Lenguaje, vol. 47, núm. 2S, 2019, pp. 437-568.

Problemática actual del procesamiento computacional anafórico: el caso de FreeLing 4.1.

Grajales Ramírez, Andrés Felipe y Molina Mejia, Jorge Mauricio.

ISSN: 2539-3804

Problemática actual del procesamiento computacional anafórico: el caso de FreeLing 4.1[1]

Andrés Felipe Grajales Ramírez
Jorge Mauricio Molina Mejía
Universidad de Antioquia
Medellín, Colombia

Resumen
En este artículo, presentamos el caso del procesamiento computacional automático de la anáfora y de la problemática que genera dicho tratamiento. Para ello, nos hemos fundamentado en el etiquetador *FreeLing 4.1*, y en su módulo para el tratamiento de correferentes textuales; analizando, en este caso, un corpus para la enseñanza del español a estudiantes extranjeros. Con el fin de establecer el análisis y los problemas suscitados por este analizador automático, se ha establecido una definición y una tipología anafórica propias, que surgen a partir de varias lecturas y teorías provenientes del campo de la lingüística textual. El artículo finaliza con el análisis estadístico que muestra los diferentes problemas generados en el reconocimiento de dichos tipos de anáforas y de su implicación en el posterior etiquetado de redes correferenciales.
Palabras clave: lingüística computacional; procesamiento del lenguaje natural; correferencia textual; resolución de la anáfora; lingüística de corpus.

[1] El presente trabajo se enmarca en el proyecto DICEELE (Dispositivo Informático basado en Corpus para la Enseñanza del Español Lengua Extranjera), proyecto del semillero de investigación Corpus Ex Machina (Facultad de Comunicaciones, Universidad de Antioquia), y financiado por el CODI (Comité para el Desarrollo de la Investigación) de la misma universidad. Convocatoria Programática Área Ciencias Sociales, Ciencias Económicas, Humanidades y Artes 2016. N° de Acta CODI: 748 de 2017-07-18. Acta N° 2016-13227.

Recibido: 31-07-2019. Aprobado: 017-10-2019.

Abstract

Current problem of anaphoric computational processing: The case of FreeLing 4.1

In this article, we present the case of automatic computational processing of the anaphora and the problem generated by such processing. To do this, we have relied on the *FreeLing 4.1* tagger, and its module for the treatment of textual coreference; analyzing, in this case, a corpus for teaching Spanish to foreign students. In order to establish the analysis and problems raised by this automatic analyzer, a definition and anaphoric typology of our own has been established, which arise from various readings and theories from textual linguistics field. The article concludes with the statistical analysis showing the different problems generated in the recognition of these types of anaphora and their involvement in the subsequent tagging of coreferential networks.

Key words: computational linguistics; natural language processing; text coreference; anaphora resolution; corpus linguistics.

Résumé

Problématique actuelle dans le traitement informatique de l'anaphore : Le cas de FreeLing 4.1

Dans cet article, nous présentons le cas du traitement automatique des anaphores et de la problématique générée par ce type de traitement informatique. Pour ce faire, nous avons travaillé à partir de l'étiqueteur FreeLing 4.1, et de son module pour le traitement de la coréférence textuelle. Grâce à celui-ci, nous avons analysé un corpus dédié à l'enseignement de l'espagnol vis-à-vis des apprenants étrangers. Ayant pour but d'établir l'analyse et les problèmes suscités par cet analyseur automatique, nous y avons établi une définition, ainsi qu'une typologie personnelle ; qui apparaissent à la suite de nos lectures et de théories provenant du champ de la linguistique textuelle. À la fin, notre article montre l'analyse statistique et les différents problèmes produits dans la reconnaissance de ce type d'anaphores, ainsi que de leur possible implication dans le postérieur étiquetage de réseaux coréférentiels.

Mots-clés : linguistique-informatique ; traitement automatique des langues naturelles ; coréférence textuelle ; résolution de l'anaphore ; linguistique de corpus.

CÓMO CITAR ESTE ARTÍCULO

Grajales, A., & Molina, J. (2019). Problemática actual del procesamiento computacional anafórico: el caso de FreeLing 4.1. *Lenguaje, 47*(2S), 437-568. doi: 10.25100/lenguaje.v47i3.8356

INTRODUCCIÓN

Dentro de la lingüística computacional, y más precisamente en lo concerniente al PLN (Procesamiento del Lenguaje Natural), existen varios fenómenos que han sido estudiados y que en la actualidad merecen nuestra atención, tanto a nivel lingüístico como informático. Se trata entre otros de: el reconocimiento de entidades nombradas (*Named Entity Recognition* – NER, por su sigla en inglés), el análisis de sentimientos (*sentiment analysis*) y, muy particularmente, el estudio de la anáfora. El primer tema y el último se encuentran muy ligados, ya que para que se puedan reconocer y analizar las anáforas, un buen sistema deberá, en gran medida, poder reconocer las llamadas "entidades nombradas", además de pronombres y otras unidades que funcionen como referentes en las llamadas "cadenas correferenciales". Este último tema, el del reconocimiento automático de las anáforas, el cual vendrá de la mano de la llamada "resolución anafórica" (*anaphora resolution*), es el que trataremos a lo largo de este artículo.

Podemos constatar, en primer lugar, que, en lenguas como el inglés y el francés, el tema ha sido bastante estudiado, de tal forma que en la actualidad existen muchos trabajos que abordan el problema del reconocimiento de unidades anafóricas de forma automática. En este sentido, tenemos los trabajos para el inglés de Liddy (1990), Garside, Fligestone y Botley (1997), Denber (1998), Mitkov (1999, 2001, 2002), Mitkov, Evans, Orasan, Ha y Pekar (2007), entre los más importantes. En lo que respecta al campo del francés, existen los trabajos de Depain-Delmotte (1997), Salmon-Alt (2002), Amsili, Denis y Roussarie (2005), Longo y Todiraşcu (2010) y Landragin (2011, 2014, 2018), entre otros. La mayor parte de estos trabajos se enfrentan a los diversos problemas que implican el reconocimiento automático de los correferentes textuales, como es el caso de la anáfora. Entre estos problemas podemos encontrar el de la ambigüedad del lenguaje, el de la falta de reconocimiento de algunas entidades nombradas y la ausencia o no reconocimiento de muchos pronombres, etc.

En lo que concierne al español, hemos podido constatar que son varios los trabajos que abordan esta problemática, teniendo, entre otros, las tesis doctorales de Ferrández (1998) "*Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos*", Martínez-Barco (2001) "*Resolución computacional de la anáfora en diálogos: estructura del discurso y conocimiento lingüístico*", Morales (2004) "*Resolución automática de la anáfora indirecta en el español*", y Tordera (2010) "*Lingüística computacional y anáfora*". También la tesis de maestría de Olivas (2006) "*Sistema de construcción de redes semánticas con detección de anáfora*". Tenemos, por otra parte, los artículos de Salguero y Soler (2010), Carrión (2014), entre otros. Todos estos trabajos, como lo veremos más adelante (*cf.* "Lingüística

computacional y resolución de la anáfora" y "metodología"), propenden por el estudio de la anáfora desde diversas perspectivas y proponen diferentes soluciones ante su posible resolución.

El objetivo primordial del presente artículo consiste en identificar y analizar los diferentes problemas existentes en el tratamiento computacional de la anáfora en español, utilizando para ello herramientas de uso libre. Para tal fin, se partió del estudio de un analizador automático del lenguaje natural ya construido. Se trata del etiquetador morfosintáctico *FreeLing 4.1*[2], y para efectuar el análisis se empleó un corpus dedicado al aprendizaje de ELE (Español Lengua Extranjera). El estudio de este analizador se realizó por medio de la comparación de las etiquetas que arroja el sistema con las nuestras realizadas en un etiquetado manual.

Con el fin de lograr el objetivo antes señalado partiremos del trabajo de Grajales (2019), trabajo en el cual el autor tuvo que considerar el diseño de una nueva tipología de anáforas. Este trabajo surge dentro del marco del proyecto DICEELE (Molina, Grajales y Pemberty, 2019), y busca ayudar en la solución al etiquetado correferencial del corpus para la enseñanza de ELE de este proyecto. Del trabajo de Grajales (2019) se retomarán aquí la tipología por él creada, el análisis de la herramienta *FreeLing 4.1*, y algunos elementos de programación, como lo es el caso de un algoritmo desarrollado a partir del lenguaje *Python*, que sirvió durante el proceso de etiquetado semiautomático de las anáforas del corpus de muestra (*cf.* anexo A). No obstante, en el presente artículo no se incluirán los resultados de dicha problemática, puesto que nos centraremos de manera más específica en el análisis de las anáforas a partir de la herramienta antes mencionada.

Anáfora

En un principio, es importante que definamos la concepción de anáfora con la cual hemos trabajado en este proyecto. Como es común en el ámbito de la lingüística, especialmente en la que trabaja el texto y el discurso, aún no existe un consenso a partir del cual podamos definir lo que puede ser o no una anáfora dentro de un corpus[3]. Con base en diferentes autores que abordan el fenómeno anafórico (Cuenca, 2010; Ferrari, 2014; Peña, 2006; Saiz, 2002), consideramos anáfora al mecanismo que se utiliza para retomar un elemento que ha sido mencionado con anterioridad en un texto. Esto significa que la anáfora solo se presenta cuando el referente está dentro del contexto lingüístico (*endofórico*) y no fuera de él (*exofórico*). Dentro de las relaciones endofóricas también encontramos la catáfora, que se diferencia de la anáfora en que el elemento fórico se conecta con un elemento del contexto

[2] http://nlp.lsi.upc.edu/freeling/index.php/node/1.

