autoridades-cifran-en-26-los-fallecidos-en-el-tiroteo-en-una-iglesia-de-texas/

B2_NCUE_per_028:http://ciudadseva.com/texto/muerte-del-cabo-cheo-lopez/

C1_ACOL_esp_023:
https://elpais.com/elpais/2012/10/11/opinion/1349955088_315730.html

C1_ARES_esp_006:https://historiasbizarrasybizantinas.wordpress.com/2015/06/07/resena-de-la-metamorfosis-de-franz-kafka/

C1_NCUE_arg_032:http://ciudadseva.com/texto/continuidad-de-los-parques/

C1_NCUE_ecu_031: http://linnguagem.com.br/downloads/espanhol/Conciencia-breve.pdf

C1_NREL_esp_024:http://elpais.com/elpais/2015/11/19/opinion/1447948386_644145.html