[3] Por ejemplo, se suele considerar la anáfora, en unos casos, como un tipo de deixis y, en otros, como un elemento opuesto a la deixis en cuanto esta última solo se refiere a lo extralingüístico (García, 2011).

subsiguiente. Sin embargo, para el análisis del tratamiento computacional no tuvimos en cuenta la catáfora, a pesar de que sí hace parte del proyecto DICEELE.

Utilizamos esta definición debido a que en el proceso de identificación de anáforas de nuestro corpus nos dimos cuenta de que no podíamos partir de una sola concepción sobre este fenómeno, pues muchos casos que usaban el mecanismo no se tomarían en cuenta. Por ejemplo, con base en Cuenca (2010), no consideraríamos como anáforas los casos de pronombres de primera persona, pues estos pertenecerían a la deixis y no a la correferencia textual. Igualmente, las anáforas adjetivales que mencionan Saiz (2002) y Olivas (2006) no se tomarían en cuenta si nos restringimos a las definiciones que las consideran tipos de elipsis, tal como aparece en Cuenca (2010).

Suele hacerse, además, la distinción entre anáfora directa y asociativa. Por un lado, las anáforas directas son aquellas que toman un referente específico del texto. Por otro lado, en las anáforas asociativas el referente es generado por el contexto, sea por asociación semántica del léxico o por el conocimiento del mundo (Ferrari, 2014). En los siguientes ejemplos[4], se observan ambos casos. En el primero, el pronombre 'la' directamente se relaciona con 'la película' (lo cual aporta a la cohesión del texto), mientras que, en el segundo, el sintagma 'los espectadores' se relaciona con 'la película' por medio de un campo semántico (aportando a la coherencia textual):

- **La película** está en proceso de filmación. **La** veremos en cartelera para el próximo año.
- **La película** está en proceso de filmación. **Los espectadores** serán testigos de algo único.

Dicho esto, a partir de los datos encontrados en el corpus sobre la anáfora y de autores que la clasificaran (Ferrari, 2014; Olivas, 2006; Saiz, 2002), construimos una tipología que nos permitió clasificar y tener nuestra propia concepción sobre lo que podíamos considerar en el proyecto[5] como elementos anafóricos. La Tabla 1 resume esta concepción en los tipos de elementos gramaticales que pueden ser considerados, por una parte, como antecedente anafórico, es decir, a lo que puede referirse una anáfora y, por otra parte, los elementos que pueden ser una anáfora:

[4] Ambos extraídos del trabajo de Grajales (2019, p. 14).

[5] Para los ejemplos, véase el Anexo B.

Tabla 1. Tipos de correferentes según Saiz (2002), Olivas (2006), Ferrari (2014) y la información del corpus

Tipos de antecedente	Subtipo	Tipos de anáfora	Subtipo
Nominal	Sintagma nominal **(SN)** Nombre propio **(NP)**	**Pronominal**	Sujeto implícito (**imp**), pronombre personal (**PP**), posesivo (**pos**), relativo (**rel**), demostrativo (**dem**), indefinido (**ind**)
Verbal	Sintagma verbal (**SV**)	**Nominal**	Nombre propio (**NP**), nombre común (**NC**), sintagma nominal (**SN**)
Enunciado	Frase(**fras**), oración (**orac**), secuencia de oraciones (**sec-orac**)	**Verbal**	**Proverbo**
		Adverbial	Adverbio temporal (**temp**), locativo (**loc**), de modo (**mod**)
		Adjetival	Adjetivo (**adj**), superficial numérica (**num**), posesivo (**pos**), elipsis (**elip**)
		Encapsuladora	**(resumativa)**

Lingüística computacional y resolución de la anáfora

Como se ha dicho en la introducción de este artículo, el interés de la lingüística computacional y del PLN en lo que respecta a la resolución de la anáfora es de suma importancia. Esto se debe, principalmente, al hecho de que las anáforas hacen parte de aquellos elementos que permiten que un determinado texto escrito posea o no claridad, en términos de cohesión textual (Adam, 2011; Mitkov, 2001) y/o discursiva (Lundquist, 2013); evitando, además, las posibles repeticiones de una palabra o grupo de palabras, como lo vimos en el apartado anterior.

El interés de la lingüística computacional por este fenómeno de índole textual se halla, por lo tanto, en su reconocimiento y posible marcado de manera automática; para ello, se deben confrontar varios de los problemas que implica dicho reconocimiento. Por lo cual, el análisis de la resolución de la anáfora, de manera informática, se hace indispensable antes de abordar cualquier posible etiquetado correferencial en un corpus determinado.

Para algunos autores como Mitkov (2001), por ejemplo, la anáfora es un elemento de suma importancia dentro de la lingüística computacional y el PLN. Según este autor, "la resolución de la anáfora es una tarea esencial para las interfaces en lengua natural, la traducción automática, la generación automática de resúmenes, la extracción de información y en otro gran número de aplicaciones del PLN" (p.

110)[6]. También se pueden aplicar, por ejemplo, a los sistemas informáticos para el ALAC (Aprendizaje de Lenguas Asistido por Computadora), para el diseño de actividades didácticas (Molina, 2015). Pero a su vez, autores como Mitkov (2001), Landragin (2011) consideran que se trata de un elemento que presenta una serie de dificultades en lo que respecta a su resolución a nivel informático.

Según Mitkov (2001), la resolución de la anáfora necesita de extensos corpus etiquetados de manera manual que sirvan para el aprendizaje de máquina, lo que implica una fuerte demanda de recursos tanto de expertos con conocimientos lingüísticos como en el campo del texto escrito (o de ambos a la vez); en otros términos, esto conlleva a contar con un considerable esfuerzo humano en el reconocimiento de este tipo de unidades textuales. No obstante, como lo indica el mismo Mitkov (2001), el costo que esto implica ha llevado a muchos investigadores a optar por soluciones menos precisas desde el punto de vista del lenguaje; lo que ha conducido irremediablemente a lo que el autor considera como "*knowledge-poor anaphora resolution strategies*" (p. 110)[7].

En el caso de Landragin (2014), este autor propone la idea de que, en lo que tiene que ver con el procesamiento informático de la anáfora y de las cadenas correferenciales, estas demandan, en primer lugar, estructuras de datos complejas que permitan cubrir textos enteros; y, en segundo lugar, la necesidad de construir todo un complejo entramado de redes correferenciales (anáforas y catáforas), lo que implica, para el autor, la modelización de los lazos establecidos entre referentes y correferentes. En este sentido, se hace necesario un trabajo manual o semiautomático, en los términos de Mitkov (2001), con el fin de tener corpus anotados que permitan el posterior entrenamiento de un sistema de reconocimiento anafórico.

En lo que respecta al estudio de la anáfora en español, y de su posible resolución, como lo pudimos observar en la introducción, existen diversos trabajos que se interesan por esta problemática. De estos estudios cabe destacar, en primer lugar, una tesis doctoral que aborda el tema y que tiene más de veinte años de haber sido presentada; se trata del trabajo elaborado por Ferrández (1998), y cuyo análisis se enmarca en el tratamiento computacional de las anáforas pronominales y de tipo adjetival, empleando, para ello, gramáticas de unificación de huecos. Por su parte, Morales (2004) nos presenta otra tesis doctoral que estudia el tema de la resolución

[6] Traducción nuestra de: "*anaphora resolution is a key task in natural language interfaces, machine translation, automatic abstracting, information extraction, and in a number of other NLP applications*" (Mitkov, 2001, p. 110).

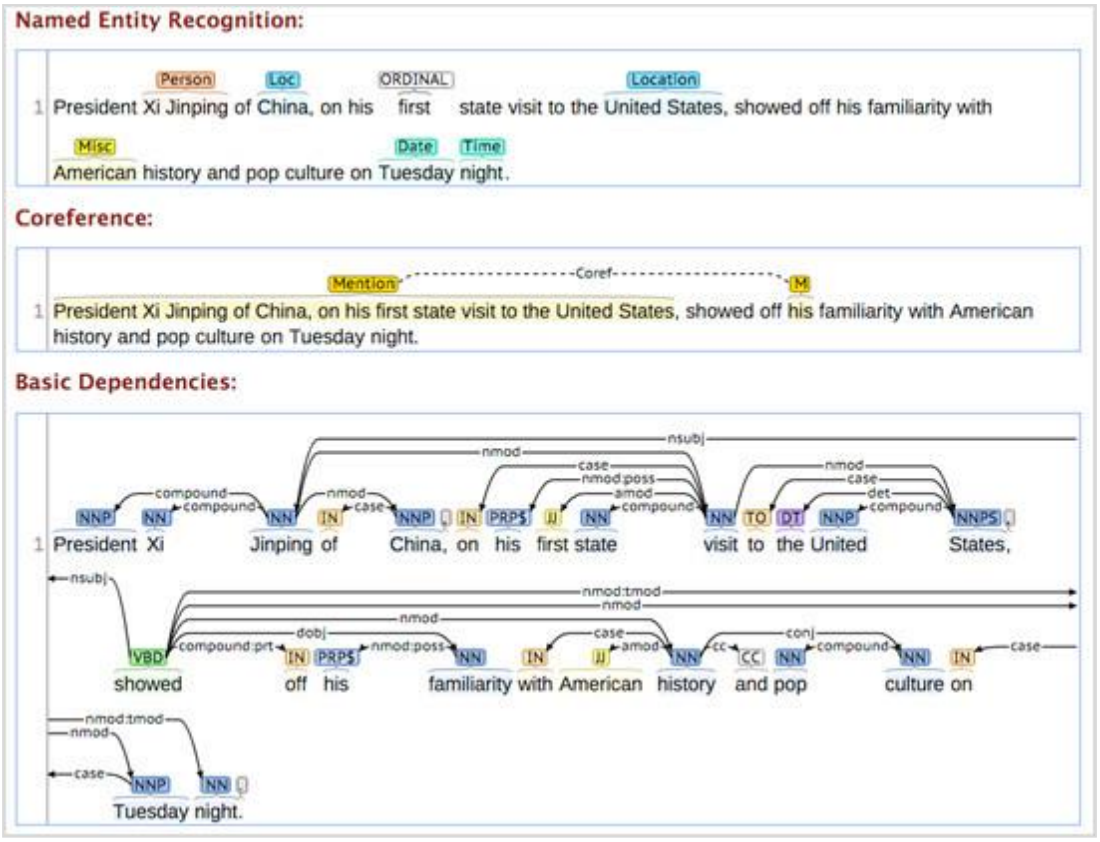
[7] Traducción nuestra: "Estrategias de resolución de la anáfora pobres en el conocimiento" (Mitkov, 2001, p. 110).

automática de la anáfora indirecta en el español; y, finalmente, Tordera (2010) presenta su tesis en la que trata de manera global el tema de la anáfora dentro de la lingüística computacional. Tenemos, por otra parte, los artículos de Salguero y Soler (2010), y Carrión (2014), que se interesan, el primero en la resolución de anáforas pronominales por medio de procedimientos abductivos (es decir, que necesitan de un razonamiento de tipo explicativo). En el segundo caso, tenemos el trabajo de Carrión (2014), que propone la utilización de una herramienta (*FunGramKB*)[8] que ha sido enriquecida con conocimientos de índole cultural. Lo que permitiría, según la autora, mejorar el reconocimiento de los referentes textuales y, por ende, del reconocimiento anafórico subsecuente.

Para terminar esta sección, tenemos el caso de algunos diseñadores de herramientas, como el *Stanford Parser* (etiquetador de la Universidad de Stanford) (Manning *et al*., 2014), que se han preocupado por poder permitir el reconocimiento de entidades nombradas, de pronombres y de sus respectivas redes de correferencia textual, a partir de un análisis anafórico (Figura 1). Con el fin de realizar dicho proceso, el mencionado analizador permite el reconocimiento de nombres simples y compuestos (comunes y propios), de pronombres personales y complementos, etc. El análisis de dichas unidades lingüísticas es generado a partir de un proceso de etiquetado que proviene del verbo y de las relaciones establecidas alrededor de este; esto se logra gracias a la fundamentación de este sistema en la gramática de dependencias[9]. En este sentido, el sistema permite el reconocimiento y desambiguado de estas unidades, lo que permite reconocer de manera más fiable las posibles anáforas, como puede observarse en la Figura 1.

[8] http://www.fungramkb.com/.

[9] La gramática de dependencias se fundamenta, principalmente, en los trabajos de Tesnière (1959) "*Éléments de syntaxe structurale*", y Mel'čuk (1988) "*Dependency Syntax: Theory and Practice*".



546

Figura 1. Muestra del análisis de la correferencia y las entidades nombradas, a partir de una gramática de dependencias, *Stanford Parser* (Manning *et al*., 2014)

Otras herramientas, como es el caso de *FreeLing*, para el español, serán presentadas en la siguiente sección. En este sentido, nos dedicaremos a analizar los diferentes problemas que presenta esta herramienta en lo relativo a la resolución de las anáforas, así como otros aspectos metodológicos concernientes a nuestro estudio.

METODOLOGÍA

Freeling 4.1

Recordemos que el objetivo principal de este trabajo fue identificar los problemas que podemos encontrar hoy en día a la hora de procesar computacionalmente la anáfora. Consideramos que el conocimiento que tuviésemos sobre la anáfora no sería suficiente para definir las dificultades que tendría un ordenador para procesarla. Por tal razón, decidimos buscar un *software* que analizara el lenguaje natural y que fuera capaz de procesar relaciones anafóricas en español. En el proceso de revisión bibliográfica nos dimos cuenta de que muchos eran los trabajos relacionados con la resolución de la anáfora (por ejemplo, Martínez-Barco, 2001; Morales, 2004; Olivas, 2006), pero eran pocos los que desembocan en el desarrollo

de *software*, por lo cual se vio entorpecida nuestra búsqueda. Por último, debíamos ser cuidadosos con la elección del programa, pues la problemática que surgiese de nuestro análisis dependería del *software* que escogiéramos, puesto que cada uno analiza la información con diferentes recursos[10].

Teniendo esto en cuenta, para este análisis seleccionamos el analizador de lenguaje *FreeLing* y su herramienta para analizar correferencias. Este programa es una librería de uso libre que puede ser utilizada fácilmente desde diferentes sistemas operativos y es capaz de procesar correferencias para el español, lo cual dio fin a nuestra búsqueda. Además de esto, la versión 4.1 de *FreeLing* trabaja con un sistema llamado *RelaxCor* (Sapena, Padró y Turmo, 2013), el cual se implementó en su programación para mejorar el análisis de correferencias. Este *software*, también desarrollado por el grupo TALP, ha participado en competencias para evaluar su efectividad en el procesamiento de correferencias en las cuales ha alcanzado muy buenos resultados[11]. Por estas razones decidimos seleccionar estas herramientas para el análisis, pues resultaron ser las más adecuadas y accesibles y con ellas procesamos nuestra muestra del corpus.

El corpus DICEELE

Como lo hemos mencionado anteriormente, el proyecto DICEELE se basa en el procesamiento de corpus para identificar los fenómenos textuales que pueden trabajarse en un curso 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 diferentes textos estuvo guiada por ciertos criterios específicos, teniendo en cuenta que se trata de un corpus especializado para la enseñanza de 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:** El corpus DICEELE se rige por brindar material para la enseñanza y el aprendizaje de la lengua que represente el uso real de esta, según los lineamientos de la lingüística de corpus. Por lo tanto, no posee textos adaptados o didactizados por profesores.
- **Textos clasificados según el MCER dentro de los niveles B1, B2 y C1:** Consideramos que a partir del nivel B1 los textos dejan de estar limitados a

[10] Esto puede observarse en Ruiz (2012) y Carrión (2014), los cuales analizan los alcances para el procesamiento de correferencias por parte de *FunGramKB* el cual utiliza información cultural para la resolución de la ambigüedad correferencial.

[11] 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).

situaciones inmediatas y campos semánticos desconectados y comienzan a hacer referencia a discursos más complejos (Molina *et al.*, 2019). Se procuró que los textos estuvieran previamente clasificados o que hubieran sido utilizados para la enseñanza en instituciones o portales en línea de ELE[12].

- **Textos cortos de diferentes géneros:** Cada nivel del MCER trabajado debía poseer una cantidad similar de textos informativos, narrativos, argumentativos y explicativos. Al mismo tiempo, cada género tendría también variedad en cuanto al tipo de texto. Es decir, los textos narrativos del nivel C1, por ejemplo, no debían ser exclusivamente cuentos sino también fábulas, mitos, relatos, etc. La extensión de estos también debía ser similar.
- **Textos de distintas fuentes:** También se procuró que los textos no fueran extraídos de una misma página web o de un solo libro de texto, por ejemplo, para privilegiar su variedad. Así mismo, se recolectó material de distintos países hispanohablantes con el fin de no centralizar una sola variedad del español. Cabe aclarar que se seleccionaron textos que no infringieran los derechos de autoría.

Otra de las características del corpus DICEELE es que posee un etiquetado morfosintáctico (*PoS-tagging*) efectuado con el etiquetador de *FreeLing*. Este consiste en la asignación de una etiqueta que contiene información acerca de la categoría gramatical, las flexiones, las conjugaciones, etc., de cada palabra del texto. Para ser más exactos, inicialmente, el *PoS-tagging* divide el texto en unidades de análisis llamadas *tokens*. En este sentido, algunos autores como Baker, Hardie y McEnery (2006) definen los *tokens* como "una sola unidad lingüística, la mayoría de las veces una palabra, aunque dependiendo del sistema de codificación utilizado, una sola palabra se puede dividir en más de un *token*, por ejemplo, he's (he + 's)" (p. 159)[13]; además de las palabras, también se incluyen, en muchos sistemas, los signos de puntuación (Brezina, 2018). Otros autores consideran, por su parte, que dentro de la frecuencia de *tokens* también se incluyen secuencias de unidades multipalabra (*multiword sequences*) (McEnery y Hardie, 2011). Posteriormente, el sistema se encarga de asignar a cada uno de los *token* una etiqueta de carácter morfosintáctico. En términos computacionales, este proceso de *tokenización* facilita en cierta medida

[12] Sin embargo, se realizaron clasificaciones para algunos textos 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.

[13] Nuestra traducción de: "*A single linguistic unit, most often a word, although depending on the encoding system being used, a single word can be split into more than one token, for example he's (he + 's)*" (Baker *et al*., 2006, p. 159).

el PLN puesto que separa cada uno de los elementos de un texto y les asigna un identificador único o *id*, el cual utilizamos para el posterior etiquetado manual de anáforas. El formato de este etiquetado se presenta en XML, como podemos observar en la Figura 2 para los *tokens* t3.7 y t3.8:

```
<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 2. Ejemplo de dos *tokens* del corpus ("en" / "Alemania") etiquetados en XML

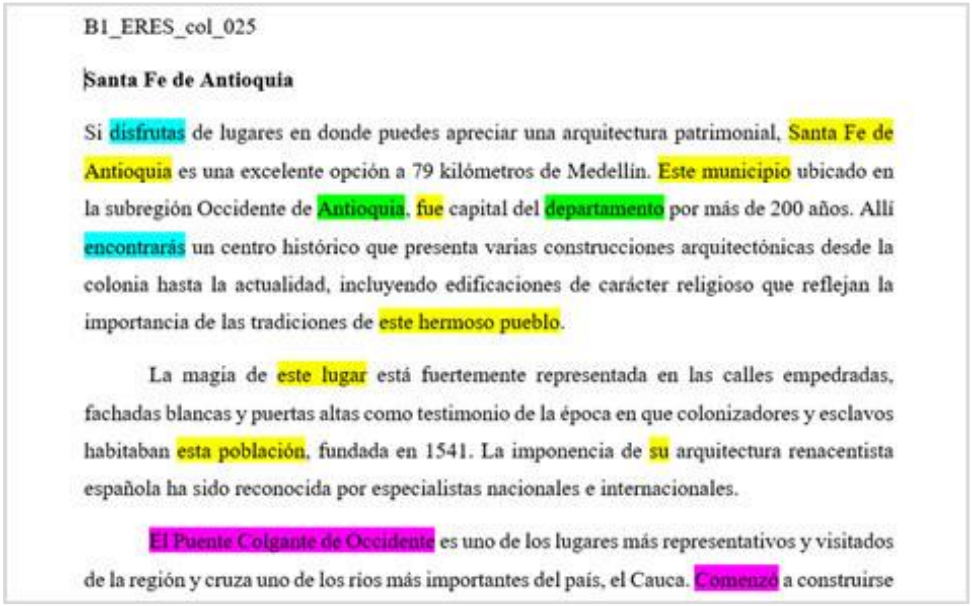
De esta manera, para el análisis del tratamiento que *FreeLing* realizaría de nuestro corpus, seleccionamos una muestra que respetara la variedad y las características del corpus. La muestra que se conformó para esta investigación se halla compuesta por 15 textos clasificados en los niveles B1, B2 y 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 dichos textos varía entre 300 y 800 palabras y, en total, la muestra está compuesta por 6094 palabras y 120 párrafos. Con el fin de poder obtener un mejor manejo de los datos, los archivos que contienen los textos fueron nombrados con información codificada acerca del nivel, el género y el tipo de texto, su país de origen y la 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.

Etiquetado manual de anáforas

Con la intención de comparar el procesamiento de la correferencia por parte del *software*, debíamos realizar un etiquetado manual de todas las anáforas presentes en la muestra seleccionada. Este debía garantizar la correcta identificación de los elementos con el fin de poder responder a nuestras intenciones investigativas. Para ello, usamos como base los principios expuestos por Navarro (2007) que permiten realizar un etiquetado efectivo: rapidez, consistencia y profundidad. Con base en esto, la anotación de anáforas debía ser lo más simple posible: procurar la ayuda del ordenador, crear un acuerdo entre anotadores (en caso de ser varios) y reflejar datos relevantes de la lengua a partir de 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). Así, a partir de lo investigado sobre la anáfora, identificamos y seleccionamos esta noción dentro del corpus. Utilizamos como fundamento la tipología expuesta anteriormente tanto para la selección de anáforas por etiquetar como para su correcta clasificación.

El proceso consistió en una primera fase de identificación, en la cual se iban buscando las relaciones entre posibles anáforas y antecedentes. Una vez ubicado el antecedente que correspondía con una anáfora, revisábamos si las expresiones anafóricas subsiguientes también se referían al mismo antecedente; en caso de que no fuera así, buscábamos otro posible antecedente para esta. Cada vez que hallábamos una nueva relación entre anáfora y antecedente, la marcábamos con un nuevo color, con el fin de poder diferenciarla de las demás. De esta manera, constituimos las cadenas de correferencia de cada uno de los textos. Recordemos que estas cadenas son el conjunto de anáforas que correfieren con un mismo referente (Sapena *et al*., 2013). Por ejemplo, en la Figura 3, el primer elemento que encontramos resaltado con color amarillo 'Santa Fe de Antioquia' es el referente de los siguientes elementos del mismo color, los cuales corresponden a expresiones anafóricas:


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 3. Selección manual de cadenas de correferencia en un el documento B1_ERES_col_025

Finalmente, con el fin de ayudar en la representación de las relaciones anafóricas presentes en el corpus, empleamos el lenguaje de balizado XML. El lenguaje XML resulta muy útil para este tipo de tareas pues permite anotar cualquier noción y convertirla en información explotable para una computadora. El primer

paso para etiquetar las anáforas en XML era localizar los *ids* que identificaban los elementos marcados como correferentes en la fase anterior. De esta forma, se podía utilizar la información ya anotada para apoyar la construcción de las nuevas etiquetas correferenciales. En otras palabras, buscamos los atributos *tag* y *form* del etiquetado morfosintáctico previo (véase Figura 2) para extraer las etiquetas morfosintácticas y las palabras correspondientes a cada correferencia según el *id*, con el fin de ahorrar tiempo en la anotación manual mediante un algoritmo desarrollado en lenguaje *Python* que reescribiera dicha información (ver anexo A).

El diseño del etiquetado anafórico en XML fue desarrollado con base en los trabajos de Molina (2015) y Navarro (2007), los cuales utilizan un esquema de anotación similar al empleado en la *Message Understanding Conference* (MUC), como lo muestra Mitkov (2002). Estas etiquetas 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 *FreeLing* que establece otro tipo de etiquetado agrupando las cadenas de correferencia y reescribiendo la información a partir de los *ids*. (Figura 4). Esto facilitó, en gran medida, el trabajo de comparación de etiquetado, pues poseían la misma estructura:

```
2660 <chain id="1">
2661   <REF ANT="nom" type="NP" id="2" words="Santa Fe de Antioquia "/>
2662   <COREF ANA="nom" type="SN" id="4" words="Este municipio "/>
2663   <COREF ANA="adv" type="loc" id="7" words="Allí "/>
2664   <COREF ANA="nom" type="SN" id="9" words="este hermoso pueblo "/>
2665   <COREF ANA="nom" type="SN" id="10" words="este lugar "/>
2666   <COREF ANA="nom" type="SN" id="11" words="esta población "/>
2667   <COREF ANA="pron" type="pos" id="12" words="su "/>
2668   <COREF ANA="nom" type="NP" id="16" words="Santa Fe de Antioquia "/>
2669   <COREF ANA="nom" type="SN" id="17" words="El pueblo "/>
2670 </chain>
```

Figura 4. Resultado de una cadena etiquetada en XML como en *FreeLing*

De esta manera, una vez etiquetadas todas las relaciones anafóricas de nuestra muestra, podríamos compararlas con la anotación del *software* que escogimos. Además, estas etiquetas también nos permitieron caracterizar el corpus a partir de los tipos y cantidades de correferentes. En total, trabajamos con **788 expresiones anafóricas** que se referían a **202 referentes**, constituyendo así **198 cadenas de correferencia**. La totalidad de estos elementos representaría un 20 % de todas las palabras de la muestra, teniendo en cuenta que muchas de las unidades correferenciales se componen por *multiword tokens* o secuencias con más de una palabra (locuciones, sintagmas, enunciados, párrafos, etc.)[14]. Estos datos generales

[14] Algunos autores como Constant *et al*. (2017) proponen, por su parte, la noción de "*Multiword Expressions*" o "expresiones multipalabra", denominación que estaría en concordancia con el uso de

sugieren la importancia de estos elementos en la cohesión de los textos y la posibilidad de revisarlos detalladamente con el fin de comprender su aplicación en el PLN y en las clases de ELE.

PROBLEMAS ENCONTRADOS EN EL PROCESAMIENTO

El primer paso para identificar los errores automáticos fue procesar todos nuestros textos por el *software FreeLing 4.1* y su análisis de correferencias. Seguidamente, basados en la previa anotación manual de las relaciones anafóricas, señalamos cada uno de los errores y aciertos que cometía el sistema, tal como lo sugiere Navarro (2007): "comparar las anáforas detectadas y anotadas automáticamente con las anáforas detectadas, anotadas y validadas por humanos" (p. 166)[15]. Finalmente, consignamos esta información en nuevas etiquetas de los archivos XML, donde ya teníamos las anáforas clasificadas, con el fin de reunir los datos, por medio del *software AntConc* versión 3.5.8 (Anthony, 2018), sobre los errores en cada tipo de anáfora.

Así pues, examinamos los resultados de errores del tratamiento correferencial por medio de *FreeLing*, describiendo tanto los casos más representativos de problemas como los más particulares, pues todos ellos representan la imposibilidad de procesar la anáfora correctamente con la tecnología actual. A continuación, referiremos las causas que provocan los errores en el procesamiento de la anáfora y los porcentajes de error para cada tipo y subtipo de esta. También explicaremos algunos casos para ejemplificar el comportamiento del sistema que lo llevó a no identificar correctamente una unidad.

En primer lugar, observemos la Tabla 2, que resume los resultados de comparar la anotación manual con la automática y muestra sus porcentajes:

Tabla 2. 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 %

tecnologías para el Procesamiento del Lenguaje Natural, y su empleo en la detección y tratamiento computacional de locuciones complejas y grupos nominales, por ejemplo.

[15] En el marco del trabajo realizado por Grajales (2019), mencionado en la introducción del presente artículo, la validación de las anáforas encontradas fue efectuada por parte del profesor Jorge Molina (dedicado al tema de la correferencia textual, como puede verse en Molina, 2015).

A partir de estos datos, nos damos cuenta de que el porcentaje de error, en general, es bastante elevado, por lo cual podemos inferir, a primera vista, que vale la pena profundizar en las razones de estos resultados. Por otra parte, observamos que los resultados son mejores para los referentes/antecedentes que para los correferentes/anáforas. Creemos que esto se debe a que la mayoría de los referentes en la muestra son de tipo nominal (92 %), los cuales, por un lado, son semánticamente más plenos, pero, por otro lado, estos sistemas suelen basarse en el reconocimiento de entidades nombradas (NER), estas últimas se contraponen a los hechos o ideas, como veremos a continuación.

Reconocimiento de referentes

En cuanto al tratamiento de referentes/antecedentes, hallamos que es evidentemente más efectivo para el tipo nominal que para los otros dos, como observamos en la Tabla 3.

Tabla 3. 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 %

Esto puede explicarse desde la construcción del sistema *RelaxCor*, con el cual funciona *FreeLing 4.1*. Una anáfora suele referirse a una entidad o a un hecho, como se ve en los tipos de antecedentes que expusimos anteriormente. Los hechos o ideas –que corresponden a los enunciados y antecedentes verbales– son unidades que no poseen una estructura sintáctica fija y que solemos delimitar por la comprensión del contexto, por lo cual, estas unidades son más difíciles de identificar que las entidades nominales. Además, la lectura de un artículo que explica el funcionamiento de *RelaxCor* (Sapena *et al*., 2013) nos sugiere que este sistema no está programado para buscar este tipo de referentes: "en este artículo no nos ocupamos de la correferencia que involucra hechos y solo nos enfocamos en la correferencia entre entidades"[16] (p.

[16] Traducción nuestra de: "*In this article, we do not deal with coreference involving events, and focus only on entity coreference*" (Sapena *et al*., 2013, p. 848).

848). Por lo tanto, con base en los datos, concluimos que *RelaxCor* solo reconoce entidades nombradas.

Por último, aunque nuestro objeto de estudio sea la anáfora, no podemos dejar a un lado el análisis del antecedente, pues de su correcta identificación depende buena parte de la resolución anafórica. Por ejemplo, las anáforas por repetición (las cuales utilizan las mismas palabras que el antecedente) presentan un porcentaje de acierto del 98 %, puesto que el sistema los relaciona directamente.

En lo siguiente, observaremos la frecuencia de error por cada tipo y subtipo de anáfora y comentaremos algunos ejemplos de casos que no fueron resueltos por *FreeLing*, a partir de los cuales deducimos y resumimos la problemática para el tratamiento anafórico.

Reconocimiento de anáforas pronominales

El tipo de anáfora pronominal[17] se considera el más común en español, además se le atribuye la dificultad de su resolución a que posee poca carga semántica (Saiz, 2002). En general, la cantidad de anáforas pronominales que logró reconocer y relacionar el sistema (218/491) se acerca al porcentaje general de reconocimiento de anáfora (Tabla 4). Sin embargo, entre los subtipos de estas unidades, gran parte de estas fueron poco reconocidas. Es decir, las anáforas de pronombre personal (la más común) y de relativo presentan valores más altos que las otras categorías. Esto puede deberse a que fue más común encontrarlas muy cerca de su antecedente o del correferente anterior en la cadena correspondiente. De esta manera, el pronombre se le asigna al posible antecedente más cercano, según su sistema de reglas.

Tabla 4. 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 %
Pronominales en total	47 %	53 %

[17] En nuestra muestra, el número de anáforas pronominales representa el 62 % del total de correferentes (788). El subtipo de anáfora pronominal más común resultó ser la de pronombre personal, seguida de la de pronombre implícito, posesivo, relativo, *resumativo* y demostrativo.

Otra de las razones por las que la anáfora de pronombre personal resultó ser más identificada que las demás fue porque, frecuentemente, los mismos PP solían repetirse. Esto es, por ejemplo, una vez se utilizaba el pronombre '*él*' para un referente, el mismo volvía a usarse varias veces en el texto, por lo que el sistema los relacionaba por simple repetición. Sin embargo, esto también genera un problema de identificación, dado que si aparece un nuevo referente al cual también se le aplica este pronombre, se genera un caso de ambigüedad para la máquina. Por otro lado, el alto porcentaje de acierto del pronombre relativo '*que*' se puede atribuir a que la mayoría de las veces aparece junto a su antecedente; cuando no lo está, el sistema se lo asigna a otra unidad más cercana.

Dicho lo anterior, revisemos el siguiente ejemplo en el cual no fue reconocida una anáfora pronominal por medio del *software*. El siguiente caso fue seleccionado porque representa uno de los problemas más comunes para la identificación de la anáfora pronominal. Por límites de extensión, en este artículo no expondremos ejemplos de todos los problemas que encontramos, sino los más representativos o complejos. Al final, resumiremos la problemática que surgió de nuestro proyecto.

- *Ahora mismo nos hacen los presupuestos en Alemania y* la ropa *en China.* La ropa *nos cae bien, pero* […] *Nos daría igual que los presupuestos nos los hiciera* ***Merkel*** *si* ***ella*** *pusiera también la pasta*…[18]

El fragmento anterior parece un caso de sencilla resolución para el lector humano, pues el referente y el correferente están prácticamente contiguos y el PP '*ella*' indica, la mayoría de las veces, la presencia de una anáfora. No obstante, el análisis automático falló en esta identificación y relacionó el pronombre con otra unidad anterior con la cual encontraba más posibilidades de concordancia de género: '*la ropa*'. Dado esto, realizamos una prueba en la cual etiquetábamos solamente esta oración: la resolución fue efectiva ya que no tenía otra opción. Lo que deducimos de este caso, es que era necesario conocer que '*Merkel*' pertenece al apellido de la canciller alemana Angela Merkel, por lo cual es un nombre femenino y puede correferir con el PP femenino '*ella*'. Es probable que *FreeLing* no posea exactamente esta información de tipo cultural en su base de datos para resolver la correferencia[19].

[18] Fragmento extraído del documento C1_ACOL_esp_023. Disponible en: https://elpais.com/elpais/2012/10/11/opinion/1349955088_315730.html.

[19] 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).

Reconocimiento de anáforas nominales

En cuanto al tipo de anáfora nominal, en comparación con los resultados para la anáfora pronominal, el tratamiento de este fue más acertado, como puede apreciarse en la Tabla 5[20].

Tabla 5. 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 %

En primer lugar, destaca que el porcentaje de nombres propios identificados es el más alto de todos los subtipos de la muestra, no solo de los nominales. Esto puede explicarse porque la mayoría de correferentes como nombres propios se hacen por medio de repetición (parcial y exhaustiva). Es decir, cuando nos encontramos con una anáfora de nombre propio es bastante probable que sea el mismo NP utilizado en el referente (repetición exhaustiva), o una parte de este (repetición parcial). En segundo lugar, el porcentaje de SN identificados por el sistema también presenta buenos resultados. Este resultado también puede deberse a que gran parte de los elementos nominales de nuestra muestra utilizaban palabras del referente.

A continuación, exponemos un ejemplo de cadena etiquetada en XML. Observemos las unidades anafóricas resaltadas en negrita dentro del atributo *words*:

[20] Recordemos que, en nuestra muestra, las anáforas nominales representan el 33 % de la totalidad de anáforas (casi la mitad de las pronominales). Dentro de estas, la más frecuente es la de tipo sintagma nominal, seguida de los nombres propios y los nombres comunes.

- <chain id="1">
  <REF ANT="nom" type="NP" id="2" words="Santa Fe de Antioquia"/>
  <COREF ANA="nom" type="SN" id="4" words="**Este municipio**"/>
  <COREF ANA="nom" type="SN" id="9words="**este hermoso pueblo**"/>
  <COREF ANA="nom" type="SN" id="10" words="**este lugar**"/>
  <COREF ANA="nom" type="SN" id="11" words="**esta población**"/>
  <COREFANA="nom" type="NP" id="13" words="Santa Fe de Antioquia"/>
  <COREF ANA="nom" type="SN" id="17" words="**El pueblo**"/>
  </chain>[21]

En este ejemplo, los elementos que destacamos no fueron reconocidos por el analizador computacional. El único elemento identificado correctamente en esta cadena fue el que correspondía con el mismo referente '*Santa Fe de Antioquia*' por repetición. Para darse una resolución efectiva de los demás elementos, el sistema tendría que tener acceso a la información semántica y cultural que se relaciona con '*Santa Fe de Antioquia*' como *municipio*, o *pueblo* de Antioquia, el cual puede relacionarse de manera hiponímica con *lugar* y sinonímicamente con *población*, *poblado* o *localidad*, por ejemplo. En nuestra lectura como hablantes de la lengua, esta es la información que nos permite relacionar el referente de tales anáforas. Por el contrario, la ausencia de esta información representa un obstáculo para el *software*.

El caso del ejemplo anterior representa el principal problema para la resolución de correferencia nominal. Identificamos muchos casos similares en los que la información de este tipo era necesaria para que fueran reconocidas debidamente. Pongamos algunos ejemplos de elementos nominales que no fueron debidamente identificados por el sistema: '*el rey de copas*' y '*el verde paisa*' como epítetos para '*Atlético Nacional de Medellín*'[22], los cuales precisarían de conocimiento cultural para su resolución. De igual manera, con '*el jefe de estado*' para '*Nicolas Maduro*', '*el país*' para '*Venezuela*'[23], '*la novela*' para '*La Metamorfosis*' y '*el libro*', usado como sinónimo para '*la novela*'[24], '*esos juguetes*' como hiperónimo para '*las canicas*, *el*

[21] Cadena de correferencia tomada del documento B1_ERES_col_025, disponible en: https://medellin.travel/destinations/santa-fe-de-antioquia.

[22] Elementos tomados de cadena de correferencias del documento B1_INOT_col_013, disponible en: http://www.elcolombiano.com/deportes/futbol/atletico-nacional-vs-independiente-del-valle-minuto-a-minuto-ME4655187.

[23] Elementos tomados del documento B1_INOT_ven_024, disponible: http://promar.tv/2018/03/06/maduro-despues-de-5-anos-me-siento-orgulloso-porque-he-sido-leal-a-chavez/.

[24] Tomados del documento C1_ARES_esp_006, disponible en: https://historiasbizarrasybizantinas.wordpress.com/2015/06/07/resena-de-la-metamorfosis-de-franz-kafka/.

camión y la pistola de hojalata'[25], fueron casos en los que se necesitaba de información semántica y contextual para su resolución. También, sucedía algo similar con el procesamiento correferencial de siglas como, por ejemplo, '*EE. UU.*' para '*Estados Unidos*'[26].

Reconocimiento de otros tipos de anáfora

Encontramos en nuestra muestra que la cantidad de casos de anáforas nominales y pronominales (entre ambas, el 95 % de las unidades) era mucho más grande que el resto de los tipos de correferencia. Por lo tanto, revisaremos en general los problemas de resolución que se relacionaron con anáforas adjetivales, verbales y adverbiales (los cuales solo representan el 5 % de anáforas en la muestra). Este grupo de anáforas representa un alto grado de dificultad para su procesamiento, como se puede observar en la Tabla 6:

Tabla 6. 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	0 %	100 %
	Con verbo estar	0 %	100 %
	Verbales en total	0 %	0 %
Adverbiales	Locativas	0 %	100 %
	De modo	0 %	100 %
	Adverbiales en total	0 %	100 %

En primer lugar, estos datos nos llevan a sugerir que, debido a su poca frecuencia[27], no se le prestó mucha atención a la resolución de estos elementos desde la programación inicial del sistema. Como lo mencionamos anteriormente, es

[25] Tomados del documento B1_NCUE_esp_021, disponible en: http://ciudadseva.com/texto/el-nino-al-que-se-le-murio-el-amigo/.

[26] Tomados del documento B2_INOT_esp_026, disponible en: https://www.practicaespanol.com/las-autoridades-cifran-en-26-los-fallecidos-en-el-tiroteo-en-una-iglesia-de-texas/.

[27] Trabajamos con 26 anáforas adjetivales, 8 anáforas adverbiales y 4 anáforas verbales. Recordemos que esta clasificación de la anáfora se basa en las nociones expuestas por Saiz (2002), Olivas (2006) y Ferrari (2014), entre los cuales se desarrollan trabajos de resolución anafórica.

probable que la construcción de *RelaxCor* solamente toma como anáforas aquellas que se refieran a entidades nominales, como vimos en la identificación de antecedentes. Así pues, la anáfora verbal, que se relaciona directamente con hechos y no con entidades, ni siquiera se tomaría en cuenta como un tipo de anáfora. En segundo lugar, ninguna de las anáforas adverbiales pudo ser reconocida. De la misma manera que con los antecedentes relacionados con hechos, el cero en porcentaje de acierto nos indica que el sistema tampoco fue diseñado para identificar adverbios como anáfora, por lo cual fueron descartados.

Por último, observamos que existe un pequeño porcentaje de acierto para la anáfora adjetival. Cabe aclarar aquí en qué consiste este tipo de anáfora. Según Olivas (2006), esta anáfora se da cuando a un referente nominal se le omite el núcleo nominal y solo permanece un adjetivo haciendo este la relación anafórica: "Wendy no [le] dio a ningún niño **una camisa verde de manga corta**, pero a Sue le dio **una roja**" (p. 27). Sin embargo, podemos observar que podría considerarse más bien 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. El pequeño porcentaje de acierto en este tipo de anáfora nos sugiere que la noción de anáfora adjetival no se tomó en cuenta, al igual que las dos anteriores. Esto se demuestra al revisar los casos en los que aparentemente se dio resolución:

- <chain id="6">
  <REF ANT="nom" type="SN" id="18" words="un público "/>
  <COREF ANA="adj" type="elip" id="19" words="**parte de** "/>
  <COREF ANA="adj" type="elip" id="22" words="**otra parte** "/>
  </chain>[28]

En esta cadena, el último elemento '*otra parte*' fue identificado por el procesador y relacionado con el elemento anterior '*parte de*'. Sin embargo, podemos deducir que esta relación solo se realizó por el hecho de utilizar la misma palabra '*parte*', como lo hemos visto anteriormente. Además, el primer elemento 'un público' no fue reconocido debidamente como un referente (dado que no detectaba elementos que pudieran correferir con él), por lo que la anáfora adjetival 'parte de' no se relaciona con este y la cadena correferencial no es efectiva. Por lo tanto, es muy probable que los autores del *software* consideraran estos como casos de elipsis.

En conclusión, podemos inferir que la teoría lingüística que se adopta antes de la programación del sistema influye notablemente en la identificación de lo que puede o no considerarse como anáfora. Como hemos visto, los verbos anafóricos se remiten a hechos, los cuales no son tomados en cuenta por el sistema como antecedentes; los adverbios no se incluyeron como tipos de anáfora, probablemente,

[28] Cadena extraída del documento B2_INOT_esp_026.

porque estos no se refieren a entidades nominales, sino, más bien, a otros adverbios o a complementos circunstanciales (Olivas, 2006; Saiz, 2002); por último, la anáfora adjetival puede considerarse como un tipo de elipsis.

Conclusiones

La revisión de todos estos tipos de problemas y su clasificación nos llevó a recapitularlos en una problemática más general que recoge las principales dificultades para el tratamiento computacional de la anáfora. Debemos tener en cuenta que solo se tomó como referencia el procesamiento correferencial de un *software* y que estos problemas podrían no aplicarse por completo a otros sistemas, como por ejemplo *FunGramKB* (Carrión, 2014; Ruiz, 2012). Por otra parte, los textos analizados fueron extraídos de un corpus especializado y representativo, por lo cual podríamos hacer algunas consideraciones sobre el tratamiento de la anáfora en español.

Tenemos, en primer lugar, como hemos destacado frecuentemente, que la programación inicial del sistema influye en gran medida en la concepción de lo que puede considerarse como anáfora en el análisis computacional. Esta programación descarta categorías gramaticales de difícil resolución o que, simplemente, no fueron tomadas como anáfora desde la teoría por su poca frecuencia. Vemos que también un factor relevante para la correcta resolución anafórica es la necesidad de utilizar información enciclopédica o cultura, que muchas veces no posee el sistema, como observamos en el ejemplo de Merkel, con los epítetos o con las siglas. Por último, observamos que el acierto en categorías como pronombres personales y nombres propios era más alto que la media general. En cuanto a los pronombres personales, vale la pena mencionar que, en la mayoría de los casos, estos debían estar muy cerca para su debida resolución. Cuando se encontraban otros elementos entre el pronombre y el referente, el sistema solía equivocarse con más frecuencia, ya que, debido a la poca carga semántica del pronombre, la resolución debe basarse en otra información, ya sea contextual o, incluso, pragmática. En lo que respecta a los nombres propios, notamos que su identificación se debía a los muchos casos de repetición. Estos casos nos indican otro problema: el sistema se basa, en gran parte, en la repetición de unidades, lo cual entorpece muchas veces el análisis correferencial.

Al mismo tiempo, se hallaron en el análisis de nuestro proyecto otros problemas que no pudimos ejemplificar en el cuerpo de este artículo. Primeramente, observamos que para resolver cadenas de correferencia hace falta mucha información de diferente tipo (morfosintáctica, sintáctica, semántica, cultural, etc.). Sin embargo, notamos que, a pesar de que el sistema tiene acceso a cierta

información, en algunos casos puede ser errada o incompleta. Tal suceso puede provocar errores en la identificación anafórica. Por ejemplo, si un pronombre personal ha sido etiquetado morfosintácticamente por el sistema como un artículo, este automáticamente se descarta como posible anáfora. Por lo tanto, los errores en etiquetados previos desembocan en dificultades para el procesamiento anafórico.

Por último, se hallaron en nuestro análisis otras causas de error menos frecuentes, pero especiales por su misma particularidad. Así, tenemos los casos en los que un elemento anafórico reúne dos o más referentes; por ejemplo, cuando un pronombre plural (*ellos, nosotros*, etc.) recoge referentes que han sido mencionados anteriormente (*Andrés*, *Jorge*, *el niño*). Este comportamiento anafórico no ha sido identificado por la máquina y probablemente no se tuvo en cuenta desde el modelo computacional. Ocurrieron otros casos similares, sobre todo en textos narrativos, en los cuales se suele cambiar la voz del enunciador, por lo cual, por ejemplo, una primera persona pasa a ser una tercera. Estos, al igual que la ambigüedad léxico/sintáctica del lenguaje natural, aunque representan un pequeño porcentaje entre todos los casos problemáticos, impiden que se procese la anáfora de una manera completamente correcta con la tecnología actual.

REFERENCIAS

Adam, J.-M. (2011). *La linguistique textuelle* (3ª ed.). París: Armand Colin.

Amsili, P., Denis, P., y Roussarie, L. (2005). Anaphores abstraites en français : représentation formelle. *TAL*, *46*(1), 15-40. Recuperado de https://tal.revuesonline.com/article.jsp?articleId=8142.

Anthony, L. (2018). AntConc (Versión 3.5.8) [Programa para computadora]. Tokyo: Waseda University. Disponible en http://www.laurenceanthony.net/software/antconc/.

Baker, P., Hardie, A., y McEnery, T. (2006). *A Glossary of Corpus Linguistics*. Edimburgo: Edinburgh University Press.

Brezina, V. (2018). *Statistics in Corpus Linguistics*. Cambrigde: Cambridge University Press. doi: 10.1017/9781316410899.

Carrión, M. (2014). Resolución de anáforas que requieren conocimiento cultural con la herramienta FunGramKB. *Revista de Lingüística y Lenguas Aplicadas*, *9*, 1-13. doi: 10.4995/rlyla.2014.2003.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., y Todiraşcu, A. (2017). Multiword Expression Processing: A Survey. *Computational Linguistics*, *43*(4), 837-892. doi: 10.1162/COLI_a_00302.

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

Denber, M. (1998). *Automatic Resolution of Anaphora in English*. Reporte técnico de Eastman Kodak Co. Recuperado de

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.904&rep=rep1&type=pdf.

Depain-Delmotte, F. (1997). La notion de cohérence textuelle et le traitement de l'anaphore. *Bulletin de Linguistique Appliquée et Générale*, (22), 129-154.

Ferrández, A. (1998). *Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos* (Tesis de doctorado). Universidad de Alicante, Alicante, España.

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

García, M. (2011). La distinción deíxis/anáfora y su aplicación a las formas de persona del español. *Revista de Filología Española*, *91*(1), 65-88. doi: 10.3989/rfe.2011.v91.i1.216.

Garside, R., Fligelstone, S., y Botley, S. (1997). Discourse Annotation: Anaphoric Relations in Corpora. En R. Garside, G. Leech, y A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 66-84). Londres: Longman.

Grajales, A. (2019). *Problemas del tratamiento computacional de la anáfora: Análisis en un corpus para la enseñanza del Español como Lengua Extranjera* (Tesis de pregrado). Universidad de Antioquia, Medellín, Colombia.

Instituto Cervantes. (2007). *Plan curricular del Instituto Cervantes*. Recuperado de: http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/.

Landragin, F. (2011). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus*, *10*, 61-80. Recuperado de http://corpus.revues.org/2010.

Landragin, F. (2014). Anaphores et coréférences : analyse assistée par ordinateur. En M. Fossard, y M.-J. Béguelin (Eds.), *Nouvelles perspectives sur l'anaphore : Points de vue linguistique, psycholinguistique et acquisitionnel* (pp. 29-54). Berna: Peter Lang.

Landragin, F. (2018). Étude de la référence et de la coréférence : rôle des petits corpus et observations à partir du corpus MC4. *Corpus*, *18*. Recuperado de: http://journals.openedition.org/corpus/3422.

Liddy, E. (1990). Anaphora in Natural Language Processing and Information Retrieval. *Information Processing and Management*, *26*(1), 39-52. doi: 10.1016/0306-4573(90)90008-P.

Longo, L., y Todiraşcu, A. (2010). RefGen : un module d'identification des chaînes de référence dépendant du genre textuel. *TALN 2010*.

Lundquist, L. (2013). *Lire un texte académique en français*. París: Éditions Ophrys.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., y McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. En *Proceedings of*

the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55-60). doi: 10.3115/v1/P14-5010.

Martínez-Barco, P. (2001). *Resolución computacional de la anáfora en diálogos: estructura del discurso y conocimiento lingüístico* (Tesis de doctorado). Universidad de Alicante, Alicante, España. Recuperada de http://hdl.handle.net/10045/1851.

McEnery, T., y Hardie, A. (2011). *Corpus Linguistics*. doi: 10.1017/CBO9780511981395.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. Nueva York: State University of New York Press.

Mitkov, R. (1999). *Anaphora Resolution: The State of the Art*. Reporte técnico. Recuperado de https://pdfs.semanticscholar.org/e782/00b1e3ba2a72de1ca9b9b2c5efa775151bfa.pdf.

Mitkov, R. (2001). Outstanding Issues in Anaphora Resolution. En A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 110-125). Nueva York: Springer.

Mitkov, R. (2002). *Anaphora Resolution*. Londres: Longman.

Mitkov, R., Evans, R., Orasan, C., Ha, L., y Pekar, V. (2007). Anaphora Resolution: To What Extent Does It Help NLP Applications? En A. Branco (Ed.), *Anaphora: Analysis, Algorithms and Applications*. Berlín: Springer.

Molina, J. (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 de doctorado). Université Grenoble Alpes, Grenoble, Francia.

Molina, J. Grajales, A., y Pemberty, J. (2019). Hacia un dispositivo informático basado en Corpus para la Enseñanza del Español Lengua Extranjera (DICEELE). *Revista Internacional de Tecnología, Conocimiento y Sociedad, 7*(1), 1-13. doi: 10.18848/2474-588X/CGP/v07i01/1-13.

Morales, R. (2004). *Resolución automática de la anáfora indirecta en el español* (Tesis de doctorado). Instituto Politécnico Nacional, México D. F., México. Recuperada de https://tesis.ipn.mx/bitstream/handle/123456789/11287/394.pdf?sequence=1&isAllowed=y.

Navarro, B. (2007). *Metodología, construcción y explotación de corpus anotados semántica y anafóricamente* (Tesis de doctorado). Universidad de Alicante, Alicante, España. Recuperada de http://hdl.handle.net/10045/7736.

Olivas, O. (2006). *Sistema de construcción de redes semánticas con detección de anáfora* (Tesis de maestría). Instituto Politécnico Nacional, México D. F., México. Recuperada de https://tesis.ipn.mx/xmlui/handle/123456789/1189.

Peña, G. (2006). *La anáfora y su funcionamiento discursivo: una aproximación contrastiva* (Tesis de doctorado). Universitat de València, Valencia, España. Recuperada de http://hdl.handle.net/10550/15244.

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

Saiz, M. (2002). *Influencia y aplicación de papeles sintácticos e información semántica en la resolución de la anáfora pronominal en español* (Tesis de doctorado). Universidad de Alicante, Alicante, España. Recuperada de http://hdl.handle.net/10045/1841.

Salguero, F. J., y Soler, F. (2010). Resolución abductiva de anáforas pronominales. En D. Fernández, E. Gómez-Caminero e I. Hernández (Eds.), *Estudios de Lógica, Lenguaje y Epistemología* (pp. 47-61). Sevilla: Fénix Editora.

Salmon-Alt, S. (2002). Le projet *ANANAS* : Annotation Anaphorique pour l'Analyse Sémantique de Corpus. *TALN 2002*, 163-172.

Sapena, E., Padró, L., y Turmo, J. (2013) A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution. *Computational Linguistics*, *39*(4), 847-884. doi: 10.1162/COLI_a_00151.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. París: Klincksieck.

Tordera, J. (2010). *Lingüística computacional y anáfora* (Tesis de doctorado). Universitat de València, Valencia, España. Recuperada de http://hdl.handle.net/10803/39087.

SOBRE LOS AUTORES

Andrés Felipe Grajales Ramírez

Estudiante de último semestre del pregrado Filología Hispánica de la Facultad de Comunicaciones de la Universidad de Antioquia. Bajo la figura de Joven Investigador del CODI para el proyecto DICEELE, ha desarrollado su tarea investigativa en el semillero de investigación "Corpus Ex Machina" en las áreas de la lingüística computacional, la lingüística de corpus y la enseñanza/aprendizaje de lenguas asistidos por computador.

Correo electrónico: afelipe.grajales@udea.edu.co

Orcid: 0000-0003-4721-2468.

Jorge Mauricio Molina Mejía

Profesor de cátedra del área de lingüística, coordinador del Grupo de Estudios Sociolingüísticos y del semillero "Corpus Ex Machina" adscritos a la Facultad de Comunicaciones de la Universidad de Antioquia. PhD en Informática y Ciencias del Lenguaje (Université Grenoble Alpes, Francia). Desarrolla sus tareas de docencia e investigación en las áreas de la lingüística computacional, la lingüística de corpus, la lingüística textual y la enseñanza/aprendizaje de lenguas asistidos por computador.

Correo electrónico: jorge.molina@udea.edu.co
Orcid: 0000-0002-1430-6364.

ANEXOS

Anexo A

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

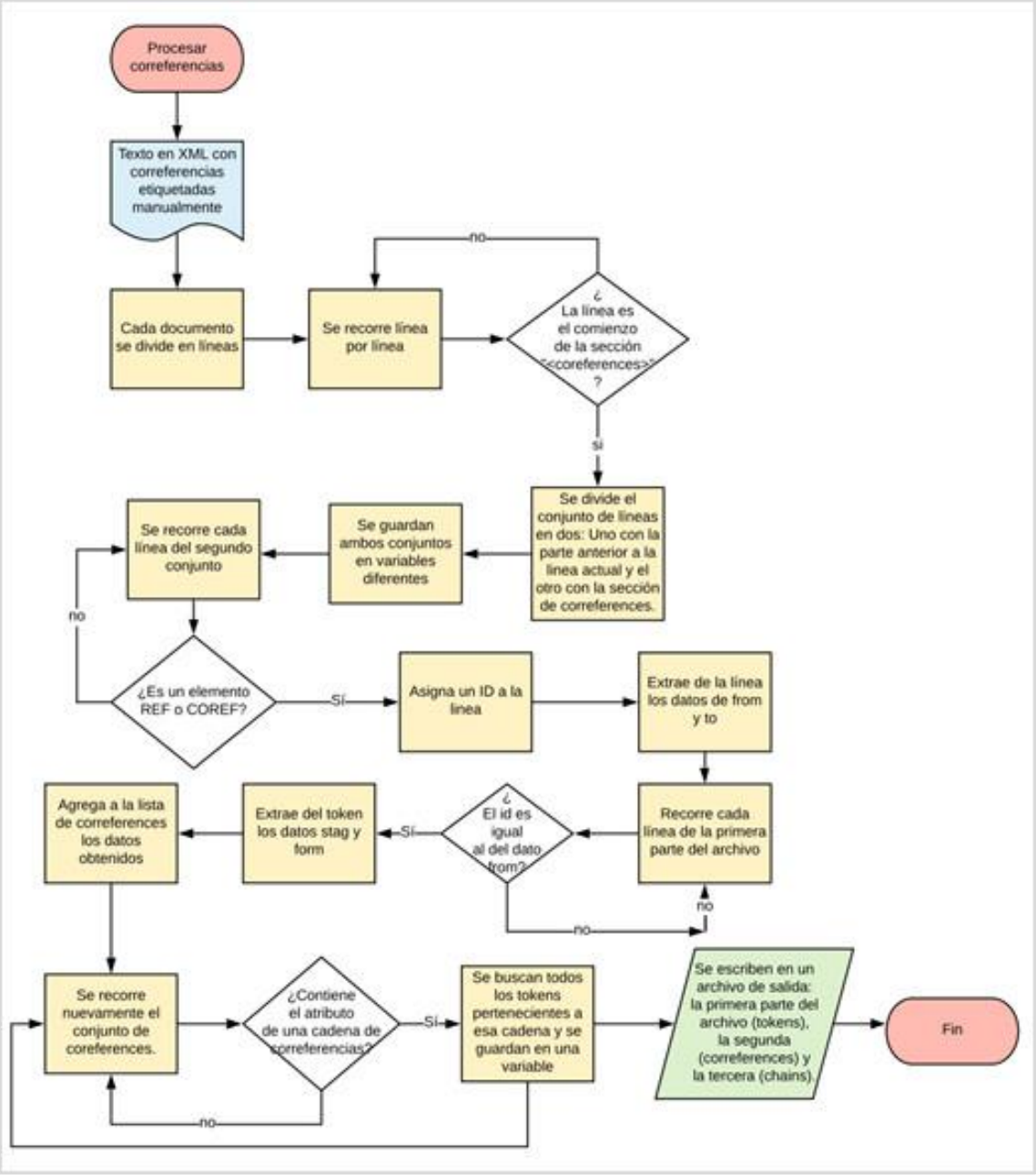


Figura 5. Diagrama realizado por José Luis Pemberty Tamayo, C*orpus ex Machina*

Anexo B

Ejemplos de los distintos tipos de correferentes encontrados en la muestra del corpus DICEELE

1) Antecedente nominal

- **a.** De sintagma nominal: ***la bailaora Rocío Molina****… la malagueña* (B2_IENT_esp_008).
- **b.** De nombre propio: *Ha muerto* ***Eusebio López****… Ya sé que usted no* ***lo*** *conoce* (B2_NCUE_per_028).

2) Antecedente verbal

- **a.** De sintagma verbal: *En las económicas te dejan* ***hablar*** *(a menos que se te ocurra* ***hacerlo*** *cerca del Congreso), pero…* (C1_ACOL_esp_023).

3) Antecedente con enunciados

- **a.** De frase: *No se creerá* ***lo que me acaba de ocurrir****. Me* ***lo*** *tendrá que contar el martes próximo* (C1_NREL_esp_024).
- **b.** De oración: ***Nota que se va convirtiendo en una pesada carga para su familia****. De* ***esto*** *podemos inferir* (C1_ARES_esp_006).
- **c.** De secuencia de oraciones: *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)*** *…* ***todo este palmarés*** (B1_INOT_col_013).

4) Anáfora pronominal

- **a.** De sujeto implícito: *no* ***había venido*** [el amante] *para repetir las ceremonias de una pasión secreta* (C1_NCUE_arg_032).
- **b.** De personal: *el martes próximo, dijo* ***ella*** [mi psicoanalista], *por hoy hemos terminado (*C1_NREL_esp_024).
- **c.** De posesivo: *Pero* [*Claudia*] *estaba arrimada a* ***su*** *puerta* (C1_NCUE_ecu_031).
- **d.** De relativo: *el estorbo de* [*los abogados de oficio*], ***que*** *no existen* (C1_ACOL_esp_023).
- **e.** De demostrativo*: por el cambio radical que experimenta* [*Gregor*]. ***Este*** *observa cómo* (C1_ARES_esp_006).
- **f.** De indefinido: *Así miraban allá en el Pacífico…* ***Todos*** [los muertos] *lo mismo…* (B2_NCUE_per_028).

5) Anáfora nominal

- **a.** De nombre propio: *la partida física del* [*presidente*] *…* ***Chávez*** *es más futuro que pasado* (B1_INOT_ven_024).
- **b.** De nombre común: *Vamos,* ***capitán*** [don Pedro]*… Hágale* (B2_NCUE_per_028).
- **c.** De sintagma nominal: [*la búsqueda*] *de respuestas a problemas comunes.* ***Esta búsqueda*** (B1_ACOL_chi_015).

6) Anáfora verbal

a. Con proverbo: *a menos que se te ocurra* ***hacerlo*** [hablar] *cerca del Congreso* (C1_ACOL_esp_023).

7) Anáfora adverbial

a. De lugar: *enderezo para* [*la casa de Cheo*] *y* ***ahí*** *me encuentro con…* (B2_NCUE_per_028).

b. De modo: [*Lo olíamos allá en el Pacífico…, el olor de los muertos*] … *Ahora Cheo López comenzaba a oler* ***así*** (B2_NCUE_per_028).

8) Anáfora adjetival

a. De adjetivo: *En* [*las dictaduras militares*] *te callan la boca y te aplican la picana. En* ***las económicas*** *te dejan hablar* (C1_ACOL_esp_023).

b. Superficial numérica: *Nadie en* [*la primera habitación*], *nadie en* ***la segunda*** (C1_NCUE_arg_032).

c. Posesiva: *para* [*el bienestar de los alemanes*], *no para* ***el nuestro*** (C1_ACOL_esp_023).

d. De elipsis: *y* ***la de este miércoles*** [la final] *en la Copa Libertadores con título continental* (B1_INOT_col_013).

9) Anáfora encapsuladora

a. Resumativa: [*desde las 3:00 de la tarde empezaron a ingresar al estadio y dos horas antes del inicio del juego ya estaba a reventar, para presenciar toda la emoción que genera una final, que incluyó un homenaje a algunos de los campeones de la Libertadores de 1989, entre ellos, el más ovacionado fue el 'loco' René Higuita*] … ***Con todo ese ambiente previo*** (B1_INOT_col_013).